Evolutionary Bioinformatics

Predicting Stability of Asexual Genomes by Global Computing

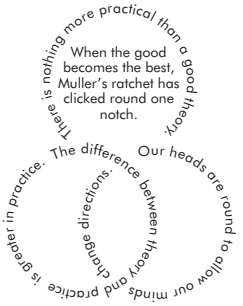
Laurence Loewe

There are problems in evolutionary biology that can only be solved by massive computing power. However, extensive use of computers creates problems of its own, usually new to biologists. This thesis contains the journey of a biologist, who was not content with the computing limitations of his day. He set out to create the first globally distributed computing system for evolutionary biology on the Internet: evolution@home.

Its first public simulator addresses Muller's ratchet, a population genetic process that might cause extinctions of asexual genetic systems like mitochondria, non-recombining bacteria and others. This could not only contribute to extinctions of species, but also to the uncultivable majority of microbes.

You are invited to join the quest of how organisms manage to stay alive.

A copy of this dissertation can be downloaded from http://www.evolutionary-research.net/



Fachgebiet Mikrobielle Ökologie
Department für Biowissenschaftliche Grundlagen
Technische Universität München

Evolutionary Bioinformatics: Predicting Genetic Stability of Asexual Genomes by Global Computing

Laurence Loewe

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: **Univ.-Prof. Dr. rer. nat.** Ludwig Trepl
Prüfer der Dissertation: **1. Univ.-Prof. Dr. rer. nat.** Siegfried Scherer,
2. Univ.-Prof. Dr. rer. nat. Eberhard Bertsch, **Ruhr-Universität Bochum**

Die Dissertation wurde am 12. November 2002 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 9. Dezember 2002 angenommen.

Zusammenfassung

Diese Arbeit untersucht die Stabilität asexueller (nicht rekombinierender) Genome mit mehreren Methoden. Die Auswirkung der beobachteten, hohen Mutationsraten auf die langfristige Stabilität von Mitochondrien wurde mit theoretischen Überlegungen zu Muller's Ratsche untersucht. Weil analytische Näherungen für die Berechnung der Geschwindigkeit der Ratsche nur in einem eingeschränkten Bereich der drei Parameter Selektionskoeffizient, Mutationsrate und effektive Populationsgröße möglich sind, wurden Individuen-basierte Simulationen durchgeführt. Diese sollten die analytischen Voraussagen überprüfen und erweitern.

Die Ergebnisse zeigen, dass Muller's Ratsche über einen Zeitraum von 20 Millionen Jahren in der Tat eine Bedrohung sein könnte. Damit scheint die bekannte Bedrohung[1] durch hohe Mutationsraten im Zellkern ein mitochondriales Gegenstück bekommen zu haben. Vielfältige biologischen Prozesse die dieses Genomzerfallsparadox lösen könnten, werden diskutiert.

Sieht man diese möglichen Lösungen, wird schnell klar, dass unter anderem viele weitere Simulationsmodelle mit einem enormen Rechenzeitbedarf nötig sein werden, um mehr herauszufinden. Daher wird ein Design für ein Software Framework vorgeschlagen, welches die Rechenkraft von global verteiltem Rechnen über das Internet für Individuen-basierte Modelle nutzt. Der erste Simulator in einer langen Reihe zukünftiger Simulatoren wurde implementiert, um das erste Global Computing System für die Evolutionsbiologie zu starten: *evolution@home* (siehe http://www.evolutionary-research.net). Die errechneten Ergebnisse (>28 000 Simulationen, >16 Jahre Prozessorzeit von >200 Teilnehmern) helfen beim Lösen der Diskrepanz von den hohen, beobachteten mitochondrialen Mutationsraten zwischen Generationen und den niedrigen, erschlossenen Raten aus phylogenetischen Stammbäumen. Neue Details zu Muller's Ratsche wurden ebenfalls beobachtet.

Der gleiche theoretische Ansatz kann auch auf (nicht rekombinierende) Mikroben übertragen werden, um abzuschätzen, welche Auswirkungen die Entstehungsrate nachteiliger Mutationen auf ihre Genome hat. Weil die Schlussfolgerungen ganz entscheidend von dieser Rate abhängen, wurde ein System entwickelt, mit dem sehr große Mengen (>6000) an Wachstumskurven analysiert werden können. Dieses System wurde benutzt, um die nachteilige Mutationsrate von *E. coli* in einem stationäre-Phase Mutationsakkumulationsexperiment nach der Bateman-Mukai Technik abzuschätzen. Dabei wurden auch besonders genaue Messungen der maximalen Wachstumsrate einer Übernachtkultur gemacht und mögliche evolutionäre Effekte beim Einfrieren auf -70°C in Glyzerin beobachtet. Detaillierte Analysen zeigen, dass Muller's Ratsche durchaus einen bedeutenden Teil zu der Mehrheit der unkultivierbaren Bakterien beisteuern könnte.

1. Eyre Walker & Keightley (1999) "High genomic deleterious mutation rates in hominids", Nature 397:344-347. - Kondrashov (1995) "Contamination of the Genome by Very Slightly Deleterious Mutations: Why Have We Not Died 100 Times Over?" J. theor. Biol. 175:583-594.

Abstract

The aim of this work is to study the stability of asexual (non-recombining) genomes. Towards this end several approaches are taken. To investigate the effects of observed, high mutation rates on long-term viability of mitochondria, Muller's ratchet theory is employed. As analytic approximations allow only predictions for a limited range of the three parameters, effective population size, mutation rate and selection coefficient, individual-based simulations are performed to check and extend the analytic solutions.

Results show that Muller's ratchet might indeed be a threat to mitochondria on a 20 million year timescale and thus appears to complement the threat that is known to come from the high deleterious mutation rates in the nuclear genome[2]. A variety of biological processes promoted to solve this genomic decay paradox are discussed.

Reviewing potential solutions shows the enormous need for further simulation models and more computing time to address these issues. Therefore, the design for a software framework is proposed that uses the power of global computing to investigate individual-based models of evolution. The first simulator in a long series of future simulators is implemented and used to start the first global computing system for evolutionary biology, *evolution@home* (see http://www.evolutionary-research.net). Its results (>28 000 simulations, >16 years CPU time from >200 participants) are used to investigate potential solutions for the discrepancy between high, short-term intergenerational mutation rates observed of human mitochondria and low, long-term mutation rates inferred from phylogenies. Further details of Muller's ratchet are observed for the first time.

The same set of tools that allows quantification of Muller's ratchet in mtDNA is applied to (non-recombining) microbes, to investigate consequences of deleterious mutations in their genomes. As conclusions critically depend on estimates for deleterious genomic mutation rates, a system is developed for analysis of large numbers of growth curves. This system is used to quantify the deleterious mutation rate of a population of *E. coli* in a stationary phase mutation accumulation experiment according to the Bateman-Mukai technique. Further observations include unprecedentedly accurate measurements of the maximal growth rate in an overnight culture and potential evolutionary effects of freezing in glycerol at -70°C. Detailed analysis shows that Muller's ratchet might contribute significantly to the majority of uncultivable bacteria.

2. Eyre Walker & Keightley (1999) "High genomic deleterious mutation rates in hominids", Nature 397:344-347. - Kondrashov (1995) "Contamination of the Genome by Very Slightly Deleterious Mutations: Why Have We Not Died 100 Times Over?" J. theor. Biol. 175:583-594.

Preface

This world is imperfect. And so is this thesis that deals largely with long-term consequences of this imperfection: Slightly deleterious mutations are one of the most interesting themes in contemporary evolutionary biology[1].

This work is truly interdisciplinary. As a biologist I see the unifying potential of evolutionary theory which makes me love to learn about the fascinating results from the many fields that touch it. Ultimately, I dream of things like predicting the structure, function and ecological meaning of newly mutated proteins to model their evolution. Such complex approaches will be needed to tackle one of the hardest problems in modern biology: the distribution of mutational effects. However, experiments alone are unlikely ever to produce enough results for good general estimates of it. Thus I am convinced that we need good individual-based computer models incorporating as much experimental evidence as feasible to actually observe it and understand its implications. But as one measurement is no measurement, one simulation is no simulation and in this area, one model is no model. Hence, we need to understand a whole armada of models and compare their results. Only then we will be able to gain confidence in their predictions.

Such an approach needs to minimize the programming needed for implementation and analysis of new models. And it needs enormous computing resources. Thus, I invested heavily in the design of a software framework that eventually will achieve exactly this: easy implementation of individual-based models that can be publicly distributed to many computers over the Internet. Success of global computing has shown that this is the way to go and thus *evolution@home* has started, the first global computing system for evolutionary biology.

The current implementation of *evolution@home* is very far from reasonable optimisation and certainly in some hole on the adaptive landscape[2]. Nevertheless, it exists and has reached a state where it makes sense to summarise the lessons of the past years, describe the current design of the framework and some of the biological results that have come out of computations so far – besides results from experimental work. I hope that further work will build on the foundation that is laid in this thesis. As any text of this size, it carries its own share of slightly deleterious mutations. It even carries some major deletions owing to restrictions in time and space. Although I wish it were not so, I cannot prevent it and if I tried, I would keep this work from ever being completed, only to demonstrate once again that perfectionism destroys life. I hope you enjoy reading the following pages and find some perfect use for this imperfect work[3].

Laurence Loewe

1. Lynch et al. (1999) "Perspective: Spontaneous deleterious mutation", Evolution 53:645-663.
2. Gavrilets (1997) "Evolution and speciation on holey adaptive landscapes", Trends Ecol. Evol. 12:307-312.
3. Schild H & Rammensee HG (2000) "Perfect use of imperfection", Nature 404:709-710.

Content Overview

I. Introduction **1**
1 Why do we need Evolutionary Bioinformatics? . . . 2
2 Global computing for bioinformatics . . . 6
3 Mutation rates in conflict . . . 13
4 Experimental evolution in microorganisms . . . 21
5 Aim of this work . . . 26

II. Muller's Ratchet Reviewed **27**
6 Theory . . . 28
7 Observations . . . 51

III. Methods and Software Development **57**
8 The vision behind evolution@home and *EEPSLION* . . . *58*
9 Lessons from development . . . 59
10 Current Design 5 of the software framework *EEPSLION* . . . *76*
11 Perspectives for future developments . . . 151
12 Design, description and tests of Muller's ratchet "Simulator005" . . . 156
13 Analytic predictions of the rate of Muller's ratchet in *Mathematica* . . . *168*
14 Fitness measurements in bacteria . . . 176

IV. Results Part 1: Experiments **187**
15 Accuracy of fitness measurements . . . 188
16 Freezing can affect growth rate of bacteria . . . 197
17 Deleterious mutation rate and effect in the bacterial stationary phase . . . 208

V. Results part 2: Extensions of Muller's ratchet theory **219**
18 A simple heuristic equation for predicting the rate of Muller's ratchet . . . 220
19 Directions for future work on theoretical models of Muller's ratchet . . . 237
20 Compensatory mutations and Muller's ratchet . . . 253

VI. Results Part 3: Biological Consequences of Muller's Ratchet **263**
21 Muller's ratchet in mtDNA might cause extinctions in mammals . . . 264
22 Muller's ratchet might threaten the Amazon Molly and ancient asexuals . . . 281
23 Muller's ratchet might threaten endosymbionts like *Buchnera* . . . *288*
24 Muller's ratchet and microbial diversity . . . 300

VII. Discussion **319**
25 Origin of the majority of uncultivable microbes . . . 320
26 Potential solutions for the paradoxes . . . 332
27 Practical relevance of Muller's ratchet . . . 345

VIII. Appendix **A-1**

Content in Detail

Zusammenfassung iv
Abstract v
Preface vi
Content Overview vii

I. Introduction 1

To set the stage for this work, the need for evolutionary bioinformatics and global computing is presented together with two intriguing questions from recent research in evolutionary biology: How can the mutation rate paradox be resolved? How do bacteria evolve under natural circumstances?

1 Why do we need Evolutionary Bioinformatics? 2

Evolutionary questions are among the most complex problems known to modern science. It is argued that high-quality computer simulations are an important key to answers. However, extensive simulation efforts meet the same typical problems again and again. To allow concentration on biological issues, a framework for easy implementation and investigation of individual-based evolutionary models would be of great help. This work is largely about developing the design for such a framework. Historical aspects are explained.

2 Global computing for bioinformatics 6

Global computing via the Internet emerges as a new way of massive parallel multiprocessing with potentially enormous CPU-power. After a short review of the his-

tory of global computing and the opportunities it offers for bioinformatics, the structure of problems well suited to such an approach are discussed. Besides existing frameworks for global computing the anatomy of successful projects is described to help decide whether benefits outweigh the extra effort of going global.

2.1 Structure of problems addressable by global computing 7
2.2 Anatomy of successful projects 9
2.3 Existing global computing frameworks 10
2.4 Costs of global computing 11
2.4.1 On the participants' side 11
2.4.2 On the operators' side 12
2.4.3 On the developers' side 12

3 Mutation rates in conflict 13

It is very common to infer mutation rates from phylogenies assuming a palaeontologically calibrated molecular clock. However, when sample sizes are large enough, mutation rates observed in known pedigrees are much higher than phylogenetically inferred rates. It is shown here, that archaeological evidence can be used to infer mutation rates that are similar to those observed in pedigrees. Mutational hot spots and selective removal are among the potential solutions discussed for this mutation rate paradox. Neither appears to solve this conflict completely. The potential threat from Muller's ratchet has been noticed, but was never really quantified in detail.

3.1 Mutation rates observed in pedigrees 14
3.2 Inferred rates from archaeology 15
3.3 Potential solutions 17
3.3.1 Mutational hot spots 18
3.3.2 Selective removal 18
3.3.3 Perspectives 19

4 Experimental evolution in microorganisms 21

Microorganisms are excellent for studying evolution, as they evolve extremely fast, even on ecological timescales. Their short generation times allow experimentation with processes that are impossible to observe in other species. This approach has been extensively used to investigate adaptive evolution and deleterious mutation rates in serial transfer and chemostat experiments. However, despite high interest in stationary phase mutations, the deleterious mutation rates associated with prolonged starvation have never been observed.

4.1 Adaptive evolution in serial transfers 21
4.2 Deleterious mutation rate estimates 24
4.3 Open problems 25

5 Aim of this work 26

II. Muller's Ratchet Reviewed 27

If too many slightly deleterious mutations occur in a population, Muller's ratchet can lead to genomic decay and mutational meltdown, driving the population to extinction. After reviewing general circumstances where this might occur, current methods for predicting the rate of the ratchet are discussed. Empirical evidence of the operation of the ratchet is presented as well as a list of features that might help recognize its operation in a species.

6 Theory 28

Muller's ratchet was discovered when investigating advantages of recombination that might have led to the evolution of sex. After hiding in theoretical population genetics for a long time, the last decade saw the discovery of its relevance for conservation biology and numerous other problems.

6.1 Definition of Muller's ratchet 29
6.2 Definition of important details 35
6.3 Predicting the rate of Muller's ratchet 41
6.4 Mutational meltdown and other long term consequences 44
6.5 When is Muller's ratchet dangerous? 47
6.5.1 Significance of Muller's ratchet for conservation biology . . 49

7 Observations 51

Numerous observations underscore the importance of Muller's ratchet for our understanding of nature. Some of the most important are listed here.

III. Methods and Software Development 57

A central part of this work is design of "EEPSLION", a framework that fosters development of evolutionary models that can be integrated into the "evolution@home" global computing system. The first simulation model to be part of evolution@home is described, as are analytical methods for predicting the rate of Muller's ratchet and a method for precise measurements of fitness in bacteria.

8 The vision behind evolution@home and *EEPSLION* 58

This chapter describes the vision behind the first global computing system for evolutionary biology, "evolution@home", as well as the motivation behind the corresponding software framework, "EEPSLION".

9 Lessons from development 59

As all non-trivial software, EEPSLION has undergone several rounds of the design-implement-use cycle. The most important lessons up to now are discussed.

9.1 Design 1: Special language (1997) 60
9.2 Design 2: Class library (1998) 61
9.3 Design 3: Framework (1999) 62

9.4 Design 4: Global computing (2000) 66
9.5 Why use C++? 71
9.6 Importance of standards and tools. 73

10 Current Design 5 of the software framework *EEPSLION* 76

An overview of the current Design 5 of EEPSLION is presented. Reasons for important design decisions are discussed.

10.1 Specifics of distributed computation of individual-based models . . . 76
10.1.1 Parallelisation of different models is natural 76
10.1.2 Parallelisation of different single runs is recommended . . . 76
10.1.3 Parallelisation of time series analysis is recommended 77
10.1.4 Parallelisation within a single run will rarely outweigh development costs 78
10.1.5 Parallelisation of multi-level populations is desirable 79
10.1.6 Vectorization is hardly feasible 80
10.1.7 Work units have extremely variable sizes 80
10.1.8 Even incomplete runs might be used 81
10.2 Overview over *EEPSLION* 81
10.3 *EEPSLION* server suite design 85
10.3.1 Functionality of EEBasicLibrary 87
10.3.2 GridSearchGenerator 91
10.3.3 Professional releases of simulators (1) 93
10.3.4 GridFolderStructureGenerator 95
10.3.5 GridPrioritySetter 95
10.3.6 RunFileProducer and Webserver 96
10.3.7 The simulate command 98
10.3.8 ResultsCollectorRaw 102
10.3.9 ServerNameServer 103
10.3.10 Professional releases of simulators (2) 104
10.3.11 ResultsSortIn 104
10.3.12 SingleRunResultAnalyzer 105
10.3.13 MultiRunResultsCollector 106
10.3.14 HighScoreGenerator 106
10.3.15 MultiRunResultsArranger 107
10.3.16 Visualisers and Datamining 108
10.3.17 Comparison with conventional global computing frameworks 110
10.4 HFF results database design 111
10.5 General simulator and security 114
10.5.1 The simplest design possible 115
10.5.2 The ideal evolution@home simulator 116

10.6 Describing the world in *EEPSLION* *119*
10.6.1 EEBioObject 120
10.6.2 The evolve() methods 122
10.6.3 Fast access to objects in the world 123
10.6.4 EEWorld 123
10.6.5 EEHabitat 124
10.6.6 EELocation 125
10.6.7 EEPopulation 125
10.6.8 EEIndividual 126
10.6.9 EEIdenticalCohort 127
10.6.10 EEGenome 127
10.6.11 EEChromosome 129
10.6.12 EEGene 129
10.6.13 EESequence 129
10.6.14 EEDistributionMutationalEffects 130
10.6.15 EEMemoryManager 130
10.6.16 EEController 131
10.6.17 RND generators 132
10.7 Types of Data and their analysis 133
10.7.1 Definitions of types of parameters 133
10.7.2 How to automate analysis of time series 135
10.7.3 Computation of Multi-run-results 136
10.7.4 Types of simulation projects 137
10.8 First experiences with evolution@home and the Participant's Free Choice System 138
10.8.1 Lower computing time boundary 139
10.8.2 Upper computing time boundary 140
10.8.3 The Participant's-Free-Choice System 140
10.8.4 Biases in participant choices 141
10.8.5 Intermediate results 142
10.8.6 Benchmarking computer systems 144
10.8.7 Prediction accuracy 146
10.8.8 Definition of the error of magnitude 147
10.8.9 Related work 148
10.8.10 Lessons and Conclusions 150

11 Perspectives for future developments 151

Eepslion is build to grow with its applications. Here, current status, next steps and potential for future developments are discussed.
11.1 Current state of development 151
11.2 The next steps 152
11.3 Long term perspectives 153

12 Design, description and tests of Muller's ratchet "Simulator005" 156

This is the scientific part of the documentation of the first simulator released as part of evolution@home. It is called Simulator005 (=S005) and simulates two Muller's ratchet - processes, one in the foreground ("ratchet") and one in the "background". The mutational parameters of these processes can be chosen freely and quite some details are observed about the foreground ratchet process.

12.1 Why was this simulator built? 156
12.2 Input and output parameters 158
12.3 Simulation Model 160
12.4 Validation of Simulator005 165
12.5 Future of Simulator005 167

13 Analytic predictions of the rate of Muller's ratchet in *Mathematica* 168

Diffusion theory and quantitative genetics methods allow for approximations of the rate of Muller's ratchet. As they involve complex calculations, Mathematica is needed for actual computation. Here an integrated method is presented that automatically picks the correct method and presents the best analytical prediction currently available.

13.1 Diffusion theory 168
13.2 Quantitative genetics theory 172
13.3 Building a comprehensive method 175

14 Fitness measurements in bacteria 176

One potential measure of fitness is the growth rate of a bacterial culture in liquid medium. It can be readily determined by measuring optical density. However, to increase accuracy of estimations of the true growth rate, many replicates are needed. To analyse about 6000 growth curves with a total of more than 1.5 million OD values, a special software module for Excel VBA was developed. It allows fast high precision fitness measurements. Equipment, software, bacterial strains and other experimental details are described in this chapter.

14.1 Experimental setup 176
14.2 Software 178
14.3 Bacterial strains and media 182

IV. Results Part 1: Experiments 187

In order to get better estimates of deleterious mutation parameters, a system was developed that allows fast and easy measurements of maximal growth rate in bacteria. This system was used to investigate effects of freezing in glycerol on growth rate and to compare fitness effects of mutations that are neutral from a molecular biology perspective. Finally, a stationary phase mutation accumulation experiment was conducted and analysed with the Bateman-Mukai *technique. The surprisingly high mutation rate is comparable to values known from stationary phase adaptive mutation experiments.*

15 Accuracy of fitness measurements 188

Accuracy of fitness measurements depends on (i) the number of replicates for a given sample and (ii) the variance within the measurements of the sample. Besides describing the detection limit of this approach, a list of potential problems for high-precision growth rate measurements is given. Finally a number of specific mutations are tested for their neutrality. The result suggests that common techniques in molecular biology have significant impact on fitness from an evolutionary perspective.

15.1 Maximal accuracy 188
15.2 Potential problems in growth speed measurements 190
15.3 Neutrality in molecular biology versus neutrality in evolution 194

16 Freezing can affect growth rate of bacteria 197

Although it is well known that most individual bacterial cells do not survive freezing in glycerol, it is generally believed that properties of surviving cells are not affected. Here it is shown that freezing in glycerol at -70°C can lead to increase or decrease of doubling time and this can not be avoided just by allowing one or two nights of recovery under assay conditions. Thus, freezing appears to trigger complex evolutionary processes with unpredictable effects. Potential mechanisms are discussed.

16.1 Experimental design 197
16.2 Variable effects of freezing can be quite stable 202
16.3 Potential causes 202
16.3.1 Freezing might exhibit strong selective pressure 202
16.3.2 Freezing induces mutations 204
16.3.3 Stationary phase events 205
16.3.4 Conclusions 206

17 Deleterious mutation rate and effect in the bacterial stationary phase 208

Mutation accumulation experiments measure important parameters for prediction of long-term evolution. Here, the first stationary phase mutation accumulation experiment is reported. The deleterious mutation rate (0.07 mutations / genome / day) appears to be surprisingly high, but is comparable to extrapolations from adaptive mutation experiments. The selection coefficient (-3% / mutation) is comparable to that found in other mutation accumulation experiments. Some implications of the findings are discussed.

17.1 Experimental design 211
17.2 Results 211
17.3 Related work 211
17.4 Mutation rate paradox for bacteria 217

V. Results part 2: Extensions of Muller's ratchet theory 219

This work has led to new observations and insights that extend understanding of traditional models of Muller's ratchet. Its database of currently >28000 simulation results with >16 years CPU time is expected to be fruitful for further research on models of Muller's ratchet.

18 A simple heuristic equation for predicting the rate of Muller's ratchet 220

Predictions of the rate of Muller's ratchet usually involve mathematical theory and equations beyond the reach of most biologists. The simple equations presented here allow average biologists to estimate orders of magnitude for the rate of the ratchet with their pocket calculator. Thus it becomes easy to check whether the ratchet might play a significant role in a given situation and further investigations have merit. These equations were developed to predict computational complexity of simulations with Simulator005.

18.1 The *Equation150* system 221
18.2 Meaning of *Equation150* elements 222
18.3 Quality of predictions 227
18.4 Complete parameter space overview for Muller's ratchet 231
18.5 The *Equation172* system 235
18.6 Future developments 236

19 Directions for future work on theoretical models of Muller's ratchet 237

Simulator005 observes a variety of parameters (some for the first time) that play a role in theoretical models of Muller's ratchet. The resulting database (over 28000 simulations with over 16 years CPU-time) is an excellent opportunity to improve theory by comparing various predictions with actual simulations on a large scale. This chapter suggests a number of theoretical aspects that might benefit from further investigation using the results computed by Simulator005.

19.1 Observations of different phases during a click of Muller's ratchet . 237
19.2 Observations of multiple clicks of Muller's ratchet 241
19.3 Observation of the variability of the rate of Muller's ratchet 244
19.4 The 'neutral slowdown' of Muller's ratchet 249
19.5 Other projects 252

20 Compensatory mutations and Muller's ratchet 253

While back-mutations do indeed play no role in an infinite sites genome, they can become important when the ratchet operates in a finite sized genome over long periods of time. Then the probability increases so that an already mutated site is hit by a compensatory back-mutation. The resulting compensatory mutation rate is further increased by the fact that many deleterious mutations can be repaired by more than one specific mutation on a molecular level. Relevant biological data is discussed.

20.1 A simple model for compensatory mutations on a molecular level in asexual genomes 255
20.2 Simulation details 257
20.3 Simulation results 258
20.4 What are realistic values for *Repairways* and *Warranty* in biological genomes? 260

VI. Results Part 3: Biological Consequences of Muller's Ratchet 263

Although recombination is almost universal, some genetic systems on earth seem to come along without it. The simulations presented here predict extinction times for various non-recombining evolutionary lines due to Muller's ratchet. For some of these systems, recombination appears to be more important than previously thought; for others there is no simple solution for the resulting genomic decay paradox.

21 Muller's ratchet in mtDNA might cause extinctions in mammals 264

Mitochondrial DNA is a very common asexual genetic system in otherwise sexual species. The observation of relatively high mitochondrial mutation rates in human pedigrees has led to the question of how such a genetic system could survive the resulting onslaught of Muller's ratchet. Here, the threat from Muller's ratchet is quantified in detail for the first time. It is concluded that for many mammals and humans a range of biologically realistic parameter combinations should have led to extinctions within 20 million years. Some implications for endangered species and evolution of the human line are discussed.

21.1 Measures for quantifying the threat of Muller's ratchet in principle 267
21.2 Biological parameter values 269
21.3 Extinction times 272
21.4 Conclusions 277

22 Muller's ratchet might threaten the Amazon Molly and ancient asexuals 281

There are species that are known to have lived without sex for extended periods of time. These have long been suspected of being threatened by Muller's ratchet. This is the first report that quantifies this threat in detail for the small unisexual fish Poecilia formosa *(Amazon Molly). It is shown that for some biologically realistic mutation rates, Amazon Molly is indeed threatened by Muller's ratchet. However, further data on mutation rates will be needed to finally settle the issue. Similarly, collecting the same data as for Amazon Molly will allow quantifying the threat from Muller's ratchet in other ancient asexuals too.*

22.1 Amazon Molly biology 281
22.2 Extinction times for the Amazon Molly 285
22.3 Assessing the threat of Muller's ratchet for other ancient asexuals is easier now 286

23 Muller's ratchet might threaten endosymbionts like *Buchnera* 288

The enteric bacteria of the Genus Buchnera have lived endosymbiontically in aphids for more than 100 million years. As they are sequestered in the cells of their hosts, they cannot exchange genetic material with bacteria from other hosts. Consequently, it has been suspected that Muller's ratchet should have driven these bacteria to extinction long ago. Molecular signatures of genomic decay have been found in their genomes (loss of genes and excessive accumulation of non-synonymous mutations). However, Buchnera is still alive and nobody knows why. Here the threat from Muller's ratchet to Buchnera is quantified to facilitate further investigation of this intriguing question.

23.1 Biological ratchet parameters 289
23.2 Extinction times 298
23.3 Conclusions 298

24 Muller's ratchet and microbial diversity 300

Bacteria can reproduce completely without recombination and can have a significant fraction of mutators in their population. Whenever these mutators become ecologically successful and do not recombine for longer periods of time, Muller's ratchet degrades their genomes. Here, Muller's ratchet is quantified (i) for free-living, asexual bacteria, (ii) for ancient asexual bacteria that have been trapped in sediments and might have grown cryptically and (iii) for RNA viruses. It is concluded that irregular sex as known in modern bacteria and RNA viruses must have played an important role in the evolution of microbes from very early times. However, when bacteria have learned how to survive in the long-term with occasional recombination, an important driving factor for the evolution of regular recombination disappeared. Thus, the mystery of the origin of sex might be bigger than previously thought.

24.1 Ratchet parameters in microbes 300
24.2 Muller's ratchet in free bacteria 305
24.3 Muller's ratchet in ancient bacteria 310
24.4 Muller's ratchet in the deep hot biosphere 314
24.5 Muller's ratchet in RNA viruses 314
24.6 Conclusion 317

VII. Discussion 319

By now the reader may ask why bacteria or humans are still alive. Good question. After reviewing the situation in cultivable and uncultivable bacteria, potential solutions for the mutation rate paradox and for the genomic decay paradox are listed. As the simulations in this work consider only a small fraction of all biological mechanisms known to be relevant and as many key parameters are not well known, the solution for these paradoxes requires considerable further work. Finally, some practical consequences are discussed.

25 Origin of the majority of uncultivable microbes 320

This chapter reviews reasons why a significant part of the uncultivable majority of bacteria is probably generated by genomic decay due to Muller's ratchet and a newly defined 'eco-ratchet'. It might well be that only bacteria that managed to escape this decay by employing some of the specific anti-decay mechanisms discussed have a high probability of being cultured by taxonomists. Despite their potentially decaying genomes, uncultivable bacteria are likely to play important, highly adapted roles in the global ecosystem. If irregular recombination indeed saved many bacteria from mutational meltdown, an important selective force in the evolution of sex has to be reevaluated: why should complicated regular sex evolve, if simple bacterial mechanisms achieve the same result?

25.1 How probable are high deleterious stationary phase mutation rates? 320
25.2 Adapting to one environment may be bad for life in another 322
25.3 The eco-ratchet and the fitness-barrel 324
25.4 What could keep bacteria from genomic decay? 327
25.5 Conclusions 329

26 Potential solutions for the paradoxes 332

This chapter discusses most known potential solutions for the mutation rate paradox and for the genomic decay paradox. As the final solution could also be a combination of those listed, an enormous amount of empirical work will be necessary to decide on these issues. However, in most cases some hints indicate how probable it is that a proposed solution contributes to the actual solution.

26.1 Mutation rate paradox 332
26.2 The fitness-balance and the genomic decay paradox 335

27 Practical relevance of Muller's ratchet 345

Discussions about Muller's ratchet and related topics are usually regarded as a highly esoteric topic from theoretical population genetics understood by only very few people. However, these topics can have very practical consequences that are much easier to understand. Some examples are given here.

27.1 Microbiology 345
27.2 Cloning 346
27.3 Conservation biology 347
27.4 Sustainability and environmental mutagens 347

VIII. Appendix A-1

Further reference material is listed here.

Introduction to Evolutionary Bioinformatics A-2

Current bioinformatics mainly compares sequences and predicts their structure to understand their immediate function. Nevertheless, grasping their adaptive evolutionary significance is the key to much deeper understanding. Here evolutionary bioinformatics is defined as all computer-based analyses that use evolutionary models explicitly or implicitly to either understand data or the models themselves. While analytical models of evolution are important for solid foundations, mathematical tractability severely limits the amount of detail they can describe. Individual-based models overcome this barrier at the costs of higher computational complexity and often a more limited analytical understanding. As every model is an arbitrary abstraction of reality, we need to check many models in order to make reasonable predictions. Perspectives and pitfalls of this approach are discussed. At its heart evolutionary bioinformatics does just that: compare many models of evolution in order to understand the real causes that shaped the data we observe today.

27.1 Why do we need bioinformatics? A-3
27.2 Evolution as a unifying theory A-6
27.3 Evolutionary bioinformatics: One definition and three viewpoints A-10
27.4 Why we need good *in silico* experiments A-12
27.4.1 Ways to understand biology A-12
27.4.2 How to make simulations that are junk A-14
27.4.3 Building good models is an art A-15
27.4.4 Advantages of good *in silico* experiments A-18
27.4.5 General limitations of *in silico* biology A-19
27.5 General simulation approaches A-21
27.6 Types of evolutionary simulations A-23
27.6.1 Evolutionary computation A-23
27.6.2 Artificial Life A-25
27.6.3 Simulations of biological evolution A-26
27.7 How to remove bottlenecks in simulating evolutionary models . . . A-27
27.8 Blessing and curse of interdisciplinary research A-31

Mutation rates paradox table for mtDNA A-32
Posters from International Conferences A-38
Glossary A-43
Abbreviations A-45
References A-48
Curriculum vitae A-72
Acknowledgements A-76

I. INTRODUCTION

To set the stage for this work, the need for evolutionary bioinformatics and global computing is presented together with two intriguing questions from recent research in evolutionary biology: How can the mutation rate paradox be resolved? How do bacteria evolve under natural circumstances?

1 Why do we need Evolutionary Bioinformatics?

Evolutionary questions are among the most complex problems known to modern science. It is argued that high-quality computer simulations are an important key to answers. However, extensive simulation efforts meet the same typical problems again and again. To allow concentration on biological issues, a framework for easy implementation and investigation of individual-based evolutionary models would be of great help. This work is largely about developing the design for such a framework. Historical aspects are explained.

Evolution as a unifying theory

One has not to go as far as Dobzhansky[1] to recognize the unifying power of the modern theory of evolution: it is currently the only theory that may unify biology. While the last decades have seen an enormous rise in quality of testing different hypotheses about neutral evolution[2], we are still far from reaching the same quantitative rigour for discerning hypotheses about adaptive evolution of specific structures or species. How many mutations of what effect were needed in the ecological setting of a given species to lead to the adaptations observed? Figure 1 depicts the extraordinary wide range of data needed for testing different adaptive hypotheses in detail. A detailed understanding of these processes is also important for prediction of ecological processes, a topic of considerable interest[3].

Limitations for investigating evolution

As argued in the introduction to evolutionary bioinformatics in the appendix of this work, high-quality computer simulations based on sound models and using biologically meaningful parameters are pivotal for more detailed understanding of evolution. While many biologists are concerned with accumulating more data about existing organisms and their contexts, those few who want to analyse such knowledge in simulations[4] encounter similar technical obstacles again and again[5]:

1. Dobzhansky (1973) "Nothing in biology makes sense except in the light of evolution", American Biology Teacher 35:125-129.
2. Huelsenbeck & Rannala (1997) "Phylogenetic methods come of age: testing hypotheses in an evolutionary context", Science 276:227-232. - Golding, (ed, 1994) "Non-neutral evolution: Theories and molecular data", New York, Chapman & Hall. - Kimura (1983) "The neutral theory of molecular evolution", Cambridge, Cambridge University Press.
3. Clark et al. (2001) "Ecological forecasts: an emerging imperative", Science 293:657-660. - Carpenter (2002) "Ecological futures: Building an ecology of the long now", Ecology 83:2069–2083.
4. See appendix for an overview over various simulation approaches.

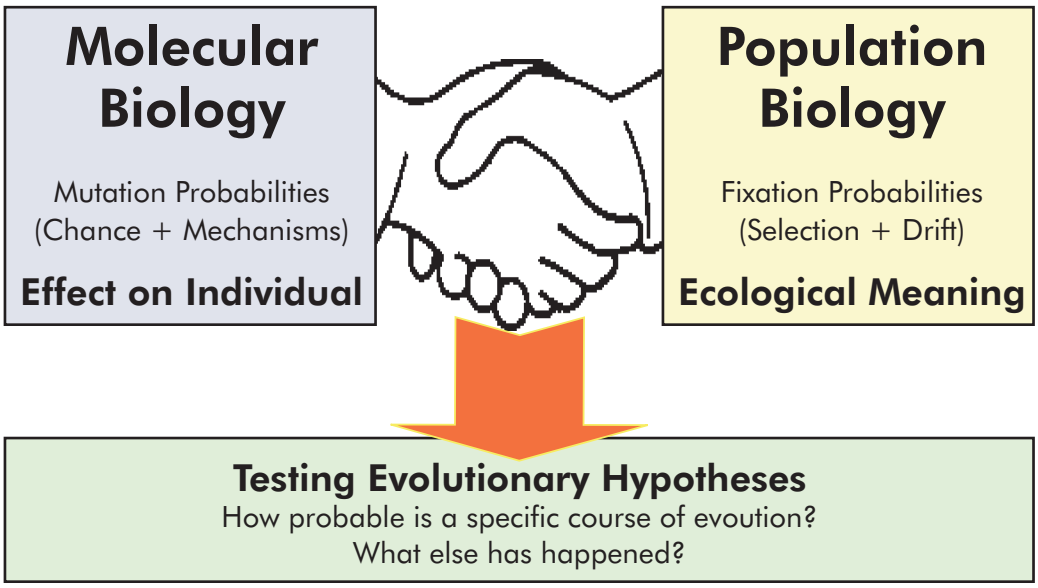


Figure 1 Evolution has a great potential to unify biology.

- **Basic simulation software development**. As a suitable framework for implementing evolutionary models is missing in most cases[6], many biologists start with programming from scratch. The need is so large that there are even books that help to address corresponding problems[7].
- **Computing time**. Once a non-trivial ecological or evolutionary model is implemented, a rigorous analysis of the model often needs more computing time than most biologists can afford. This is especially disturbing, as those parameter combinations that apply to natural settings are often very computing-intense.

Origins of this work, software side

This work started with the same needs. However, instead of just developing an one-way solution to investigate the model I was interested in, I thought much about the design of a software framework[8] that can facilitate easy implementation of new evolutionary models and that can distribute corre-

5. See appendix for a more detailed list of these obstacles.
6. See appendix for an overview over existing frameworks.
7. Eg. Wilson (2000) "Simulating ecological and evolutionary systems in C", Cambridge, Cambridge University Press. - Ebert (1999) "Plant and animal populations - Methods in demography", San Diego, Academic Press.
8. Gamma et al. (1995) "Design Patterns", Reading, Massachusetts, Addison-Wesley.

sponding simulations to many computers easily. I found that long-term computing needs in evolutionary biology can be expected to be so large that it merits building a global computing system for evolutionary biology[9]. Thus, I started evolution@home, to allow everybody from the general public with an Internet computer to participate in evolutionary research. However, I very seriously underestimated the complexity of the framework-building enterprise. As any non-trivial software development, it has seen several rounds of the design–implement–test cycle, and from this experience[10], it can be expected that future work will lead to further significant improvements in design. As of this writing, the current state of the framework is as follows. The largest part of the design phase of the current cycle has been completed and the result of that is reported in Chapter 10 of this work. However, large parts of this new design could not be implemented due to time constraints. It will be continued as soon as possible.

Origins of this work, evolution side

As the need for such a framework came from some real-world questions in evolutionary biology, I tested my various framework designs on those evolutionary problems that I was interested in. From the beginning, a key question was how the discrepancy between mutation rates observed in short-term human pedigrees and long-term phylogenies could possibly be reconciled[11]. While this is still not clear, Chapter 3 reviews the current state of the discussion and the increasing number of cases where a similar conflict appears to exist. One of the questions in that debate is how any evolutionary line could possibly survive the threat from Muller's ratchet[12], if mitochondrial mutation rates are as high as had been observed. While general answers like "this is indeed a problem" could be easily given from the literature[13], no detailed quantification is available as of today. The only possibility of quantifying this threat from Muller's ratchet in detail was to conduct extensive computer simulations of Muller's ratchet, as all analytical approximations worked only for a limited range of parameters and were routinely checked by complex computer simulations themselves.

Why global computing?

Thus, the idea was born to use global computing to assess the threat of Muller's ratchet to the stability of asexual genomes. Although the current design of the evolutionary problem-solving framework has not yet been im-

9. See Chapter 2 for a review of opportunities for global computing in bioinformatics.
10. See Chapter 9 for a more detailed account on the individual cycles.
11. Loewe & Scherer (1997) "Mitochondrial Eve: the plot thickens", Trends Ecol. Evol. 12:422-423.
12. See Part II for a review of Muller's ratchet, as no reasonably comprehensive and up to date review has been published yet.
13. Gabriel et al. (1993) "Muller's Ratchet and mutational meltdowns", Evolution 47:1744-1757.

plemented, enough code from the earlier designs could be reused for production of the first publicly released simulator of evolution@home: Simulator005. In retrospect, it was the correct decision to start evolution@home. In the 19 months of its (little advertised) public operation, it has accumulated more than 28 000 single simulation results with a combined total of more than 16 years CPU time. To get such an amount of computing time I could have applied to the Leibniz Rechenzentrum for a project[14] to run on their Hitachi super-computer (4th largest computer in the world in 2000). However, then I would have had to know from the beginning the exact set of tasks to compute and I would now have to apply for a new project, whenever I have a new question. With evolution@home, things are much easier. Investigation of new parameter combinations requires only scheduling on the web. If no new parameter combinations are scheduled, continued efforts increase statistical power by computing additional stochastically independent repeats. Thus, continuation of this research is many orders of magnitude easier[15]. Once implementation of the new fully automated design has been completed, computing power can be expected to rise considerably, as only few are willing to deal with the current tedious semi-automated details of evolution@home. Despite the lack of reasonable advertising, about 200 non-anonymous participants supported this current work by successful submission of results.

Results

As a result of these efforts, the detailed threat from Muller's ratchet can be quantified much more precisely now, not only in mitochondrial DNA, but also in a number of other asexual genetic systems that have been suspected of being threatened by the ratchet[16]. As a side effect, a number of interesting observations could be made that relate to more theoretical aspects of Muller's ratchet. Interest in deleterious mutation rates has also led me to measure this rate in the stationary phase of bacteria[17]. Together with simulation results these rates might suggest that Muller's ratchet contributes significantly to the majority of uncultivable microbes. The origin of this hypothesis is a nice example of why we need evolutionary bioinformatics: Some secrets can only be uncovered by simulation.

14. Only projects with more than 50 000 hours (5.7 years) CPU-time are allowed on this machine. As researchers from all over Germany compete for time on this machine, a national committee of reviewers has to decide on what is allowed to run there.
15. Super-computers might still be used for selected, extremely complex simulations, eg. with RAM requirements that are beyond the 32bit address-range of modern PCs.
16. See Results Part 3. For an introduction into Muller's ratchet theory, see Part 2.
17. See Results Part 1 for the first Bateman-Mukai type mutation accumulation experiment in the stationary phase.

2 Global computing for bioinformatics[18]

Global computing via the Internet emerges as a new way of massive parallel multiprocessing with potentially enormous CPU-power. After a short review of the history of global computing and the opportunities it offers for bioinformatics, the structure of problems well suited to such an approach are discussed. Besides existing frameworks for global computing the anatomy of successful projects is described to help decide whether benefits outweigh the extra effort of going global.

The idea of distributing computing tasks is old. People have been thinking how to do it since 1973, when the first Ethernet network was installed in the Xerox Palo Alto Research Center. Behind it stands the observation that one fast single CPU will always lose price per performance comparisons against many cheap CPUs - if there is a good way of distributing work. However, it took until the Internet matured in the 1990s, for the idea of distributing tasks to many other computers on the Internet to be born and to let them calculate solutions in their idle cycles. Thus, global computing was born. Global computing is a special form of grid computing, a rapidly evolving field that aims to connect many computing nodes to address complex computational problems[19].

Table 1 Short history of global computing.

1995	George Woltman: Use free computing resources on the Internet for *GIMPS*, the "Great Internet Mersenne Prime Search", the first semi-automated global computing project.
1997	Independent start of the first fully automated global computing projects: *PrimeNet* (automated GIMPS) and *distributed.net* (brute force cracking of encryption keys)
1999	Start of SETI@home, currently the largest project with >3.8 million participants, >1 000 000 CPU years and > 1.7×10^{21} FLOPs in 38 months
2000	Global computing goes commercial: *United Devices, Entropia, Parabon*, and others start business.
2001	First companies go out of business.
2002	More than 60 active and 20 completed global computing projects as listed by Kirk Pearson (http://www.aspenleaf.com/distributed/)

18. An extended version of this chapter can be found in Loewe (2002) "Global computing for bioinformatics", Brief. Bioinformat. vol 3 issue 4 (Dec) pages 377-388.

Table 2 Some global computing projects that touch bioinformatics.
For links to these and other projects see http://www.aspenleaf.com/distributed/

Linear regression	exhaustive linear regression of clinical trial data (Parabon)
Sequence analysis	HMMER, BLAST, SIM4 (United Devices) - TurboBLAST (Entropia) - multiple sequence alignments (Übero)
Protein structure prediction	folding@home, genome@home, Parabon, Folderol
Docking drugs to protein structures	THINK search for cancer drugs (United Devices) AutoDock in FightAIDS@home (Entropia)
Individual-based models	evolution@home

Grid computing[19] usually demands faster network connectivity than is generally available on the Internet. It can be defined as follows:

Definition

> *Global computing is massive parallel distributed computing using the Internet for data transfer to build a centrally controlled meta-computer from the idle CPU cycles of PCs of voluntary participants from the general public.*

A short history of global computing can be found in Table 1. Global computing projects can be operated at two levels:

- **semi-automated** projects involve occasional manual interaction on the participants' side to get new computing tasks and submit results.
- **fully automated** projects install a program that automatically gets computing tasks and submits results. It sometimes even automatically updates itself to newer versions.

Bioinformatics in global computing

While sometimes hopes run high of easily solving complex problems by global computing, in reality it is not always feasible to start a global computing project. However, a number of applications in bioinformatics have been found to be well suited to such an approach. Table 2 lists some known global computing projects that touch bioinformatics.

2.1 Structure of problems addressable by global computing

There are well defined characteristics of problems that can be efficiently addressed by global computing:

19. Foster & Kesselmann (1998) "The Grid: Blueprint for a new computing infrastructure", Morgan Kaufmann Publishers. See Loewe (2002), ibid. for further references.

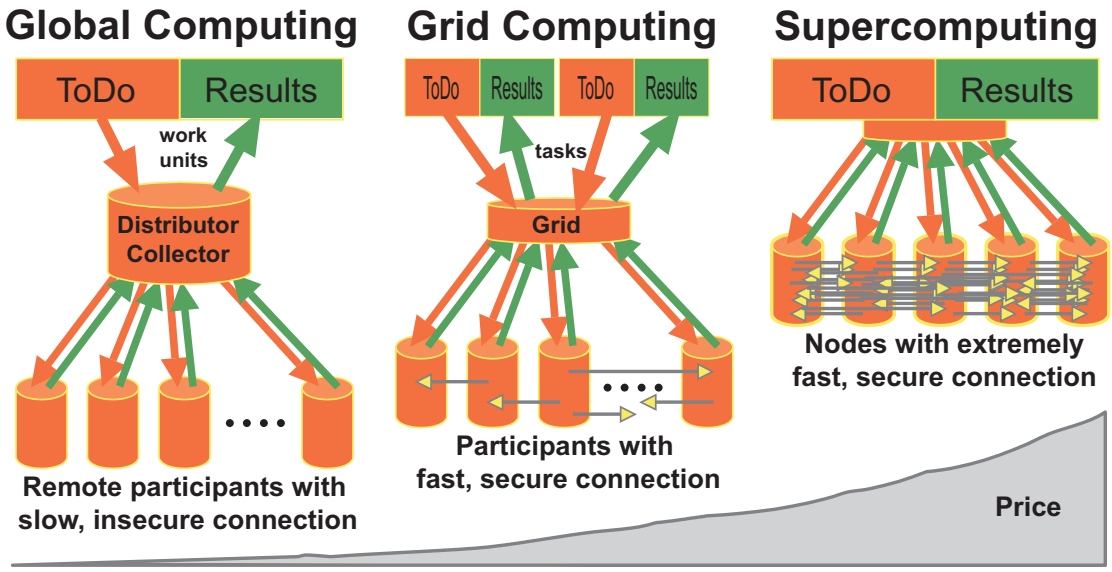


Figure 2 Structure of problems addressable by global computation in contrast to grid computing and supercomputing. Please note the arrows that symbolise communication between different work units on different CPUs.

- They express trivial parallelism (embarrassingly parallel or multi-parameter applications)[20] and can be packed into independent work units that do not need to talk to each other.
- Work units have a high computation per bandwidth demand and are of reasonable size.

If network demand is too high or different work units need to communicate, then a typical grid solution should be employed. If work units need to communicate extensively, then a super-computer should be used. The price of the latter solutions increases as the network is what you pay for. Figure 2 depicts this relationship graphically.

Examples of problem structures that fit this description extremely well are different drugs that need to be fitted to all known protein structures or repeats of individual-based simulations of evolution that just start with different parameters.

20. Fedak et al. (2001) "XtremWeb: A generic global computing system", pp. 582-588 (http://www.lri.fr/~fedak/XtremWeb/Gcpd.ps) in: Buyya & Mohay (eds) 1st International Workshop on Global and Peer-to-Peer Computing on Large Scale Distributed Systems at the 1st IEEE/ACM International Symposium on Cluster Computing and the Grid - CCGrid2001, Brisbane, Australia, IEEE Press.

Table 3 Ideal worker software features

Readily available	Cross-platform code included in major share-ware collection CDs. Easy distribution without registration over the web. Small installer size (download-times and space on the local hard disk).
Fully automated	Get work units and submit results without participant's interaction. Automated computing code download. Automated upgrading of the worker itself.
Non-interfering	Lowest background priority possible. No use of virtual memory (may block the rest of the system). Immediate stop without losing results at participant's wish. Stop automated online connections at participant's wish. Screen-saver only version for less interference.
High security standards	Secure execution environment shields participant's computer from potentially malicious code of a computing application (Sandbox). Secure results transmission keeps participants from cheating. Saves intermediate results to minimise losses in case of system crash. System service version for computing after logout.
Nice graphical user interface	Show beautiful picture of what is being computed. Show progress and local computing statistics and allow easy interruption. Easily edit preferences that describe the commitment of the participant. (Allow commercials? What project? High-score identity? Work unit size in terms of CPU-time, RAM, disk space, network bandwidth, etc.?)

2.2 Anatomy of successful projects

What then are the details of successful global computing projects? These involve a surprising number of features that have little to do with informatics. This reflects the fact that such projects usually depend on participants from the general public for their computing resources[21]. To run a successful global computing project you need:

- A good way to pack your problem into independent work units of appropriate size (CPU-time hours to weeks, transmitted data as small and as rare as possible, depends on the network.).
- A nice worker software that does not interfere in practice with the user's daily work. See Table 3 for features of the ideal worker software.
- A central server that distributes work units.
- A central server that collects and analyses results.
- A nice website that explains the problem you are trying to solve, distributes the worker software and publishes computing statistics.
- Offer a motivation to participate. See Table 4 for examples.
- Organise a good advertisement campaign.

21. This does not apply to companies that can afford to buy a Pentium-farm including administrators.

Table 4 Possible motivations for participants

Interest in the specific problem
Contribute to global progress in general
Like high-score races with meaningful side effects
Get famous by finding something significant
Get the chance to win some money
Commercial selling of computing time
Execution of orders of their companies on its computer
Like beautiful screen savers
Want to have some fun
Compute because their friends are doing it

As can be easily seen from this list, the reward of global computing power comes only at the cost of dealing with numerous problems that have no primary connection with the original content of the research. This raises the question about existing frameworks for global computing.

Table 5 Some frameworks for global computing. Commercial products are marked with *. For more frameworks and for frameworks for grid computing, see collections in http://www.aspenleaf.com/distributed/distrib-devel.html, http://dmoz.org/Computers/Computer_Science/Distributed_Computing/Platforms/ and http://www.GridComputing.com/

Platform name	Developer	Website
MetaProcessor*	United Devices	http://www.ud.com
DCGrid*	Entropia	http://www.entropia.com
Frontier*	Parabon Computation	http://www.parabon.com
COSM	Beberg at Mithral	http://cosm.mithral.com
XtremWeb	Cappello et al.	http://www.xtremweb.net
BOINC	SETI@home II	http://boinc.ssl.berkeley.edu
Fida	Ding et al.	http://odinn.chem.washington.edu
Models@home	Kriege & Vriend	http://www.cmbi.nl/models
P2P Accelerator Kit	Intel	http://www.intel.com/ids/p2p

2.3 Existing global computing frameworks

To solve the general problems every global computing project faces, a number of frameworks have been developed (Table 5). Deciding which framework will be employed is no easy task, as a wrong decision can severely limit long-term development of a project. Individual-based simulations have a special set of features that prohibit simple application of existing frameworks. As no global computing framework will help with pre-processing and post-processing of tasks (the most complex part of going global with individual-based simulations) and a semi-automated mode of operation is compar-

atively easy to implement, evolution@home does not use any of these frameworks. The final decision on this issue will have to be made when transitioning to full automation. The following features will be important for evolution@home:

Flexible scheduling

Most platforms show little support for easy priority setting for specific work units. This, however, is very important for investigating individual-based models by simulation, as there is no fixed number of tasks that have to be run to complete the project. Rather, the principal investigator determines the range of parameters of interest, and after evaluating results, refines or increases the range of parameters to compute. This involves more interactivity in the scheduling of tasks than blind brute force parallelisation.

Variable sizes and incomplete work units

Simulations that are too simple for scheduling on remote PCs need to be sorted out for simulation on a few local PCs with high bandwidth capable of transferring large amounts of results data. On the other hand, simulations that are too complex need to be sorted out too, lest they block computing power for more important intermediate parameters. As actual computing times can be very difficult to predict, support for preliminary analysis of incomplete results should be available (ie. evolution was valid, but ended before the end of the run).

Users like to choose their commitments

Currently there is also no global computing framework that allows users to choose computational complexity of their work units. This primarily comes from the relatively fixed size of work units in other projects. However, as individual-based models encompass a wide range of computational complexities a framework that supports them may easily offer this feature.

2.4 Costs of global computing

When deciding whether a particular problem should be solved by global computing or other means, the price tag plays an important role.

2.4.1 On the participants' side

Participation in global computing is free of charge. However, every participant has to pay for his connectivity and electrical energy. The latter is the equivalent of about one light bulb per PC, while the former is not very expensive in many projects. Finally, there are no administration costs, if one's system is fine. Uninterrupted computation, however, may expose weaknesses in a system that were not apparent before (cooling problems in summer, hard disk failures, RAM problems, etc.). Therefore, participation is only re-

commended if participants have working air conditioning in summer and a reasonable backup strategy (for most ordinary PCs the cheapest solution is to regularly copy everything to a dedicated backup hard disk). As system administrators of larger facilities have to provide that anyway, additional costs of making a large number of PCs participate in this way is low from the hardware perspective.

The software perspective depends on the project. With fully automated worker software that even updates itself, administration is reduced to a minimum. At the other extreme, semi-automated projects require regular interaction which is no longer trivial if the number of PCs increases. Therefore a worker software that is as simple as possible is important for global computing projects.

2.4.2 On the operators' side

Costs on the operators' side depend heavily on the framework and on what one includes in these costs. If one defines these costs in terms of distributing worker software and work units on one side and collecting results on the other, then this is extremely cheap: United Devices manages about half a million participants with just 2 system administrators. However, if one includes analysing the data, then costs may rise considerably: unclear objectives and little automation result in exploding costs. However, a global computing project should not start at all if its goals are not precisely clear.

Besides the results-oriented part of administration, no global computing project can be active without a website. Keeping it up to date and organising the publicity needed, demands human resources too. And finally one should not forget the costs of maintaining a high-scores list as people want to see that their contributions count. When all this operational work is fully automated, operator costs are low. However, to automate this, one has to invest in development.

2.4.3 On the developers' side

This is the biggest hurdle, as no automated way exists for porting an arbitrary application to a global computing platform. As no existing framework supports the features needed for effective global computation of individual-based models of evolution, I decided to invest in development in order to be able to investigate those evolutionary issues that I am interested in. These will be explained in the next chapter.

3 Mutation rates in conflict[22]

It is very common to infer mutation rates from phylogenies assuming a palaeontologically calibrated molecular clock. However, when sample sizes are large enough, mutation rates observed in known pedigrees are much higher than phylogenetically inferred rates. It is shown here, that archaeological evidence can be used to infer mutation rates that are similar to those observed in pedigrees. Mutational hot spots and selective removal are among the potential solutions discussed for this mutation rate paradox. Neither appears to solve this conflict completely. The potential threat from Muller's ratchet has been noticed, but was never really quantified in detail.

Conventional knowledge

Mitochondrial DNA is one of the most often used systems for phylogenetic analyses due to experimental ease. However, recent observations have made theoretical analysis extremely difficult. Traditionally, substitution rates[23] for use in molecular clocks were constructed by dividing half the number of observed differences between two groups by a palaeontologically known date. Such phylogenetic rates for the hypervariable regions of the D-loop[24] have been in the range of 2.5% - 26% substitutions per site per million years (= /bp/Myr)[25] with 12%/Myr being one of the faster values[26].

Mutation rates in coding region

Studies comparing complete mitochondrial genomes[27] or genes from the coding region and from the hypervariable control region (D-loop)[28] find

22. This chapter and the corresponding part of the appendix are being prepared for publication. The initial idea for such an approach has already been published in Loewe & Scherer (1997) "Mitochondrial Eve: the plot thickens", Trends Ecol. Evol. 12:422-423.
23. Do not confuse with divergence rates (approximately 2x substitution rate).
24. See "Overview of the mitochondrial genome and known mutations." on page 265.
25. Vigilant et al. (1991) "African populations and the evolution of human mitochondrial DNA", Science 253:1503-1507. - Pesole et al. (1992) "The evolution of the mitochondrial D-loop region and the origin of modern man", Mol. Biol. Evol. 9:587-598. - Tamura & Nei (1993) "Estimation of the Number of Nucleotide Substitutions in the Control Region of Mitochondrial DNA in Humand and Chimpanzees", Mol. Biol. Evol. 10:512-526. - Nei (1992) "Age of the Common Ancestor of Human Mitochondrial DNA", Mol. Biol. Evol. 9:1176-1178.
26. Stoneking et al. (1992) "New approaches to dating suggest a recent age for human mtDNA ancestor", Phil. Trans. R. Soc. Lond. B 337:167-175.
27. Lopez et al. (1997) "Rates of nuclear and cytoplasmic mitochondrial DNA sequence divergence in mammals", Mol. Biol. Evol. 14:277-286. -
28. Luikart et al. (2001) "Multiple maternal origins and weak phylogeographic structure in domestic goats", Proc Natl Acad Sci U S A 98:5927-5932. - Giuffra et al. (2000) "The origin of the domestic pig: independent domestication and subsequent introgression", Genetics 154:1785-1791. - Hiendleder et al. (1998) "Analysis of mitochondrial DNA indicates that domestic sheep are derived from two different ancestral maternal sources: no evidence for contributions from urial and argali sheep", J Hered 89:113-120.

that phylogenetically inferred substitution rates in the rest of the mitochondrial genome is between one sixth[29] and one half[30] of the D-loop rate in humans, while experimental observations in the worm *Caenorhabditis elegans*[31] and in humans[32] found no differences. Mutation rates in mtDNA have practical importance for confidence levels of forensic identifications[33].

3.1 Mutation rates observed in pedigrees

High rates are no statistical fluke

Thus a great stir was generated by the first report of a pedigree mutation rate[34] that found 2 intergenerational substitutions in 81 generational events as this equals to about 250%/Myr. Depending on the populations sampled, subsequent studies found similar[35] or lower rates[36]. But what initially appeared to be a statistical fluke has now received good statistical support: In a total of 2076 human generational events 30 mutations in the 673 bp of the D-loop have been observed, amounting to a rate of about 107%/Myr with large confidence intervals[37]. Thus these observed rates appear to be roughly tenfold higher than phylogenetically inferred rates. Direct observations of (complete) mitochondrial genomic mutation rates in *C. elegans* lead to even higher estimates of 890%/Myr[38] and suggest high mutation

29. Ingman et al. (2000) "Mitochondrial genome variation and the origin of modern humans", Nature 408:708-713.
30. Horai et al. (1995) "Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs", Proc Natl Acad Sci U S A 92:532-536.
31. Denver et al. (2000) "High direct estimate of the mutation rate in the mitochondrial genome of Caenorhabditis elegans", Science 289:2342-2344.
32. Howell et al. (1996) "How rapidly does the human mitochondrial genome evolve?" Am. J. Hum. Genet. 59:501-509.
33. Ivanov et al. (1996) "Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij Romanov establishes the authenticity of the remains of Tsar Nicholas II", Nature Genet. 12:417-420. - Parson et al. (1998) "Population data for 101 Austrian Caucasian mitochondrial DNA d-loop sequences: Application of mtDNA sequence analysis to a forensic case", Clinical-Medicine 111:124-132. - Gill et al. (1994) "Identification of the remains of the Romanov family by DNA analysis", Nature Genet. 6:130-135.
34. Howell et al. (1996) "How rapidly does the human mitochondrial genome evolve?" Am. J. Hum. Genet. 59:501-509.
35. Parsons et al. (1997) "A high observed substitution rate in the human mitochondrial DNA control region", Nature Genet. 15:363-368. - Parsons & Holland (1998) "Mitochondrial mutation rate revisited: hot spots and polymorphism - Response", Nature Genet. 18:110-110.
36. Soodyall et al. (1997) "The founding mitochondrial DNA lineages of Tristan da Cunha Islanders", Am J Phys Anthropol 104:157-166. - Jazin et al. (1998) "Mitochondrial mutation rate revisited: hot spots and polymorphism", Nature Genet. 18:109-110. - Siguroardottir et al. (2000) "The mutation rate in the human mtDNA control region", Am. J. Hum. Genet. May 66:1599-1609.
37. The most complete estimate by Parsons & Holland (1998) is 135%/Myr (95%CI=72%-198%). If Siguroardottir et al. (2000) data (32%/Myr, 95% CI = 6.5%-95%) are added, then 1371 (27) +704 (3) = 2076 generational events (30 mutations) result in 30/2076/673bp/20a*1000000 = 107%/Myr.

rates in mitochondrial DNA may be more widespread. If that is indeed the case, then it is not clear what mutation rate should be applied to analyses of the very recent past: high recently observed pedigree rates or old, remotely inferred phylogenetic rates?

3.2 Inferred rates from archaeology

Archaeological knowledge can suggest ...

If one entertains these two possibilities for a while, interesting inferences can be made in archaeological settings. Where sometimes counter-intuitive archaeological conclusions have been made on the bases of phylogenetic mutation rates, observed pedigree rates open new molecular perspectives for conclusions that archaeologists have sometimes reached independently before. The complex table given in the appendix extends an earlier report[39] and lists instances where interesting alternative conclusions can be reached by allowing mutation rates to be in the pedigree order of magnitude. Then these alternative archaeological hypotheses are used to estimate 'archaeological' substitution rates with values between long-term phylogenetic and short-term pedigree rates.

... that higher mutation rates might be more widespread

If all three types of data are plotted over the timeframe they were observed in, the resulting graph shows that molecular evolution appears to be slower when studied over longer timescales (Figure 3). This contradicts the simple expectation that evolutionary substitution rate equals intergenerational mutation rate, if mutations are neutral[40]. While molecular clocks are known to display considerable rate heterogeneity[41], the discrepancies reported here are systematic. This study is the first to use archaeological inferences to estimate the extent of the problem in mtDNA for an intermediate time frame. The discrepancy has certainly become more than a statistical

38. Denver et al. (2000) "High direct estimate of the mutation rate in the mitochondrial genome of Caenorhabditis elegans", Science 289:2342-2344 report 1430%/Myr including indels and 890% (± 220%) for base substitutions only. This depends on a generation time of 4 days which is close to the maximal generation time of 3 days reported in Vassilieva et al. (2000) "The fitness effects of spontaneous mutations in Caenorhabditis elegans", Evolution 54:1234-1246. The longest possible lifetime of14-20 days is also given by Vassilieva et al. (2000). To account for a potentially longer generation time *in natura* the molecular clocks may be halved to 715% and 445%, respectively. However, even if generation time were 20 days, the slowest possible clock in *C. elegans* mtDNA would still tick at 220% - fast by any standard.
39. Loewe & Scherer (1997) "Mitochondrial Eve: the plot thickens", Trends Ecol. Evol. 12:422-423.
40. For details see Loewe & Scherer (1997) ibid. or Kimura (1983) "The neutral theory of molecular evolution", Cambridge University Press.
41. Cutler (2000) "Understanding the overdispersed molecular clock", Genetics 154:1403-1417. - Gillespie (1991) "The Causes of Molecular Evolution", New York, Oxford University Press. - Scherer (1990) "The protein molecular clock: time for a reevaluation", Evol. Biol. 24:83-106.

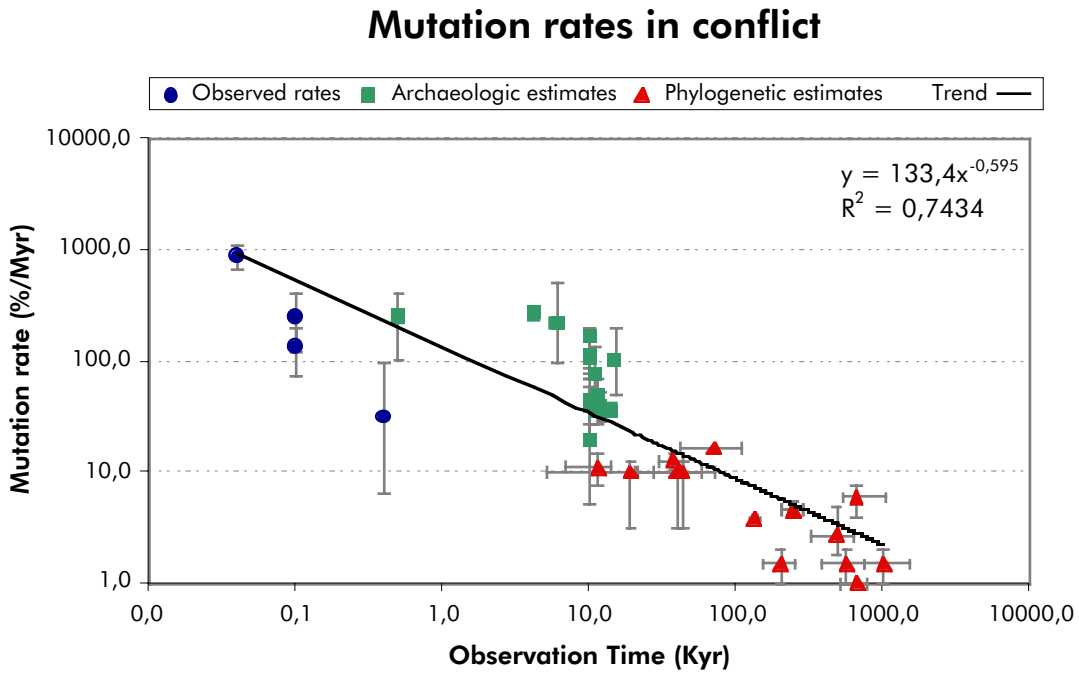


Figure 3 Mutation over the timeframe they were observed in.
Observed rates and phylogenetic estimates are from the literature, archaeological estimates are from this work (for details, see the long table in the appendix). Tom Parsons introduced this type of plot at the first international workshop on human mitochondrial DNA, 25 - 28 October 1997, Washington, D.C. (See Gibbons, 1998, "Calibrating the mitochondrial clock", Science 279:28-29). Here, the first attempt is made to replace the question marks in the original figure.

fluke and careful analyses are needed, as important new principles might stand behind it. To put it in one sentence:

Definition of the mutation rate paradox

> *The mutation rate paradox consists of two conflicting mutation rates that both should apply to the same evolutionary line: a long-term, low, inferred (phylogenetic) rate and a short-term, high rate supported by independent evidence from recent observations (e.g. sequences in pedigrees or inferences from archaeology) .*

While this work has an emphasis on mtDNA, similar problems can be found in other systems too[42].

42. For example a standard clock would lead us to believe that the last 500 000 years spanned a longer period of time than the previous 700 million years in the male fertility gene *Odysseus* in *Drosophila* according to Ting et al. (1998) "A rapidly evolving homeobox at the site of a hybrid sterility gene", Science 282:1501-1504. See also Strauss (1999) "Can mitochondrial clocks keep time?" Science 283:1435, 1437-1438 and the discussion of bacterial mutation rates observed in this work.

3.3 Potential solutions

Heterogeneity of rates in space and time

Careful comparison of the pedigree rates and their confidence intervals[37] in different populations shows considerable heterogeneity between various evolutionary lines - potentially a reflection of their lifestyle. No mutations could be observed in the extremely conservative population of Tristan da Cunha[43], mutation rates in the fast-paced US[44] are high and Icelanders[45] have some intermediate position. While such potential relationships await further analysis, considerable rate heterogeneities among populations of the same species have been observed repeatedly[46]. However, they do not easily solve the paradox from an evolutionary perspective (unless one attributes *all* these observations to a recent rise in mutation rates).

Complex replication details of mtDNA

As the rates observed are intergenerational rates and segregation of mutations in mtDNA is usually very rapid[47], complex multi-level population genetics[48] is no simple solution either. Solutions proposed include mutational hot-spots and long-term loss of mutations due to selective removal or drift effects[49].

43. Soodyall et al. (1997) "The founding mitochondrial DNA lineages of Tristan da Cunha Islanders", Am J Phys Anthropol 104:157-166.
44. Parsons et al. (1997) "A high observed substitution rate in the human mitochondrial DNA control region", Nature Genet. 15:363-368.
45. Siguroardottir et al. (2000) "The mutation rate in the human mtDNA control region", Am. J. Hum. Genet. May 66:1599-1609.
46. Zhang & Ryder (1995) "Different Rates of Mitochondrial DNA Sequence Evolution in Kirk's Dik-dik (Madoqua kirkii) Populations", Mol. Phylogenet. Evol. 4:291-297. - Arctander et al. (1996) "Extreme genetic differences among populations of Gazella granti, Grant's gazelle, in Kenya", Heredity 76:465-475.
47. Jenuth et al. (1996) "Random genetic drift in the female germline explains the rapid segregation of mammalian mitochondrial DNA", Nature Genet. 14:146-151. - Koehler et al. (1991) "Replacement of bovine mitochondrial DNA by a sequence variant within one generation", Genetics 129:247-256.
48. Jenuth et al. (1997) "Tissue-specific selection for different mtDNA genotypes in heteroplasmic mice", Nature Genet. 16:93-95. - Otto & Orive (1995) "Evolutionary consequences of mutation and selection within an individual", Genetics 141:1173-1187. - Otto & Hastings (1998) "Mutation and selection within the individual", Genetica 103:507-524. - Bergstrom & Pritchard (1998) "Germline bottlenecks and the evolutionary maintenance of mitochondrial genomes", Genetics 149:2135-2146.
49. Proposed by Parsons et al. (1997) "A high observed substitution rate in the human mitochondrial DNA control region", Nature Genet. 15:363-368. Discussions: Loewe & Scherer (1997) "Mitochondrial Eve: the plot thickens", Trends Ecol. Evol. 12:422-423. - Siguroardottir et al. (2000) "The mutation rate in the human mtDNA control region", Am. J. Hum. Genet. May 66:1599-1609 - Donnelly (1991) "Comment on the growth and stabilization of populations", Stat. Sci. 6:277-279. - See also Chapter 19 of this work.

3.3.1 Mutational hot spots

Hot spots are no complete solution

The most obvious solution is that a few sites with high mutation rates flip-flop back and forth every other generation[50]. Thus, a high number of inter-generational mutations are generated, but the resulting long-term substitution rate is low, as many mutations hide behind multiple hits. Investigations of this have shown that some sites clearly show this behaviour[51]. A number of attempts have been made to find the hot-spots of the control region from polymorphism data in populations[52]. However, leaving aside some clear cases, considerable disagreement can be found among studies on the exact identity of the hot-spots. This may be due to an imprecise definition. Compared to those sites in the control region that have never been observed to change, many sites will be hot-spots, whereas only few sites actually really flip-flop back and forth every few generations. Not many studies have tried to reconcile pedigree rates and phylogenetic rates via rate heterogeneity[53]. The overall emerging picture is that hot-spots do explain a part of the mutation rate paradox, but they do not seem to explain it completely.

3.3.2 Selective removal

Selection can be a dangerous solution

Another hypothesis is that many of those excess mutations are not neutral, but actually slightly deleterious. Thus, selection will remove them in the long term[54]. A high slightly deleterious mutation (SDM) rate is supported

50. Pääbo (1996) "Mutational hot spots in the mitochondrial microcosm", Am. J. Hum. Genet. 59:493-496.
51. Parsons et al. (1997) ibid. - Koehler et al. (1991) "Replacement of bovine mitochondrial DNA by a sequence variant within one generation", Genetics 129:247-256. - Howell & Smejkal (2000) "Persitent heteroplasmy of a mutation in the human mtDNA control region: Hypermutation as an apparent consequence of simple-repeate expansion/contraction", Am. J. Hum. Genet. 66:1589-1598.
52. Meyer et al. (1999) "Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA", Genetics 152:1103-1110. - Misawa & Tajima (1997) "Estimation of the amount of DNA polymorphism when the neutral mutation rate varies among sites", Genetics. vol.147 pp.1959-1964. - Wakeley (1996) "The excess of transitions among nucleotide substitutions: New methods of estimating transition bias underscore its significance", Trends Ecol. Evol. 11:158-163. - Aris-Brosou & Excoffier (1996) "The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism." Mol. Biol. Evol. 13:494-504. - Wakeley (1993) "Substitution Rate Variation Among Sites in Hypervariable Region 1 of Human Mitochondrial DNA", J. Mol. Evol. 37:613-623. -
53. Meyer et al. (1999) "Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA", Genetics 152:1103-1110. - Schneider & Excoffier (1999) "Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA", Genetics 152:1079-1089.
54. Parsons et al. (1997) "A high observed substitution rate in the human mitochondrial DNA control region", Nature Genet. 15:363-368.

by the observation that the fraction of non-synonymous substitutions is much higher within species than between[55].

However, selection is no simple solution. While it is clear that heavily deleterious mutations are selectively removed, the people who carried the mutations in the pedigree studies appeared to be "normal" (whatever that means). If mutation effects are small, two problems can occur:

Muller's ratchet can be a problem

- If effects are in a certain range, they can drive Muller's ratchet that might drive the population to extinction. It is currently not known how a population could survive a pedigree-rate-type mutational pressure[56] and detailed quantifications are not available up to now.
- If effects are very small, they do drive Muller's ratchet almost deterministically, but their combined effect is so small that they could not drive the population to extinction in 4 billion years. However, as they have nearly no effect, such mutations are hardly removed by selection in the long term[57].

Thus, only a very small range of subtle selection coefficients would actually solve the mutation rate paradox. This work quantifies this range of selection coefficients for the first time. However, it seems improbable that mutation effects are always of the right size to solve the problem.

3.3.3 Perspectives

Complex multi-factor solutions

As mutational hot-spots and selective removal do not appear to solve the problem on their own, we are left looking for other solutions. Perhaps the generation time effect[58] or linked loci under selection[59] or complicated multi-level population genetics[48] or some yet little known peculiarities of drift[49] may contribute towards a solution that comprises a little bit of everything. Until a solution is found, we can think about the implications of the

55. Hasegawa et al. (1998) "Preponderance of slightly deleterious polymorphism in mitochondrial DNA: Nonsynonymous/synonymous rate ratio is much higher within species than between species", Mol. Biol. Evol. 15:1499-1505.
56. Howell et al. (1996) "How rapidly does the human mitochondrial genome evolve?" Am. J. Hum. Genet. 59:501-509. - Gabriel et al. (1993) "Mullers Ratchet and mutational meltdowns", Evolution 47:1744-1757.
57. Kondrashov (1995) "Contamination of the Genome by Very Slightly Deleterious Mutations: Why Have We Not Died 100 Times Over?" J. theor. Biol. 175:583-594.
58. Martin & Palumbi (1993) "Body size, metabolic rate, generation time, and the molecular clock", Proc Natl Acad Sci U S A 90:4087-4091.
59. Although Ohashi & Tokunaga (2000) "Sojourn times and substitution rate at overdominant and linked neutral loci", Genetics Jun 155:921-927 did not find a significant influence of overdominant selection on the nucleotide substitution rate at a linked neutral locus, other studies of hitch-hiking events or background selection may find such an influence.

mutation rate paradox for inference of recombination in mitochondria by homoplasy[60], traditional phylogenetic studies[61] and the mystery of humanity's missing mutations[62]. However, one conclusion can be drawn already. As simple as mitochondrial DNA is in the laboratory, as complex it appears to be in thorough theoretical analyses.

Building a database of mutation rate paradoxes

Therefore authors of phylogenetic studies can facilitate further investigation of these issues, if they include all data necessary to make the various currently conflicting conclusions. Researchers are invited to send such analyses by email to

MutRatParDB@evolutionary-research.net

to help build a database of instances, where the mutation rate paradox occurs. Thus, evolutionary studies might contribute to better understanding of a genetic system of great importance to human health[63].

60. Eyre-Walker (2000) "Do mitochondria recombine in humans?" Philos Trans R Soc Lond B Biol Sci 355:1573-1580. - Denver et al. (2000) "High direct estimate of the mutation rate in the mitochondrial genome of Caenorhabditis elegans", Science 289:2342-2344.
61. Schneider & Excoffier (1999) "Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA", Genetics 152:1079-1089.
62. Gibbons (1995) "Human evolution: The mystery of humanity's missing mutations", Science 267:35-36.
63. Wallace (1999) "Mitochondrial diseases in man and mouse", Science 283:1482-1488. - Wallace et al. (1999) "Mitochondrial DNA variation in human evolution and disease", Gene 238:211-230.

4 Experimental evolution in microorganisms

Microorganisms are excellent for studying evolution, as they evolve extremely fast, even on ecological timescales. Their short generation times allow experimentation with processes that are impossible to observe in other species. This approach has been extensively used to investigate adaptive evolution and deleterious mutation rates in serial transfer and chemostat experiments. However, despite high interest in stationary phase mutations, the deleterious mutation rates associated with prolonged starvation have never been observed.

The advantages of short generation times and large population sizes in small volumes, combined with the possibility of freezing, make microorganisms excellent systems for the study of evolution. While the former allows investigation of rather weak selective processes, the latter allows us to construct an artificial palaeontological record in the freezer[64] and the usually assumed 250 000 generations that separate us from our common ancestor with chimps may shrink to less than 30 years in a large scale bacterial evolution experiment. Thus, such studies do contribute much to our understanding of evolution[65].

4.1 Adaptive evolution in serial transfers

The study of adaptive evolution in microorganisms is a very active field[66]. One of the main conclusions that comes out of this work is that bacteria will adapt relatively fast to a new environment by generation of advantageous mutations[67], if population size is large enough[68]. However, once they feel comfortable, the rate at which they come up with new adaptive mutations slows down considerably. In the longest bacterial evolution experiment up to date (20 000 generations[69]), all major adaptations in cell size and growth rate appeared to have been completed after about 2000 - 5000 generations[70]. Thus, the sustained long-term rate of innovation in terms of uncon-

64. Lenski & Travisano (1994) "Dynamics of adaptation and diversification: A 10,000-generation experiment with bacterial populations", Proc. Natl. Acad. Sci. U.S.A. 91:6808-6814.
65. Travisano (2001) "Evolution: Towards a genetical theory of adaptation", Curr Biol 11:R440-442.

66. Bennett et al. (1990) "Rapid evolution in response to high-temperature selection", Nature 346:79-81. *** Bennett et al. (1992) "Evolutionary adaptation to temperature: I. Fitness responses of Escherichia coli to changes in its thermal environment", Evolution 46:16-30. *** Bennett & Lenski (1993) "Evolutionary adaptation to temperature: II. Thermal niches of experimental lines of Escherichia coli", Evolution 47:1-12. *** Bennett & Lenski (1997) "Evolutionary adaptation to temperature: VI. Phenotypic acclimation and its evolution in Escherichia coli", Evolution 51:36-44. *** Bronikowski et al. (2001) "Evolutionary adaptation to temperature. VII. Effects of temperature on growth rate in natural isolates of Escherichia coli and Salmonella enterica from different thermal environments", Evolution Jan 55:33-40. *** Cooper & Lenski (2000) "The population genetics of ecological specialization in evolving Escherichia coli populations", Nature 407:736-739. *** Cooper et al. (2001) "Evolution of thermal dependence of growth rate of Escherichia coli populations during 20,000 generations in a constant environment", Evolution 55:889-896. *** de Visser et al. (1997) "An experimental test for synergistic epistasis and its application in Chlamydomonas", Genetics 145:815-819. *** Elena & Lenski (1997) "Long-term experimental evolution in Escherichia coli .7. Mechanisms maintaining genetic variability within populations", Evolution. 51:1058-1067. *** Elena et al. (1998) "Distribution of fitness effects caused by random insertion mutations in Escherichia coli", Genetica 103:349-358. *** Hall (1994) "On alternatives to selection-induced mutation in the Bgl operon of Escherichia coli", Mol. Biol. Evol. 11:159-168. *** Lenski (1988) "Experimental studies of pleiotropy and epistasis in Escherichia coli: I. Variation in competitive fitness among mutants resistant to virus T4", Evolution 42:425-432. *** Lenski (1988) "Experimental studies of pleiotropy and epistasis in Escherichia coli: II. compensation for maladaptive effects associated with resistance to virus T4", Evolution 42:433-440. *** Lenski (1988) "Dynamics of interactions between bacteria and virulent bacteriophage", Adv. Microb. Ecol. 10:1-44. *** Lenski et al. (1991) "Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2000 generations", Am. Nat. 138:1315-1341. *** Leroi et al. (1994) "Temperature acclimation and competitive fitness: An experimental test of the beneficial acclimation assumption", Proc. Natl. Acad. Sci. U.S.A. 91:1917-1921. *** Leroi et al. (1994) "Evolutionary adaptation to temperature. III. Adaptation of Escherichia coli to a temporally varying environment", Evolution 48:1222-1229. *** Mongold et al. (1996) "Evolutionary adaptation to temperature: IV. Adaptation of Escherichia coli at a niche boundary", Evolution 50:35-43. *** Mongold et al. (1999) "Evolutionary adaptation to temperature. VII. Extension of the upper thermal limit of Escherichia coli", Evolution 53:386-394. *** Moxon et al. (1994) "Adaptive evolution of highly mutable loci in pathogenic bacteria", Curr. Biol. 4:24-33. *** Nakatsu et al. (1998) "Parallel and divergent genotypic evolution in experimental populations of Ralstonia sp." J. Bacteriol. 180:4325-4331. *** Papadopoulos et al. (1999) "Genomic evolution during a 10,000-generation experiment with bacteria", Proc. Natl. Acad. Sci. U.S.A. 96:3807-3812. *** Rainey & Travisano (1998) "Adaptive radiation in a heteregeneous environment", Nature 394:69-72. *** Riley et al. (2001) "Rapid phenotypic change and diversification of a soil bacterium during 1000 generations of experimental evolution", Microbiology 147:995-1006. *** Rozen & Lenski (2000) "Long-term experimental evolution in Escherichia coli. VIII. Dynamics of a balanced polymorphism", Am. Nat. Jan 155:24-35. *** Schneider et al. (2000) "Long-term experimental evolution in Escherichia coli. IX. Characterization of insertion sequence-mediated mutations and rearrangements", Genetics Oct 156:477-488. *** Souza et al. (1997) "Long term experimental evolution in Escherichia coli. 5. Effects of recombination with immigrant genotypes on the rate of bacterial evolution", J. evol. Biol. 10:743-769. *** Travisano et al. (1995) "Experimental tests of the roles of adaptation, chance, and history in evolution", Science 267:87-90. *** Travisano et al. (1995) "Long-term experimental evolution in Escherichia coli. III. Variation among replicate populations in correlated responses to novel environments", Evolution 49:189-200. *** Travisano (1997) "Long-term experimental evolution in Escherichia coli. VI. Environmental constraints on adaptation and divergence", Genetics 146:471-479. *** Travisano & Rainey (2000) "Studies of adaptive radiation using model microbial systems", Am. Nat. Oct 156:S35-S44. *** Turner et al. (1998) "Tradeoff between horizontal and vertical modes of transmission in bacterial plasmids", Evolution Apr 52:315-329. *** Velicer et al. (1998) "Loss of social behaviors by Myxococcus xanthus during evolution in an unstructured habitat", Proc. Natl. Acad. Sci. U.S.A. 95:12376-12380. *** Zeyl & Bell (1997) "The advantage of sex in evolving yeast populations", Nature 388:465-468. *** Zeyl & DeVisser (2001) "Estimates of the rate and distribution of fitness effects of spontaneous mutation in Saccharomyces cerevisiae", Genetics Jan 157:53-61. *** Zeyl et al. (2001) "Mutational meltdown in laboratory yeast populations", Evolution May 55:909-917. *** This is to name just a few.

ditionally advantageous mutations is still one of the unsolved problems of the field.

For a while it seemed that a phenomenon called adaptive mutation[71] would seriously disturb the Darwinian view of evolution that mutations are random and selection leads to adaptation. But now it is understood that transient and heritable mutator phenotypes explain the phenomenon[72]. They turn up mutation rates in stressful situations or have generally elevated levels of mutation. Thus they are more likely to produce the next adaptive mutation and will generate the impression that adaptive mutations are always there when they are needed. Understanding evolutionary biology of mutators is important, as they occur at high frequencies in nature and especially in pathogens[73].

However, mutators have not only advantages[74]. If a population tries to fix too many adaptive mutations at once, they interfere with each other's fixation[75] and the positive effects of a generally high mutation rate can be easily offset by the flood of slightly deleterious mutations it brings along.

---

67. Imhof & Schlotterer (2001) "Fitness effects of advantageous mutations in evolving Escherichia coli populations", Proc. Natl. Acad. Sci. USA 98:1113-1117.
68. Rainey (1999) "Evolutionary genetics: The economics of mutation", Curr. Biol. 9:R371-R373.
69. Cooper et al. (2001) "Evolution of thermal dependence of growth rate of Escherichia coli populations during 20,000 generations in a constant environment", Evolution May 55:889-896. - Cooper & Lenski (2000) "The population genetics of ecological specialization in evolving Escherichia coli populations", Nature 407:736-739.
70. See Lenski & Travisano (1994) "Dynamics of adaptation and diversification: A 10,000-generation experiment with bacterial populations", Proc. Natl. Acad. Sci. USA 91:6808-6814. and Elena et al. (1996) "Punctuated evolution caused by selection of rare beneficial mutations", Science 272:1802-1804.
71. Harris et al. (1994) "Recombination in adaptive mutation", Science 264:258-260. - Rosenberg et al. (1994) "Adaptive mutation by deletions in small mononucleotide repeats", Science 265:405-407. - Cairns & Foster (1991) "Adaptive reversion of a frameshift mutation in Escherichia coli", Genetics 128:695-702.
72. Bull et al. (2000) "Evidence that stationary-phase hypermutation in the Escherichia coli chromosome is promoted by recombination", Genetics Apr 154:1427-1437. - Bull et al. (2000) "Response to John Cairns: The contribution of transiently hypermutable cells to mutation in stationary phase", Genetics Oct 156:925-926. - Rosenberg et al. (1995) "Molecular handles on adaptive mutation", Mol. Microbiol. 18:185-189. - Rosenberg (1997) "Mutation for survival", Curr. Opin. Genet. Dev. 7:829-834. - Rosenberg et al. (1998) "Transient and heritable mutators in adaptive evolution in the lab and in nature", Genetics 148:1559-1566.
73. LeClerc et al. (1996) "High mutation frequencies among Escherichia coli and Salmonella pathogens", Science 274:1208-1211. - LeClerc & Cebula (1997) "Highly variable mutation rates in commensal and pathogenic Escherichia coli - Reply", Science 277:1834-1834. - Matic et al. (1997) "Highly variable mutation rates in commensal and pathogenic Escherichia coli", Science 277:1833-1834. - Moxon et al. (1994) "Adaptive evolution of highly mutable loci in pathogenic bacteria", Curr. Biol. 4:24-33. - Taddei et al. (1997) "Role of mutator alleles in adaptive evolution", Nature 387:700-702.
74. Giraud et al. (2001) "Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut", Science 291:2606-2608.

4.2 Deleterious mutation rate estimates

Although it is commonplace in biology that the overwhelming majority of mutations are deleterious, surprisingly little is known about the precise effects mutations usually exhibit[76]. Bacteria have been used to address this question too[77], but much more work will be needed before we come close to really understanding the distribution of mutational effects even in such 'simple' organisms as bacteria. After all, the adaptive landscape[78] is one of the most complex places to navigate through.

The fact that adaptive evolution pervades even deleterious mutation rate measurements[79] further complicates disentanglement of mutational effects. A variety of different approaches has been taken to estimate deleterious mutation rate parameters: inferences from sequences in populations[80] have been used like serial transfer experiments[77] and chemostats[81]. Currently the best estimate of the total bacterial mutation rate is about 1/300 mutations per replication per genome[82]. The largest mutation accumulation experiment up to date might be used to infer that more than 6% of these mutations have slightly deleterious effects with an average of 1.2% or less on total growth rate. Assuming a distribution of mutational effects of *E. coli*

75. de Visser et al. (1999) "Diminishing returns from mutation supply rate in asexual populations", Science 283:404-406. - Gerrish & Lenski (1998) "The fate of competing beneficial mutations in an asexual population", Genetica 103:127-144.
76. Lynch et al. (1999) "Perspective: Spontaneous deleterious mutation", Evolution 53:645-663. Drake et al. (1998) "Rates of spontaneous mutation", Genetics 148:1667-1686.
77. Zeyl & DeVisser (2001) "Estimates of the rate and distribution of fitness effects of spontaneous mutation in Saccharomyces cerevisiae", Genetics Jan 157:53-61. - Andersson & Hughes (1996) "Muller's ratchet decreases fitness of a DNA-based microbe", Proc. Natl. Acad. Sci. U.S.A. 93:906-907. - Kibota & Lynch (1996) "Estimate of the genomic mutation rate deleterious to overall fitness in E. coli", Nature 381:694-696. - Elena & Lenski (1997) "Test of synergistic interactions among deleterious mutations in bacteria", Nature 390:395-398.
78. Gavrilets (1997) "Evolution and speciation on holey adaptive landscapes", Trends Ecol. Evol. 12:307-312. - Kauffman (1987) "Towards a general theory of adaptive walks on rugged landscapes", J. theor. Biol. 128:11-46. - Rose & Lauder, (eds, 1996) "Adaptation", San Diego, Academic Press. - Schuster (1996) "How does complexity arise in evolution", Complexity:22-30.
79. [Zeyl et al. (2001) "Mutational meltdown in laboratory yeast populations", Evolution 55:909-917.
80. Hartl et al. (1994) "Selection intensity for codon bias", Genetics 138:227-234.
81. Dean et al. (1988) "Fitness effects of amino acid replacements in the beta- galactosidase of Escherichia coli", Mol. Biol. Evol. 5:469-485. - Dean (1989) "Selection and neutrality in lactose operons of Escherichia coli", Genetics 123:441-454. - Dykhuizen & Hartl (1980) "Selective neutrality of 6PGD allozymes in E. coli and the effects of genetic background", Genetics 96:801-817. - Dykhuizen & Hartl (1983) "Selection in chemostats", Microbiol Rev 47:150-168. - Dykhuizen et al. (1984) "Selective neutrality of glucose-6-phosphate dehydrogenase allozymes in Escherichia coli", Mol Biol Evol 1:162-170. - Dykhuizen et al. (1987) "Metabolic flux and fitness", Genetics 115:25-32. - Dykhuizen (1990) "Experimental studies of natural selection in bacteria", Ann. Rev. Ecol. Syst. 21:373-398. - Hartl et al. (1985) "Limits of adaptation: The evolution of selective neutrality", Genetics 111:655-674.
82. Drake et al. (1998) "Rates of spontaneous mutation", Genetics 148:1667-1686.

that is similar to *Drosophila* would lead to estimates that deviate up to an order of magnitude from that (i.e. 60% of all mutations are slightly deleterious and have a mean effect of 0.1%)[83]. However, indirect estimates of selection coefficients in natural populations have led to upper limits for average selection coefficients that are several orders of magnitude lower (for non-synonymous mutations $2x10^{-8}$; for synonymous ones $7x10^{-9}$)[84]. The true values are currently unknown and probably vary from one extreme to the other. As evolution of bacteria has been shown to happen on ecological timescales[85], the distribution of mutational effects will probably not be understood without a solid ecological foundation that predicts when which genes will be needed to what degree.

4.3 Open problems

From the plethora of open questions in the field, this work addresses the following aspects:

- What is the deleterious mutation rate in the stationary phase?
- What are the implications of bacterial mutation rates from the standpoint of Muller's ratchet? Special emphasis is placed on endosymbiotic bacteria and potential contributions to the large uncultivable majority of bacteria found in nature.
- Molecular biological methods are often used to construct strains that are supposed to be neutral, but have properties useful for a given experimental setting. An example of evolutionary consequences of such mutations will be investigated.
- Freezing and thawing is a widely used method for conservation of bacterial strains. It is checked whether this process has subtle effects on the fitness of a bacterial population.
- The Batemann-Mukai technique is a widely used method for evaluating mutation accumulation experiments[86]. Limitations of particular experimental settings using this technique in bacteria are investigated.

83. See Kibota & Lynch (1996) ibid. The mutational parameters observed are Umin = 0.0002 and average s_{max}= 1.2%. Both parameters are linked so that increasing one decreases the other.
84. Hartl et al. (1994) "Selection intensity for codon bias", Genetics 138:227-234.
85. Rainey & Travisano (1998) "Adaptive radiation in a heteregeneous environment", Nature 394:69-72.
86. Lynch & Walsh (1998) "Genetics and analysis of quantitative traits", Sunderland, Massachusetts, Sinauer Associates.

5 Aim of this work

The aim of this work is to investigate the stability of asexual genomes. Towards this end several approaches are taken:

Simulations of Muller's ratchet

To investigate the effects of the high observed mutation rates on long-term viability of mitochondria, Muller's ratchet theory is employed. As analytical approximations allow only predictions for a limited range of the three parameters, population size, mutation rate and selection coefficient, individual-based simulations are performed to check and extend the analytical solutions. A slightly different population model from those in previous simulations is used to see that these changes have little effect on predictions.

Potential solutions for the genomic decay paradox

These simple models show that Muller's ratchet might indeed be a threat to mitochondria on a 20 million year timescale and thus complements the threat that comes from the high deleterious mutation rates in the nuclear genome[87]. The variety of biological processes promoted to solve this genomic decay paradox is discussed (see Part VII).

Evolution@home framework design

Reviewing these processes shows the enormous need for further simulation models and computing time to address these issues. Therefore, a framework is designed that uses the power of global computing to investiagate individual-based models of evolution. The first simulator in a long series of future simulators is implemented and used to start the first global computing system for evolutionary biology, *evolution@home*. Its results are used to investigate potential solutions of the mutation rate paradox and to observe details of Muller's ratchet that have not been reported up to now.

Potential solutions of the mutation rate paradox

Muller's ratchet in bacteria

The same set of tools that allows quantification of Muller's ratchet in mtDNA is applied to (non-recombining) bacteria, to investigate consequences of deleterious mutations in their genomes. As conclusions critically depend on estimates for deleterious genomic mutation rates, a system is developed for analysis of large numbers of growth curves. This system is used to quantify the deleterious mutation rate of a stationary phase population of *E. coli* and potential evolutionary effects of common microbiological practices. Implications for origin of the majority of uncultiviable bacteria are discussed.

Measurements of mutation rates

87. Eyre Walker & Keightley (1999) "High genomic deleterious mutation rates in hominids", Nature 397:344-347. - Kondrashov (1995) "Contamination of the Genome by Very Slightly Deleterious Mutations: Why Have We Not Died 100 Times Over?" J. theor. Biol. 175:583-594.

II. Muller's Ratchet Reviewed

If too many slightly deleterious mutations occur in a population, Muller's ratchet can lead to genomic decay and mutational meltdown, driving the population to extinction. After reviewing general circumstances where this might occur, current methods for predicting the rate of the ratchet are discussed. Empirical evidence of the operation of the ratchet is presented as well as a list of features that might help recognize its operation in a species.

6 Theory

Muller's ratchet was discovered when investigating advantages of recombination that might have led to the evolution of sex. After hiding in theoretical population genetics for a long time, the last decade saw the discovery of its relevance for conservation biology and numerous other problems.

A popular view of evolution depicts it as a process that fixes advantageous mutations, removes deleterious mutations and improves organisms' fitness. One of the astonishing results of population genetics is, however, that not all advantageous mutations eventually get fixed and not all deleterious ones get removed[1]. Instead, random genetic drift plays an important role, when mutation effects are small[2]. Equipped with such a world view, we will now review a phenomenon that describes the accumulation of slightly deleterious mutations: Muller's ratchet.

Its general mechanism was first described by H. J. Muller in 1964, when he wondered about the advantage of recombination[3]. J. Felsenstein gave it the name Muller's ratchet[4] and a key paper for its mathematical analysis followed in 1978 by J. Haigh[5]. But it was not before M. Lynch, W. Gabriel and others started investigating mutational meltdowns[6] in the 1990s, that Muller's ratchet became more widely known.

In spite of recent interest, no simple, comprehensive review is available for biologists in general[7]. Thus, here the most important concepts regarding Muller's ratchet will be explained along with their relevance.

1. See page 48 in Li (1997) "Molecular evolution", Sunderland, Sinauer Associates Incorporated.
2. Kimura (1995) "Limitations of Darwinian selection in a finite population", Proc. Natl. Acad. Sci. U.S.A. 92:2343-2344.
3. Muller (1964) "The relation of recombination to mutational advance", Mut. Res. 1:2-9.
4. Felsenstein (1974) "The evolutionary advantage of recombination", Genetics 78:737-756.
5. Haigh (1978) "The accumulation of deleterious genes in a population - Muller's Ratchet", Theor. Pop. Biol. 14:251-267.
6. Lynch & Gabriel (1990) "Mutation load and the survival of small populations", Evolution 44:1725-1737.
7. The simplest review is still Maynard Smith (1978) "Some consequences of sex and recombination - II. Muller's ratchet", pp. 33-36 in: Maynard Smith (ed) The Evolution of Sex, New York, Cambridge University Press. - One of the later reviews that have an explicit section on the ratchet is Baake & Gabriel (1999) "Biological evolution through mutation, selection and drift: An introductory review", Annual Reviews of Computational Physics 7:203-264. - Many papers on evolution of sex touch the ratchet, eg. Barton & Charlesworth (1998) "Why sex and recombination?" Science 281: 1986-1990. - Mutational meltdown is reviewed by Lynch et al. (1993) "The mutational meltdown in asexual populations", J. Hered. 84:339-344. The rest is scattered over many original papers.

6.1 Definition of Muller's ratchet

A visualization of the principle of Muller's ratchet can be found in Figure 4 and here is a one sentence definition of it:

> *Muller's ratchet is a stochastic process that leads to the repeated loss of the fittest individual lines of descent in a finite population due to accumulation of slightly deleterious mutations (SDMs).*

To understand the Ratchet, you must understand the distribution of mutations in the individuals of a population ...

Just imagine a natural population of asexual individuals. Then sort these individuals into classes by the number of SDMs they carry in their genome. Give each such class the label $\mathbf{N}_k$, where $\mathbf{N}$ stands for the number of individuals in this class and $\boldsymbol{k}$ for the number of mutations that defines the class (all mutations have equal effects here). Once you have done that, you will see a distribution with some individuals having few mutations, others having many. Often, the majority of individuals will carry an intermediate number of mutations. With this distribution, you can tell how many mutations are found in the best individuals of the population and how many best-class individuals exist in the population. Now remember how many mutations this best class has and then relabel all classes in such a way that the best class gets the label $\mathbf{N}_0$, the class with individuals that carry one mutation gets the label $\mathbf{N}_1$, those with 2 mutations $\mathbf{N}_2$ and so forth. Finally, remember all values of the distribution in this generation to compare them with the values of future generations.

... and the way it changes.

Now let us consider the processes that change this distribution from one generation to the next:

- **Selection.** Individuals with many SDMs will have a disadvantage given by the selection coefficient $\boldsymbol{s}$ per mutation, whereas the least loaded class $\mathbf{N}_0$ has a selective advantage compared to the average individual of the population. Thus, individuals from the best class will leave more offspring than individuals in the worst class. This effect can be so strong that it effectively stops the ratchet, as all incoming mutations are removed by purifying selection (such mutations are no longer SDMs, but might be considered as 'normal' deleterious mutations). However, when mutational effects become smaller, fewer and fewer mutations are removed up to the point where selection coefficients are smaller than the reciprocal of effective population size $\mathbf{N}_e$ and drift dominates selection.

  Thus, selection increases the probability that the best class parents produce enough offspring to keep the size of the best class.

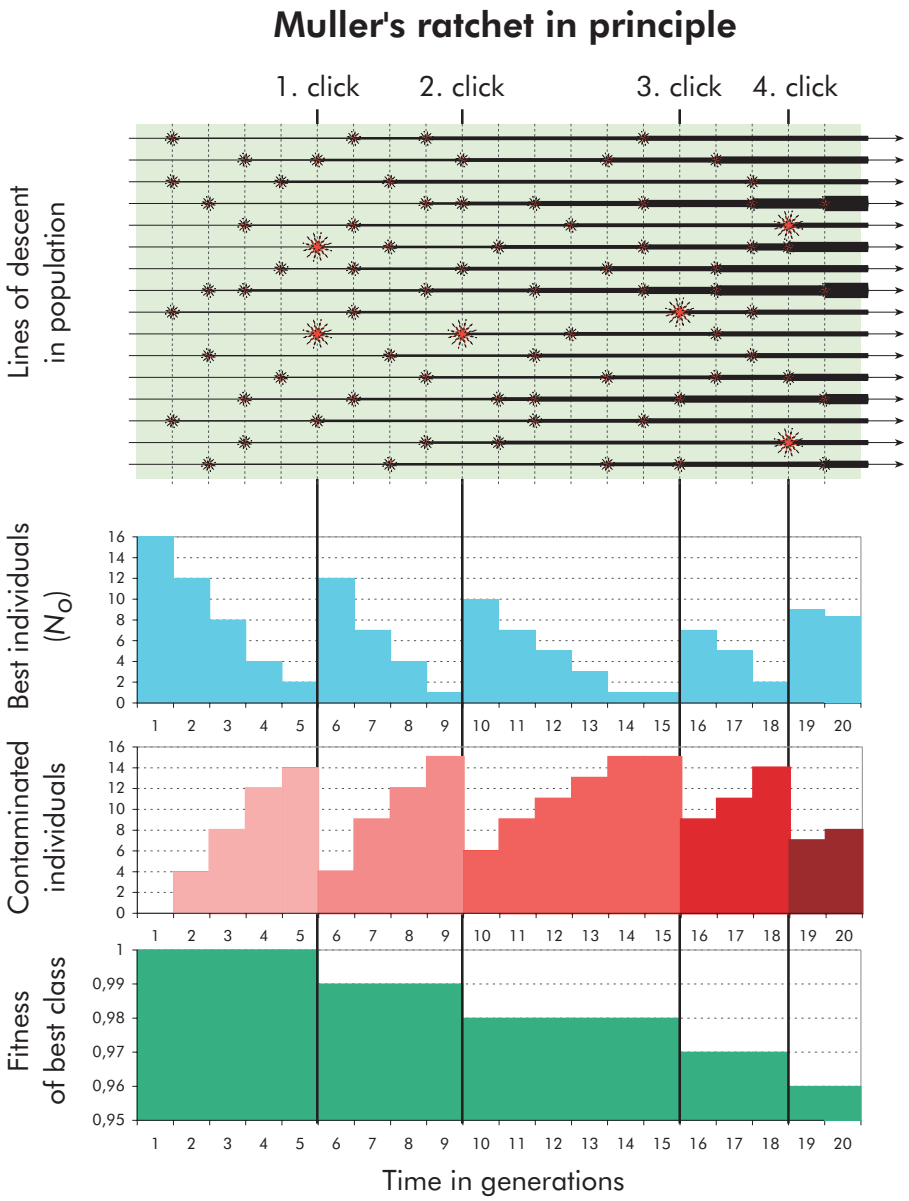


Figure 4 Visualization of the principle of Muller's ratchet.
The upper part contains an imaginary population whose lines of descent accumulate slightly deleterious mutations (SDMs, marked by small stars). Due to finite population size, eventually the best lines catch SDMs, too (larger stars). Once they have done so, the ratchet clicks, as the formerly second best lines become the best lines still available (least mutation-loaded). Over time, the best class available gets worse and worse, as contamination increases and fitness of the genome decays. Thickness of lines indicates the number of mutations they carry. For simplicity the graph omits that selection and drift change frequency of lines, too. Ratchet parameters in this example are $U = 0.25$, s=-0.01, $N_e = 16$.

- **Mutation pressure.** Occasional genome copy errors give rise to new SDMs with the deleterious genomic mutation rate U per generation. If this happens, parents are replaced by offspring in a class with more mutations.
- **Drift.** As individuals cannot be divided, slight changes in fitness of parents cannot lead to say 1.04 individuals: Either 0 or 1 or 2 or ... offspring are produced, largely depending on chance. Selection only influences general expectations. It does play an important role, when population size is large enough that a few (whole) individuals difference would result in the whole population (ie. $N_e s > 1$). Thus, drift can remove any lines of descent from the population, even the best. It is all a question of probability.

  Drift can remove the best class from the population, if it is small enough and it will do so deterministically, if its expectation in mutation-selection equilibrium is below 1.
- **Unidirectionality.** In simple ratchet models mutations are unconditionally deleterious and hit somewhere on an infinite genome. Thus, neither advantageous nor back-mutations can occur. This is an often used approximation that holds at least for a period of time after the ratchet has started to run, as all new mutations are likely to hit a new site and back-mutation rate can be neglected[8].

  Thus, in simple ratchet models as those considered here (without recombination or back-mutation), there is no process that can restore a once lost best-class.

It is hard to build one comprehensive model of all this

Analytical models of Muller's ratchet critically depend on a precise description of the changes of the distribution of mutations in individuals of a population, something extremely hard to obtain[9]. For nice examples of graphs of the evolution of this distribution over time, see the work of Gessler[9]. Before we continue, some symbols important in ratchet theory have to be introduced properly (Table 6). As mentioned above, the ratchet clicks,

8. See Maynard Smith (1978) "Some consequences of sex and recombination - II. Muller's ratchet", pp. 33-36 in: Maynard Smith (ed) The Evolution of Sex, New York, Cambridge University Press. However, a careful investigation suggests that if the ratchet runs for a long time, then mutations are likely to hit sites twice and might change the context of previous mutations. Resulting mutation effects are the topic of potentially very complicated theories of epistatic interactions. Moreover, as discussed in Chapter 20 the probability of back-mutations and compensatory mutations on a molecular level might increase significantly over long period of time.
9. See p. 247 in Baake & Gabriel (1999) "Biological evolution through mutation, selection and drift: An introductory review", Annual Reviews of Computational Physics 7:203-264.
   For a success story and nice pictures of examples for such a distribution, see Gessler (1995) "The constraints of finite size in asexual populations and the rate of the ratchet", Genet. Res. 66:241-253.

whenever the best class of individuals in a population no longer reproduces due to chance. This view of the ratchet goes back to HAIGH[10], who used a result derived by KIMURA & MARUYAMA[11] to describe the distribution of individuals in the various mutation classes N_k under mutation selection equilibrium

$$N_k = N_e \cdot e^{\left(\frac{U}{s}\right)} \cdot \left(\frac{U}{-s}\right)^k / k! \tag{1}$$

where s has to be a negative selection coefficient, as only deleterious mutations are considered[12]. From this formula, one can easily derive the well known simple formula that gives the expectation of the size of the best class under mutation-selection equilibrium[10]:

$$N_0 = N_e \cdot e^{\left(\frac{U}{s}\right)} \tag{2}$$

This formula is widely used. Recently, BAAKE & GABRIEL proposed an alternative that might be more realistic in certain cases[13]:

$$N_0 = N_e \cdot \exp\left(\frac{U(1+s)}{s}\right) \tag{3}$$

where again s has to be negative, as only deleterious mutations are considered. It should be kept in mind, that such formulae compute only a static mean under equilibrium, where as the true N_0 fluctuates strongly (see Figure 5 for a realistic example). These dynamic fluctuations need to be understood to predict the rate at which the ratchet clicks. This rate is needed to predict extinction times for species that might be endangered by the ratchet.

10. Haigh (1978) "The accumulation of deleterious genes in a population - Muller's Ratchet", Theor. Pop. Biol. 14:251-267.
11. Kimura & Maruyama (1966) "The mutational load with epistatic gene interactions in fitness", Genetics 54:1337-1351.
12. In this work selection coefficients of deleterious mutations are negative and those of advantageous mutations positive, if not stated otherwise. Typical papers on Muller's ratchet often omit the negative sign, as they consider slightly deleterious mutations only.
13. Baake & Gabriel (1999) "Biological evolution through mutation, selection and drift: An introductory review", Annual Reviews of Computational Physics 7:203-264.

Table 6 Some important definitions for Muller's ratchet theory.

Symbol	Meaning
N_e	Effective population size over time period of interest (=N for simple models).
U	Deleterious genomic mutation rate per generation, if not stated otherwise.
s	Constant selection coefficient (equal effects for all mutations), if not stated otherwise. In this work, advantageous mutations have $s>0$, deleterious ones $s<0$, if not stated otherwise (many ratchet papers use $s>0$ for SDMs).
w	Wrightian fitness = average number of offspring per individual, usually multiplicative, ie. $w=(1-s)^k$, assuming mutational effects are independent.
R_{max}	Maximal reproductive capacity, ie. maximal offspring number per generation in an empty habitat full of nutrients and without enemies.
V_{max}	Maximal viability (% offspring that can survive to maturity).
N	Current census population size, usually larger than N_e in reality.
N_k	Number of individuals in mutational class k.
k	Mutational class = number of mutations in an individual, assumes equal effects.
N_0	Size of the best (= least contaminated) class of individuals in a population.
p_0	Frequency of the best class in the population (N_0/N_e)
$click$	Advance of the ratchet, where a best class goes extinct and all k are redefined with the second best class as new best.
P_{cl}	Probability that the ratchet would click in a generation.
T_{cl}	Time between two clicks of the ratchet in general, ie. clicktime (often implicit T_{cleff}).
T_{clabs}	Absolute clicktime, does not depend on clicksize Cl_s.
T_{cleff}	Effective clicktime, = T_{clabs}/Cl_s.
Cl_s	Clicksize, ie. how many mutations are fixed with a click.
Cl_r	Click rate of the ratchet = $1/T_{cl}$.
Cl_{mm}	Expectation for the number of clicks needed to start mutational meltdown.
Cl_{hfa}	High fecundity advantage factor, describes how many more clicks are needed to bring a high fecundity species to the brink of meltdown than a low fecundity species.
T_{mm}	Time when mutational meltdown starts.
T_{ex}	Time when the ratchet has driven a population to extinction.
T_{gen}	Duration of a generation, needed to scale generation time of the ratchet to real time.
SDM	**S**lightly **D**eleterious **M**utation (selection too weak to be completely purifying)
VSDM	**V**ery **S**lightly **D**eleterious **M**utation, can no longer be selected against, even with recombination, as $s < 1/4N_e$[a]
SAM	**S**lightly **A**dvantageous **M**utation (weak selection)
VSAM	**V**ery **S**lightly **A**dvantageous **M**utation, can not be selected as it brings no advantage that selection could notice, $s < 2/N_e$[b]

a. Approximate limit by Kondrashov (1995) "Contamination of the Genome by Very Slightly Deleterious Mutations: Why Have We Not Died 100 Times Over?" J. theor. Biol. 175:583-594.

b. Approximate limit by Kimura (1995) "Limitations of Darwinian selection in a finite population", Proc. Natl. Acad. Sci. U.S.A. 92:2343-2344.

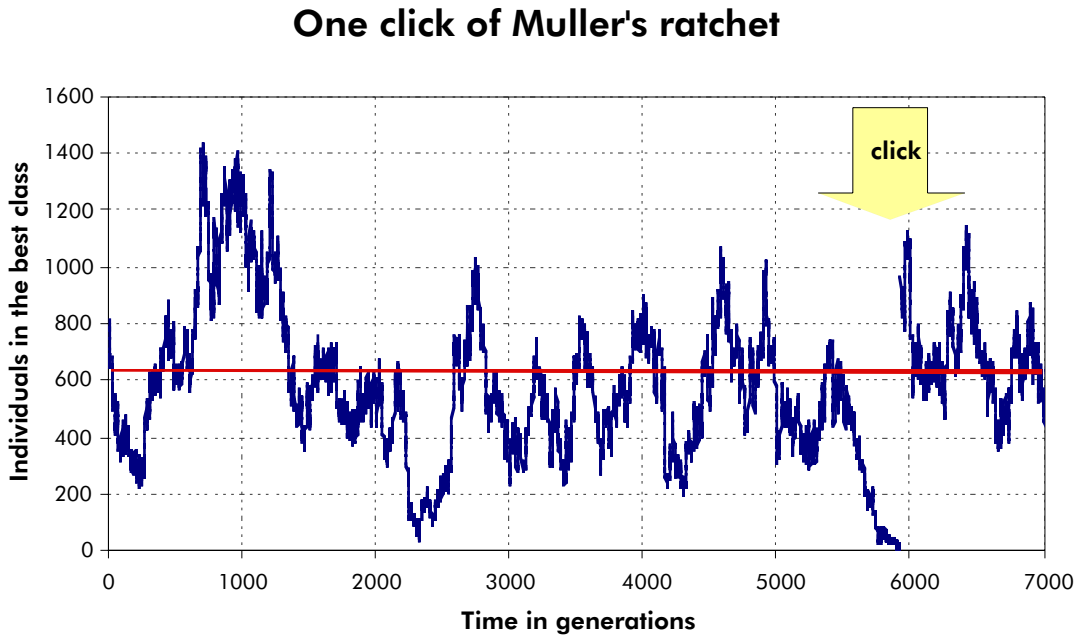


Figure 5 Detailed observation of one click of Muller's ratchet.
Fluctuations in the size of the best class can be enormous, especially if the ratchet clicks slowly. This simulation was performed using the professional release of Simulator005 of the evolution@home global computing system (see http://www.evolutionary-research.net).
The parameters used for this simulation are $U = 0.05$, s=-0.0075, $N_e = 500\,000$. Analytic theory predicts $N_0 = 636$ (see red line) and T_{click} between 4800 and 5900 generations (mean of observation in other simulations = 3912), see Stephan & Kim (2002) "Recent applications of diffusion theory to population genetics", pp. 72-93 in: Slatkin & Veuille (eds) Modern developments in theoretical population genetics, Oxford, Oxford University Press.

Error thresholds and Muller's ratchet

Muller's ratchet is not the only theory that can predict extinctions from a combination of N_e, U and s. The other theory that predicts an upper limit for the mutation rate, beyond which selection can no longer control accumulation of mutation is the theory of error thresholds[14]. It was developed by Eigen & Schuster[15] and has been used to predict maximal length of self-replicating molecules and viral genomes[16]. Comparing both theories[17]

14. Baake (1995) "Diploid models on sequence space", Journal of Biological Systems 3:343-349. - Baake & Wiehe (1997) "Bifurcations in haploid and diploid sequence space models", J. math. Biol. 35:321-343. - Bonhoeffer & Stadler (1993) "Error thresholds on correlated fitness landscapes", J. theor. Biol. 164:359-372. - Higgs (1994) "Error thresholds and stationary mutant distributions in multi-locus diploid genetics models", Genet. Res. 63:63-78. - Swetina & Schuster (1982) "Self-replication with errors. A model for polynucleotide replication", Biophysical Chemistry 16:329-345.

shows that, although their mathematical models are similar, Muller's ratchet works only in finite populations, as it critically depends on random drift. On the contrary, error thresholds are deterministic as they assume a lower limit of fitness and predict extinction as soon as mutation pressure would drive the best class beyond that value[18]. See also discussion of fitness definitions below.

6.2 Definition of important details

Before we can continue to review methods for predicting rate and effects of the ratchet, a number of details are important.

Effective population size

The concept of effective population size was introduced[19] by SEWALL WRIGHT to quantify that fraction of a population that contributes genes to the next generation. It will help to compare different, more realistic, population structures in nature to a standard that later became known as the WRIGHT-FISHER standard population (diploid, recombining, random mating, constant size, no further substructure). A large body of theory was developed to compute effective population size for various natural settings[20]. Comparisons to the standard can be made from different viewpoints that lead to different numbers for $\mathbf{N_e}$. For example, investigating the coalescent leads to the conclusion that strong substructure with little migration increases $\mathbf{N_e}$, because the time to the most recent common ancestor of alleles from different sub-populations increases with decreasing probability of migration[21] – compared to a random mating population. However, when investigating Muller's ratchet the contrary can be true: Selection's efficiency

15. Eigen (1971) "Selforganization of matter and the evolution of biological macromolecules", Naturwissenschaften 58:465-523. - Eigen & Schuster (1977) "The Hypercycle - A principle of natural self-organization. Part A: Emergence of the Hypercycle", Naturwissenschaften 64:541-565. - Eigen & Schuster (1978) "The Hypercycle - A principle of natural self-organization. Part C: The realistic Hypercycle", Naturwissenschaften 65:341-369. - Eigen & Schuster (1978) "The Hypercycle - A principle of natural self-organization. Part B: The abstract Hypercycle", Naturwissenschaften 65:7-41.
16. Eigen (1993) "The origin of genetic information: viruses as models", Gene 135:37-47. - Nichol (1996) "RNA Viruses: Life on the edge of catastrophe", Nature 384:218-219. - Domingo (2000) "Viruses at the edge of adaptation", Virology 270:251-253. - Domingo et al., (eds, 1999) "Origin and evolution of viruses", San Diego, Academic Press. - Boerlijst et al. (1996) "Viral quasi-species and recombination", Proc. R. Soc. Lond. B 263:1577-1584.
17. Nowak & Schuster (1989) "Error thresholds of replication in finite populations mutation frequencies and the onset of Muller's ratchet", J. theor. Biol. 137:375-396. - Wagner & Krall (1993) "What is the difference between models of error thresholds and Muller's ratchet?" J. math. Biol. 32:33-44. - Baake & Gabriel (1999) "Biological evolution through mutation, selection and drift: An introductory review", Annual Reviews of Computational Physics 7:203-264.
18. See Wagner & Krall (1993), ibid.
19. Wright (1931) "Evolution in mendelian populations", Genetics 16:97-159.

can be decreased by geographic substructure and thus $\mathbf{N_e}$ can be smaller [22]. However, on some occasions population structure can increase the power of selection, too[23]. For our purpose, it is enough to know that $\mathbf{N_e}$ is usually *much* smaller than the census population size. More details become only interesting when an actual number is needed for investigating the ratchet in a particular species.

Population model

A population model is needed to keep population size in check. Thus the state of the next (discrete) generation is usually[24] computed as follows:

1. Pick an individual from the parental generation at random with replacement.
2. Clone offspring, add mutations, and add the individual to the next generation, if it passes viability selection. To do the latter just pick a random number and compare it to the individual's fitness (0 ... 1). If the fitness is greater, the individual survives. Additional genotype independent selection may also occur.
3. Repeat this process until the offspring generation has reached the predetermined size (that might be constant or is determined by some external function[25]).

   On the fly, the distribution of individuals' fitness in the population can be determined which allows observations of the ratchet.

20. For a simple review see Kimura (1983) "The neutral theory of molecular evolution", Cambridge, Cambridge University Press. - Predicting Ne is far from easy and a current topic of research. More details on this can be found in the recent literature and in Caballero (1994) "Developments in the prediction of effective population size", Heredity 73:657-679. - Caballero (1995) "On the effective size of populations with separate sexes, with particular reference to sex-linked genes", Genetics 139:1007-1011. - Chesser et al. (1993) "Effective sizes for subdivided populations", Genetics 135:1221-1232. - Ewens (1989) "The Effective Population Sizes in the Presence of Catastrophes", pp. 9-25 in: Feldman (ed) Mathematical evolutionary theory, Princeton, New Jersey, Princeton University Press. - Kawata (1995) "Effective population size in continuously distributed population", Evolution 49:1046-1054. - Nunney (1999) "The effective size of a hierarchically structured population", Evolution 53:1-10. - Santiago & Caballero (1995) "Effective size of populations under selection", Genetics 139:1013-1030. - Sugg & Chesser (1994) "Effective population sizes with multiple paternity", Genetics 137:1147-1155. - Vucetich et al. (1997) "Fluctuating population size and the ratio of effective to census population size", Evolution 51:2017-2021.
21. Nei & Takahata (1993) "Effective population size, genetic diversity, and coalescence time in subdivided populations", J. Mol. Evol. 37:240-244.
22. Whitlock & Barton (1997) "The effective size of a subdivided population", Genetics 146:427-441.
23. Whitlock (2002) "Selection, load and inbreeding depression in a large metapopulation", Genetics 160:1191-1202.
24. p.242 in Gessler (1995) "The constraints of finite size in asexual populations and the rate of the ratchet", Genet. Res. 66:241-253. - p. 1745 in Gabriel et al. (1993) "Mullers Ratchet and mutational meltdowns", Evolution 47:1744-1757. - or p. 67 in Charlesworth & Charlesworth (1997) "Rapid fixation of deleterious alleles can be caused by Muller's Ratchet", Genet. Res. 70:63-73.
25. like in Gabriel et al. (1993) "Mullers Ratchet and mutational meltdowns", Evolution 47:1744-1757. or in Lynch et al. (1995) "Mutation accumulation and the extinction of small populations", Am. Nat. 146:489-518.

Together with a sequence of events in the life history of the individuals[26], this defines the basic set-up needed to simulate Muller's ratchet. However, this is not the only population model possible and if more realistic details are allowed, interesting results can be found[27]. It will be interesting to see how details of population models influence operation of the ratchet.

Definition of fitness

There is much truth in the popular understanding that the fittest is the one who survives best or reproduces fastest in a given environment. However, it can be complicated to link fitness to measurable life-history characteristics like number of offspring, generation length, reproductive rate or survival until maturity[28]. Currently, the only basic advice that can be given for the use of fitness measurements is to only use definitions that fit the purpose[29]. This means that ratchet investigations have to employ some definition of fitness that is not rescaled every generation and has a lower limit. If not, it will generally be impossible to observe extinctions. While important research on the ratchet has been done this way[30], one must understand potential limitations due to this issue (see discussion of mutational meltdown below). A typical definition used is Wrightian fitness, the average number of progeny per individual[31] that reach maturity in a generation.

Fitness models and epistasis

Two of the most complicated topics in biology are how genotypes are mapped to phenotypes[32] and how different mutations interact epistatically[33] to produce their combined effect on an individual (see also discussion of artificial life and evolutionary computation under "Types of evolutionary simulations" on page 23 in appendix). However, no investigation of the ratchet is possible without assuming at least some very simple models here.

26. Sequence of events per generation can be: Reproduction>Mutation>Selection p.242 in Gessler (1995), ibid. or Mutation>Reproduction>Selection p.67 in Charlesworth & Charlesworth (1997), ibid. or Juvenile Production>Mutation>ViabilitySelection>DensityDependentCulling in Lynch et al. (1995) "Mutation accumulation and the extinction of small populations", Am. Nat. 146:489-518.
27. Higgins & Lynch (2001) "Metapopulation extinction caused by mutation accumulation", Proc. Natl. Acad. Sci. USA 98:2928-2933. - Gabriel & Burger (2000) "Fixation of clonal lineages under Muller's ratchet", Evolution 54:1116-1125.
28. Brommer (2000) "The evolution of fitness in life-history theory", Biol Rev Camb Philos Soc 75:377-404. - Benton & Grant (2000) "Evolutionary fitness in ecology: Comparing measures of fitness in stochastic, density-dependent environments", Evolutionary Ecology Research Oct 2:769-789. - Murray (1997) "Population Dynamics of Evolutionary Change: Demographic Parameters as Indicators of Fitness", Theor Popul Biol 51:180-184.
29. See discussion in Stearns (1992) "The evolution of life histories", Oxford, Oxford University Press.
30. For examples see Stephan et al. (1993), ibid. & Gessler (1995), ibid.
31. See p. 5-11 in Crow & Kimura (1970) "An Introduction to Population Genetics Theory", Edina, Burgess International Group Incorporated.
32. Wilke et al. (2001) "Evolution of digital organisms at high mutation rates leads to survival of the flattest", Nature 412:331-333. - Lenski et al. (1999) "Genome complexity, robustness and genetic interactions in digital organisms", Nature 400:661-664.

The simplest and most common choice is to employ a multiplicative fitness model and assume that all mutations are independent of each other (no epistasis). Under such a model, the fitness of an individual is given by the product of all individual mutation effects

$$w = \prod_{i=1}^{k} (1 + s_i) \tag{4}$$

where s_i is the negative selection coefficient of the i-th SDM of a total of k in the genome. If only equal effects occur, (4) can be simplified to $w = (1+s)^k$. An experimental study of epistatic interactions of multiple transposon mutations in bacteria has shown that Equation (4) can reflect the mean type of interaction of mutations quite well[34]. In the same study, as many synergistic (positive) interactions could be found as there were antagonistic (negative) ones.

If one wants to be precise in modelling advantageous mutations, then Equation (4) can not be extended to incorporate advantageous mutations. Instead, combined advantageous mutation effects have to be computed by

$$w = \prod_{i=1}^{l} \frac{1}{(1 - s_i)} \tag{5}$$

where s_i is the positive selection coefficient of the i-th SAM of a total of l in the genome.

Fecundity and reproductive rates

The last detailed definition we need concerns fecundity. Here two types of models exist: the *infinite fecundity model* and the *limited fecundity model*. The former assumes that initially organisms could have infinite numbers of offspring, if they lived in a habitat that would support this. Such a model fol-

33. Wolf et al., (eds, 2000) "Epistasis and the evolutionary process", New York, Oxford University Press. Mallet (2001) "Holy Landscapes", Science 291:602-602. - Fenster et al. (1997) "Epistasis and its consequences for the evolution of natural populations", Trends Ecol. Evol. 12:282-286. - Whitlock et al. (1995) "Multiple fitness peaks and epistasis", Ann. Rev. Ecol. Syst. 26:601-629. -
Some studies that investigate Mullers ratchet and epistasis; - Butcher (1995) "Muller's ratchet, epistasis and mutation effects", Genetics 141:431-437. - Schultz & Lynch (1997) "Mutation and extinction: The role of variable mutational effects, synergistic epistasis, beneficial mutations, and the degree of outcrossing", Evolution 51:1363-1371. -Kondrashov & Crow (1988) "King's formula for the mutation load with epistasis", Genetics 120:853-856. -
Some recent experimental investigations: Elena & Lenski (1997) "Test of synergistic interactions among deleterious mutations in bacteria", Nature 390:395-398. - Orr & Irving (2001) "Complex epistasis and the genetic basis of hybrid sterility in the Drosophila pseudoobscura Bogota-USA hybridization", Genetics 158:1089-1100. - Peters & Keightley (2000) "A test for epistasis among induced mutations in Caenorhabditis elegans", Genetics Dec 156:1635-1647. - Burch & Chao (1999) "Evolution by small steps and rugged landscapes in the RNA virus phi 6", Genetics 151:921-927.
34. Elena & Lenski (1997) "Test of synergistic interactions among deleterious mutations in bacteria", Nature 390:395-398.

lows easily form the simplest implementation of a ratchet simulation (see page 36): If the process of picking parents and producing offspring is always repeated until the habitat's capacity is filled, then fecundity will have to be enormous to continually offset the effects of the ratchet. Models that incorporate mutational meltdown apply some constraints here.

Conditions for extinction

Assuming multiplicative fitness, the actual number of mutations needed for extinction can be computed easily. As soon as fitness falls below

$$w_{critical} = \frac{1}{R_{max} \cdot V_{max}} \tag{6}$$

the parental generation will no longer be able to replace itself, because the maximal expectation of offspring in any generation is given by $w \cdot R_{max} \cdot V_{max}$. Then mutational meltdown (see below) will start and extinction usually happens short after that. Under multiplicative fitness with equal mutational effects, this allows to formulate the following condition:

$$R_{max} \cdot V_{max} \cdot (1+s)^{Cl_{mm}} = 1 \tag{7}$$

where $R_{max} \cdot V_{max}$ gives the expectation of the maximal number of offspring conceivable for *any* individual of this species (if free from deleterious mutations and in an empty, ideal habitat), s is the negative selection coefficient and Cl_{mm} is the number of clicks needed for mutational meltdown to be started. Solving for Cl_{mm} yields

$$Cl_{mm} = \frac{\log\left(\frac{1}{R_{max} \cdot V_{max}}\right)}{\log(1+s)} \tag{8}$$

Clicks needed for extinction

which can be regarded as a rough estimate of extinction time, given the usual errors associated with such a calculation. To provide an easy overview, Cl_{mm} is plotted for various combinations of R_{max} and s in , assuming that V_{max} is 1. It shows that the most crucial factor in determining extinction times is the selection coefficient.

Advantage of high fecundity

How large then is the advantage of high fecundity? This can be seen from computing Cl_{hfa}, the high fecundity advantage factor that quantifies how many more clicks are needed to bring a high fecundity species to extinction as compared to a low fecundity species. From the formulae above follows, that

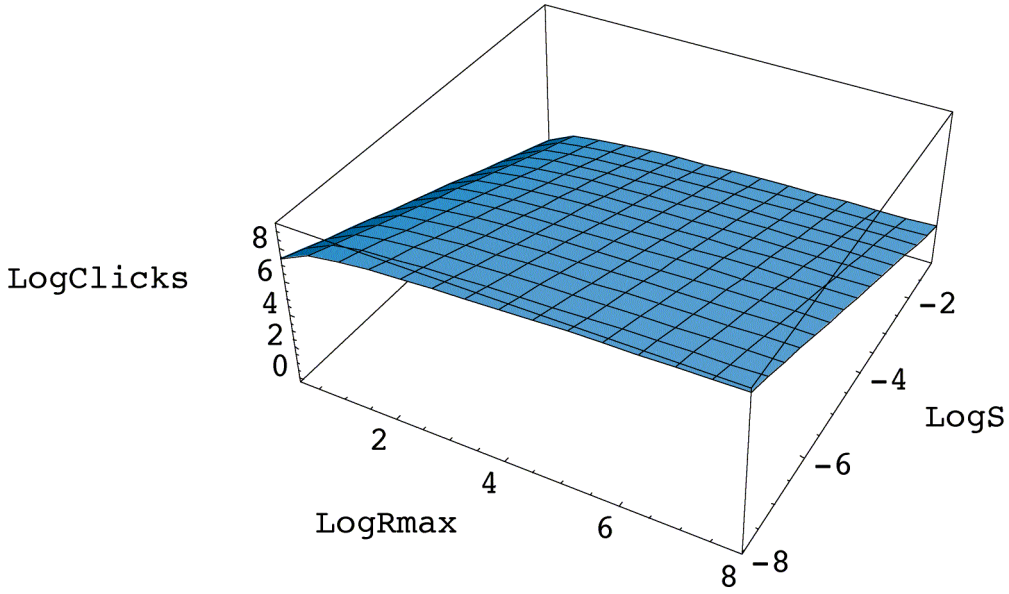


Figure 6 Clicks needed for mutational meltdown as a function of fecundity and selection coefficient. Decadic LogPlot of Cl_{mm} (the number of clicks needed to bring a population to the border of mutational meltdown) over $\log_{10}$ of R_{max} and s. Assumes $V_{max} = 1$. It is easy to see that selection coefficients have a stronger effect on time to meltdown than fecundity.

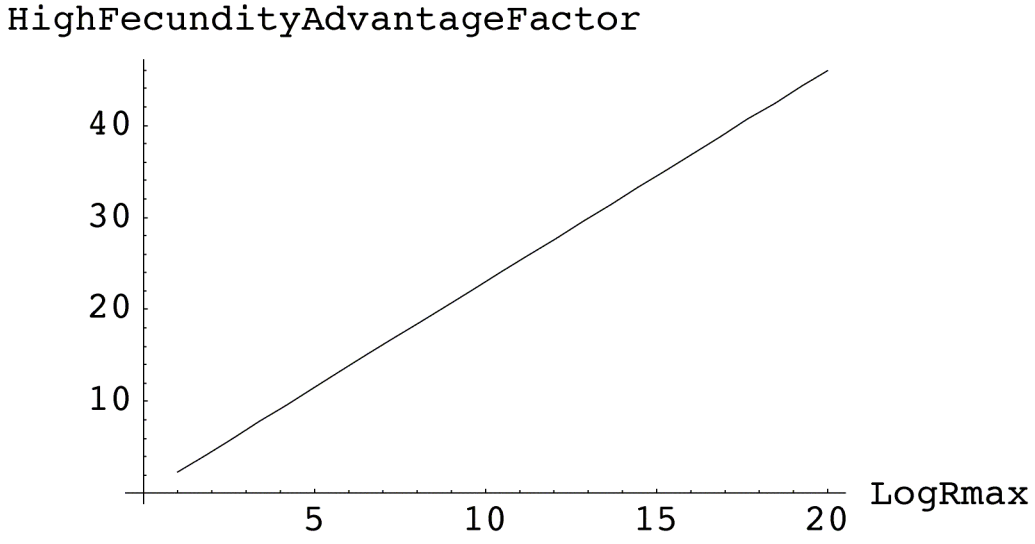


Figure 7 High fecundity delays mutational meltdown only by a small factor.
Plot of Cl_{hfa} (high fecundity advantage factor) over fecundity R_{max}. Please note the linear scale of Cl_{hfa} and the $\log_{10}$ scale of R_{max}. This picture does not change significantly for deleterious selection coefficients in the biologically interesting range . Fecundities are compared to a species with $R_{max} \cdot V_{max} \approx 2.7$. Assumes $V_{max} = 1$.

$$Cl_{hfa} = \frac{\log\left(\frac{1}{R_{max2} \cdot V_{max2}}\right)}{\log\left(\frac{1}{R_{max1} \cdot V_{max1}}\right)} \quad (9)$$

where the indices 1 and 2 are for low and high fecundity species respectively. If a very low fecundity species has $R_{max} \cdot V_{max} \approx 2.7$, then it will need $Cl_{mm} \approx 1/s$ clicks for extinction and a high fecundity species has an advantage of $Cl_{hfa} = Cl_{mm2} \cdot s$, where Cl_{mm2} gives the clicks needed for extinction of the high fecundity species. Figure 7 plots Cl_{hfa} for a range of fecundities as compared to such a low fecundity species. Figure 7 shows that high fecundity species have a several-fold longer time to extinction, because they can mask more clicks by their high fecundity. However, this might amount only to little more than one order of magnitude longer to extinction, when compared to a low fecundity species. As a rule of thumb, high fecundity delays mutational meltdown by a linear factor of $\ln(R_{max} \cdot V_{max})$, where ln is the natural logarithm.

Rule of thumb for high fecundity advantage

Thus, contrary to widespread belief, a species that is likely to suffer because of its combination of U and s will rarely be saved by high fecundity under multiplicative fitness. A similar result has been found for population size[35] and is easily confirmed by looking at Figure 8 and Figure 9.

6.3 Predicting the rate of Muller's ratchet

To estimate the threat from Muller's ratchet, we need to know its speed. There have been numerous attempts to solve this problem analytically and by simulation[36], but up to now no simple (or complex) comprehensive solution is available. Analytical approximations that work in certain parameter ranges are the best that is available. Furthermore, the best approximations involve advanced mathematical theory and the use of maths-packages like MATHEMATICA[37]. Both are beyond the reach of most wet-lab biologists. Therefore discussions of the ratchet have been confined to finding systems where the ratchet could operate in principle, while the actual threat is hardly ever quantified for a particular system in detail[38].

In such a situation a two-step strategy is helpful. First, very rough approximations are used to check whether it is worth further investigating the

35. Bernardes (1996) "Mutation load and the extinction of large populations", Physica A 230:156-173.

ratchet in a particular system. Then more refined approaches are taken to actually compute an estimate for the threat of the ratchet.

Historically, the first step in this sense has been to use $\mathbf{N_o}$ as an extremely rough indicator for the operation (not the speed) of the ratchet[39]. If $\mathbf{N_o}$ is plotted over a wide range of the parameters $\boldsymbol{U}$ and $\boldsymbol{s}$, three qualitative different regions can be distinguished (see Figure 8 and Figure 9):

- $\boldsymbol{U < s}$, on the top of the hill in Figure 8. Here $\mathbf{N_o}$ is almost as large as $\mathbf{N_e}$ and the ratchet practically never turns, even if populations are relatively small (as long as they are not so small that they go extinct due to demographic stochasticity[40]).
- $\boldsymbol{U >> s}$, in the valley of Figure 9. Here $\mathbf{N_o}$ is so small that the Ratchet turns at a very high speed. The only question is whether the mutations that accumulate can erode fitness fast enough to cause extinction.

Everything in between. This is the range of parameters where actions of the ratchet are hard to predict.

---

36. Stephan et al. (1993) "The advance of Muller's ratchet in a haploid asexual population: Approximate solutions based on diffusion theory", Genet. Res. 61:225-231. - Gessler (1995) "The constraints of finite size in asexual populations and the rate of the ratchet", Genet. Res. 66:241-253. - Bell (1988) "Recombination and the immortality of the germ line", J. evol. Biol. 1:67-82. - Butcher (1995) "Muller's ratchet, epistasis and mutation effects", Genetics 141:431-437. - Charlesworth & Charlesworth (1997) "Rapid fixation of deleterious alleles can be caused by Muller's Ratchet", Genet. Res. 70:63-73. - Gabriel et al. (1993) "Mullers Ratchet and mutational meltdowns", Evolution 47:1744-1757. - Gabriel & Burger (2000) "Fixation of clonal lineages under Muller's ratchet", Evolution 54:1116-1125. - Haigh (1978) "The accumulation of deleterious genes in a population - Muller's Ratchet", Theor. Pop. Biol. 14:251-267. - Higgs & Woodcock (1995) "The accumulation of mutations in asexual populations and the structure of genealogical trees in the presence of selection", J. math. Biol. 33:677-702. - Lynch et al. (1993) "The mutational meltdown in asexual populations", J. Hered. 84:339-344. - Lynch et al. (1995) "Mutational meltdowns in sexual populations", Evolution 49:1067-1080. - Lynch et al. (1995) "Mutation accumulation and the extinction of small populations", Am. Nat. 146:489-518. - Pamilo et al. (1987) "Accumulation of mutations in sexual and asexual populations", Genet. Res. 49:135-146. - Stephan & Kim (2001) "Recent applications of diffusion theory to population genetics", in: manuscript to be published in book. - Gordo & Charlesworth (2000) "On the speed of Muller's ratchet", Genetics 156:2137-2140. - Gordo & Charlesworth (2000) "The degeneration of asexual haploid populations and the speed of Muller's ratchet", Genetics 154:1379-1387. - Schultz & Lynch (1997) "Mutation and extinction: The role of variable mutational effects, synergistic epistasis, beneficial mutations, and the degree of outcrossing", Evolution 51:1363-1371.
37. See http://www.wolfram.com
38. A general discussion of the impact of the ratchet in mitochondria can be found in Gabriel et al. (1993) "Mullers Ratchet and mutational meltdowns", Evolution 47:1744-1757.
39. As suggested by Maynard Smith (1978) "Some consequences of sex and recombination - II. Muller's ratchet", pp. 33-36 in: Maynard Smith (ed) The Evolution of Sex, New York, Cambridge University Press.
40. Gabriel & Burger (1992) "Survival of small populations under demographic stochasticity", Theor. Pop. Biol. 41:44-71.

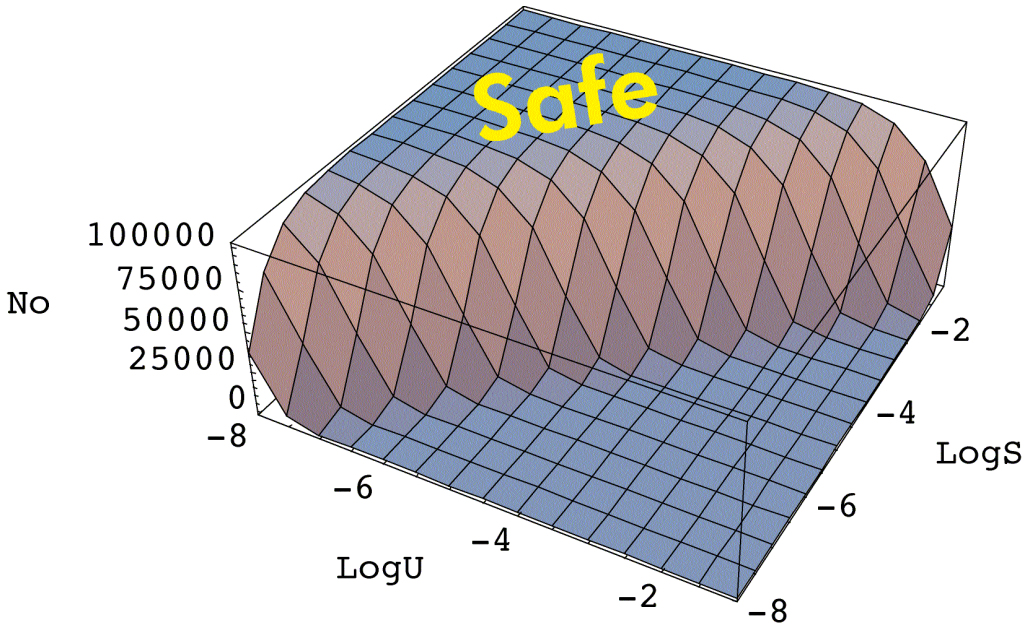


Figure 8 N_o plotted over a wide range of the parameters U and s. Please note the log scales of U and s. See Figure 9 for a better impression of what appears to be a plane here. N_e=100000.

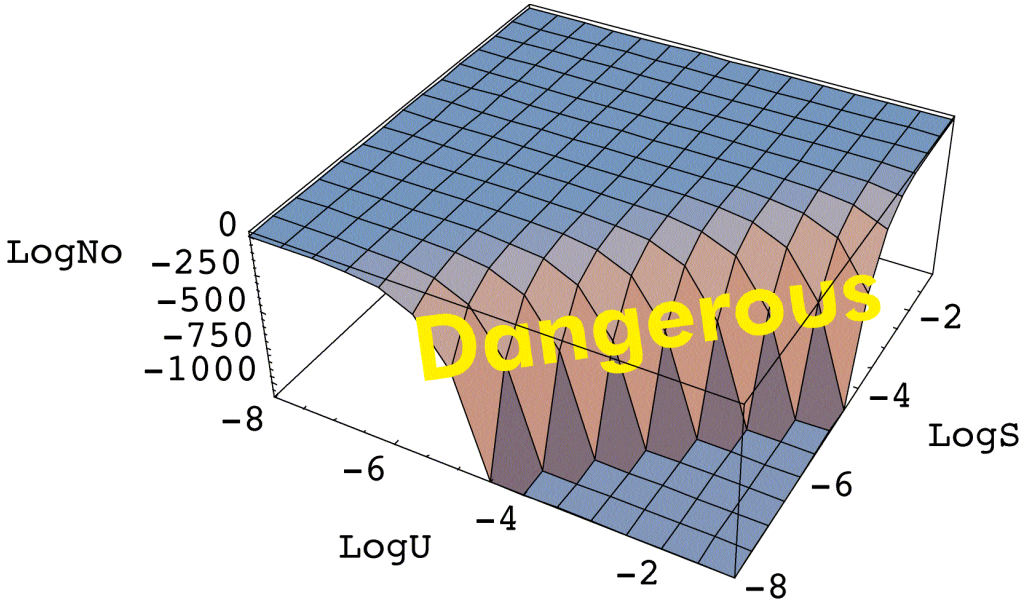


Figure 9 Log of N_o plotted over the same range of U and s as in Figure 8. N_e=100000.

These figures show that population size does not primarily determine operation of the ratchet – U and s are much more important. Recent research has shown that N_e can be a very poor indicator[41] of the operation of the ratchet. Therefore, the use of $N_e s$ was suggested[41].

The best analytic approaches

Currently, the two best approaches for predicting the rate of the ratchet are the diffusion theory approximation developed by Stephan et al. (1993)[42] and the quantitative genetics approximation of Gessler (1995)[43]. While the former is well suited when N_e is large (>>1), the latter is best when N_e is small (<<1). For details, see also Chapter 13. However, up to now an integrated analytical approach is missing and numerical problems can seriously complicate use of the approaches above for certain parameter combinations. The need to get a fast and easy prediction for the rate of the ratchet led to the simple heuristic equation presented in Chapter 18. To determine the exact rate of the ratchet, however, one has to check these approximations by computer simulations. This is frequently done, as simulations also allow incorporation of a bit more biological reality, something that can easily become impossible with analytical models.

6.4 Mutational meltdown and other long term consequences

Simulations have been used to investigate potential long term consequences of the operation of Muller's ratchet under a variety of settings[44]. After genomic decay due to accumulation of SDMs has operated long enough, mu-

41. Gordo & Charlesworth (2000) "The degeneration of asexual haploid populations and the speed of Muller's ratchet", Genetics 154:1379-1387. - Gordo & Charlesworth (2000) "On the speed of Muller's ratchet", Genetics 156:2137-2140. - Gordo & Charlesworth (2001) "The speed of Muller's ratchet with background selection, and the degeneration of Y chromosomes", Genet. Res. 78:149-161.
42. Stephan et al. (1993) "The advance of Muller's ratchet in a haploid asexual population: Approximate solutions based on diffusion theory", Genet. Res. 61:225-231. - Stephan & Kim (2002) "Recent applications of diffusion theory to population genetics", pp. 72-93 in: Slatkin & Veuille (eds) Modern developments in theoretical population genetics, Oxford, Oxford University Press. - See also explanations in Gordo & Charlesworth (2000) "The degeneration of asexual haploid populations and the speed of Muller's ratchet", Genetics 154:1379-1387.
43. Gessler (1995) "The constraints of finite size in asexual populations and the rate of the ratchet", Genet. Res. 66:241-253.
44. Gabriel et al. (1993) "Mullers Ratchet and mutational meltdowns", Evolution 47:1744-1757. - Lynch et al. (1995) "Mutational meltdowns in sexual populations", Evolution 49:1067-1080. - Lynch et al. (1995) "Mutation accumulation and the extinction of small populations", Am. Nat. 146:489-518. - Schultz & Lynch (1997) "Mutation and extinction: The role of variable mutational effects, synergistic epistasis, beneficial mutations, and the degree of outcrossing", Evolution 51:1363-1371. - Higgins & Lynch (2001) "Metapopulation extinction caused by mutation accumulation", Proc. Natl. Acad. Sci. USA 98:2928-2933.

tational meltdown was found to occur[45]. It accelerates the demographic decline of a population on its way to extinction by fixing increasingly deleterious mutations by drift. Thus, if we consider a newly arisen, SDM free, obligate asexual population as an example, the extinction process can be partitioned in three phases[46]:

1. Initially, all individuals have a similar genotype, as they all were derived from a recent founder individual which defines the best class. The number of best class individuals is well above $\mathbf{N}_o$, and the ratchet has not yet started clicking. Before mutation-selection-drift balance has not emerged, the mean number of mutations per individual rises comparatively fast. **Perfection** / **Emergence of mutation-selection-drift balance**
2. This fast rise of the mean of mutations per individual is slowed down, when eventually mutation-selection-drift balance emerges. The degree of this slowdown depends on the rate of the ratchet for that population. After the slowdown, mutations are being steadily accumulated by the ratchet. However, eventually the point is reached where deleterious mutation effects can no longer be compensated for by fecundity, because $\boldsymbol{w} \cdot \boldsymbol{R}_{max} \cdot \mathbf{V}_{max}$ has fallen below 1. **Long term genomic decay**
3. At this point the population can no longer sustain itself, as fewer offspring are produced than parents exist. Thus, population size starts to decline for demographic reasons and that would be enough to cause extinction. However, as selection becomes more and more ineffective in smaller populations, deleterious mutations are being accumulated even faster. As a consequence, fitness declines further and demographic decline of population size is accelerated, which again leads to fixation of more deleterious mutations and so on. This positive feedback loop is called mutational meltdown and continues until extinction. **Mutational meltdown** / **Extinction**

If predictions of the effective click time of the ratchet $\boldsymbol{T}_{\boldsymbol{cleff}}$ are combined with $\mathbf{Cl}_{\boldsymbol{mm}}$, the number of mutations needed for meltdown to start, then an approximate prediction of the time when meltdown starts $\boldsymbol{T}_{\boldsymbol{mm}}$ can be obtained by further multiplication with generation time $\boldsymbol{T}_{\boldsymbol{gen}}$:

$$\boldsymbol{T}_{\boldsymbol{mm}} \cong \mathbf{Cl}_{\mathbf{m}} \cdot \boldsymbol{T}_{\boldsymbol{cl}} \cdot \boldsymbol{T}_{\boldsymbol{gen}} \qquad (10)$$

45. First reported in Lynch & Gabriel (1990) "Mutation load and the survival of small populations", Evolution 44:1725-1737. - Many more papers followed, see footnote [44]
46. Lynch et al. (1993) "The mutational meltdown in asexual populations", J. Hered. 84:339-344. See Figure 1 in there for a qualitative description of the temporal pattern.

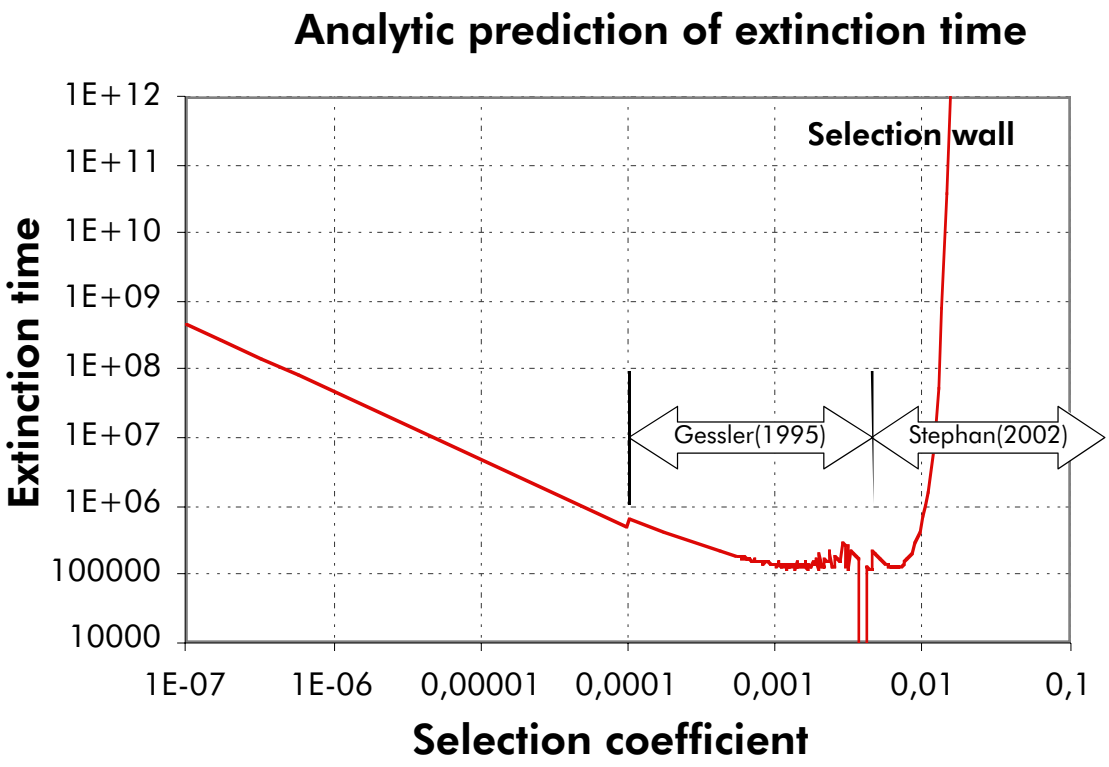


Figure 10 U-shaped plot of extinction time over selection coefficient for a population with N_e = 50000, R_{max} = 10 and U = 0.05 under multiplicative fitness.
Clicktimes used in this plot were computed with *Mathematica* according to the analytic methods of Gessler (1995) and Stephan & Kim (2002), as described in Chapter 13. Values with T_{ex} below 10^5 are due to unpredictable, complicated numerical errors that occurred occasionally. Deleterious selection coefficients are given as positive values for easier plotting. Please note the '**selection wall**' and its sharp transition to the most dangerous region of the ratchet: Beyond the wall, no operation of the ratchet is possible, because selection coefficients are so deleterious that purifying selection becomes too strong. The straight line below the range of the method of Gessler assumes a clicktime of $1/U$. Time is measured in generations.

This estimate is approximately equal to T_{ex}, as the absolute duration of mutational meltdown itself can usually be neglected, when compared to the time until it starts. Plotting T_{ex} over s for a given combination of U and N_e, yields an U-shaped relationship[47], which will be called the **U-shaped plot of extinction time over selection coefficients** (see Figure 10). It will be

47. To my knowledge, the U-shaped plot of extinction time over selection coefficient was introduced by Lynch et al. (1993 "The mutational meltdown in asexual populations", J. Hered. 84:339-344) as a qualitative figure and first used for simulation results by Gabriel et al. (1993) "Mullers Ratchet and mutational meltdowns", Evolution 47:1744-1757.

heavily used in the Chapters 21-24 of this work to quantify the threat of Muller's ratchet in a variety of systems.

The U-shaped plot

It indicates the range of selection coefficients that poses the highest risk for extinction. If mutation effects are larger, the ratchet will turn slower and thus extinction will be delayed or prevented. If mutation effects are much smaller, ratchet speed will no longer be limited by selection, but the steps towards extinction get smaller, as many more mutations are needed to cause extinction. We have hardly any detailed information on selection coefficients in the dangerous range. Thus the best one can currently do to quantify the threat from Muller's ratchet is to predict the details of this U-shaped plot for a set of parameter values that are as realistic as possible (U, N_e, R_{max}, etc). Then a discussion of related information on selection strength will help to clarify under what circumstances extinction may turn out to be a problem.

The J-shaped plot

This approach is aided the **J-shaped plot** of extinction time over population size[48] that can be used to investigate the effect of population size on extinction times for a fixed combination of U and s (see Chapter 21-24 for examples).

6.5 When is Muller's ratchet dangerous?

Muller's ratchet is not always dangerous, but if it is, the following factors contribute to the threat.

High SDM rate

The most prominent factor is a high deleterious mutation rate. This comprises two things: (i) a high overall mutation rate and (ii) a high frequency of mutation effects in the dangerous range of s. In principle, the ratchet becomes a problem when enough significantly large SDMs are being generated to cause extinction and selection can no longer remove them fast enough (see discussion of the U-shaped plot above). As a practical consequence, decreasing mutational effects by benign environments is not desirable under all circumstances[49]: If this moves s from the save (purifying selection) to the dangerous (slowly accumulating) range, then this will turn the ratchet. It would have to be moved down to the irrelevant range, where the ratchet turns, but has no significant long-term effects. If that cannot be achieved, no change would be better.

48. To my knowledge it was first employed by Lynch et al. (1995) "Mutation accumulation and the extinction of small populations", Am. Nat. 146:489-518.
49. Gabriel et al. (1993) "Muller's ratchet and mutational meltdowns", Evolution 47:1744-1757.

Wide distribution of mutational effects

In the simplest model of the ratchet with equal mutation effects, epistatic interactions might stop the ratchet[50]. The reason for this result is that synergistic epistasis increases the effects of mutations with each additional mutation until selection becomes quasi-truncating[51] - one of its most powerful forms. Thus the moment will come where **s** is so large that it has the power to stop the ratchet. All this, however, depends critically on the assumption that mutation effects are strictly equal. If they are allowed to vary, they lose their potential to stop the ratchet[52], as the distribution will always produce some mutations in the dangerous range - a more likely reflection of biological reality.

Not enough beneficial mutations

A powerful way to offset the ratchet are beneficial mutations. If mutations in some quantitative trait can compensate for loss due to SDMs, then they easily stop the ratchet if they appear frequently enough[53]. Together with compensatory mutations on a molecular level and unconditionally advantageous mutations, such mutations might lead to a situation where selection is very busy, but no adaptation occurs[54], as only the status quo is maintained. This is similar to the red-queen scenario[55] that characterizes host-parasite arms races: You have to constantly evolve to stay in the same place. However, when measuring beneficial mutation rates, the interesting quantity is not so much how many will appear in the beginning, but rather the sustained beneficial mutation rate in the long term.

Low fecundity, small populations

As can be seen from the discussion above, small population size and low fecundity increases the threat from Muller's ratchet. However, if the ratchet does not turn for other reasons, these two factors will not start it and if the ratchet does turn for other reasons, high fecundity and large population size will rarely change its course. However, there is a small band of **U/s** combi-

50. Kondrashov (1994) "Muller's Ratchet under epistatic selection", Genetics 136:1469-1473.
51. Crow & Kimura (1979) "Efficiency of truncation selection", Proc. Natl. Acad. Sci. U.S.A. 76:396-399.
52. Butcher (1995) "Muller's ratchet, epistasis and mutation effects", Genetics 141:431-437.
53. Wagner & Gabriel (1990) "Quantitative variation in finite parthenogenetic populations: What stops Muller's ratchet in the absence of recombination?" Evolution 44:715-731.
54. Hartl & Taubes (1996) "Compensatory nearly neutral mutations: Selection without adaptation", J. theor. Biol. 182:303-309.
55. Van Valen (1973) "A new evolutionary law", Evol. Theory 1:1-30. - Maynard Smith (1978) "The Evolution of Sex", New York, Cambridge University Press. - Clarke et al. (1994) "The red queen reigns in the kingdom of RNA viruses", Proc. Natl. Acad. Sci. U.S.A. 91:4821-4824. - Kawecki (1998) "Red queen meets Santa Rosalia: Arms races and the evolution of host specialization in organisms with parasitic lifestyles", American Naturalist. vol.152 pp.635-651. - Martens & Schon (2000) "Parasites, predators and the Red Queen", Trends Ecol. Evol. Oct 15:392-393. - Stenseth & Maynard Smith (1984) "Coevolution in ecosystems: Red Queen evolution or stasis?" Evolution 38:870-880.

nations in parameter space, where population size indeed does make a critical difference.

Not enough recombination

Finally, sex (ie. regular recombination) is one potent agent that can stop Muller's ratchet. This is the only mechanism that can potentially restore an ancient genotype that is better than the current average. While thinking about this, H. J. Muller discovered the ratchet phenomenon and it still plays an important role in theories about the evolution of sex[56] today.

Genomic decay

However, if $U > 1$, not even sex might stop Muller's ratchet, because selection will be overwhelmed by mutation pressure, as *each new individual* is expected to carry one *additional new* deleterious mutation. This appears to be the case in large mammalian nuclear genomes[57] and it is not clear yet how this apparent genomic decay can be reconciled with their existence over millions of years. Perhaps the potential solutions discussed for the survival of obligate asexuals have broader implications than previously thought.

6.5.1 Significance of Muller's ratchet for conservation biology

The last decades have seen the rise of conservation biology[58]. Muller's ratchet has obvious implications for the conservation of endangered species. As species are usually protected by law only when their population sizes have

56. Just a few exemplary references to this vast field: Butlin (2002) "Opinion - evolution of sex: The costs and benefits of sex: new insights from old asexual lineages", Nat Rev Genet 3:311-317. - Rice (2002) "Experimental tests of the adaptive significance of sexual recombination", Nat Rev Genet 3:241-251. - Otto & Lenormand (2002) "Resolving the paradox of sex and recombination", Nat Rev Genet 3:252-261. - Maynard Smith (1978) "The Evolution of Sex", New York, Cambridge University Press. - Kondrashov (1988) "Deleterious mutations and the evolution of sexual reproduction", Nature 336:435-440. - Kondrashov (1993) "Classification of hypotheses on the advantage of amphimixis", J. Hered. 84:372-387. - Kondrashov (1997) "Evolutionary genetics of life cycles", Ann. Rev. Ecol. Syst. 28:391-435. - Crow (1994) "Advantages of sexual reproduction", Developmental Genetics 15:205-213. - Bell (1982) "The Masterpiece of Nature: The Evolution and Genetics of Sexuality", Berkeley, University of California Press. - Bernstein & Bernstein (1991) "Aging, Sex, and DNA Repair", San Diego, Academic Press. - Halvorson & Monroy, (eds, 1985) "The Origin and Evolution of Sex", New York, Alan R. Liss Inc. - Michod & Levin Bruce, (eds, 1988) "The Evolution of Sex: An Examination of Current Ideas", Sunderland, Sinauer Associates Incorporated.

57. Eyre-Walker & Keightley (1999) "High genomic deleterious mutation rates in hominids", Nature 397:344-347. - Crow (1993) "Mutation, mean fitness, and genetic load", pp. 3-42 in: Futuyma & Antonovics (eds) Oxford Surveys in Evolutionary Biology. 9, Oxford, Oxford University Press. - Crow (1997) "The high spontaneous mutation rate: Is it a health risk?" Proc. Natl. Acad. Sci. U.S.A. 94:8380-8386. - Crow (2000) "The origins, patterns and implications of human spontaneous mutation", Nat Rev Genet 1:40-47. - Nachman & Crowell (2000) "Estimate of the mutation rate per nucleotide in humans", Genetics 156:297-304. - Kondrashov (1995) "Contamination of the Genome by Very Slightly Deleterious Mutations: Why Have We Not Died 100 Times Over?" J. theor. Biol. 175:583-594.

decreased dramatically, mutational meltdown might complicate conservation efforts. This is not limited to asexual species, as very small populations of sexual species are capable of mutational meltdown too[59].

Thus the investigation of genetic reasons for species declines should be advanced. This includes detailed investigations of situations where genetic reasons potentially can or actually do play a role in extinction, as well as careful investigation of all biological factors that might keep species from extinction (see Chapter 26 for a list). It is expected that a better understanding of these processes will lead to better management decisions, at least in some cases.

However, potential genetic reasons should not detract from primary reasons for species declines like habitat fragmentation and other ecological causes. In many cases it is scientifically clear what has to be done, but lack of political determination maintains the destructive status quo. Thus, save the rain forests.

58. See Meffe & Carroll, (eds, 1997) "Principles of conservation biology". 2nd, Sunderland, MA, Sinauer Associates. - Primack (1998) "Essentials of Conservation Biology". 2nd, Sunderland, MA, Sinauer Associates. -
59. Lynch et al. (1995) "Mutational meltdowns in sexual populations", Evolution 49:1067-1080.

7 Observations

Numerous observations underscore the importance of Muller's ratchet for our understanding of nature. Some of the most important are listed here.

What is an observation of Muller's ratchet? The ultimate product of Muller's ratchet is an extinction. The world is full of examples for extinct species. Furthermore, the ratchet is driven by slightly deleterious mutations. It is biological commonsense, that deleterious mutations are more frequent than advantageous ones and numerous observations confirm that, so why look any further?

General ways to observe the ratchet

Unfortunately, things are not that easy. We can not exclude the plethora of other (mainly ecological) reasons for extinction and the mere observation of SDMs is not necessarily proof that they accumulate in a ratchet-like manner. Nevertheless observations of the ratchet are possible:

- Observation of high deleterious mutation rates with effects in the dangerous range[60] (whatever that is in the study system) allows us to infer that the ratchet operates.

  When is a mutation rate high enough to invoke the ratchet? Although not really valid, many people use $U>1$ for sexual species and $U>0$ for obligate asexuals, while not bothering about the rest. For a more refined approach to spread, predictions of the rate of the ratchet have to become easier for most biologists, an important motivation for this thesis.
- Observation of the effects of the ratchet, namely accumulation of deleterious mutations, decreased fecundity and extinction, while being able to exclude other causes for the latter two.

60. Hartl (1989) "The physiology of weak selection", Genome 31:183-189. - Hartl et al. (1994) "Selection intensity for codon bias", Genetics 138:227-234. - Akashi (1995) "Inferring weak selection from patterns of polymorphism and divergence at "Silent" sites in Drosophila DNA", Genetics 139:1067-1076. - Akashi (1997) "Distinguishing the effects of mutational biases and natural selection on DNA sequence variation", Genetics 147:1989-1991. - Akashi & Schaeffer (1997) "Natural selection and the frequency distributions of "silent" DNA polymorphism in Drosophila", Genetics 146:295-307. - Akashi et al. (1998) "Mutation pressure, natural selection, and the evolution of base composition in Drosophila", Genetica 103:49-60. - Akashi (1999) "Within- and between-species DNA sequence variation and the 'footprint' of natural selection", Gene 238:39-51. - Akashi (1999) "Inferring the fitness effects of DNA mutations from polymorphism and divergence data: Statistical power to detect directional selection under stationarity and free recombination", Genetics 151:221-238.

Thus it is difficult, but not impossible, to observe the ratchet in populations[61].

Slightly deleterious mutation rates

No small amount of work has been directed at estimating deleterious mutation rates[62]. However, techniques in the lab are only capable of observing moderately small differences in relatively limited time spans (How after all do you measure a difference of 10^{-4} or less in reproductive rate?). Thus information on selection coefficients is extremely limited, whereas estimates of the total mutation rate usually can be made with higher confidence. It will not be attempted to review knowledge on deleterious mutation rates here, as this has been done elsewhere[63]. Generally, when unusually high mutation rates are found, they are often connected to Muller's ratchet, especially in asexual genetic systems[64].

Viruses

A number of studies have investigated Muller's ratchet in viruses[65] by serial passage experiments[66]. Although many used a bottleneck size of 1, it can be generally concluded that Muller's ratchet is active in many viruses, especially RNA viruses with their high mutation rates[67]. However, viruses have found ways to escape extinction: they often show enormous fitness increase when they occur in large populations[68], have significant compensatory mu-

61. Observation of accumulation of SDMs in a single line of descent is a very special case of the ratchet and is thus not a 'full' observation of the ratchet in itself. See Baake & Gabriel (1999) "Biological evolution through mutation, selection and drift: An introductory review", Annual Reviews of Computational Physics 7:203-264.
62. Drake et al. (1998) "Rates of spontaneous mutation", Genetics 148:1667-1686. - Lynch et al. (1999) "Perspective: Spontaneous deleterious mutation", Evolution 53:645-663. - Denver et al. (2000) "High direct estimate of the mutation rate in the mitochondrial genome of Caenorhabditis elegans", Science 289:2342-2344. - Parsons & Holland (1998) "Mitochondrial mutation rate revisited: hot spots and polymorphism - Response", Nat. Genet. 18:110-110. - Vassilieva et al. (2000) "The fitness effects of spontaneous mutations in Caenorhabditis elegans", Evolution 54:1234-1246. - and many other works.
63. See Drake et al. (1998), ibid. and Lynch et al. (1999), ibid.
64. Howell et al. (1996) "How rapidly does the human mitochondrial genome evolve?" Am. J. Hum. Genet. 59:501-509.
65. Morse, (ed, 1994) "The evolutionary biology of viruses", New York, Raven Press. - Domingo et al., (eds, 1999) "Origin and evolution of viruses", San Diego, Academic Press. - Domingo (2000) "Viruses at the edge of adaptation", Virology 270:251-253.
66. Novella et al. (1995) "Size of genetic bottlenecks leading to virus fitness loss is determined by mean initial population fitness", J Virol 69:2869-2872. - Chao (1990) "Fitness of RNA virus decreased by Muller´s ratchet", Nature 348:454-455. - Escarmis et al. (1996) "Genetic lesions associated with Muller's ratchet in an RNA virus", J. Mol. Biol. 264:255-267. - Burch & Chao (2000) "Evolvability of an RNA virus is determined by its mutational neighbourhood", Nature 406:625-628. - Duarte et al. (1992) "Rapid fitness losses in mammalian RNA virus clones due to Muller's ratchet", Proc Natl Acad Sci U S A 89:6015-6019.
67. Nichol (1996) "RNA Viruses: Life on the edge of catastrophe", Nature 384:218-219. - Domingo & Holland (1997) "RNA virus mutations and fitness for survival", Annu. Rev. Microbiol. 51:151-178. - Drake (1993) "Rates of spontaneous mutation among RNA viruses", Proc. Natl. Acad. Sci. U.S.A. 90:4171-4175.

tation rates[69] and some even recombine[70] in order to escape the ratchet. Nevertheless, mutational meltdown can drive them to extinction[71]. Should it turn out to be possible to *specifically* increase mutation rates in viruses, then these concepts might even contribute to the design of antiviral strategies[72].

Microorganisms

In microorganisms loss of fitness in serial transfer experiments has been observed[73] if population size is *very* small or if selection is relaxed on the genes investigated. If population size increases a bit, the first thing observed in serial transfer experiments is adaptation to the current environment[74]. The latter has been found to be limited when mutation rate gets so high that several different advantageous mutations compete in the same population[75]. Thus an important way the ratchet operates might be through the cost of adaptation in changing environments[76]. Experiments in yeast[77] have also resulted in the observation of mutational meltdowns[78].

68. Novella et al. (1995) "Exponential increases of RNA virus fitness during large population transmissions", Proc. Natl. Acad. Sci. USA 92:5841-5844. - Novella et al. (1999) "Exponential fitness gains of RNA virus populations are limited by bottleneck effects", Virology 73:1668-1671.
69. Burch & Chao (1999) "Evolution by small steps and rugged landscapes in the RNA virus phi 6", Genetics Mar 151:921-927. - Elena et al. (1998) "Evolutionary dynamics of fitness recovery from the debilitating effects of Muller's ratchet", Evolution 52:309-314. - Novella et al. (1996) "Repeated transfer of small RNA virus populations leading to balanced fitness with infrequent stochastic drift", Mol Gen Genet 252:733-738.
70. Chao (1988) "Evolution of sex in RNA viruses", J. theor. Biol. 133:99-112. - Chao et al. (1997) "The advantage of sex in the RNA Virus f6", Genetics 147:953-959. -
71. Fraile et al. (1997) "A century of tobamovirus evolution in an australian population of Nicotiana glauca", J. Virol. 71:8316-8320.
72. Lee et al. (1997) "Negative effects of chemical mutagenesis on the adaptive behavior of vesicular stomatitis virus", J. Virol. 71:3636-3640.
73. Andersson & Hughes (1996) "Muller's ratchet decreases fitness of a DNA-based microbe", Proc. Natl. Acad. Sci. U.S.A. 93:906-907. - Kibota & Lynch (1996) "Estimate of the genomic mutation rate deleterious to overall fitness in E. coli", Nature 381:694-696. - Zeyl & DeVisser (2001) "Estimates of the rate and distribution of fitness effects of spontaneous mutation in Saccharomyces cerevisiae", Genetics Jan 157:53-61. - Funchain et al. (2000) "The consequences of growth of a mutator strain of Escherichia coli as measured by loss of function among multiple gene targets and loss of fitness", Genetics 154:959-970. - Cooper & Lenski (2000) "The population genetics of ecological specialization in evolving Escherichia coli populations", Nature 407:736-739. - Wloch et al. (2001) "Direct estimate of the mutation rate and the distribution of fitness effects in the yeast Saccharomyces cerevisiae", Genetics 159:441-452.
74. Zeyl et al. (2001) "Mutational meltdown in laboratory yeast populations", Evolution 55:909-917. - Lenski & Travisano (1994) "Dynamics of adaptation and diversification: A 10,000-generation experiment with bacterial populations", Proc. Natl. Acad. Sci. U.S.A. 91:6808-6814.
75. de Visser et al. (1999) "Diminishing returns from mutation supply rate in asexual populations", Science 283:404-406. - Gerrish & Lenski (1998) "The fate of competing beneficial mutations in an asexual population", Genetica 103:127-144.
76. Giraud et al. (2001) "Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut", Science 291:2606-2608.
77. Zeyl & Bell (1997) "The advantage of sex in evolving yeast populations", Nature 388:465-468. - Zeyl & DeVisser (2001) "Estimates of the rate and distribution of fitness effects of spontaneous mutation in Saccharomyces cerevisiae", Genetics Jan 157:53-61.

Another line of evidence for the ratchet in microbes comes from endosymbiotic bacteria. The genus *Buchnera*, a close relative of *E. coli*, lives endosymbiotically in aphids. As it cannot recombine, it has been suggested, that Muller's ratchet should operate in these bacteria. As they are supposed to be about 100 million years old, their survival is an 'evolutionary scandal'[79]. Comparison of *Buchnera*'s genes with those of free living relatives like *E. coli* shows, that Buchnera has accumulated more non-synonymous substitutions than those bacteria that could horizontally exchange genes[80]. As non-synonymous substitutions change amino acids, they are presumably slightly deleterious.

Ancient asexual species

Investigation of ancient asexual species has similarly been often related to Muller's ratchet. Currently, a number of ancient asexual lines are known with varying degrees of certainty about their asexuality[81]. Most of the work has traditionally focused on establishing their abstinence from sex over such long periods of time. However, for some species like Bdelloid rotifers this can now be safely concluded[82]. Thus the time has come to investigate how these species have managed to survive for so long, although they have absolutely no recombination to remove their spontaneous deleterious mutations.

Y chromosomes

As Y chromosomes do not recombine, they might be deteriorated by Muller's ratchet[83]. This has been observed experimentally on some occasions[84]. The fact that gene density in the human Y chromosome is less than on other chromosomes[85] adds further to the evidence.

78. Zeyl et al. (2001) "Mutational meltdown in laboratory yeast populations", Evolution 55:909-91 - Funchain et al. (2000) "The consequences of growth of a mutator strain of Escherichia coli as measured by loss of function among multiple gene targets and loss of fitness", Genetics 154:959-970.
79. Hurst & McVean (1996) "Evolutionary Genetics:..and scandalous symbionts", Nature 381:650-651.
80. Moran (1996) "Accelerated evolution and Muller's rachet in endosymbiotic bacteria", Proc. Natl. Acad. Sci. U.S.A. 93:2873-2878. - Baumann et al. (1995) "Genetics, physiology, and evolutionary relationships of the genus Buchnera: intracellular symbionts of aphids", Annu Rev Microbiol 49:55-94. - Funk et al. (2001) "Intraspecific variation in symbiont genomes: Bottlenecks and the aphid-buchnera association", Genetics 157:477-489. - Lambert & Moran (1998) "Deleterious mutations destabilize ribosomal RNA in endosymbiotic bacteria", Proc. Natl. Acad. Sci. U.S.A. 95:4458-4462. - Rispe & Moran (2000) "Accumulation of deleterious mutations in endosymbionts: Muller's ratchet with two levels of selection", Am. Nat. Oct 156:425-441.
81. Judson & Normark (1996) "Ancient asexual scandals", Trends Ecol. Evol. 11:41-46. - Wuethrich (1998) "The asexual life", Science 281:1981-1981.
82. Welch & Meselson (2000) "Evidence for the evolution of Bdelloid Rotifers without sexual reproduction or genetic exchange", Science 288:1211-1215.
83. Charlesworth & Charlesworth (2000) "The degeneration of Y chromosomes", Philos Trans R Soc Lond B Biol Sci 355:1563-1572. - Charlesworth (1991) "The evolution of sex chromosomes", Science 251:1030-1033. - Gordo & Charlesworth (2001) "The speed of Muller's ratchet with background selection, and the degeneration of Y chromosomes", Genet. Res. 78:149-161.
84. Rice (1994) "Degeneration of a nonrecombining chromosome", Science 263:230-232.

Mitochondrial DNA

Mitochondrial DNA is another asexual genetic system that nearly all sexual species need[86]. Since the early days after discovering mutational meltdowns, it has been speculated how mtDNA escapes Muller's ratchet[87] and comparisons of thermodynamical stability of tRNA genes on mitochondria with those in the nucleus show decreased stability of mitochondrial genes[88]. However, the actual threat of the ratchet to mitochondria has never been quantified in detail.

High mutation rates in humans

Although sometimes overlooked, nuclear mutation rates of U>1 can drive Muller's ratchet too[89]. After MULLER'S seminal paper on "our load of mutations"[90], a number of reports confirmed such high mutation rates in humans[91] (and, thus, probably also in other mammals). Currently, some form of epistatic interaction is invoked to answer the question why we are still alive today[92] – among an array of other potential solutions (see Chapter 26). Further work is necessary to solve these issues in detail[93].

Observations of mutational meltdown

Occasionally, experimental observations of a mutational meltdown are possible. Such observations have been made in bacteria[94], yeast[95] and in *C. elegans*[96]. Currently, it is unclear how a mutational meltdown could be observed in nature, as exclusion of other reasons for extinction is difficult and obtaining complete genomes with a mechanistic analysis of the effects of the mutations that accumulate before extinction of a rare species is not possible.

85. Lander et al. (2001) "Initial sequencing and analysis of the human genome", Nature 409:860-921.
86. Except for organisms like yeast. These do not die without mitochondria. They just show the "petite" phenotype. Contamine & Picard (2000) "Maintenance and integrity of the mitochondrial genome: a plethora of nuclear genes in the budding yeast", Microbiol Mol Biol Rev 64:281-315.
87. Gabriel et al. (1993) "Muller's ratchet and mutational meltdowns", Evolution 47:1744-1757.
88. Lynch (1996) "Mutation accumulation in transfer RNAs: Molecular evidence for Muller's ratchet in mitochondrial genomes", Mol. Biol. Evol. 13:209-220. - Lynch (1997) "Mutation accumulation in nuclear, organelle, and prokaryotic transfer RNA Genes", Mol. Biol. Evol. 14:914-925.
89. Maynard Smith (1978) "Some consequences of sex and recombination - II. Muller's ratchet", pp. 33-36 in: Maynard Smith (ed) The Evolution of Sex, New York, Cambridge University Press.
90. Muller (1950) "Our load of mutations", Am. J. Hum. Genet. 2:111-176.
91. Eyre-Walker & Keightley (1999) "High genomic deleterious mutation rates in hominids", Nature 397:344-347. - Nachman & Crowell (2000) "Estimate of the mutation rate per nucleotide in humans", Genetics 156:297-304. - Lee et al. (1996) "Number of lethal equivalents in human populations: how good are the previous estimates?" Heredity 77 (Pt 2):209-216.
92. Crow & Kimura (1979) "Efficiency of truncation selection", Proc. Natl. Acad. Sci. U.S.A. 76:396-399. - Crow (1997) "The high spontaneous mutation rate: Is it a health risk?" Proc. Natl. Acad. Sci. U.S.A. 94:8380-8386.
93. Lynch et al. (1999) "Perspective: Spontaneous deleterious mutation", Evolution 53:645-663.
94. Funchain et al. (2000) "The consequences of growth of a mutator strain of Escherichia coli as measured by loss of function among multiple gene targets and loss of fitness", Genetics 154:959-970.
95. Zeyl et al. (2001) "Mutational meltdown in laboratory yeast populations", Evolution 55:909-917.
96. Vassilieva et al. (2000) "The fitness effects of spontaneous mutations in Caenorhabditis elegans", Evolution 54:1234-1246. - Denver et al. (2000) "High direct estimate of the mutation rate in the mitochondrial genome of Caenorhabditis elegans", Science 289:2342-2344.

One potential approach[97] has been used to document extinction of Australian TMV, an RNA virus.

Reading ratchet footprints

However, indirect evidence might be used to infer operation of the ratchet in nature. As explained in Chapter 6, the ratchet leads to a decrease in either the maximal viability or to a decrease in maximal reproductive rates. These life-history parameters can be accessed experimentally. Thus, if a species shows unusually low reproductive rates, even if it is brought into a "perfect" environment (empty habitat, full of nutrients, no whatsoever limitations) Muller's ratchet might be the cause. If, then a sexual sister species has high reproductive rates, one can predict that the non-synonymous substitution rate is higher in the species with the low reproductive rates. When this actually can be observed, evidence for operation of Muller's ratchet is pretty conclusive. Such a degree of certainty has currently been reached in some endosymbiotic bacteria[80]. However, to see the actual meltdown of such a population in nature is highly improbable, as it would have to happen in a very short timeframe of observation and exclusion of ecological reasons for extinction may be still difficult.

Conclusion

Hints at the operation of Muller's ratchet can be observed in a wide range of biological systems. Precise estimations of the threat of Muller's ratchet, however, are rare. One reason for this might be the difficulty of obtaining good predictions for the rate of the ratchet, as only the simplest models are accessible via complicated mathematical theory and extensive computer simulations will usually not be conducted by those people who are closest to the observation of Muller's ratchet. Furthermore, computer simulations of the ratchet can easily become prohibitive, as realistic parameter combinations are often those that need months of computing time. Therefore it is desirable, to build a database of ratchet simulation results that allows easy access for many researchers[98].

As Muller's ratchet is supposed to operate under a variety of circumstances, and species exist that apparently survived that onslaught for many millions of years, it is not clear how precisely this genomic decay paradox should be solved. Welcome in an active area of research.

97. Fraile et al. (1997) "A century of tobamovirus evolution in an australian population of Nicotiana glauca", J. Virol. 71:8316-8320.
98. Just as the evolution@home project is about to do. See http://www.evolutionary-research.net

III. Methods and Software Development

A central part of this work is design of "EEPSLION", a framework that fosters development of evolutionary models that can be integrated into the "evolution@home" global computing system. The first simulation model to be part of evolution@home is described, as are analytical methods for predicting the rate of Muller's ratchet and a method for precise measurements of fitness in bacteria.

8 The vision behind evolution@home and *EEPSLION*

This chapter describes the vision behind the first global computing system for evolutionary biology, "evolution@home", as well as the motivation behind the corresponding software framework, "EEPSLION".

The need for global computing in evolutionary biology

As the previous chapters have shown, there is a big need for *in silico* experiments to understand adaptive evolution in general and the mutation rate paradox and Muller's ratchet in particular. To further understanding, many models with countless parameter combinations and even more individual stochastic simulations are to be computed. To facilitate code development for this enormous task "*EEPSLION*" is developed, a global computing softwareframework for individual-based simulations of evolution. It powers "*evolution@home*", the first global computing system for evolutionary biology.

Vision: remove bottlenecks in modelling

The vision behind this combination is to remove the three bottlenecks that plague so many simulations of evolution: implementation, computation and analysis of results. Thus, a computing biologist can concentrate on implementing the model and no longer needs to worry about simulation basics like parameter handling, flow control, and the like. Once the model is implemented and stable, simulation tasks can be distributed via the evolution@home global computing system to share the computational load with many others who are willing to help. Finally a system for scheduling new runs, storing results, retrieving data, and making plots will facilitate analyses. If this gigantic database uses XML-files, data remain easily accessible and different simulation models can be easily compared. Finally, this system facilitates export of data to the web to publish the results of eventually thousands of years computing time. Although this kind of research currently heavily depends on programming biologists, one might even envision a code generator bench with a graphical interface that further simplifies implementation of evolutionary models.

Vision: communicate science

All global computing projects depend heavily on public support, which is a great opportunity to communicate the science behind the models that are computed on people's machines. This is achieved by an accompanying web site that helps build a computing community interested in evolution – and in the highscores table that lists all participants' contributions.

9 Lessons from development

As all non-trivial software, EEPSLION has undergone several rounds of the design-implement-use cycle. The most important lessons up to now are discussed.

Development cycles

All non-trivial software is imperfect. That comes from the complexity of the systems and the limitations of the human mind. Thus in an effort to solve complex problems by software, development cycles are unavoidable: You often can see only at the end of the current road whether you have reached your destination or whether you have still some way to go. This is a common experience of all who develop complex software[1]. Often there is no other possibility than starting with a vague idea of how things should work. Then, as more experience is gained, a more precise definition of the problems allows for better solutions. Thus with each new cycle the system comes closer to what the final user actually expects.

Importance of problem domain knowledge

Detailed expertise from the problem domain plays a pivotal role in the search for solutions. To gain this kind of knowledge, considerable language barriers between biologists and computer scientists have to be overcome. This is needed when a framework for the simulation of evolution is to be developed. Still, often only actual work with the system will show the next steps in development. It was no different with the development of EEPSLION.

The cathedral and the bazaar

Generally, there are two extreme approaches to developing complex software. One is very hierarchical, thoroughly analysing the problem, building a solid design and then implementing it. This approach, usually taken by companies that develop a large application, has also been compared to a cathedral: a good plan and many people are needed to build it[2]. The other way is very informal, quick, and easy, at least for small projects. Much of the open source movement uses this approach that has been compared to a bazaar. While the latter works best when many people can make many small contributions that are usable units in themselves, the former is usually

1. Booch (1994) "Object-oriented analysis and design. With applications". 2nd, Benjamin / Cummings Publishing Company. - Jacobson et al. (1998) "Object-oriented software engineering". 11th, Harlow, England, ACM press. - Bruegge & Dutoit (2000) "Object-oriented software engeneering: Conquering complex and changing systems", Upper Saddle River, NJ, Prentice Hall.
2. Raymond (2001) "The cathedral and the bazar" http://www.tuxedo.org/~esr/writings/cathedral-bazaar/

adopted for large complex software projects[3]. Why am I discussing this here? Not because I started an open source programming project[4], but simply because in the development of EEPSLION I had constantly to choose between two options:

- Improve design of the framework for future applications.
- Develop code for the next application.

Over the years I developed a pragmatic approach using both programming paradigms. While paying careful attention on how to design the framework for maximal long-term usability, the actual implementations were rather bazaar-like exercises in Extreme Programming[5] to meet various deadlines.

Thus over the years EEPSLION experienced 4 design-implementation cycles and completion of this thesis will be a milestone in the 5th design phase. Before the current design is presented in the next chapter, some lessons from the previous four designs are reviewed. One general pattern emerges behind all these lessons: each implementation actually offered more possibilities than were conceivable at the time of its design. However, actual work with it quickly showed limitations that required fundamental design changes. These fundamental issues are discussed in the next sections.

9.1 Design 1: Special language (1997)

Since the beginning, removal of bottlenecks in modelling evolutionary processes has been a key vision of the whole enterprise.

The first design aimed at implementing a new programming language for the simulation of evolution. The picture in mind was that biologists could use this language to implement evolutionary models easily. The language should also allow non-specialists to understand the model, simply by reading its implementation code in this new language.

A prototypic implementation of this language was done by Stephan Peter in 1997[6] using FLEX and BISON, two powerful tools for building compilers. The program he developed would read code describing a simple evo-

3. Stevens (2001) "C Programming: It's good work when you can find it", Dr. Dobb's J. May:121-124.
4. EEPSLION might one day become open source, when the design has reached enough stability for release. Currently, however, it is not open source.
5. Bleul (2001) "Programmier-Extremisten: Softwareprojekte auf den Kern reduziert", c't Heft 3:182-185. - Elting & Huber (2001) "Schnellverfahren: Mit Extreme Programming immer im Plan?" c't Heft 3:186-191.
6. Peter (1997) "Prototypischer Entwurf und Implementierung einer Simulationssprache für evolutive Vorgänge in der Biologie", Lehrstuhl für Systemanalyse, Fachbereich Informatik, Universität Dortmund.

lutionary process and automatically translate it into C++ code that could be compiled with standard tools. By running the resulting binaries, evolution was simulated while recording important quantities that had to be analysed later with an external visualisation application like gnuplot. It was great to have all the details of a simple evolutionary model[7] and play around with them in search of better understanding.

Biologists don't program

Besides being only deterministic[8], one fundamental flaw pervaded the design: biologists usually do not program. Those who do usually already use some standard programming language and will be very reluctant to learn a new one. Thus the goal to make implementation of evolutionary models easy for biologists in general was not achieved. If someone who wanted to use this new tool would have to know how to program anyhow, it would have been easier to build a class library for some widely distributed, well-accepted general purpose language. This makes development much easier, because one does not have to reinvent the wheel and add the general purpose programming features to a new language. Thus from now on the design of models will have to be left in the hands of computing biologists[9].

Realizing this required a fundamental change in the design, and efforts were refocused on building a good class library.

9.2 Design 2: Class library (1998)

The next design goal was to build a C++ class library that facilitated stochastic individual-based simulations of evolution in general and of Muller's ratchet in particular. The design assumed that a good system of classes that describe biological reality in object oriented language was the only important thing.

Such a class library[10] was implemented by myself (like all the following designs described here) and used to run some initial simulations of Muller's

7. After the model and simulator ID terminology was introduced (see current design chapter), this model was given the name "S001" for historic reasons. It is currently not under active development and has never been published.
8. This first system was so focused on adaptive evolution, that it neglected stochastic effects like random drift. Thus only deterministic simulations were possible and analyses of the first and last stages of an adaptive event were as impossible as analysis of very small selection coefficients. Thus not only random number generators, but also all repeat-management infrastructure was lacking.
9. Which is a severe constraint, as such people are rare, see Marshall (1996) "Hot property: Biologists who compute", Science 272:1730-1732.
10. After the model and simulator ID terminology was introduced (see current design chapter), this model was given the name "S002" for historic reasons. It is currently not under active development and was never published. However, parts of this code were used for development of Simulator005.

ratchet in human mitochondria[11]. It was great to interactively explore individual parameter combinations and their effects.

However, while working with the system a major practical drawback was discovered. It is not enough to have classes that model biology. It is important to have a good simulation infrastructure, too. If this is lacking, trivial issues like parameter input, scheduling and repeating simulations with only slight changes in one parameter become a major administration task.

Need for a simulation framework

Thus the object-oriented class library for individual-based models had to be complemented by an appropriate framework for conducting the actual simulations. This is more than just adding another few classes. As Erich Gamma and others have pointed out[12]:

- Development of **applications** is difficult. You just need to think about how to solve one specific, usually non-trivial problem.
- Development of a **class library** is more difficult, as you have to make it useful for application to many different problems of the same underlying type.
- Development of a **framework**, however, is the most difficult. It means to develop a problem-solving environment that is not only applicable to many different problems of a field, but also facilitates integrated solutions, either by already providing a solution or by allowing development of such a solution. While good applications and class libraries also mature over the years, frameworks build a growing foundation upon which many detailed solutions can be developed.

The decision was made to develop such a framework.

9.3 Design 3: Framework (1999)

The goal was to develop a framework for individual-based simulations of evolution that helps with the practical simulation aspects too. It should be script-controlled and facilitate bundling of different evolutionary models into one code basis to facilitate comparisons.

Such a framework was developed, although actually only one model was put into the bundle. The resulting code[13] was used for simulations of Muller's ratchet that incorporated advantageous mutations and compensa-

11. Loewe & Scherer (1998) "Muller's ratchet in human mitochondrial DNA". Poster presented at the Sixth Annual International Meeting of the Society for Molecular Biology and Evolution, 17-20 June 1998, University of British Columbia, Vancouver, Canada.
12. pp. 30-32 in Gamma et al. (1996) "Entwurfsmuster. Elemente wiederverwendbarer objectorientierter Software". 1, Bonn, Addison-Wesley.

tory mutations[14]. The ease that allowed scheduling of whole series of simulations by relatively simple scripts was an incredible experience compared to manual scheduling of simulations earlier.

Need for n-dimensional parameter space searches

However, several drawbacks were observed while working with the resulting application. Although it is great to have a facility that schedules a series of simulations where everything is constant except one parameter, practical work consisted largely of scheduling series of such 1-dimensional parameter grid searches to obtain a feeling of what happens in n-dimensional parameter space. Therefore it was decided to design a facility that schedules searches of at least 5 parameter dimensions at once. It was found later that a good design would allow scheduling of n-dimensional searches as well.

Need for more computing time

This brings us to the next problem that begged for a solution: computing time. Scripts with a few 1-dimensional parameter searches could easily block the computer for months, if the "wrong" parameter was picked. This is a problem, as the "wrong" parameters could not always be easily predicted. For example to observe actual extinction times, a script was written to examine three different population sizes under a given mutation rate for various selection coefficients. Figure 11 shows the resulting extinction times along with the corresponding computing times.

The pragmatic solution many computing biologists apply[15] is just not to consider such computer-intensive parameter values like population sizes of 1 million individuals. This, however, keeps *in silico* biology under tight computing time constraints and severely limits the types of questions that can be addressed. Furthermore, part of the constraints is artificially imposed by the way simulations are scheduled.

To keep computers busy can keep you busy

The simplest way of scheduling individual simulations is just to use a big loop that automatically starts one simulation after another, collects the results requested on the fly and presents them to the user at the end. This is

13. After the model and simulator ID terminology was introduced (see current design chapter), this code was given the name "S003" for historic reasons. It is currently not under active development and was never published. However, parts of this code were used for development of Simulator005.
14. Loewe & Scherer (1999) "How many beneficial mutations stop Muller's ratchet?", Poster presented at Evolution '99, 22-26 June 1999, University of Wisconsin, Madison, Wisconsin. - Loewe & Scherer (1999) "How many beneficial mutations are needed to stop Muller's ratchet?", Talk delivered at Seventh Congress of the European Society for Evolutionary Biology, 23-28 August 1999, Barcelona, Spain. - Loewe (2000) "How many beneficial mutations are needed to stop Muller's ratchet in mtDNA?", Talk delivered at Spacial Ecology Programme: Workshop on Population Extinction, 5-7 November 2000, Tvärminne Zoological Station, Finland, Division of Population Biology, University of Helsinki.
15. For an example see Mooij & Boersma (1996) "An object-oriented simulation framework for individual-based simulations (OSIRIS): Daphnia population dynamics as an example", Ecol. Model. 93:139-153.

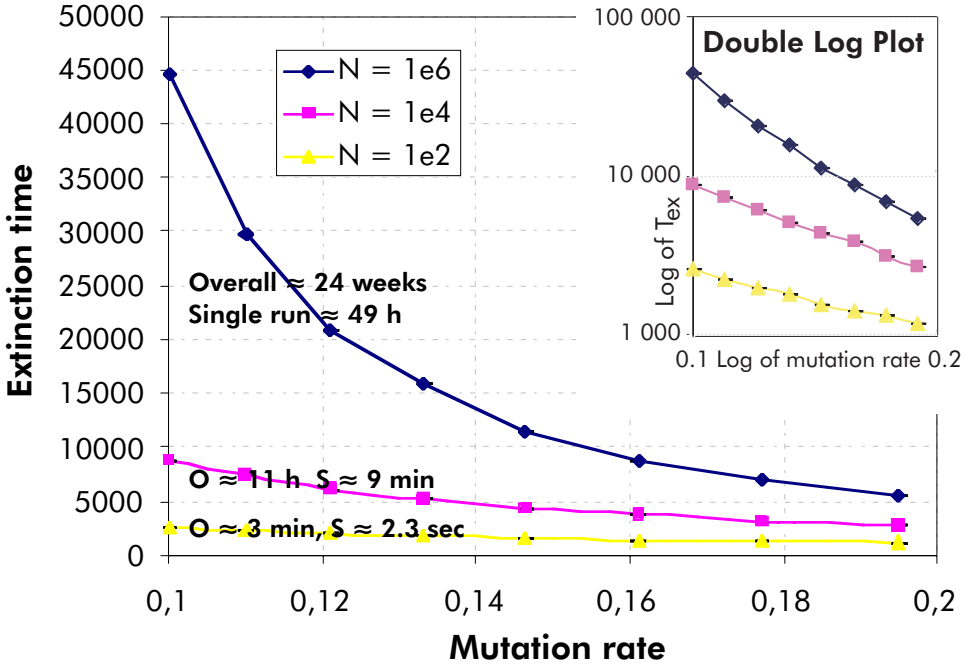


Figure 11 Only slight variation of parameters can lead to dramatic differences in runtime. Computing times of individual simulations (within the same population size) were proportional to extinction times observed. The overall computing time for each population size is given along with the corresponding average computing time for a single run. This graph comprises 240 single simulations in total and shows extinction times (in generations) of three populations (100, 10 000, 1000 000 individuals) for various mutation rates (per generation) assuming a selection coefficient of -1%. Each parameter combination was simulated 10 times. The coefficient of variation for extinction times was 4%-7% with a mean of 5.3% and no observable trends.
Simulations by Simulator003 on a PowerMac G3@350MHz with 1MB 2nd-level cache (speed 1:2).

fast and fine for tasks in the seconds or minutes range and can be tolerated over a weekend or so. However, scheduling parameter searches that need months to complete becomes a torture if one is keen to know the results. Such a situation frequently occurs, despite a reasonable length of individual runs. In the example of Figure 11, the first 3 parameter combinations with 10^6 individuals needed more computing power than all the other 21 parameter combinations. Furthermore, they are repeated 10 times, just like all the others. While this is easy to code (just nested loops), this is clearly inappropriate for efficient work. To use computing resources efficiently, simulations that need seconds can be repeated more often, while it is often enough to wait for one single run of a complex simulation to decide what

should be simulated next. Only publication quality results demand more repeats, but these are usually generated when most of the initial exploratory work has already been completed. Thus most *in silico* experiments with individual-based models are limited to 'reasonable' computing times or demand that most of these complex runs are scheduled manually, interrupted, preliminarily analysed, rescheduled and so on. Porting the code to a super-computer does not really help, as the single nodes of super-computers are (very roughly) as fast as CPUs in modern day PCs, except that they have to be shared with many others. Furthermore, system administrators will not be very happy if you exceed your computing time quota, a potentially difficult situation for simulations of processes with poorly understood timing. Thus some form of parallelisation had to be supported by the framework.

Need sophisticated scheduling of simulations

While parameter space searches for individual-based models belong to the embarrassingly parallel problems, and are thus easily parallelised in principle, actual implementation work should not be underestimated, as the problem of dynamic scheduling and re-scheduling of parameter combinations is not solved by mere parallelisation. All this brings us back to design questions in framework development: we need a sophisticated system for dynamic scheduling of parameter combinations that is practical enough for daily work.

Need for global computing

Furthermore we need much more computing power. While one might argue about whether it was actually reasonable to devote 6 months of computing time to the exact parameters simulated in Figure 11, one can not argue about the fact that currently all individual-based simulations of evolution hit the computing time barrier for some interesting parameter combinations. This situation appeared to be a perfect task for global computing. As each parameter combination can be simulated completely independently of all others, it should be worth the effort to distribute parameter combinations over the Internet and to collect results. Later success of evolution@home proved these expectations to be right. Two other lessons were learned during this cycle.

***One* model *one* simulator doctrine**

It unnecessarily complicates things if more than one model is packed into one code-basis that leads to one executable. As each model has its own evolution (repeated development and testing), it makes work much easier if one follows the *one model one simulator doctrine*[16].

One simulation one result doctrine

Design 3 allowed me to choose the parameters that should be reported when scheduling simulations. This is an advantage compared to having to choose which parameters to report at compile time or compared to writing everything to disk and afterwards drowning in Megabytes of irrelevant data. However, when a simulation was scheduled to report, say, extinction time as in Figure 11, then sometimes it would be interesting afterwards to look at eg. clicktimes or some other output parameter. This, however, was not possible with Design 3. It would require re-running the whole set of simulations. While no problem for simple runs, this is clearly intolerable for complex simulations. Therefore it is desirable to bundle all potentially interesting output parameters of a single run into one results file that is stored for future analysis. While this *one simulation one result doctrine* requires more effort before a simulator can be released to compute lasting results (find and define all interesting parameters), this disadvantage is more than compensated for by the flexibility of later analyses. However, such a feature imposes a new demand on the framework: manage collected results.

9.4 Design 4: Global computing (2000)

Importance of physical space for large designs

To conceptually remove these shortcomings a fundamental redesign of the whole framework was required. This took more than 2-3 months uninterrupted work in a dedicated 70m^2 room[17] that allowed for enough space to lay out enough A2 sheets describing the various design aspects to get an overview over the multitude of relationships involved in scheduling and analysis of individual-based simulations of evolution. Many of the foundations of the current design presented in the next chapter were laid in that period. Enough physical space to lay out all ideas on large sheets of paper and thus make fast and easy connections between diverse topics, proved to be critical in the design of this complex system.

Design 4 was never implemented to the degree that it could actually be used as a whole for simulations[18]. Nevertheless, enough was implemented

16. This was the decision I made after analysing the various possibilities that allow comparison of different models. The 3rd design actually contained some infrastructure for including additional evolutionary models, however, this was never used as in the long term it is much more feasible to have many clear defined applications each implementing one clear defined model with a clear defined release number and not a few 'mega'applications that contain some complex collection of models.
17. See acknowledgements in appendix for a picture.
18. After the model and simulator ID terminology was introduced (see design chapter), this codebasis was given the name "S004" for historic reasons. It is currently not under active development and was never published. However, parts of this code entered into the current Simulator005.

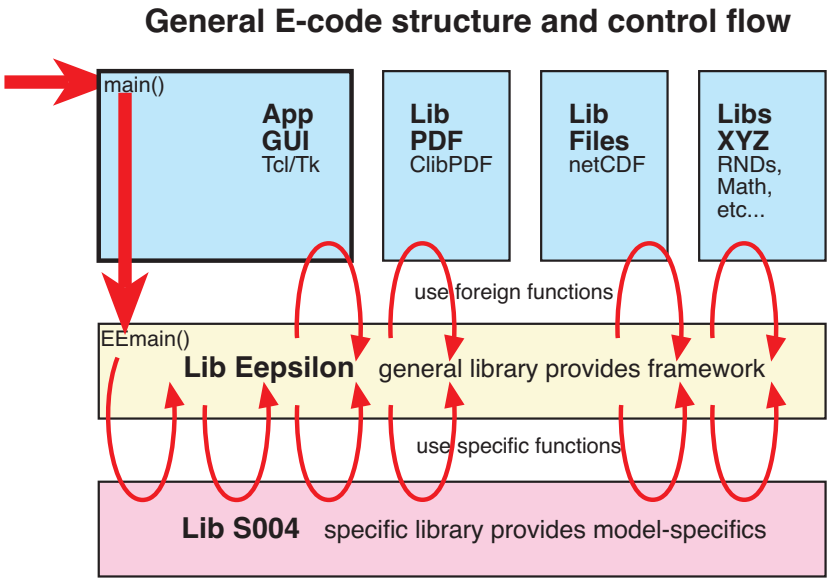


Figure 12 Libraries linked together to build a simulator in Design 4.
Please note that this is a very rough overview over the *internal* library structure of *the one* program that does it all (from server to client to analyser etc., see Figure 13 for an rough overview of functions.) Only simulators for public distribution would be shrunk down by leaving out the server parts.

to see its shortcomings. These will be discussed as they confer important lessons before we go on to discuss the current design in the next chapter. For the remainder of this discussion it may be helpful to have a short look at two of the important overview figures that were developed in that time. Figure 12 shows the various libraries that are linked into an application that can perform any of the modes of operation depicted in Figure 13.

Do not link everything into one application

It is nice to have a compact application like the one produced by the collection of statically linked libraries in Figure 12. While it can be easily installed, moved or deinstalled, things get difficult if functionality of that application grows beyond a certain complexity. Then the overhead to keep it all in one application becomes unbearable, especially if several instances of the same application have to work together, each on a different task to build a large data-processing network like in Figure 13. Thus the question arises, why not divide the functionality into different applications that act like modules performing only one very special task each? This task would be well defined and can be easy, but makes sense only as a part of the larger network. Keeping track of further development of the system is much easier with such an approach[19]. Therefore, the idea of linking everything into one application binary was abandoned. Rather, a network of applications would be employed, as in the current Design. Extremely complex software products like modern operating systems or SAP take the same approach[20].

Do not store all results in one file

Another point where Design 4 needed to be "decomposed" was the collection of results. As can be seen in Figure 13, all results would eventually end up in one big single-run results file (srr-file). While this is no problem for small simulation tasks, extensive gridsearches can produce enormous amounts of data and easily surpass the 2GB file limit of most operating systems. Now, while one could go to a 64bit operating system that does not know 2GB limits, another problem still persists:

As there is no order in the big srr-file of Design 4, fast searching for duplicates becomes a problem. This is needed to exclude double submissions from further analysis, as they do not constitute independent observations that contribute to statistical quality of the results. While one could find a way to work around that, such a design has still another flaw: the various pieces of information regarding a particular parameter combination are widely scattered in the various files of Design 4. This means that, given a particular parameter combination, it is not easy to

- find all (or at least one) repeats of that run
- find the multi-run-result that summarises all repeats
- find the corresponding run-file that led to their computation
- find additional analyses (in form of extra pdf-files, eg. a time series)
- find the place where scheduling priorities are set for this file (in order to eg. ask for more repeats of this combination)

While ways could be found to search for all these individually, it would be much more elegant if all information concerning one parameter combination could be found in one folder. Thus manual evaluations become possible and setting correct priorities is facilitated by an up to date overview of all results available for the corresponding parameter combination. Furthermore, it is not likely that results for an individual simulation run come even close to current limits of file size. The sum of all files (whose number easily explodes for extensive parameter space grid searches) is thus no longer limited to maximal size of an individual file and as the corresponding collection of folders can be spread over several hard disk drives, it can grow quite easily

19. Having version numbers for each program and making a general system release from time to time is simpler than having one big program with various internal modules compiled into it. Slight changes in one part would require a new release of the whole system although nothing else has changed. This can generate more confusion than necessary in such a complex system.
20. Size of executables of SAP for example has grown into Gigabytes, as one module after another was added in a similar approach. The whole SAP system is organized into modules that consist of transactions. Each binary is capable of conducting one (or a few) transactions. Each transaction is uniquely identified by an ID. It performs a relatively simple operation upon some well defined interfaces. This simple concept has allowed SAP to grow to its current size.

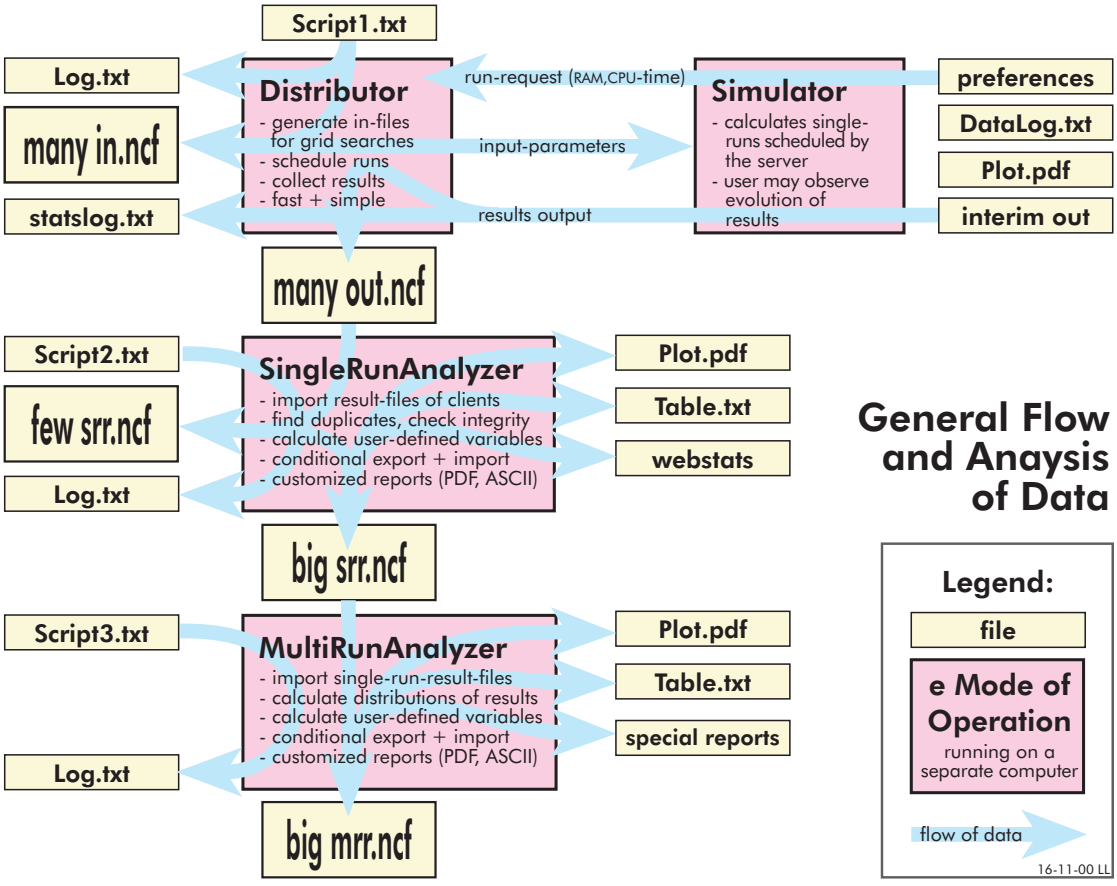


Figure 13 General flow and analysis of data of Design 4. The various modes of operation would run in different instances of the program controlled by the corresponding scripts.
Abbreviations: **in** = file with input parameters = run-file; **out** = file with output parameters = individual results-file; **srr** = single run results file (here a big collection); **mrr** = multi run results file (here a big collection); **ncf** = netCDF-file, a special file format for storing large multidimensional arrays of data; **pdf** = Adobe Portable Document Format; **txt** = ascii (tab) text-file; **statslog** = elementary CPU statistics: who computed what single run; **webstats** = published html summary on who calculated how much; **Log** = computer generated file that records the programs actions; **Script** = small manually generated file that tells the program what to do; **big** = so large that the 2GB file limit of most operating systems could become a problem; **many** = millions and more, depends on details of grid search.

according to current needs. However, to keep that manageable, a number of automated tools for administration of such a database of folders are needed.

Databases

Why not use one of the many relational databases available? At the time of writing the first draft of this chapter (Oct 2001) it was found that databases that were readily available were not built to cope with such large

amounts of data and learning how to operate a database and integrate it into the automated work flow only to hit some barrier that is probably much less than the current 2GB file size limit is not worth the effort. Those databases that were built for handling large amounts cost huge licence fees and needed a specialised system administrator due to their complexity. While typical databases have a large overhead of functionality not needed for management of simulation results, none of them substitutes the need for a carefully designed workflow or specifically designed applications that perform analyses. Therefore the decision was made to use the folder-based system described in the current Design 5. This situation had changed by the time of proofreading this chapter (Oct 2002), because SAP-DB was released into the open source community and the power of MySQL has increased significantly[21]. Their ability to handle Terabytes of data at no licensing costs calls for an earnest re-evaluation of this issue before starting to implement Design 5.

Tcl/Tk and visual results analysis

One final comment on the use of Tcl/Tk. It was developed as a glue to connect different applications under UNIX and is an excellent script language with a comfortable graphical user interface (GUI)[22]. It is available for Windows, MacOS and Unix. One of its strengths is that it can be used from C++ by embedding a statically linked Tcl/Tk library in the code that is called at runtime with Tcl code to be executed on the spot. This makes it easy to implement simulators that are simple to install. Therefore, it was chosen to be included in Design 4 code to provide graphical capabilities, Internet connectivity and the ability to evaluate regular expressions (needed for results analysis). One of the major developments for evolution@home in 2002 was discovery and integration of the open source R-system[23] into work flow. This had become unavoidable, as efficient analysis of more than

21. SAP (2002) "SAP-DB Homepage" http://www.sapdb.org/ - For a glowing review see eg. Banke (2002) "Open-Source-RDBMS von SAP: Erfaphrungswerte", iX Aug.:78-80. - See also http://www.mysql.com/
22. Ousterhout (1994) "Tcl and the Tk Toolkit", Reading, MA, Addison-Wesley. - Ousterhout (1995) "Tcl und Tk: Entwicklung grafischer Benutzerschnittstellen für das X Window System", Bonn, Addison-Wesley. - Welch (2000) "Practical programming in Tcl and Tk". 3rd, Upper Saddle River, New Jersey, Prentice Hall. - Harrison & with other contributors (1997) "Tcl/Tk Tools". 1, Cambridge, O'Reilly & Associates. - Harrison & McLennan (1998) "Effektiv Tcl/Tk programmieren". 1, Bonn, Addison-Wesley. - Nelson (2000) "Tcl/Tk Programmer's Reference". 1, Berkeley, CA, Osbourne/McGraw-Hill. - Raines (1998) "Tcl/Tk kurz und gut". 1, Cambridge, O'Reilly Verlag. - Raines & Tranter (1999) "Tcl/Tk in a nutshell. A desktop quick reference". 1, Sebastopol, CA, O'Reilly & Associates. - Webster & Francis (2000) "Tcl/Tk für Dummies. Gegen den täglichen Frust mit Tcl/Tk". 1, Bonn, MITP-Verlag.
23. R-system (2002) "R: A Language for Data Analysis and Graphics" http://www.r-project.org/ - Maindonald & Braun (2002) "Data analysis and graphics using R: An example-based approach", Cambridge, Cambridge University Press. - Ihaka & Gentleman (1996) "R: A Language for Data Analysis and Graphics", Journal of Computational and Graphical Statistics 5:299-314.

22000 simulation results was no longer possible with Excel, the original simple ad-hoc choice for results analysis. While this change took several months of additional work, it brought a level of automation and power in analyses that one could only dream of from inside of Tcl/Tk (or Excel). The fact that additional modules for R facilitate data exchange with professional databases suggests a very powerful solution for future simulation data handling.

Nice, if small can become horrible, if big

Furthermore, practical work with Tcl scripts has shown that one major drawback is its lack of compile-time checks. While the C++ compiler would check for all C++ errors, it can not check the Tcl scripts included as strings. These either produce an error, when interpreted at run-time or even worse, cause errors to appear later. One particularly annoying behaviour is that you do not need to declare variables in Tcl, as this is done automatically. Thus, each typographical error in variables of Tcl scripts led to errors that are sometimes hard to find – not to mention the additional compile cycles. This is a good example for an experience made over and over again in development: systems that are comfortable when small often make trouble when large.

As the one big overall application that has everything compiled into it is being split in Design 5, Tcl loses its central position as general entry point (see Figure 12). Currently, Tcl/Tk is not used in evolution@home and wxWindows[24] appears to be a better choice for development of graphical user interfaces for simulators. Before discussing the current Design 5 of EEPSLION, a few more decisions should be explained.

9.5 Why use C++?

The choice of development language and platform fundamentally influences the system. What then would be the best language for development of a framework for the simulation of evolution? The following criteria were important in the choice:

Important features

- **Object-oriented**. This is the best approach for building individual-based models. After all, the idea for object-oriented programming and inheritance etc. came from looking at biology.
- **Modern**. It should offer the world of modern programming, as it is hard to predict what constructs will be needed over time. Therefore, no artificial restrictions (like no pointers, no templates) should apply.

24. wxWindows community et al. (2002) "wxWindows Home: Cross-Platform GUI Library" http://www.wxWindows.org/

- **Tested by time**, to avoid peculiarities of an immature system.
- **Widely used** by others. Thus, a reasonable number of libraries and development tools are available already.
- **Cross-platform**. As development of EEPSLION is a long term enterprise, it will see a considerable number of systems come and go over the years. Therefore, its foundation language should be one that is as independent as possible from any particular system.
- **Fast**. Suitable for high demand scientific number-crunching.
- **Super-computer option**. While most of the work will be done on ordinary PCs, the possibility of investigating a small set of extraordinarily complex simulations on super-computers should not be easily thrown out. Only a few code changes should be needed for running on a supercomputer.

It is obvious from this list that all script languages (that have nice GUIs like Tcl) and interpreted languages will not be fast enough to meet the criteria. While Java (i) offers nice GUIs, (ii) is modern, (iii) reasonably stable, (iv) has recently increased in performance by Just-In-Time compilers, and (v) is platform independent, it does not offer as many options as C++. It is also not available for most current super-computers. On the other side, one of the traditional super-computing languages, Fortran, is not the language of choice for modern object-oriented programming.

C++ offers most possibilities

Therefore it was decided that standard ANSI/ISO C++[25] would be the best long-term choice for development of EEPSLION. It is regularly used for scientific, engineering and even biological applications[26] and offers more possibilities than anyone could ever employ. Admittedly, some of them are dangerous programming constructs in the sense that using them can produce errors that are hard to find. However, a programmer can decide to refrain from using such possibilities or learn how to use them properly[27], when there is no other possibility of solving a particular problem. The availability of the Standard Template Library makes otherwise complex tasks

25. Stroustrup (1994) "Design and Evolution of C++", AT&T Bell Labs. - Stroustrup (1998) "Die C++ Programmiersprache". 3., aktualisierte und erweiterte, Bonn, Addison-Wesley. - Isernhagen (2000) "Softwaretechnik in C und C++: Modulare, objektorientierte und generische Programmierung". 2. korrigierte, München, Carl Hanser Verlag. - Lippman (1993) "C++ Primer". 2, Reading, Massachusetts, Addison-Wesley. - Schader & Kuhlins (1996) "Programmieren in C++: Einführung in den Sprachstandard". 4. neubearbeitete und erweiterte, Berlin, Springer Verlag.
26. Smith (1999) "C++ Toolkit for Engineers and Scientists". 2nd, Heidelberg, Springer. - Wilson (2000) "Simulating ecological and evolutionary systems in C", Cambridge University Press.
27. Hyman & Vaddadi (2000) "Effektive C++-Techniken". 1, Bonn, Galileo Press. - Meyers (1995) "Effektiv C++ programmieren". 2. korrigierte, Bonn, Addison-Wesley.

very easy[28] and a reasonable number of books describe important algorithms in C++ or in its related ancestor language C[29].

As most modern programming languages enable one to call code from another language, the decision for C++ does not mean that other languages are principally inaccessible; they might still be employed, if needed.

9.6 Importance of standards and tools.

Codewarrior is an excellent IDE

For efficient development an integrated development environment (IDE) is needed. It allows one to easily edit code, compile, find compile-time errors and debug run-time errors just as the widely distributed Visual C++ product of Microsoft for Windows. However, as code should be as platform independent as possible, the Metrowerks Codewarrior was employed. It offers the opportunity for easily recompiling code for other platforms. This feature was used to develop under Codewarrior 5 and 6 on a PowerMac G3 and then cross-compile console-based code for the Windows platform.

Redefinition of numerical types

One of the strengths of Java is its cross-platform definition of numerical types. In C++ the exact number of bytes used by an `int` depends on the implementation. As this can lead to potential problems in code portability, EEPSLION redefines basic numerical C++ types according to the number of bits they represent. Thus possibilities for future portability of the code are increased.

Use of prefixes

To write one-way code is easy (write, use once and never understand again). However, when code needs to be usable over many years, it should be written in a way that allows other people to read and understand it - a fact more well known in industry than in university. Readable code means extensive explanations have to be included and *only* speaking names with prefixes are to be applied. The prefix indicates the type of a variable and the speaking name should explain its meaning. Speaking names are as long as needed and as short as possible, but without weird abbreviations. Furthermore, all global names get the additional prefix "`EE`" to easily identify them and avoid name conflicts with libraries that might be included later. While it takes some time to get used to give each variable the right prefix, writing code this way is not much slower in practice. However, maintenance is much easier. The current system of prefixes can be found in Table 7.

28. Josuttis (1996) "Die C++ Standardbibliothek", Bonn, Addison Wesley.
29. Sedgewick (1992) "Algorithmen in C++", Bonn, Addison-Wesley. - Press et al. (1992) "Numerical recipies in C". 2nd, Cambridge, Cambridge University Press.

Use of libraries

Finally, it is not feasible to implement everything from scratch. For some problems, excellent external libraries have been developed for C or C++ and either can be included in C++ code. To use libraries, however, a trade-off has to be made between the cost of employing a library[30] and the cost of implementing from scratch. This is discussed where needed.

30. Find potential candidates, learn enough to see whether they do what you want and check whether they place fundamental constraints on your code (eg. platform dependence etc.). Hear what others say and solve licence issues.

Name prefixes in EEPSLION-code

current as of 01-08-01

variability +		visibility +		category +	
"c"	constant	---classproperty---		o	object
""	variable or enum	s	static	p	pointer
		""	nonstatic or enum	a	array
		+		r	reference
		p	private	""	enum
		h	protected		
		u	public		
		---parameter------			
		r	by reference		
		v	by value		
		---variables--------			
		l	local		
		m	module		
		g	global		

General rules:

- EE is prefix for every *global* class, function, variable, enum, macro, ...
- e is prefix for *very few global key variables* (like eCEO, eWorld and few others)
- all others have as prefix: *variability-visibility-category-typecode*
- Use Get/Put for access functions (not Set!)

Classnames:

baseclass	EESpeakingClassName
derived class	EEParentChildSpeakingClassName
function	EEtypImportantGlobalFunctionName

Simulatorspecific library

EE000 =	simulatorspecific, needed by framework
EE### =	specific to S###, unknown to framework

EE000SpeakingClassName
EE000typImportantGlobalFunctionName

Examples

EEenu	global enumeration list name
EEeni	global enumeration list item
croxxx	parameterobject, by reference, constant
lacha	local array of characters = c-string

typecodes

ε **framework types**

mod	EEModeOfOperation
pro	EEProgressIndicator
rnd	Random Number Derivate
fil	file related
err	error related)
ipr	EEInterpreter
cmd	command related
par	EEParameter
pdc	EEParameterDataComplex
his	EEHistogramm (observed)
dis	pdc distribution
tim	pdc timeseries
snap	SNAPshot (of current world)
spsn	SPatial SNapshot
gra	plot related
bio	EEBioObject
bcs	Bioobject Current State
wrl	EEWorld (base or derived)
edt	EEDate (=f64)
gps	EEGlobalPositioningSystem
pos	EEPosition (in a gps-level)
ha	EEHabitat (base or ..)
loc	EELocation (base ...)
pop	EEPopulation (...)
popv	EEPopulationVector (...)
popl	EEPopulationList (...)
ind	EEIndividual (...)
indl	EEIndividualList (...)
indv	EEIndividualVector (...)
coh	EECohort (...)
gno	EEGenome (...)
chro	EEChromosome (...)
gen	EEGene (...)
seq	EESequence (...)
dme	EEDistributionMutationEffects

libraries & templates

---STL-types----------	
str	string
stm	stream (any type)
vec	STL-vector
lst	STL-list
set	STL-set
map	STL-map
ite	STL-iterator
---other libraries-------	
cpp	ANSI standard
net	netCDF type
pdf	ClibPDF type
ttk	Tcl/Tk related

elementary types

voi	void (function type)
enu	enumeration list
eni	enumeration item
boo	boolean
cha	char (7 or 8 bits)
i8	signed char
u8	unsigned char
i16	signed 16bit integer
u16	unsigned 16bit integer
i32	signed 32bit integer
u32	unsigned 32bit integer
...	
vli128	128bit very long int
f32	32bit (single) float
f64	64bit (double) float
---avoid the following:	
gui	GUI related
usr	any user type
lib	library xyz
eee	framework special
sim	simulator specific
xxx	unspecified type
uni	union type

No changes, only additions
write down here:

Table 7 Name prefix rules for making EEPSLION code readable.

10 Current Design 5 of the software framework *EEPSLION*

An overview of the current Design 5 of EEPSLION is presented. Reasons for important design decisions are discussed.

Before describing the design of *EEPSLION* (or for short ε), some general problems of parallelisation of individual-based models are discussed. As the efficiency of parallelisation critically depends on the structure of the problems, certain foundational decisions are unavoidable.

10.1 Specifics of distributed computation of individual-based models

Parallelisation of computation of individual-based models can occur at several levels. Some of these parallelisations are natural, some easy, some possible with a little additional work, or with much work, but some cannot be recommended in most cases.

10.1.1 Parallelisation of different models is natural

The most natural way of parallelisation in the simulation of individual-based models is to keep the models separate and run them on different computers. This may seem trivial, but when comparison of different models is an important goal of research, then one might succumb to the temptation of building code that tries to analyse several models on the same computer. As reported in the last chapter, the current design of ε adheres to the **one model one code basis one simulator** doctrine. Thus different models will *always* run on different computer. Extra effort to change this natural work flow does not pay off in most cases.

10.1.2 Parallelisation of different single runs is recommended

The next level of parallelisation does not come for free, but can be attained relatively easily, at least on a small scale: distribute the computation of different parameter combinations to different computers. While this can always be done manually, situations that require *many* such simulations can

easily become very challenging, as manual work can become prohibitive and a more automated solution is needed. To parallelise this level is one of the core goals of ε.

10.1.3 Parallelisation of time series analysis is recommended

One of the computationally expensive things in simulations of evolution can be their analysis. As described below, several types of data can be generated during a simulation and it is important to get a clear understanding of what data is to be analysed at what stage. For example, if each simulation records a precise time series of certain events or distributions, then one has two possibilities of analysing this:

1. Collect all the data and analyse them later.
2. Pre-analyse all the data of a simulation during the simulation and collect only results of this pre-analysis.

Poorly organized chaos ...

While the first possibility is cheap during code development, it may lead to undesirable effects afterwards. If all data of all simulations are collected, one will certainly drown in data, unless only a few, very simple simulations are conducted. Besides the need for a good system for organising the flood of data, it is often just physically impossible to collect all the data[31]. And even if it were possible, then what would be done to analyse them? Probably for each timepoint a distribution would be computed and the behaviour of each time series of distributions would be quantified somehow. All this would be done on a central computer, after simulations have been completed.

versus

... predetermined scheme of analysis

The second possibility just shifts the work and leads to a much more pleasant life after simulations have been completed: Each simulation con-

31. Just imagine a simulation with 1000000 individuals that runs for 1000000 generations, really no out of range parameters for evolutionary questions. If one is only interested in all fitness values, then this simple approach would lead to ca. 8 MB of double precision float values per generation, that mount up to 8 Terabytes, for the whole simulation. If 16-digit ascii representations were written to file, then even up to 20 TB would results - just for one simulation. Analysis of a reasonable set of such simulation results would exceed even the capacities of many super-computers of today.
Fortunately not every individual value may be really important, so the researcher decides to keep only the distributions of all fitness values for one generation. Such a distribution may be characterized at least by a mean, the number of observations, a variance, the minimum and maximum values and perhaps by some more numbers. Assuming only 5 floats, this makes about 100 MB in ascii text. Now just assume the researcher wants to check 3 parameters over a range of 5 orders of magnitude, something not uncommon in biology, then for each order of magnitude at least one value should be checked. Thus $5^3 = 125$ simulations result. As stochastic simulations should be repeated at least 3 times, 375 * 100 MB = 37 GB results would have to be analysed from this little sneak preview of parameterspace.

tains a high level of data analysing expertise packed into code, that analyses the data on the fly and keeps only those key values that are of interest. Thus Terabytes of data can be collapsed into a few important values. However, this does not come for free. It requires that good questions be asked about the model in the code and a sufficient number of exemplary tests be run to develop the model until it records all data relevant for later analysis. This is not always easy and might delay the start of simulations. However, if one manages to conquer the poorly organised parameter chaos before a distributed simulation project is started, then it is easy to use the distributed approach for distributed data analysis – as long as one does not need a central supplementary copy of the original data.

An important part of ε is to supply methods that help with the organisation and analysis of all these data in order to distribute the load to those CPUs that generate the data.

10.1.4 Parallelisation within a single run will rarely outweigh development costs

When thinking about complex simulations, many computer scientists often think about sharing the load between several processors[32]. However, for simulations of evolution all parallelisation beyond a SMP node does not make sense, except for the special case of hierarchical global computing for the simulation of multi-level population models (see below). Why is that?

Time can not be parallelised

All simulations of evolution have the nasty property that the later stages of evolution depend on the former. Therefore one can not take the simulated world history and divide it between several processors. If this were possible, one could already predict the details of evolution over time and would not be running simulations of evolution at all. Thus the only way where parallelisation might indeed help within a single run is to help large populations reach the next moment in simulated time faster.

SMP parallelisation is highly model dependent

However, this is not easily achieved and cannot be well automated from a framework perspective, as the computational complexity of the various levels of a biological model can vary significantly. For example, a simulation with only one big population of individuals could be parallelised in such a way that each processor computes the next timestep for a part of the population. However, in a spatially explicit model with many locations it might

32. SMP (symmetric multiprocessing) or massively parallel systems, or networks of workstations or one of the many other parallel computing concepts

be more feasible to divide the locations between CPUs. Thus SMP parallelisation is highly model-dependent.

Therefore, the design decision was made not to consider parallelisation within a single run at all from a Design 5 framework perspective. If a specific model needs it, it can be included in that simulator just as anything else that is needed for that particular model.

How to use additional CPUs

This does not mean that the additional CPUs of a SMP system can not contribute to the simulation of evolution. As the need for simulating different parameters well exceeds even the largest computational capacities, each processor can work on a different parameter combination. Thus it is up to the simulator part of the framework to detect additional CPUs and launch simulations there, if the user permits this. Alternatively, the user can install a simulator without such sophisticated capabilities on each CPU. Compared to the amount of time needed for SMP parallelisation, this is probably the most efficient overall approach.

10.1.5 Parallelisation of multi-level populations is desirable

Multi-level population dynamics pervade biology (eg. replicating mtDNA molecule drift and selection in replicating mitochondria drift and selection in replicating cells drift and selection in replicating individuals). Complex, little understood evolutionary processes occur here.

Hierarchical global computing will come

Such phenomena might be parallelised by hierarchical global computing. This means that each subordinate population within one of the higher-level entities is computed on a different slave CPU and only the results of that small evolutionary process are submitted to the higher-level master-CPU that waits for all low level results to come in, before the next high-level timestep can be started. This, of course, is only feasible with computers that have a high speed standing Internet connection.

Such parallelisation is desirable, as it significantly expands the range of biological questions that can be answered. However, it will take some time until the basics of ε are well enough developed for such an advanced approach to be integrated in the framework. Therefore, this topic is postponed to some future work.

10.1.6 Vectorization is hardly feasible

Often, when talking about computing speed, the topic of vector processors comes up. These impressive highly specialised machines achieve their legendary speed by overlapping a large number of similar instructions in long pipelines. Thus, they complete a relatively large number of operations with each CPU clock cycle. To use these advanced properties, code must be vectorised, something that is either done through a special vectorising compiler or manually.

Besides the fact that only few vector super-computers exist (where many other people want to compute, too) and the fact that vectorisation of code is by no means trivial, individual-based models (IBMs) often make this approach virtually impossible by their immanent characteristics[33]:

- IBMs usually contain large numbers of conditional statements. As vectorisation would typically compute all branches or at least clear the pipeline, performance is seriously slowed down.
- IBMs often move individuals around. Thus data locality is often destroyed and vectorisation less efficient.
- Vectorising compilers do not vectorise loops, if another (non-vectorised) function is called from that loop. This constitutes a problem, as the most important loop to be vectorised, would be the transition of many individuals to the next moment in time. Each individual, however, would typically call a random number generator (containing if-statements) and then decide what it would do.

While one might find workarounds for these issues in particular cases, it is probably rarely worth the effort. Thus evolutionary biologists do not compete with engineers for those costly moments on a vector machine.

10.1.7 Work units have extremely variable sizes

One of the striking characteristics of IBMs is their tremendous variability in terms of computational complexity. Simple parameter combinations might complete in split seconds and need less RAM than a modern CPU has in its cache. The most complex ones, however, could run for so many years, that it would be actually faster to wait for Moore's Law to make computers faster and start such a simulation on some future computer, than to start it

33. Haefner (1992) "Parallel computers and individual-based models: An overview", pp. 126-164 in: DeAngelis & Gross (eds) Individual-based models and approaches in ecology, New York, Chapman & Hall.

on a system today and run it for more than hundred years on a computational dinosaur of the future. Questions about evolution have a nasty habit of involving many individuals with many genes over large periods of time.

Between these two extremes, *in silico* evolutionary biology has to find its way. Researchers have no other choice than to stick to those parameter combinations that are computable. All others have to be left out, regardless of whether they are interesting or not. This arbitrary way of deciding when and where to stop simulation, as well as the extreme heterogeneity of computing times, place special demands on the framework that schedules such tasks.

10.1.8 Even incomplete runs might be used

One of the points that is often decided more or less arbitrarily is when to stop a simulation, especially for runs with many individuals. However, as simulations usually do not just wait for some important event to happen, but rather record time series, a new possibility arises: incomplete simulations might be used. If each time series is correctly reported in the results, then an incomplete run might convey an initial impression for its parameter combination. Based on such information further scheduling decisions can be made.

Developing check-point functionality is not trivial

This has practical relevance at least for situations, where a simulator is not developed to the point where it offers a check-point feature that writes the whole state of the model to disk to continue evolution after an interruption. While support for the use of incomplete runs is relatively easy from a framework perspective, implementation of check-pointing feature is a considerable challenge: (i) a framework should help with the implementation for a great variety of different models, (ii) a framework has to make sure everything is indeed correct after restoration and (iii) a global computing framework has to check for cheaters, who might write one advanced check-point file to disk, and then repeatedly start computations from there to gain a prominent place in the high scores. As simulations can be run without that feature, development is postponed due to time constraints.

10.2 Overview over *EEPSLION*

Now that the basics have been explained, the journey into Design 5 of EEPSLION can start. Occasional reference will be made to the design patterns explained by Gamma and his colleagues[34]. EEPSLION is an acronym that explains what it is (see Table 8):

Table 8 Meaning of the name of ε.

E	Evolutionary
E	Ecological
P	Process
S	Simulation
L	Language
I	Interpreter
O	Organising
N	Network

ε is a network that organises interpreters for script-languages that facilitate the simulation of processes from ecology and evolution. In short, a network of scriptable applications. Before we go into the details of how these applications interact, we shall look at the overall work flow used by a computing evolutionary biologist who works with ε (Figure 14).

Evolutionary bioinformatics at work

Everything starts with some idea about evolution. This idea is developed to the state where it can be formulated into a precise individual-based model. Then this model gets an ID. This ID is centrally allocated by the Evolutionary-Research initiative[35] to make sure that each model developed for evolution@home has a unique ID[36]. This greatly facilitates comparison of different models. Thus,

Definition of SimulatorID ReleaseID

> *One model with one fixed list of input and output parameters defines one* **simulator**. *Any given simulator in* ε *and in evolution@home is identified by its* **SimulatorID** *and* **ReleaseID**.

The model is then implemented with the help provided by ε. Several rounds of testing, preliminary results analysis and code improving bring the simulator to the point where it can be released. This point is reached when *all* parameters that are to be (*and ever will be*) included in results have been defined and validated[37]. It will not be possible to change definition of parameterspace after the first release of a simulator, as this endangers comparability.

34. Gamma et al. (1995) "Design Patterns", Addison-Wesley. - Gamma et al. (1996) "Entwurfsmuster. Elemente wiederverwendbarer objectorientierter Software", Bonn, Addison-Wesley.
35. This is the initiaitive that develops EEPSLION and runs evolution@home. See http://www.evolutionary-research.net
36. While development of various models will be opened to other computing biologists after the framework has become stable enough, the allocation of IDs will remain under central control. Under no circumstances shall the same ID be used for different models.

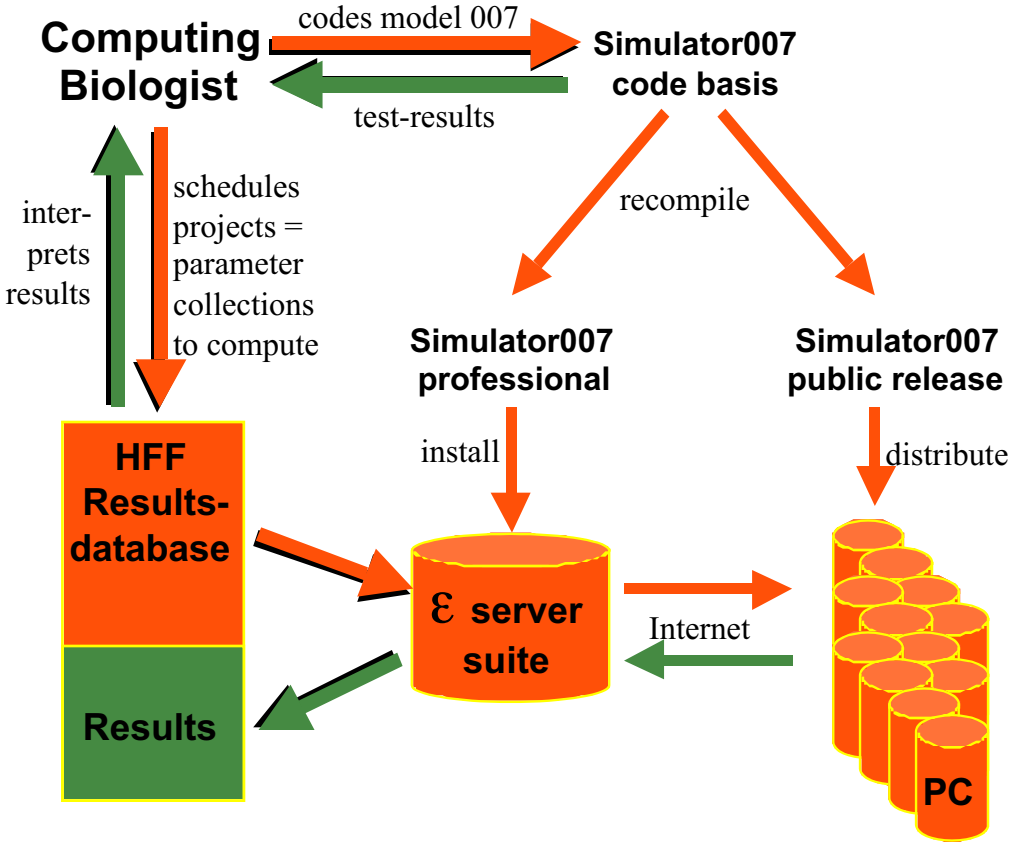


Figure 14 Typical work flow[38] during investigations of evolution with EEPSLION.

Releasing a simulator means to make two compiles: one that includes all the functionality needed in the server suite and one that is as thin as possible for public distribution via the web. While the latter will be installed on many PCs (either automatically or manually), the former supplies all model-specific information to the server suite. This includes the definition of parameter space, capabilities that check whether a certain result is correct and predictions of run-time for simulations that are being scheduled.

Once everything is installed, the server suite just waits for the computing biologist to schedule a *project* by supplying a script that defines a collection of parameters that are to be computed. The server suite takes such a script,

37. Validation is done by comparing test results to known or analytical results, using simple, extreme parameter values, repeatedly checking the code and by finding explanations for strange results. Strange simulation results should only be accepted if code has been repeatedly doublechecked and debugged in detail. The use of a comfortable modern debugger is an absolute must here!

prepares all the input (run-files) for the simulators, sets priorities and hands the corresponding tasks out over the web. Simulators on participants' PCs compute results and return them to the server suite that further analyses them before long-term storing. Thus,

Definition of ProjectID

> *One scheduling script that defines a potentially large number of parameter combinations according to some rules, defines one evolution@home computing* **project** *for one simulator. Each project in evolution@home is identified by its ProjectID and the SimulatorID of the simulator it was designed for.*

From time to time the computing biologist will return to interpret the results that have been accumulating. Then he might change priorities for the distribution of individual parameter combinations and might schedule new projects. In the (hopefully rare) event of a simulator release containing bugs that affect the accuracy of results, a new release of the simulator with the necessary corrections has to be published as quickly as possible. Strict controls and rigid tests before release, however, should make sure that this is the exception. New releases of simulators with additional features or other improvements might be made at any time, as long as they do not change the definition of model or parameter space. Each public release of a simulator gets a new, unique **ReleaseID** included in every result computed. This allows filtering for all results computed with a given release in the worst case of bugs that threaten results integrity.

After many cycles of scheduling, results-interpretation and analyses, the computing biologist reports results to the general public together with his non-computing colleagues who helped with details in the biological background. Such a report should not only be in print but also on the web in technical and non-technical versions to give computing participants the opportunity to understand what they contributed to.

While the simulator continues to run until every useful information is extracted from it, the computing biologist starts again by implementing a new model in a new simulator. Start and stop of simulators as well as their IDs are centrally coordinated by the Evolutionary-Research initiative. It also collects suggestions for improvement of the workflow, as those that actually do the work often have the best ideas on how to improve a given workflow[38].

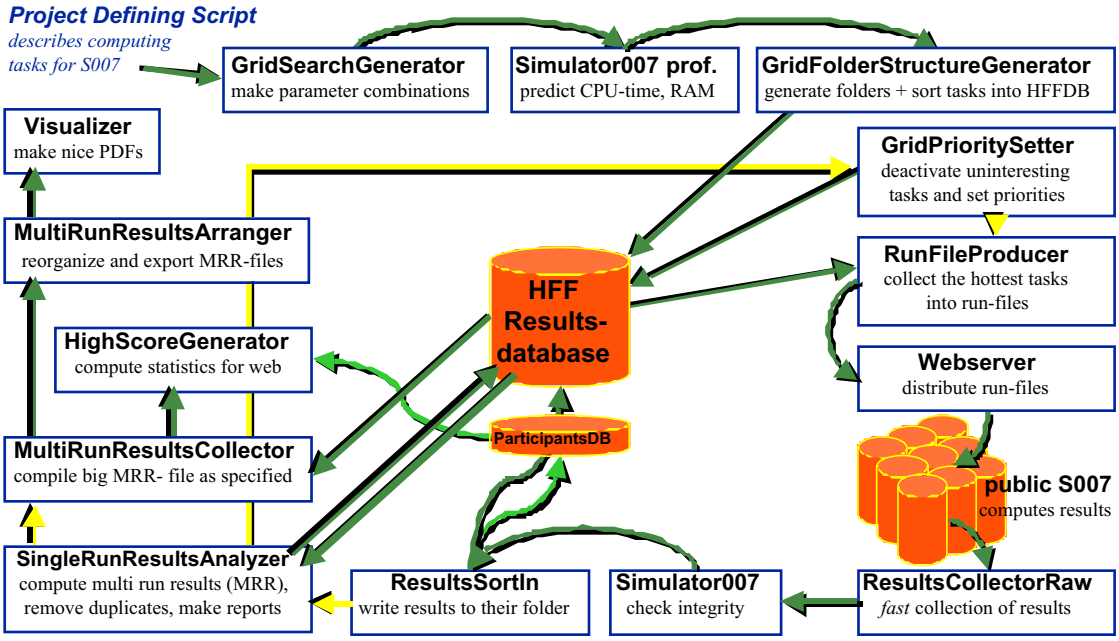


Figure 15 A rough high-level overview over the various modules in the server suite of the EEPSLION global computing framework for evolutionary biology.
Green arrows denote flow of data, yellow arrows regular update-notifications. HFF = Hierarchical Folder File; DB =Data Base; prof. = professional release; MRR = Multi Run Result; S007 = SimulatorID used for illustration. One could add a MultiRunAnalyzer, an incremental data-mining application, that looks for interesting conditions to appear, as data come in. However, the same functionality can also be distributed among the modules shown. See text for details.

10.3 *EEPSLION* server suite design

A central component in the whole enterprise is the server suite where priorities are set, tasks are handed out and results are collected and analysed. A very rough high-level overview over its various modules can be found in Figure 15. The core of the server suite, the HFFDB, is a database distributed over many files that are stored in hierarchically nested folders. It is explained in the next section. Please note that while HFFDB is part of Design 5, corresponding functionality of the modules presented here is also re-

38. For a typical overview over workflow in simulation studies see pp. 106-109 in Law & Kelton (1991) "Simulation modelling and analysis". 2nd., New York, McGraw-Hill. - For more general references on workflow see Peccoud (1995) "Automating molecular biology: A question of communication", Bio/Technology 13:741-745 for a similar situation in molecular biology and Trammell (1996) "Work flow without fear", Byte April:55-60 for work-flow problems in general.

quired, when using a typical relational database. Each of the modules seen in Figure 15 is a standalone application specialised in its task. All these applications might be run sequentially on one CPU except for `Webserver` and **`ResultsCollectorRaw`** who have to run on a high availability 24h/d level in parallel. The following iterative tasks can be distributed, if work load becomes too heavy:

How to parallelise workload

- `RunFileProducer` can run on a node that has reasonably fast access to the hard disks of the HFF database
- `Webserver` and `ResultsCollectorRaw` can be completely different systems that just have a reasonably fast connection to the server-suite for regular transmission of new run-files and incoming results. All experiences in developing scalable webservers can be used to make these as fast as necessary to serve the requests of all participating computers.
- `Simulator007` and `ResultsSortIn` can be distributed to as many nodes as necessary, as long as they have reasonably fast access to the HFFDB. They are embarrassingly parallel as each incoming single run result is checked for integrity and sorted into its corresponding location in the HFFDB.
- `SingleRunResultsAnalyzer` and `MultiRunResultsCollector` might run on a different node each for accelerated computation. If they need too long to process all parameter combinations of all projects, each of them might be distributed to several nodes that take care of all parameter combinations from one project each.
- `HighScoreGenerator` is not easily parallelised. However, if that ever becomes a problem, then it might be exchanged towards lazy evaluation (ie. an ordinary database generating dynamic webpages with those high-scores that are actually requested).
- `MultiRunResultsArranger` and `Visualizer` are needed only when the computing biologist looks for results. As many instances as helpful might run anywhere to rearrange, sort, filter and visualise multi-run-results.

Thus the ε server suite has considerable scalability to cope with growing simulation projects. Just a glance at Figure 15 is enough to explain why it is more feasible not to pack all these different functionalities into one application that changes between various modes of operation as in Design 4.

10.3.1 Functionality of EEBasicLibrary

Before we go on to discuss the various modules in more detail, the functionality of `EEBasicLibrary` should be reviewed. This library is used by every module of the server suite, while the public releases of simulators uses a shortened version of it. Among the common tasks facilitated by this library are:

- **Console related functions.** Output is never written just to `cout`, but to global functions including a verbosity level. These global functions check whether a particular message should be recorded and thus either record it or discharge it. Recording is done automatically on screen and logfile at the corresponding verbosity levels. These levels are recorded in the preferences of each module to allow individual configuration (eg. make critical tasks report more than modules that never show any problems). Interactive input functions are provided that are more robust than `cin` against bad input[39]. They automatically check for proper range and type of numbers, so that eg. safe casts can be made from interactive input.
- **Error reporting functions**. ε has a differentiated error handling system that makes extensive use of C++ exception handling constructs. A class hierarchy based on `EEException` differentiates various types of errors and warnings. The constructor of these classes demands that a description of the exception be included in form of a string. When such an exception is caught, it is either thrown again after adding the name of the throwing function or it is passed to the global `EEError` or `EEWarning` functions. These record the first error and warning, then their total number and terminate the code in case a loop was entered that generates huge amounts of problems.
  Extensive use of this error reporting infrastructure helps track errors during development (you get the stack without calling the debugger) and contributes to overall reliability of the simulators (whenever strange things are seen, they are reported). The number of errors recorded during a simulation is an integral part of the result.
- **Timer related functions**. These take relative start and stop times to measure computing times and provide easily readable timestamps for inclusion in filenames (eg. `2001Oct17_11h50m29`). Universal time

39. If cin expects a number but you happen to type in text by mistake, this can seriously geopardize your code execution on some platforms.

is used and such filenames sort according to date in a usual operating system folder items list window.

- **Execution framework.** As all modules are essentially controlled by commands from scripts (interactive input is rarely used), a main loop is set up that looks for scripts or interactive commands and executes them. The class that manages all interpreter-related functionality is called `EEController` and has only one global object called `eCEO`[40] according to the singleton pattern. However, each module has to instantiate the global key object `eCEO` with its own `EEController` class, that is linked to its own `EEInterpreter`. `eCEO` is expected to know all interpreters that belong to the module and therefore has a complete overview over the definition of parameterspace. These additional interpreters (linked to other classes) have to be made known to `eCEO` explicitly in the module code.

  In its main loop, `eCEO` waits for commands (either from file, from commandline or from a GUI that generates commandlines) and hands them over to the interpreters that search through their command database and execute the corresponding function or return a command-not-found message. If no interpreter can be found that knows the command, a corresponding error message is produced. Reading and writing particular parameters is done correspondingly. Except for interactive sessions, errors are counted and reported at the end of a script to determine easily whether everything went OK. Access to global parameterspace is mediated in the same way. Thus definition of new commands and parameters becomes easy, all definitions can be ordered in an appropriate way. This way of organising scripts is the "*Language Interpreter*" part of the name of ε (the "*Organising Network*" refers to the many modules that have to interact in server suites). As scripts can start heavy computing, but do not contain the core loops, execution speed should be OK.

  However, massive access to input and output parameters of a simulation might be too slow. Therefore simulators construct a global array of pointers called `eParameter`. It does nothing else but contain a pointer to the actual value of the parameter that is either read or written in the simulation code. It is accessed by its ParameterCodeID `eParameter[ EEeniSpeakingNameOfTheParameter]` .

40. eCEO stands for EEPSLION Central Execution Officier. It takes control after main() has started.

- **Interpreter related functions**. Operations that can be executed by modules are implemented by corresponding functions. These can be called by the user via script commands and their parameters. To avoid the long-term fallacies of a list of if-statements for calling the corresponding functions, a system of classes was devised that facilitates the process of implementing new commands and writing proper online documentation on the fly.
  Key to this is the class `EEInterpreter` that contains a vector of `EECommand` objects and a vector of `EEParameter` objects. After instantiation, it constructs these vectors from the corresponding `EECommandInitializationData` and `EEParameterInitializationData` objects that contain the corresponding definitions as constant arrays of numbers and chars in the source code and describe name and other properties of commands and parameters. `EEInterpreter` facilitates easy iteration over all commands and parameters and checks for uniqueness of names. Thus a search function for online help can be easily implemented by iterating over all texts that describe commands, their parameters or problems known to be related to them. All these texts and more information is included in the definition of the corresponding initialisation data.
  `EEInterpreter` itself, as provided in `EEBasicLibrary` contains a minimal implementation that can be expanded by derived classes according to need. Singleton objects of classes derived from `EEInterpreter` can be linked to other classes derived from `EEWorld`, `EEModeOfOperation` and `EEBioObject` via static class variables. Thus, these classes get interpreters that manage all related commands and parameters that belong to them.
  `EEBasicLibrary` expects that the module it is linked to provides all derived interpreters and makes them known to eCEO.
- **XML related functionality**. Many modules have to read a variety of parameter values and the scripts they read from constitute a part of the documentation of a project. Therefore, the comma separated, line separated or other similar simplistic approaches for value input were abandoned. Instead of constructing a new system, where start and end of each value is marked by a speaking name that documents the value, XML is employed[41]. This new emerging standard way to

41. Beveridge (2000) "Transporting Data with XML: The C++ problem" http://www.xmlmag.com/upload/free/features/xml/2000/03sum00/jb0300/jb0300.asp

share data[42] is receiving wide and enthusiastic support from the software industry, as can be seen by the large number of XML parsing libraries freely available. Basically, there are two approaches to read an XML file and construct a corresponding representation with C++ objects: (i) the DOM (Direct Object Model) approach that constructs the whole XML tree in memory allowing arbitrary navigation and (ii) the SAX (Simple API for XML) event-driven approach that makes callbacks to user-defined functions whenever a significant item has been parsed (start of value, end of value, etc.). The latter approach was chosen[43] for ε due to its simplicity, smaller RAM requirements and higher speed. Realising the advantages of XML, it was decided to use XML for storing run-files and results too. At the risk of using more disk space and transmission bandwidth, this confers the huge advantage that all values can actually be read in the raw files without the need for some special viewer. As XML is cross-platform compatible and likely to enjoy long-term future support, this appears to be a safe choice.

This does not mean that the earlier chosen format for storage of results, netCDF, is no longer needed (see discussion of visualisers below). Its primary advantage lies in the cross-platform portability of multidimensional data. It is therefore an excellent choice for preparing data for visualisation. However, as it will be needed only in those modules that prepare data for visualisation, it is not included in the `EEBasicLibrary`.

- **Database access tools**. Scheduled runs and results are stored in a large database. Design 5 planned this to be a Hierarchical Folder-Files DataBase (many files organised in many hierarchically nested folders, see "HFF results database design" on page 111). Whether this approach is used or a relational database is employed, many modules need automated access to it. Therefore `EEBasicLibrary` includes methods for
  - Automated iteration over all parameter combinations
  - Retrieving the folder (ie. all data) of a given parameter combination

42. W3C (2000) "Extensible Markup Language (XML) 1.0" http://www.w3.org/TR/REC-xml. For additional information on XML see http://www.w3.org/ or search the Internet, its full of xml.
43. Doyle (2000) "Balance Software: Expat XML parser port to Macintosh" http://www.balancesoftware.com/people/brian/code/Expat_and_ParseXMLFile.sit.hqx

- Adding new parameter combinations and their folders
- Consistency checks

- **Calculator functionality**. Sometimes it might be desirable to include a simple regular expression that uses the values of parameters given to compute some critical limit used for making a decision. On other occasions, while analysing results, it might be helpful to compute synthetic parameters from combining some of the basic input and output parameters to facilitate discovery of new functional relationships. All these possibilities can not be hardwired into the code of either the simulator or the various models of the server suite. Therefore one of the basic functionalities is a evaluator for regular expressions. This allows the computing biologist to provide a formula in a script, while the program substitutes all variable names by their corresponding values, hands this over to the evaluator function and retrieves the result. Tcl/Tk appears to be a competent choice for this. Alternatively other, smaller evaluators are available on the web.

After these major functionalities provided by `EEBasicLibrary` have been reviewed, we can go on to discuss the workflow of a simulation project in the server suite.

10.3.2 GridSearchGenerator

After installing a new simulator that is being distributed on the web, the computing biologist will provide a script that defines the first simulation project.

Repeated optimisations are not considered here

Generally there are two types of parameter searches that can be performed by such a system. (i) A **grid search** of parameterspace can be made. This means that input parameters in predetermined distances are selected for computation of results. (ii) A **repeated optimisation** can be performed. Here a certain start setting is given and a particular output parameter is expected to be optimised towards a goal value. One or more input parameters are allowed to vary in order to achieve this goal. Such an approach would be needed to perform population viability analyses[44], inverse modelling and similar things. They are not considered here any further as it suffices to say that they require a `RepeatedOptimizationGenerator`. This module would have a combination of functionality of `GridSearch-`

44. Brook et al. (2000) "Predictive accuracy of population viability analysis in conservation biology", Nature 404:385-387. - Coulson et al. (2001) "The use and abuse of population viability analysis", Trends Ecol. Evol. 16:219-221.

Generator, GridFolderStructureGenerator and GridPrioritySetter. It further needs an appropriate interface to RunFileProducer and to ResultsSortIn, as priorities are set according to the results received. The future will hopefully lead to the development of the corresponding modules.

Projectscripts

Now, here comes a new script defining a new grid search. It is written in XML, and contains the following information:

- ProjectType + ProjectID
- RequiredSimulatorID + MinimalReleaseID
- UpperLimitForStepsPerDimension (against infinite loops)
- UpperLimitForPointsInThisGrid (against poor planning)
- A list of input parameters, with a definition of the size of their respective dimension.

The latter can be provided in several forms:

- constant value
- start value + value to add between steps + end value[45]
- start value + factor to multiply between steps + end value[46]
- start value + value to add + factor to multiply + end value[47]
- start value + factor to reach new order of magnitude + number of linear steps within each order of magnitude + end value[48]

After reading the script GridSearchGenerator computes the number of parameter combinations it contains and asks for approval to generate all combinations. If the computing biologist accepts, all combinations are written to a file in a format that can be read by the professional version of the simulator. A word of caution should be added here. Evolutionary biology contains several parameters that are unknown to a large extent, as they are hard to assess in nature. As the number of parameters and potentially interesting values is exceedingly great, it is easy to quickly block the local hard disk with a file that contains billions of parameter combinations. The art in

45. Optimal to check linear relationshipts
46. Optimal for log relationships
47. Byproduct of the algorithm
48. This is best for making nice plots with log relationships. Being used to orders of magnitude of ten, it is not nice to schedule some obscure intermediate values, but rather eg. population sizes of 10, 100, 1000 and so on. While this could be easily done with the factor to multiply, a problem appears, if more than 1 point per order of magnitude should be sampled. The simple multiply algorithm can do it, but it produces nasty numbers. Thus an algorithm was developed, that schedules only values, that are follow a log relationship for their orders of magnitude and a linear relationship within their orders of magnitude. For example start=10, factor = 10, steps = 2, end = 1000 would generate the following values for scheduling: 10, 50, 100, 500, 1000. This greatly helps in scheduling gridsearches that span several orders of magnitude.

designing simulators for evolution@home is, to reduce the number of parameters so that things can be handled and still make biological sense, without being too easy. The art of designing projects is to start with regions of parameterspace that appear to be the most interesting, then continue with 'normal tasks' and finally stop when it is clear that further results do not contribute to biology.

10.3.3 Professional releases of simulators (1)

After all parameter combinations have been written to file, in principle computation could start. However, practical issues prohibit this. The long list generated by `GridSearchGenerator` is ordered according to parameter values but completely mixed up regarding CPU time and RAM complexity. To filter for those runs that are extremely cheap (seconds) and those that are not feasible to schedule at all (many years), a prediction of computational complexity is needed. This is impossible for the framework, as the variety of models in simulators is too great. Thus, before public release of a simulator, a CPU time and RAM complexity prediction system has to be developed - a price one has to pay, if complex runs are not to be generally avoided.

How to predict computational complexity

A fundamental tension is encountered here. On one side, *some* prediction system is needed. On the other side, if one could really predict the runtime for some types of models, then they are so well understood, that no further simulations are necessary. For example project 1 for Simulator005 will run until either 500 clicks of Muller's ratchet have been observed or until a certain generation is reached. Many of the interesting simulations reach the click limit *long* before the generation deadline. Thus generation deadlines are often not a good indicator at all and one has to rely on a simple and easy prediction of click time to predict runtime. However, if this were trivial, no simulations of Muller's ratchet would be needed. Thus a fast and simple formula was needed to estimate at least the order of magnitude of clicktime for most cases. This led to the development of the simple equation described in Chapter 18 after checking well over 150 heuristic candidates for prediction quality. Please note that neither manual interaction nor complicated maths packages can be involved for prediction of computing time, or else these predictions alone would require a global computing project. Inaccurate predictions are often better than no predictions, and as ε shall support feedback on computing time properties from incomplete runs, the inability to find a reasonable simple equation for prediction will not prohibit the start

of computations (try this with your local super-computing centre;-). As with Simulator005 and its *Equation172* (see Chapter 18), the need for a simple prediction formula can be expected to stimulate interesting biological research in a number of cases.

Life with bad predictions

To work around bad predictions, the following action is taken on the assumption that the only problem would be runs that are too long. While the most probable CPU time can only be predicted by the simulator, the framework can predict a most probable maximal CPU time. This can be done easily with the following formula

$$T_{CPUmax} = P_{INDS} \cdot IND_{max} \cdot TS_{max} \tag{11}$$

where T_{CPUmax} is the maximal absolute time in seconds this CPU needs for computation of the simulation, P_{INDS} is the performance of this CPU measured in average *IND*ividuals per *S*econd[49], $IND_{max}$is the maximal total number of individuals for one moment of time averaged over the whole simulation[50], and TS_{max} is the maximal number of timesteps[51]. If nothing better is available, runtime predictions can be made on this basis to get started.

To allow users to get more reliable run times, they can set an upper limit for the uncertainty they are willing to live with. It is expressed in terms of a factor that describes how much the upper run time limit of the user might be exceeded if the simulation requests that. If this limit is reached, the simulation is terminated by the framework and an 'ordinary intermediate result' is generated[52]. As the actual computing requirements of results are evaluated and written to a place where the `RunFileProducer` can find them, the quality of the predictions increases dramatically, when the first result for a parameter combination comes in.

Thus all the server suite expects the professional release of the simulator to do at the current moment is to make a good prediction of the computational complexity in terms of $IND_{max} \cdot TS_{max}$. More precise scaling to the participant's computer and a nice human readable value for an estimat-

49. This scales predictions to a particular CPU – helpful for the human eye, but not for precise predictions. Therefore this term is only added by the simulator on a participant's computer after *observing* actual performance.
50. If this is too difficult to predict due to strongly fluctuating population sizes, just use the maximum or some value that is certainly above the true one. As a maximal CPU time shall be predicted it is important, that true computing time is *always* shorter.
51. This can be confidently said, as *all* simulations are terminated by the framework after this upper limit of simulated time is reached.
52. The need for this feature was detected only after the first 5 releases of Simulator005.

ed absolute time are added by the framework. Furthermore, a definitive upper limit for RAM complexity is needed to make very sure that the use of virtual memory on the participant's computer can be excluded with absolute certainty[53]. Finally the maximal amount of disk space needed on the participant's computer should be predicted as well as the frequency and size of data transfer over the Internet, assuming that each result is sent back immediately[54]. All these predictions are made and written to a big file that is read by the `GridFolderStructureGenerator`.

10.3.4 GridFolderStructureGenerator

The only task of the `GridFolderStructureGenerator` is, to generate a new entry in the data base for each new parameter combination and to store all complexity prediction data in it. In the HFFDB, it needs to generate one folder for each parameter combination. The folder contains a corresponding file with plain scheduling information, the results that are about to come and everything else concerning a parameter combination. As the HFFDB is characterised by a hierarchy of folders, this module generates that complex hierarchy in a way that allows automated traversing and easy manual access. Before it can do that, it needs to be told the path to the folder that belongs exclusively to this project. As explained in the HFF section (page 111), this might be on another hard disk drive, but it has to be somewhere in the current file system. It is assumed that the hard disk where the folders are to be installed has not only enough space for the folders and runfiles, but will also be able to contain all single-run and multi-run results accumulated for this project. Thus, the computing biologist in charge needs to think about this before writing the scheduling script.

10.3.5 GridPrioritySetter

One last step is feasible before actual runfiles can be produced. Grids defined the way described above are rectangular. They might include parame-

53. If virtual memory is used, even modern preemptive multi-tasking operating systems can effectively be blocked, as the typical simulator accesses all his memory for computation of each timestep. Thus excessive paging is requested and paging has one of the highest priorities on a operating system, even if the request for paging came from the lowest priority application. There is no way around this, except the following:
Simulations that exceed the RAM committed by the participant are refused, if encountered in a run file. This assumes that participants choose reasonable upper limits as explained on the website. Cross platform determiniation the free non virtual memory of a system is difficult.
54. Predictions of disk space and communication needs are not yet made by Simulator005.

ter combinations that are not feasible to run (due to complexity) or that are simply uninteresting. To remove them from being scheduled, their priorities have to be changed, but to do so manually would be extremely tedious. Here the `GridPrioritySetter` comes in. It allows one to define a (non rectangular) set of parameter combinations, that is either deactivated or set to a higher priority than the rest. Here the use of regular expressions comes in. Just by combining the input parameters to a condition that contains as many (simple math) formulas as necessary, they allow for incredible flexibility here. A differentiated system of priorities allows one, for example, to give repeats of hot parameter combinations a higher priority than initial computation of tasks from higher complexity classes. However, usually, priorities are determined only among tasks within the same complexity class.

This module greatly simplifies administrative tasks of the computing biologist in charge.

10.3.6 RunFileProducer and Webserver

Now everything is ready for the production of the actual run files. A run-file is a text file that contains all information needed for a participant's simulator to start the simulations described in the run-file. As it is sometimes convenient to have more than one simulation to compute in store, run-files contain a number of simulations depending on computational complexity. Most importantly, run-files are ordered according to their computational complexity with the hottest simulations in the most prominent positions.

Sort interesting tasks

The `RunFileProducer` iterates over all parameter combinations of all projects of this simulator, collects the most important[55] tasks and removes tasks that are too small to be published . These are treated separately, as they either require a high bandwidth connection or a local computer. Published tasks are sorted into different complexity classes according to RAM, CPU time complexity, bandwidth and disk space requirements. Then, within each class runs are sorted according to priority. For a list of RAM and CPU time categories of complexity see Table 9 or the current evolution@home run-files on the web. A number of static html-pages can be generated that follow a standard structure published on the website to allow manual download of tasks for users who prefer the semi-automated mode of operation.

55. All runs that have been repeated more than the average are automatically deactivated, whereas tasks that have been computed only rarely get high priorities.

Table 9 Upper limits of various run-file categories. Feasible categories for disk space and bandwidth needed still have to be developed. [u] = unpublished

CPU-time	RAM needed
1 sec [u]	100 KB[a]
10 sec [u]	200 KB
1 min [u]	500 KB
5 min [u]	1 MB
15 min [u]	2 MB
30 min	5 MB
1 hour	10 MB
2 hours	20 MB
4 hours	30 MB
8 hours	40 MB
16 hours	50 MB
1 day	75 MB
2 days	100 MB
3 days	200 MB
4 days	300 MB
1 week	400 MB
2 weeks	500 MB
3 weeks	750 MB
1 month	1 GB
2 months	1.5 GB
3 months	2 GB
4 months	3 GB
5 months	4 GB
6 months	larger
9 month	expand as needed
12 months	
18 months	
2 years [b]	
5 years [b]	
10 years [b]	
50 years [b]	
100 years [b]	
larger [b]	

a. The first 3 values are actually only interesting to make use of L2 cache features.
b. Moore's Law prohibits starting of such tasks. They are published to underscore the fact that evolutionary biology needs more CPU power.

Publicly released, fully automated simulators read their run-files from the same website like semi-automated participants. If the website is static, simulators use a published list of categories for access.

Dynamic webservers are possible

However, the `RunFileProducer` can also be implemented as a web-front-end to a database that generates all run-files on the fly. Then it will still generate human readable pages for manual download, but the automated simulator will use the actual performance of the participant's computer to make the complexity choice. Detailed checks will have to show whether the managing database can be the large results database, should it be relational. If this is not possible, due to performance or security issues, a smaller database of hot tasks can be generated at regular (ie. daily) intervals from the large database. If multiple hot databases are employed, the solution becomes significantly scalable[56]. The small databases can be thought of as a fast efficient cache. It might use MySQL and PHP or FastCGI and needs its own server.

Until such a dynamic solution is developed, run-files are served as static files. Since each parameter combination has to be computed several times in any case, this extremely fast and simple method of distributing tasks is the ideal choice, as it allows simple and cheap delegation of web services requiring high availability to Internet service providers.

10.3.7 The simulate command

The decision has been made to make communication lines public, so that each participant can see what he is computing. Thus the simulations scheduled and the results submitted are to be human-readable text files. This makes it possible to change them by simple editing. To keep participants from manipulating scheduled tasks or results, checksums are included to validate the accuracy of the information transmitted. The algorithm for their computation is not published (which does not mean that a real hacker will not find it).

The most readable representation considering one simulation would of course be XML. However, this inflates text to such an extent, that an overview of a large number of simulations is no longer easily achieved on a web

56. If participation in SETI@home is considered as current upper limit for global computing with a good advertisement campaign, about continous 500000 participants have to be handled. If they contact the server once a day on average, about 6 requests per second have to be handled continually. However, peaks will probably be more than one order of magnitude higher than that.

page[57]. The following potential convention gives an overview of the data that can be used to start a task and to describe its complexity details.

```
&simulate
RequiredSimulatorID_
MinimalReleaseID_
ProjectType_
ProjectID_
ParameterCombinationID_
RunID[58] :
ListOfInputParameters[59] :
ListOfAddidtionalDataFiles[60] :
ListOfTimeSeriesRequested[61] :
ListOfSnapshotDatesRequested[62] :
WorldHistoryEndDate[63]
```

57. An example for the convention presented reads like this:
&simulate S005 9 GS 1 35234 5345dkfjkwl042ofsJRLE94: 1000000 10 0.0004 -0.000002 5 0.25 +0.0002 0 400 : : ratch : : 10000000 0 1 235342352 requires 46 MB RAM + 0 MB on disk + offline + 0.005 Mbit may generate 10 KB results and may need 1.2 days for 2.46e5 MFLOPS @ 236 MFLOPS#
Appropriate resizing of the browser window allows a fast and easy overview. See the run-files for Simulator005 with a similar structure (http://www.evolutionary-research.net). Should this format not be used, any new format will have to transport similar data.
58. `RunID` can be used to make distributed tasks unique by either a simple running number or by including a secure long key that uniquely identifies this simulation from all other simulations and all potential hacker guesses of the key. Such a key can be used to easily check for duplicates at a very early stage in results processing; It will also allow identification of early and late stages of a simulation, should more than 1 stage be submitted.
59. Length of the `ListOfInputParameters` is fixed for a given simulator
60. `ListOfAddidtionalDataFiles` is either empty or takes as many arguments as the simulator expects. Additional data files might contain a complex dataset like a metapopulation structure or geographical information for a particular species. Such data clearly belongs to the input category, but inclusion in any input parameters list is not feasible, as the numbers of parameters would then become exceedingly large. Moreover, such realistic data ususally belong to a particular natural setting that is under investigation. As the number of such realistic settings is always limited by the number of biologists who could gather such data, it is likely to stay small. Inclusion of the ID makes it possible to reuse once downloaded data for the next simulation of just a slight variation on the same set of data. Thus, less bandwidth is needed.
For artificially produced datasets, the approach is completely different. As they are usually generated from a few numbers that describe their properties, these are not transmitted via file, but rather these numbers are included in the list of input parameters and the whole complex structure is generated by the simulator on the participants computer only.
61. `ListOfTimeSeriesRequested` allows collection of particular time series to be distributed. As collection of all potential time series in each simulation is prohibitive due to the amount of data generated, the easiest solution is to restrict time series collecton to the professional release only. However, this seriously limits the computing time of simulations for whom time series can be collected. Thus this more flexible approach is used.
62. `ListOfSnapshotDatesRequested` allows collection of selected, detailed snapshots to be distributed. What is true for time series is even more true for snapshots.
63. `WorldHistoryEndDate` contains the number of timesteps after which the simulation is forced to end.

RNDSeedProposed[64]_
SecurityAlgorithmID_
SecurityCheck[65] **_requires_**
RAMExpectationMB[66] **_MB RAM +_**
DiskspaceNeededMB[67] **_MB on disk +_**
OfflineComputationPossible[68] **_+_**
BandwidthClassRequestedMBit[69] **_Mbit_**
may generate
ResultsGeneratedKB **_KB results and may need_**
ComputingTimeExpectation[70] **_for_**
ExpectedWorkUnitsMFLOP[71] **_MFLOP @_**
PerformanceAssumedMFLOPS[72] **_MFLOPS#_**

64. The standard RNDSeedProposed is 0 which means that the simulator should generate his own seed. Usually the number of seconds since a certain date is used. Thus parallel runs of the same parameter combination started at the same time use the same seed and are therefore useless repetitions (except for checking the integrity of a result). Overall probability that this might happen is low, as long as many fewer than 24*3600 participants start the same parameter combination per day.
However, on a multiprocessor system that computes the same parameter combination on several machines in parallel, additional care is necessary. Either these runs are started with more than 1 second delay each, or an additional distinctive feature is used to generate diverse seeds.
Alternatively a central seed assignment can be implemented, if an integrated PrioritySetter-RunFileProducer-Webserver solution is used to dynamically generate the run-files distributed. Such run-files, however, must not be used more than once and thus rely heavily on a smoothly working automatic.
65. SecurityCheck is a number computed by the unpublished algorithm "SecurityAlgorithmID" to make manipulations of the simulate command as difficult as possible.
66. RAMExpectationMB makes sure that computation refuses to start when the participant did not dedicate enough memory on that computer. The amount dedicated is set in the preferences and helps the simulator avoid virtual memory slow-down that can paralyze the whole system.
67. DiskspaceNeededMB is diskspace for storing temporary files (eg. intermediate state of the simulated world) *and* final results for this simulation.
68. OfflineComputationPossible is either "offline" or "online". The latter means at least a modem connection that is started automatically upon request from the simulator; online features are only interesting for hierarchical global computing
69. BandwidthClassRequestedMBit is a recommendation for the network speed needed to allow comfortable working with a simulation of this dimension. If actual performance is lower, transmission takes longer, but is not impossible. However, deviations that are too large should be avoided. This number gives an estimate of the CPU time/bandwidth ratio and simulators might exclude simulations that need to much bandwidth.
In case of "offline" simulations, this float number indicates the bandwidth in Mbit per second needed to transmit the results of this simulation in 20 seconds.
In case of "online" hierarchical global computing, the speed becomes a continually requested maximum speed throughout the simulation. It no longer indicates the amount of data actually transmitted, but shows the bandwith needed to ensure reasonable operation. Deviations are unavoidable, but this parameter helps select a general setting that works.
70. ComputingTimeExpectation is included solely to make the following two numbers readable for humans: 2 digits and the right unit (sec, min, hours, days, weeks, months, years).

71. `ExpectedWorkUnitsMFLOPS` contains the best estimate available of how much computational work is needed to complete this simulation. It is either estimated from the input parameters and performance knowledge gathered during test simulations before finalising simulator development or it is measured in the first simulation(s) with this parameter combination.
The only appropriate work unit in the simulation of evolution is 1 individual making 1 step forward in time, as all other measures vary more than this. Unfortunately the number of equivalent FLOPS varies greatly between simulators or even within a simulator, if different species with different computational complexity are modelled. Thus it is part of each simulator development to develop a standard simulation with an expected average species mix that can be used to measure performance of a given CPU in terms of average individuals per second. As this number is absolutely no meaning except a nice label, it is scaled to approximate MFLOPs needed. Then a system has to be developed that estimates computational compexity of a simulation in terms of these average individuals (ie. the smallest FLOP package possible) from input parameters. This value is used to find a simulator that is willing to compute this run.
The simulator records (i) the actual total of individuals in this simulation, regardless of their type and (ii) the total time needed for the simulation. These values are returned together with (iii) the standard FLOPS performance of this CPU, (iv) the peak FLOPS performance as measured in a loop over a block of 10 register to register 64 bit multiplications, (v) maximal memory bandwidth in MB/sec (vi) the total average performance of this CPU in INDS and (vii) the overall computing time.
From these data the server suite extracts the different high scores categories (highest total contribution in FLOPS, Individuals or CPUTime; various highest performance measures etc...) and updates expected work unit sizes.
72. `PerformanceAssumedMFLOPS` is a benchmark result measured by computing a number of average individuals per second for one timestep and scaling to MFLOPS.
As Gustafson & Todi (1998, ref see below) point out, such benchmarks usually capture only a small part of the performance spectrum and properties like main memory bandwidth often play a more important role than CPU speed. To make absolute precise computing time predictions, all details of performance like in the HINT benchmark would have to be uncovered. As this is hardly feasible an arbitrary decision has to be made on what benchmark to use to compare performance.
This is archived by (i) implementing a standard benchmarking routine in each simulator that (ii) runs a standard world history with (iii) a parameter combination that produces an 'average' standard mix of all kinds of individual types that differ in computational complexity. (iv) This test simulation is scaled to fill the whole RAM dedicated to simulations or at least 10 MB, as this is usually well above any cache size, so that limiting effects of memory bandwidth are integrated. (v) As the result of this benchmark is used to decide what simulations are started on this CPU, it should be run while the CPU is working on its usual set of tasks. The resulting number of average standard MFLOPS should represent a minimum guaranteed computing speed for simulations of this RAM complexity. Applying this to simulations that need less RAM leads to an overestimation of computing times that are already likely to be small. However, this is not likely to be such a problem as significant underestimation of already very complex tasks.
If such precision is insufficient due to frequent computation of small tasks or far from average compositions of the individuals mix, two possibilities might increase accuracy of prediction: (i) Each simulator might use the average of the performance he actually computes to predict CPU time or (ii) it refuses to accept simulations that have never been computed before. For all others the expected work units in MFLOPS are actually derived from observations in the first run(s) of this parameter combination. These are much more accurate than extimations might ever be.
Gustafson & Snell (1995) "HINT: A new way to measure computer performance". Proceedings of the 28th Hawaii International Conference on System Sciences, January, Wailea, Maui, Hawaii.
Gustafson & Todi (1998) "Conventional benchmarks as a sample of the performance spectrum"
HINT website: http://www.scl.ameslab.gov/HINT

Like all commands in ε, the simulate command is encapsulated by & and # to allow retrieval from files like html-files, extensively commented scripts and log-files. The bold and underlined characters are expected from the simulator as separators (required spaces are underlined too). The lists contains simulator specific details that are separated by " : ". Single values within a list are separated by spaces.

This structure can accommodate a diverse array of input parameters and facilitates an easy overview over the details of many runs. Its conciseness reduces transmission bandwidth requirements and allows download of multiple tasks to either reduce the frequency of contacting the webserver or to keep the CPU busy when the webserver is temporarily not reachable.

10.3.8 ResultsCollectorRaw

After results have been computed by various simulators around the globe they have to be collected centrally for analysis. Main requirement for this step is speed and therefore simplicity, as submission of results has to be fast from a user's point of view. To achieve maximal scalability, all further complex analysis is separated from results collection and thus, more than one CPU could be dedicated to the task of collecting raw results, if needed. To use one computer per simulator as `ResultsCollectorRaw` or even one per project is a natural way to parallelise, but even several collectors for very complex projects are conceivable. Simulators get the IP addresses of these computers either via the `ServerNameServer` or they remember them from a former submission or they are forwarded from one `ResultsCollectorRaw` to the next. It is clear in global computing that all communication goes over the Internet. No proprietary protocol will be used, as this is blocked by many firewalls. FTP and SMTP would require a special implementation just for ε, but would allow faster rejection of bad data than a WWW solution with PHP or FastCGI could probably achieve. On the other hand, a PHP system is extremely flexible. Either solution requires significant programming on the simulator's side and on the corresponding server suite connection. Advantages of an email-based solution are (i) speed: no need to load a cgi or execute a script (ii) intermediate storage: mail that cannot be delivered now will follow as soon as the connection works again (iii) different email addresses are possible for different simulators.

Security is a big issue

Whatever solution is employed, the corresponding computer will be the first line of defence against hackers. There is no ultimate security for any computer with a standing connection to the Internet, as the experiences of

Gibson show: *every* Internet server or website can be completely knocked out by a cleverly distributed denial of service attack[73]. Nevertheless it is important to invest in security to make the server as safe as possible.

It should be noted, that a change of the communication protocol between `Simulator` and `ResultsCollectorRaw` comprises only a very small part of the overall framework. It is therefore conceivable that the protocol is changed in order to employ some global computing solution that will help with the data transfer.

10.3.9 ServerNameServer

Where do fully automated simulators get the URLs of the run-files and of `ResultsCollectorRaw` from? Coding this into the Simulator itself leads to the problem that each change of such an address requires a new simulator release. Furthermore, a simulator can not know the addresses of future simulators. Thus the following solution is proposed:

Each simulator knows only one central URL (eg. `http://www.evolution-at-home.net/ServerNameServer/index.html`). This URL contains the address of an `ServerNameServer`. This `ServerNameServer` returns the URLs for the run-files and details for results submission for a given Simulator. For Simulators that are automated enough to load other simulators, it points to the URL of a simulator with urgent tasks.

Initially, it will suffice to have just the corresponding static websites. However, as the need for more automation and flexibility grows, a dynamic version can be envisioned that distributes simulator requests according to load balancing requests from the run-file distributors and results collectors.

The `ServerNameServer` is contacted by the simulator after its installation, whenever it experiences problems in getting a run-file or submitting a result and after regular intervals to be able to focus computing power on the most pressing tasks. If projects are complex enough to get their own server, they will be listed by the `ServerNameServer`, otherwise each simulator get one entry.

73. Gibson (2001) "The strange tale of the Denial of Service attacks against grc.com" http://media.grc.com:8080/files/grcdos.pdf - http://grc.com/dos/grcdos.htm

10.3.10 Professional releases of simulators (2)

While `ResultsCollectorRaw` has only time for simple general checks of results, more thorough analyses are needed before results can be accepted as true. As such analyses depend heavily on the evolutionary model computed, they have to be performed by the simulator itself. The corresponding functionality is not for public use and even if source code is published, this section should not be, in order to make more difficult for people who want to get to the high scores other than by computing or simply want to disturb the system[74].

Besides recomputing checksums included in the results, various known relationships between parameters are checked for plausibility. All results that fail the tests are filtered out from the stream of results that is on its way to the final large database. Results are marked for manual investigation, when checksums are correct but the simulator indicates a lack of plausibility. Thus, potential errors in the filter can be detected. Strange, but important results are scheduled for repetition on a dedicated internal simulator.

10.3.11 ResultsSortIn

Results that arrive here have been checked for integrity and now have to be sorted into the big results database, whether it is relational or a HFFDB. As the hierarchical folder file data base (HFFDB) is organised according to the input parameters, the proper directory can be found by travelling the corresponding path (see “HFF results database design” on page 111). Once the right directory has been found, the corresponding result is added to the file "`IncomingResults.xml`", where it waits for further processing. To improve performance, a platform specific shortcut to frequently used directories can be kept in memory (eg. as a STL map). Whether results from different projects are merged into the same HFFDB or are sorted into different locations, depends on the size of the project. Complex projects that need excessive diskspace might be configured to have their HFFDB on a separate volume.

74. If you are one of these, let me just share a personal thought. I know you are intelligent enough to produce faked results and smuggle yourself into the high scores, if you take enough time to learn how to do it. However, what do you gain? This is a non-profit public research project. If you want fame or money, you will have to look for other targets. And if you prove that you can do it, you proved only that I was right when I said that you would be able to do it.

To speed up the next steps, a log-file records all those parameter combinations, where new results arrived. Thus `SingleRunResultAnalyzer` does not have to work through the whole database.

If a certain result has been submitted by a new participant, the name, group, passwords and other information included in the results are added to the participants' database. If the corresponding name is already in use by someone else, the new participant gets an email request for resolution of the issue over a web interface.

10.3.12 SingleRunResultAnalyzer

After pre-sorting results this way, it has become comparatively easy to integrate the various single-run-results to one multi-run-result for this parameter combination: There is no need to search for other runs with identical input parameters as they are all in the same directory[75].

`SingleRunResultAnalyzer` needs to search the current directory for duplicates. As it contains repeats of only one parameter combination, the number of results in it will be relatively small. After sorting out duplicates (due to double manual submissions), all different values of the various single-runs are combined to a multi-run-result. The framework knows the various types of single-run output-parameters and has corresponding multi-run result parameters that record mean, variance, etc. for each single number in a single-run result. Thus a summary is prepared that contains as much information as possible in the most concise form for an interactive visualiser session. The complete result of this analysis is stored in the file "`MultiRunResult.xml`". It can also contain 'synthetic' parameters that are defined by external scripts that contain formulae that combine the fixed set of in- and output parameters of the simulator to new values that greatly accelerate hypothesis testing.

Furthermore, automated analysis of any time series observed during the simulations is written to a PDF file along with plots of the raw time series data. Then, a short version of the most important results is composed for eventual publication on the web and a file with all high-score relevant information is generated.

75. All results collections that are added to a single file as they come in, have a big problem here, once they exceed certain limits: If there is no sophisticated multidimensional index or another kind of costly order, all existing results (old and new!) have to be checked to find all those belonging to one parameter combination. Cheap identification of duplicates becomes impossible this way.

To avoid having to iterate over the whole HFFDB although only a few results have been added, `SingleRunResultAnalyzer` usually processes only those parameter combinations that received new results. If desired, it might change the scheduling priority once a given accuracy in terms of standard error is reached. It writes a log of such changed priorities to the `GridPrioritySetter`, who notifies the `RunFileProducer`.

After this point, one might manually inspect results in the HFFDB. Although tedious for complex searches, this is great for detailed investigations of a few special parameter combinations and should certainly be used to get a feeling about what is going on in detail.

10.3.13 MultiRunResultsCollector

To get an overview over the big picture and to explore relationships between parameters, all multi-run-results need to be collected into a more compact format. Towards this end, the `MultiRunResultsCollector` needs to iterate over all parameter combinations and compile a big multi-run-results XML file. This can be either for all results of a simulator, or for a single project, or for all results that have certain parameters in a certain range.

To generate the high-scores for a simulator or a project, `MultiRunResultsCollector` iterates over the corresponding range of parameters and compiles a large file with all computing time contributions of the various participants. This file is either a total summary or an incremental file that collects everything since the last high-score generation.

10.3.14 HighScoreGenerator

Importance of high-scores for a global computing project should not be underestimated. They motivate participants and ultimately prove to them that their contribution counts, as it has reached its destination and is important to somebody[76].

The `HighScoreGenerator` takes the long log of single-run-results computing complexities and adds them to the corresponding participants' accounts. This can be done incrementally to speed up the task. After that, all groups and all individuals within these groups are ranked according to their computing contribution in a particular 'discipline'. From this, html-files are generated that list the top 20 contributors of a discipline or all the rest, too. This is repeated for all existing disciplines (from the number of

76. In the long term, this has to be complemented by research reports that are based on the results.

simulations over total work unit complexity to the largest simulations conducted by means of RAM size or CPU time). The great diversity of different computing tasks generates many niches with many different top-positions allowing for a great variety of CPU races.

These webpages are generated at regular intervals and transmitted to a static webserver, until a completely dynamic solution becomes affordable (ie. a request to high-scores goes directly to the database and is answered by a direct list of current entries – a task that needs much more computing power and might get performance problems with a weak server and many requests).

10.3.15 MultiRunResultsArranger

As generation of complete multi-run-results files (MRR-files) is presumably quite costly, it might be worth having a `MultiRunResultsArranger` to allow merging and filtering particular sets of data. It should furthermore allow one to export results or a subset of them to various file-formats used by other visualising software like eg. tab files for Excel[77], netCDF[78] files for more professional visualisers or others as needed. As efficient visualisation requires data to be readily accessible, such potentially very complex rearrangement tasks have to be completed earlier.

Another attractive but very complex possibility would be to feed all MRRs in a large high-dimensional database that has a fast connection to visualiser and data-mining tools. Thus complex manual rearrangements would be no longer necessary. However, high-dimensional databases have to deal with the curse of dimensionality[79], a phenomenon that slows normal index algorithms down to the point where a linear search through all records of the database is faster. As potentially every input and output parameter might be used to restrict the set of data used for further analysis, complex cutting-edge algorithms for fast searching in high dimensional da-

77. Use Excel only for very small sets of data. While it can hold 256 columns and 32000 rows in theory, actual work becomes nasty at much smaller datasets.
78. See Unidata Program Center et al. (1997) "NetCDF User's Guide for C: An access interface for self-describing, portable data". Version 3, http://www.unidata.ucar.edu/packages/netcdf/, Unidata Program Center. netCDF was chosen from the trio of CDF, HDF and netCDF, the three big packages that handle multidimensional data (see FAQ at Unidata's page) , as it has the widest distribution and yet offers elegant functionality for simple analyses of multidimensional data. netCDF is also available in a CodeWarrior port by Malyshev (2000) "NetCDF, udunits and CPP for Macintosh" http://crga.atmos.uiuc.edu/~sergey/soft/index.html
79. Berchtold et al. (1998) "The Pyramid-Technique: Towards breaking the curse of dimensionality", 142-153. Proc. Int. Conf. on Management of Data, ACM SIGMOD, Seattle, Washington.

tabases[80] have to be employed - a non-trivial project in itself. The possibility to use SAP-DB[81] as a database and the R-system[86] for statistical analysis and visualisation appears to be a potential implementation for such a high-end analysis system.

10.3.16 Visualisers and Datamining

Ultimate goal of the whole simulation process is the discovery of graphical plots that represent new knowledge. New insights gained have to be tested manually by intuition, by known theory, by analytical proofs, or by independent simulations. Before this can be done, the very process of datamining is greatly facilitated by visualisers and automatic knowledge discovery modules.

Visualisers

Once all interesting data have been gathered, the most natural thing is to plot various parameters against each other, while trying to make sense of the result. To do this efficiently, the ideal visualiser ...

- allows fast selection of x and y axis,
- automatically retrieves the corresponding data,
- determines an appropriate scale automatically
- allows for selection of results based on other parameter values to get easily readable overviews[82]
- has a system that allows one to integrate several dimensions in the points plotted[83] to easily identify the causes of outliers
- allows easy PDF generation, printing and exporting
- can be scheduled to plot sets of parameters against each other in batch runs. Thus the user would only have to click away uninteresting plots and continue in-depth analysis at the interesting points.

It remains to be determined whether this functionality can be best archived with one of the big visualisers[84] or whether an independent module tightly integrated with ε is more feasible. Anyhow, the most important visualiser

80. e.g.: Berchtold et al. (2000) "Independent Quantization: An index compression technique for high-dimensional data spaces", 577-588. 16th International Conference on Data Engineering (ICDE), San Diego, CA. For more see Kriegel (2001) "Access Methods for High-Dimensional Data Spaces" http://www.dbs.informatik.uni-muenchen.de/Forschung/Index
81. SAP (2002) "SAP-DB Homepage" http://www.sapdb.org/
82. Dimensions should be either rolled up (= show all points of that dimension), integrated (see next footnote) or used for exclusion (= exclude all results whose value in this dimension is not in the selected intervall). Ideally, one would have the generated plot in the middle of the screen, a series of sliders for interactive manipulation of the excluding dimensions values below, and a list of the integrated dimensions on the right. It should be easy to change the status of each dimension and the whole setting should be recordable to faciliate building standard analysis sets with common parameter-look-and-feel.

for daily work with ε will probably involve few 3D diagrams, as skilfully made 2D plots can often contain much more information. To generate nice 3D plots for presentations, data might to be exported to a package that places more emphasis on static presentation beauty.

Excel-horror

Initial analysis of Simulator005 results with excel were possible, but extremely tedious and complicated in preparation (at least on a G3/350 Mac with 768 MB RAM, the only machine available at the time). A cursory estimation of the time needed to produce the plots needed for this thesis showed that it merits to look for a faster alternative[85], especially since Excel is limited to less than 32 000 results anyway, a barrier that is not very far away (23 000 were available at the time of the decision, 28 000 are available today).

R-system

I found the open-source R-system to be an excellent alternative[86]. It is much faster, runs with greater stability, can import even more than the 303 parameters I need[87], allows one to write readable code that leads to highly automated production of plots that allow control over more details than possible from within Excel. While it became clear quickly that R is the way to go, it took time to (i) learn how to program R for production of high qual-

83. This can be done by automated generation of plot symbols. For example the log of population size could by expressed by the size of the dot, mutation rate could be coded by a colour from blue (=low) to red (=high). If each dot has 8 lines that point from its center outwards and at the end of each line a small symbol, then the maximal number of dimensions that can be rolled up in such a visualisation is 50, as each line can have a thickness, a length and a colour while each small symbol at the end of a line can have a shape, a colour and a size. Needless to say that a capable visualiser allows to use smaller subsets for more easily understandable plots.
While learning to read these plots will take its time, once a standard coding system is established, the human eye will be very fast in identifying causes for the position of those points that can be read easily: the outlyers. In a potent visualiser the step from such an idea to a special plot showing the trend is small.
84. The largest standard visualisation package freely available is OpenDX from IBM. Commercial alternatives include IRIS Explorer, AVS and IDL. These are built to handle multidimensional data. The latest release of STATISTICA is built to handle excessive amounts of data too (Grant, 2001 "Statistica goes from strength to strength", Scientific Computing World Nov/Dez:32-34).
Other alternatives include math packages (Mathematica, Maple, Matlab), statistics packages (SPSS, Systat, StatView, MiniTab, Origin, ...), graphing packages (GnuPlot, SigmaPlot, DeltaGraph, ...) or good old limited Excel if no better solution has been installed yet.
85. Faster meant integration of the new system in the ε workflow + production of plots in similar or less time than I would have spent with Excel. Most of the latter time would consist in waiting for Excel to redraw some plots or save results, etc. You would not really want to work with a system that needs eg. about 60 minutes just to save an 85 MB file that had been imported from an 40 MB tabed file of results. A large number of similar frustrating events led to abandonment of Excel for this application.
86. R-system (2002) "R: A Language for Data Analysis and Graphics" http://www.r-project.org/ - Maindonald & Braun (2002) "Data analysis and graphics using R: An example-based approach", Cambridge, Cambridge University Press. - Ihaka & Gentleman (1996) "R: A Language for Data Analysis and Graphics", Journal of Computational and Graphical Statistics 5:299-314.
87. For Excel I had picked the 250 most important.

ity plots like those in Part VI and (ii) actually write all scripts needed for production of these plots. The current active script collection contains 23 high quality scripts with a total of more than 1 MB code with an estimated more than 15 000 lines of code. The best thing is that update of all plots is easy: To check where new results appear in an old plot, just re-run the script. All expectations have been more than satisfied with R and the overall time for system integration and results analysis was certainly not longer than I would have needed for the one-way production of Excel plots. As there are even modules available for R that facilitate retrieval of data from databases, the future of R in the ε server suite appears to be bright.

Datamining

Automated data mining for knowledge discovery techniques are potent tools that might be useful in search for understanding relationships between the various parameters[88]. As the evolution@home results database grows, these techniques will become more interesting and mutual stimulation of data mining technology and evolutionary biology might eventually turn out to be significant.

10.3.17 Comparison with conventional global computing frameworks

After explaining some details about the EEPSLION server suite it should be clear that existing frameworks for global computing (see page 10) cover only a very small part of the functionality required for efficient investigation of evolution by simulations. While they usually provide a secure frame for automated distribution of tasks and collection of results, they do not

- help to implement series of evolutionary simulation models
- generate the parameter combinations that are to be computed
- help set priorities to get hot tasks done first
- support handling of work units from seconds to months
- build or help to organise a results database
- help to analyse results.

As only an integrated solution will be able to unleash the power of global computing for evolutionary biology, currently there is no way to avoid the thorny path of building something like ε. In that process existing global computing solutions should be incorporated when

- it saves more time than it costs to integrate the new software,

88. Kriegel (2001) "Knowledge Discovery in Databases" http://www.dbs.informatik.uni-muenchen.de/Forschung/KKD - Klösgen & Zytkow, (eds, 2001) "Handbook of data mining and knowledge discovery", Oxford, Oxford University Press.

- the package provides *all* required functionality and
- no dangerous dependencies are formed by relying too much on proprietary solutions of companies that might go out of business or other software that will not be maintained in the long term.

10.4 HFF results database design

Before connections between different sets of data can be drawn effectively, an integrated data base solution is needed. This should have a structure that is simple and standardised enough to allow building of links between individual entries. While such interoperability is evolving for molecular bioinformatics data[89], evolutionary simulation data is rarely published electronically today. Thus, Design 5 makes the attempt to define a stochastic-simulation-centric database structure that might also be used to publish results on the web. Its core design decision was to put all data belonging to one parameter combination into different files in *one* folder. It uses XML[90] to facilitate future exchange of data with other similar databases and its design helps the server suite to manage simulation tasks and results.

The files within each such folder are described in Table 10. This structure greatly facilitates manual inspection and priority setting by having all information in one place. For systematic changes in many folders, the corresponding tools of the ε-server-suite automatically iterate over the relevant folders. Iterator functionality is provided by `EEBasicLibrary` (see page 90).

Hierarchical folder substructure

Implementation of such a system is possible, but not trivial, as even modern operating systems can have serious problems to handle millions of subfolders in *one* directory. Thus, a substructure is needed. To avoid arbitrary partitioning of consecutively numbered folders into a structure, where it becomes virtually impossible to find anything manually, a nested hierarchy of folders is employed. It reflects the values of the different parameter combinations that are being investigated. Thus, the folders in one level carry the name of the parameter assigned to that level with the corresponding values being investigated. Each of these folders contains all subfolders necessary to store all different values under investigation for the next parameter (one level below). Thus, all input parameter values are turned into folders. The se-

89. Karp (1996) "Database links are a foundation for interoperability", Trends Biotechnol. 14:273-279.
90. Spedding (2001) "XML to take science by storm", Scientific Computing World Sept/Oct:15-22. - W3C (2000) "Extensible Markup Language (XML) 1.0" http://www.w3.org/TR/REC-xml

Table 10 Files in the folder of a parameter combination in the HFFDB.
P = permanent files, T = temporary files , SRR = SingleRunResult, MRR = MultiRunResult, TS = TimeSeries, FSG = `FolderStructureGenerator`, MRRC = `MultiRunResultsCollector`, PS = `PrioritySetter`, RFP = `RunFileProducer`, RSI = `ResultsSortIn`, SRRA = `SingleRunResultsAnalyzer`

	Filename	processed by	Description
P	run.txt	FSG, SRRA, RFP	simulate command with run time prediction
P	RunScheduling.txt	PS, manual, SRRA	priority on the ε priority scale
P	SRRComplete.xml	SRRA	list of results from complete runs
P	SRRIncomplete.xml	SRRA	results from incomplete runs[a]
P	MRRCompleteOnly.xml	SRRA, MRRC	summary statistics of completed single runs only
P	MRRTotal.xml	SRRA, MRRC	includes incomplete runs too
P	TSid.pdf	SRRA	plots of all observed time series of type "id" from complete and incomplete runs[b]
P	HighScoresRawInfo.txt	SRRA, MRRC	who in ParticipantsDB computed how much for this parameter combination[c]
P	index.html	SRRA	nice html overview over results
T	IncomingSRR.xml	RSI - SRRA	these new results still have to be processed (check for duplicates, compute MRR, ...)
T	ErrDuplicates.xml	SRRA	saved for manual inspection until deletion[d]
T	ErrBadCheckSums.xml	RSI	"[d]
T	ErrInconsistentResults.xml	RSI	"[d]
T	LogBadResults.txt	RSI, SRRA	"[d]
T	LogGeneralErrors.txt		any module that finds any problem here

a. Incomplete results are removed, when a more complete result of the same run becomes available.
b. Each time series gets its own pdf file. If a combination is feasible, it still can be calculated later.
c. Incremental high scores information is not collected in this folder, but in a central file easily accessible to the high score generator. SRRA adds new high scores information to the central as well as to the local file after removing duplicates.
d. As duplicates, erroneous checksums and inconsistent results might point to weaknesses in the framework, these are not thrown away immediately. If their number exceeds a certain limit, the oldest ones are deleted.

quence of input parameters in the folder hierarchy can be adapted for each computing project. It should place parameters that change the least to the top, to reduce the amount of folders generated. A resulting path could look like: `Project1/dmer=0/dmeb=0/Ub=0/Sb=0/Cend=500/Tend=10000/Rmax=10000000/Ur=0.0003/Sr=-1e-5/N=1000/`. Preliminary tests under MacOS classic have worked well for nesting folders, while too many subfolders (eg. 100 000) in one directory led to unbearable slowdowns.

Implementation is not trivial

Navigation in such a hierarchy is not easy, as ANSI C++ can not determine content of a folder and either OS-specific code or a plain text file with the names of subfolders are required. Another difficulty is that absolute

path names beyond a certain limit (eg. 128 or 256 chars) are not supported by some operating systems, and sorting of new results into HFFDB might need considerable speed-up (eg. by remembering OS-specific handles of recurrently used folders in a RAM based map).

Limitations

The upper limit for parameter combinations that can be processed this way on a given hard-disk volume depends on the (adjustable) number of i-nodes generated when the volume is formatted. With a typical `newfs` UNIX command, one i-node is generated every 8 Kbytes, leading to 13 million i-nodes on a100 GB volume. Assigning input parameters to folder levels according to increasing numbers of parameter values used in the computing project makes it possible to keep administrative folder overhead at a fraction of the total number of parameter combination folders. More formally,

$$\prod_{i=1}^{P_{last}} p v_i \quad > \quad \sum_{i=1}^{P_{last}-1} \left(\prod_{j=1}^{i} f v_j \right) \tag{12}$$

where i counts input parameters from the first to P_{last} arranged according to monotone increasing v, the number of different values investigated; p is the number of i-nodes needed to store elementary files for a parameter combination (eg. 10 to 20, includes folder), f is the number of i-nodes needed for one folder in the super-hierarchy (eg. 1 = folder only, 2 = folder with list of subfolders). The left side gives all i-nodes needed to store data in files, the right side gives the i-nodes of the hierarchical folder administration overhead. Adding both sides gives the total number of i-nodes needed to store results of a particular parameter grid search computing project. Thus, a properly configured 100 GB volume might support up to 10^6 parameter combinations, if each needs less than 100 KB of data. This translates into allowing grid searches with eg. 4 parameters a 20 values, while adding a fifth with that many values is already too complex.

One volume per project

These considerations are important for the design of computing projects for a simulator, especially in light of the enormous heterogeneity of computational complexity. While it takes some time for a 100 GB disk to be filled with simulations that take many weeks each, a single PC can do the same in not many days – if simulations need less than a second. Thus, setting good priorities and project design are pivotal for governing global computing projects with evolution@home. To keep hard-disk size from limiting the HFFDB, different projects can be assigned to different volumes, known to

the server-suite by a central file that records access information to allow iteration over *all* parameter combinations.

Other solutions

The use of hierarchical databases has become rare with the success of relational solutions. The idea of using folders of the operating system is even rarer. However, the HFFDB approach has some similarity to IBM's FAP-ORES[91], and has been proposed to substitute XML repositories on some occasions[92].

The enormous heterogeneity of tasks highlights the danger of employing one of the many small scale databases readily available: their limits are easily reached and in the long term they would call for a database of databases. Total costs for the most well known industry standard professional relational database solutions (eg. Oracle, DB2, SQL-Server[93]) are prohibitive (see "Databases" on page 69). However, the increasing power of open source database management systems like SAP-DB and MySQL[94] call for a fundamental re-evaluation of the HFFDB concept. It is probably easier to learn administration of such a database and to implement the functionality described here on top of it, than to implement the HFFDB design from scratch. As these open source databases are used in professional industry production environments, they come with a large number of professional features like scalability, regular generation of backups, security concepts, and the like. The fact that these currently support up to 32 TeraBytes of data (SAP-DB) or more (MySQL) in one database suggests that such systems might be the best way available to handle simulation data in the long term (assuming one separate database for each simulator and one central high-scores database; by the time this will be too restrictive, larger open-source databases will hopefully be available).

SAP-DB looks very promising

10.5 General simulator and security

The framework part of the simulator (= worker) is pivotal for success of the whole global computing project. If potential participants like it and it is easy

91. Inouye & IBM TPF Development (1999) "Introducing Folders and Pockets - Outcomes Research (FAP-ORES)" http://www-4.ibm.com/software/ts/tpf/news/nv5n4/v5n4a13.htm
92. Bosworth (2002) "XML Magazine - END TAG XML Repositories: Do we need them?" http://www.devx.com/upload/free/features/xml/2002/01jan02/et0106/et0106-1.asp
93. See e.g. Barclay et al. (1998) "The Microsoft TerraServer™" http://research.microsoft.com/~gray/Papers/MSR_TR_98_17_TerraServer.pdf - Barclay et al. (2000) "Microsoft TerraServer: A Spatial Data Warehouse". Proceedings of the ACM SIGMOD, Austin, TX, ACM.
94. SAP (2002) "SAP-DB Homepage" http://www.sapdb.org/ - Banke (2002) "Open-Source-RDBMS von SAP: Erfaphrungswerte", iX August:78-80. - See http://www.mysql.com/

to handle, they will give it a try; if not, they will get another worker or waste their CPU cycles. Any worker design has to choose between the following extremes:

- Simplicity of implementation versus simplicity of use and
- less interesting general user interface (eg. show CPU and RAM usage only) versus interesting special user interface (eg. show the molecule computed).

Fault-tolerant systems are an important, long debated topic in informatics[95]. Design of global computing systems is a wonderful playground for development of fault-tolerance strategies, as virtually every conceivable error will attack computation due to the extremely heterogeneous and uncontrollable system background. General design of the simulator plays an important role in handling all kinds of fault.

10.5.1 The simplest design possible

In the beginning, evolution@home started with the simplest possible semi-automatic implementation. Such a simulator

- is a command line application
- allows to set upper limits for RAM and CPU time complexity
- can read special input (run-files) from its current directory
- says when it starts a new simulation (with estimated duration)
- allows for temporary interrupts of computation by some fussy procedure[96], but not to save its current state to disk
- computes results and writes them to its current directory
- allows anonymous contributions
- has a check-sum computing methods to ensure that transmitted information has not been altered
- reports potential simulation errors and intermediate results (where the latter is a speciality of the problem domain)
- records preferences with the amount of work conducted, long-term average performance, and participants' information for inclusion in future high-scores.

95. Meinl (1988) "Fehlertolerante Anwendungssysteme - Möglichkeiten und Probleme", HMD Handbuch der modernen Datenverarbeitung 25:14-23. - Echtle (1988) "Ansätze zur Fehlertoleranz von Rechensystemen", HMD Handbuch der modernen Datenverarbeitung 25:3-13.
96. It not possible from ANSI C++ to check, whether a key has been pressed without waiting for a key to be pressed. If use of an OS specific GUI framework becomes too complex, a possible work around is, to check for a file with the name "break.txt" in the folders directory. Of course it is much better to use a corresponding cross-plattform GUI framework.

Such a simulator needs to be complemented by software that can produce input-files (run-files) and can analyse results-files. Then, a web server is needed for distribution of executables and run-files, as well as appropriate email software for results submission. Disadvantages of this approach include the need for manual interaction, when all simulations in a run-file have been completed, and the text-based approach to everything. This scares many participants away and seriously limits computing power accessible (together with lack of high-scores and print media coverage).

10.5.2 The ideal evolution@home simulator

While the simple approach is a good place to start, over time the following features should be added to evolution@home simulators to make their use an enticing experience (see "Ideal worker software features" on page 9):

Start and forget

- **Full automation.** This includes not only automated download of new run-files and automated results submission with buffering techniques for situations where no Internet-connection is available. It includes also automated download of new releases of the current simulator and download of new simulators. If computing code and worker framework code are not part of the same distributed binary, they are updated separately.

Trust certificates of evolution@home

- **Security mechanisms for participants.** As any Internet connection is a potential security risk and automated code download even more so, the ideal goal of a completely secure global computing system will be never attained. However, system design should make it as hard as possible to threaten participants. Many global computing systems choose Java and its sandboxing techniques to keep potentially dangerous application code from doing any damage. This is not so easy for evolution@home, as it uses C++ to keep the super-computer option for complicated problems and native code called from Java is as insecure as original C++. Thus evolution@home will not work with some measure of trust between participants and the makers of evolution@home (participants have to trust in effectiveness of Java sandboxing too, after all.)

  To ensure that everything works properly, and downloaded code is indeed what everybody believes it is, no executables will be published, if code is not available to trusted evolution@home scientists and has been tested extensively. Then a non-published algorithm computes checksums for each downloadable simulator. Download itself uses

encrypted files to make potential attacks as difficult as possible. Before a new simulator is started, these checksums are computed again and compared to the originals downloaded separately from the website. No code will be started, if its checksum is bad (due to transmission errors or some hacking attack). All this is done automatically. Additional security might be added by Software Fault Isolation[97] and operational code authentication techniques[98].

Extremely careful users can turn off all automatic connections to the Internet and restrict computation to certain simulators only, manually request code checks from their simulator and compare checksums with those published. If the user is still unsure, certification by the central authority on the simulator should be possible to facilitate detection of hacked simulators. All in all, the system should be as transparent and configurable as possible.

Checksums and repeats

- **Security mechanisms for results.** Reliability of results might be endangered by anything from transmission errors to cheater attacks. To detect such instances, a checksum is computed for every result besides some plausibility checks. Future simulators will provide a richer and more secure mechanism[99] than the one currently employed. A nice property of stochastic simulations is that they have to be repeated several times anyhow. In a global computing setting, this is most likely to happen on different machines. The probability that all these participants manipulated content in the same way is extremely low and a false result is likely to be detected as an outlier during analysis. If nevertheless uncertainties remain, the corresponding results can be scheduled for computation on trusted computers.

Open communication

- **Transparent communication lines.** To build trust between users and the evolution@home system, communication lines should be as open as possible. Thus, run-files and results-files are all plain text XML files that are only protected against modification, but not against reading. All automation features can be turned off one by one, to configure simulators as individually and non intrusive as pos-

97. Small (1997) "A Tool for Constructing Safe Extensible C++ Systems". Third USENIX Conference on Object-Oriented Technologies (COOTS), June, Portland, Oregon.
98. Lawson (1999) "Operational code authentication" http://www.distributed.net/source/specs/opcodeauth.html - Fedack et al. (2001) "XtremWeb: A generic global computing system", pp. 582-588 in: Buyya & Mohay (eds) First IEEE/ACM International Symposium on Cluster Computing and the Grid - CCGrid 2001, Brisbane, Australia, IEEE Press.
99. Golle & Mironov (2001) "Uncheatable distributed computations" http://crypto.Stanford.EDU/~pgolle/papers/distr.ps

sible. Thus, participants might even choose semi-automated operation. Please note that evolution@home is not a commercial global computing system that has to keep results or computed models secret. Participants are allowed to see what they contribute to, so that they can check themselves that their computing time is not abused (eg. for some obscure military weapon design program[100]).

Use only *one* static address

- **Flexible server contacts.** It is the mission of evolution@home to develop a larger number of simulators, than can be managed by a few servers, and to stay active much longer than the expected hardware life-time of most servers. Thus, it should be easy for the simulator to learn about new Internet addresses for getting run-files, submitting results, downloading updates of itself and downloading new simulators. This is facilitated by only a single fixed address that points to a file on the evolution@home website. This file contains the address of the `ServerNameServer`, where a list of all servers of all active simulators can be found, including some general priority suggestions (see page 103). The simulator automatically picks one of the hottest simulator-servers to get the corresponding simulator (if necessary), download a run-file, and submit results after computation. The system makes simulators return to the other addresses they know, when some server cannot be found. Thus overall traffic at the one central pointer to the `ServerNameServer` should be low. A list of recently used servers and a server-problem-reporting-channel will make this system flexible enough to deal with temporary server problems. Related security issues will be investigated.

Users choose commitment

- **Interrupt handling.** Besides the fundamental computational complexity choices required for evolution@home participation, flexible on the spot governing of computing commitment will be possible by (i) just clicking a button to interrupt computation at any time, (ii) choosing other pre-emptive priority levels than the lowest default, (iii) saving the current state of the simulator completely to disk to allow resuming computation eg. after system shut-down and (iv) scheduling arbitrary interruptions of computation for users who want their CPU to rest once in a while (eg. to cool down in summer).

The double continuation problem

- **Save to disk.** The possibility of saving an intermediate state to disk creates a delicate problem for a global computing framework. How

100.Scheppach (2001) "Jetzt wird aus jedem Computer ein globales Supergehirn!" P.M. April:80-86.

can be made sure, that users continue only once from one intermediate state, if the intermediate state is written to file and can be easily copied and moved? While it is easy to keep 'stupid cheaters' from continuing the same simulation twice, it will not be easy to do so for 'intelligent cheaters'. The problem is that results generated from these independent continuations are not really independent, and are hard to detect with usual techniques like checksums. Moreover, they might not be identical to allow simple identification by the `SingleRunResultAnalyzer` (eg. if both are intermediate results from different stages). The most probable solution is a system the gives each distributed task a unique RunID that is stored on the distributing server and cannot be faked (long random string). If this RunID is included in every result, it becomes easy to identify multiple results that stem from one run. However, costs on the distributor's side might be too high for such a solution (store all RunIDs, even if many might never be returned; no static run-files possible).

- **Basic global computing GUI.** An easy to use graphical user interface should facilitate access to all functionality, easy setting of preferences and display progress and local computing statistics. — Nice GUI
- **Evolution GUI.** One of the strength of evolution@home simulators will be interactive observation of evolving populations and their features. This functionality is necessary for the scientists who investigate the models and can thus be provided at no extra cost to the general public. It will be possible to look at current values of all parameters, see plots of time series parameters, make snapshots that visualise the current state of the population or even watch evolution of spatial patterns. All observations can be saved to PDF files. — Another nice GUI
- **Form of executables and availability.** The publicly distributed simulator should be available as stand alone application, screen saver and system service (can run in background, even if no user is logged on). Either should have a small, easy downloadable installer that might be distributed on CD and requires only a double click for installation on a wide variety of Windows, MacOS and UNIX systems. — Wide distribution

10.6 Describing the world in *EEPSLION*

An important part of ε is to provide facilities for easy implementation of evolutionary and ecological models. This part is completely independent

from global computing and can be used for implementing models that run on one computer only. For two reasons, only the basics of this infrastructure are reviewed in this and the next two sections. First, this work does not provide enough space for a thorough review of all details, not even of those whose design has been completed. Second, an important property of ε is its ability to grow with its needs. Thus many of the opportunities of this framework will be realised only when appropriate extensions are made for the models that require them. This extensibility, a key strength of this framework, will be highlighted in the following description.

Basic worldview

EEWorld
EEHabitat
EELocation
EEPopulation
EEIndividual

The basic modelling worldview of ε is the following: Everything that is being modelled is controlled by a specific class derived from `EEWorld`. The latter implements a model specific worldhistory that describes the sequence of events from the first date of `WorldTime` to the last date simulated, and calls the worlds `evolve()` method for the appropriate number of times. The one single object of this model-specific `EEWorld` contains objects of classes derived from `EEHabitat`, which contain objects of a class derived from `EELocation`, which contain objects of classes derived from `EEPopulation`, which contain corresponding objects of a class derived from `EEIndividual` or `EEIdenticalCohort`. Individuals might contain objects of classes needed for modelling genomes. These classes provide `evolve()` methods needed for computation of the state of the world in the next moment of simulated time from the current state of the world. Both states are stored separately to ensure overall integrity, even with complicated non-linear access schemes.

TimeResolution

The time between both is given by `TimeResolution` units. The choice of a `TimeResolution` for a model is very fundamental, as a large number of details in the model's code are adapted to it. While the simplest case uses 1 generation, the underlying 64 bit float with more than 15 digits precision allows simulations of about 2 billion years with a resolution of 1 minute (if 10^6 timesteps/sec could be computed, such a single run would take more than 33 years). In most models, `TimeResolution` will be the same for all parameter combinations and remain constant throughout worldhistory, as deviations from this standard will make models exceedingly complex.

10.6.1 EEBioObject

Providing easy and flexible iteration is difficult

One of the challenging aspects of framework design is to facilitate recurring tasks by providing corresponding infrastructure. Iteration over all important elements of this model is one such task: For evolution, the `evolve()`

method of each individual in each population of each location from each habitat needs to be called once during each timestep of evolution. To allow for the flexibility expected from a framework, a simple array solution can not be employed, as it is eg. not desirable to call dummy `evolve()` methods of individuals that do not exist. Thus, a system is needed that allows easy integration and removal of arbitrary numbers of items (as a list does), while allowing use of array-type random access when needed. As not both ways to implement a collection can be used for the same collection, it should be possible to choose appropriate implementations for each level of the `EEWorld` hierarchy. This requires an interface that implements the *composite* pattern[101] and allows use of different implementations of collections in the same hierarchy. It is implemented by `EEBioObject`, the abstract base class that

- `EEWorld`,
- `EEHabitat`,
- `EELocation`,
- `EEPopulation` and
- `EEIndividual`

inherit from. `EEBioObject` declares (i) iterator related methods for access of the first, last, previous and next lower objects (relative to a given address), (ii) node related methods for access of the previous and next elements of the old and new state on the same level, (iii) appropriate constructors and destructors for lower objects, (iv) script command execution methods that allow such classes to have a corresponding `EEInterpreter` for non-public, interactive work with the simulator, (v) general evolution related methods like `evolve()`, (vi) composite related methods that eg. count lower objects, check consistency and facilitate saving and loading of the current state with its associated lower objects, and finally (vii) methods that return some general information. With the help of an array of `EEPosition`-objects that can remember the position of the current lower objects, be it in a list or in an array, ε needs to know only the corresponding methods of `EEBioObject` in order to iterate over all individuals of the modelled world. Such a general functionality is provided in the `evolve()` method of `EEWorld`, which calls the `evolve()` methods of all relevant entities of the model – without any new adjustment of iteration methods, array bounds or other error-prone code details.

EEBioObject uses the *composite* pattern

101.Gamma et al. (1995) "Design Patterns", Addison-Wesley.

A similar infrastructure can be used to iterate over all elements in order to save the current state of a model to disk or load it from there. In this case, appropriate `save()` or `load()` methods are called.

10.6.2 The evolve() methods

Virtual inheritance ensures that the correct evolve() method is called for each object. This is one of the great strengths of the use of object-oriented programming for simulation of individual-based models: The framework only knows that the corresponding objects have an `evolve()` method, without knowing any details and the compiler makes sure that the correct code is called. This results in code that is much more elegant and easy to maintain.

Some people have wondered whether the compilers overhead needed for this would not offset its benefit. However, if the same flexibility is implemented by manual branchings, it will be hardly any faster. Furthermore the Barton-Nackman Trick might be used to transfer a considerable part of this overhead from run-time to compile-time by using templates[102].

To ensure flexibility and integrity, the following methods are used to evolve an entity from one moment to the next:

- `EvolveOneStep()` will do only the minimum of what the model requires to compute a consistent state of the model after the next `TimeResolution` units of time have passed. The new state is computed without changing the old state, as any old state object might access any other old state object to compute a new state, but this must not have any effects on iteration over all individuals of the old state.
- `EvolveAndObserveOneStep()` will observe additional features of interest. This can be computationally complex, so the framework allows simulators to adjust the amount of detail observed.
- `Actualize()` will copy the new state to the old state after the new state has been computed for all individuals and other entities of the model. This requires an additional iteration over all elements of the model, just like `evolve()`.

It should be noted that evolving a population means more than evolving all its individuals: some features of the population itself might need a sophisti-

102. Striegnitz & Veldhuizen (2001) "LRZ Tutorial: Writing efficient programs with C++" http://www.lrz-muenchen.de/services/compute/hlrb/manuals/lecture_notes/HPC++1.ps.gz - Barton & Nackman (1994) "Scientific and Engineering C++: An Introduction with Advanced Techniques and Examples", Reading, Massachusetts, Addison-Wesley.

cated method for updating (eg. to implement density regulation or count clicks of Muller's ratchet). The same can be true of locations, habitats and the world itself, so derived classes of the final model on each of these higher levels need their own `evolve()` methods. As the standard iteration of ε is over all individuals in all populations in all locations in all habitats, these special `evolve()` methods might be used to implement considerable deviations from this standard structure.

10.6.3 Fast access to objects in the world

Besides being important for automated iteration over all `evolve()` methods, the standard structure above is also needed to allow constructions of standard methods of access to individuals. While some simple simulations do not need that, more complex models involving interactions between individuals need an easy way to search for other individuals in the same location (predators, prey, mates, etc.). This relates also to the modelling of migrations (access to other locations and potential individuals there). While `EEPosition` objects can remember the position within any `EEBioObject` collection, `EEGlobalPositioningSystem` manages an array of `EEPosition` objects that can remember the exact position of any simulated object and of all the collections it belongs to.

Small objects do not store their collection

The latter is important, as evolving individuals might need access to some property of their population, although they do not store a pointer to their corresponding population object. To allow this, they use an access method to get the information from the `EEGlobalPositioningSystem` object used by the currently active `evolve()` method. The same system is used to get access to properties of locations and habitats. If individuals need to find objects with particular properties that are not known at compile-time, they might use a special method searching this hierarchy of objects. While the more general of such search methods might become part of the framework, most are probably so specialised that they will remain in the simulator specific code section as a derived class of `EEGlobalPositioningSystem`. Anyhow, such searching has to be used with great care, as it might easily lead to excessive computing times.

10.6.4 EEWorld

`EEWorld` manages `TimeResolution`, `WorldTime`, the evolve-iterator of type `EEGlobalPositioningSystem` and some other infrastructure. It

provides an `EEInterpreter` that manages some general `EECommands` and `EEParameters`, common to all models. It is derived from `EEBioObject`, but still an abstract base class. To use it for a simulator, a derived class needs to implement all model-specific functionality. Employing the singleton pattern[103], the *one* object of this derived class becomes the main integrating structure of the simulator. It encapsulates all other model-specific objects. It can be reached itself under the global pointer `eWorld`. Among other details, any specific simulator needs to (i) define the model's `EEParameters`, (ii) implement model-specific commands, (iii) implement the method `RunSingleWorldHistory()`, and (iv) make all habitat objects accessible, as each world is a collection of habitats.

Some simulators will need large additional input files to define the exact structure of their models (eg. field data of a metapopulation structure or other geographical information). In such simulators `eWorld` is responsible for reading such files and distributing the data to the corresponding objects.

10.6.5 EEHabitat

Each habitat is a collection of locations. While simulations without spatial aspects will only use a minimal implementation with one location, this abstract base class might be extended to incorporate a wide range of spatial structures used in current spatial ecology research[104]. This includes 2-dimensional landscapes, 3-dimensional waterbodies, arbitrary meta-population structures, and grid-based or vector-based geographic information systems data. Corresponding classes will be incorporated into the framework when they are needed. Some models require additional functionality from their habitats (eg. seasonal variations of some properties that depend on input-parameters). This will be implemented in classes that are derived from the habitat class with the appropriate spatial structure. These derived classes also know the derived location classes of a simulator.

103. Gamma et al. (1995) "Design Patterns", Reading, Massachusetts, Addison-Wesley.
104. ESRI (1998) "ESRI Shapefile Technical Description. An ESRI White Paper" http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf - Durrett & Levin (1994) "Stochastic spatial models: a user's guide to ecological applications", Philos. Trans. R. Soc. Lond. B Biol. Sci. 343:329-350. - Dunning et al. (1995) "Spatially explicit population models: current forms and future uses", Ecol. Applic. 5:3-11. - Ball (1994) "Ecosystem modeling with GIS", Environmental Management 18:345-349. - Slothower et al. (1996) "Some guidelines for implementing spatially explicit, individual based ecological models within location-based raster GIS" http://www.sbg.ac.at/geo/idrisi/gis_environmental_modeling/sf_papers/slothower_roger/sf23.html - Czaran & Bartha (1992) "Spatiotemporal dynamic models of plant populations and communities", Trends Ecol. Evol. 7:38-42. - Gathmann & Williams (1998) "Inter-site: a new tool for the simulation of spatially realistic population dynamics", Ecol. Model. 113:125-139.

It should be noted that one world can contain different habitats. This allows construction of models that incorporate different spatial structures, eg. a metapopulation that interacts with another species from a normal 2D lattice landscape or a species that lives in two habitats of different structure. As all such habitats are part of the same planet, a spatial coordinate system is provided to allow transformation of coordinates of one habitat into coordinates of the other, should that be needed. If habitats do not overlap, migrations between such habitats need to be defined manually in the code.

Input and output parameters that are related to such habitats should be implemented at this level, together with corresponding commands for interactive exploration of habitat evolution.

10.6.6 EELocation

Each location is a collection of one or more populations, depending on the number of species in the model. Its size depends on the model. Interactions of individuals at one location have the same probability, whereas individuals at other locations need to come to this location first, before they can interact with anything at this location. Classes derived from `EELocation` need to know the particular population classes whose individuals might access the location. Besides a very simple implementation with one population only, alternatives might either offer a fixed number of populations or a variable number of populations at the same location. This might be needed for simulations where few individuals of different species migrate over large numbers of locations, as the memory needed for empty locations might considerably influence RAM complexity.

10.6.7 EEPopulation

Each population belongs to one species only and is either a collection of individuals or contains an array of identical cohorts. The standard evolve iterator walks linearly over all items of a population. This behaviour can be switched off for models that require non-linear access to individuals of a population (eg. Wright-Fisher sampling). In this case, the evolution methods of the population need to make sure that the new state of every individual is being computed. For more details on populations that consist of identical cohorts see page 127. Now some details on populations that consist of actual individuals are discussed.

Separation according to species facilitates access and counting, as there is no need to sort individuals into species at a location. This also allows different types of populations to be used for different species. In case of large spatially dispersed populations, it is important that empty populations use as little RAM as possible, to reduce memory complexity of simulations.

`EEPopulation` itself provides only a very general population interface, because there are two extreme possibilities to implement a population of individuals: a vector and a list. Main advantage of a vector implementation is its fast access to arbitrary individuals, even if the population is very large (a simple list would always have to go through about half of the population to find the random individual). However, this is bought by an extra need for RAM, as empty places in the population use the same space as if individuals were present. If this is too costly, use of a list-based population is more feasible. There all non-used individuals can be handed over to an appropriate memory management heap that allows recycling of individuals even to other locations or habitats. Thus the amount of RAM needed reflects the real number of individuals and not some artificial implementation requirements. Of course vector size can be reduced by reallocation, but considerable run-time costs prohibit extensive use of this possibility.

Thus it is easy to see that there are many different ways to implement a population and this number increases even more if all the different possibilities for density regulation etc. are considered. Therefore a large number of abstract bases classes derived from `EEPopulation` will become part of ε to allow programmers of simulators to choose the appropriate mechanism. Usually, a particular population class will only be used together with an appropriate individual's class.

10.6.8 EEIndividual

i-state *configuration* models

Models that use any type of `EEIndividual` objects belong to the class of i-state configuration models[105], as each individual's configuration is modelled by a separate data structure in memory. Each species in a simulator is implemented as one particular class of individual that is derived from a class that can interact with the proper population class and that is itself derived from `EEIndividual`. Over time quite a collection of types of individuals' classes will accumulate in ε, so that models in the future can reuse time-tested code from earlier models – given it had been coded generally enough. An important point of interaction is the implementation of density regulation, as it requires individuals' and populations' code to work together.

Among other things, the general individual's code interacts with memory management to remove no longer used individuals to corresponding heaps and to reuse such heap individuals, before a brand new individual is allocated in RAM.

To implement a model-specific individual's class, the most important part is to write the corresponding `evolve()` methods and all subfunctionality they need. These methods are so important to the model that their definition might become even part of a scientific paper.

10.6.9 EEIdenticalCohort

i-state *distribution* models

Models that use any type of `EEIdenticalCohort` objects belong to the class of i-state distribution models[105], as individuals of identical type are subsumed under a single number. While this approach allows for much larger numbers of individuals to be modelled[106], it requires enormous amounts of RAM if a large number of different states are allowed in the model, as one individual's counter needs to be present for each state, even if no individual is present. Usually, such models are implemented with vectors. They can allow significant analytical tractability (see appendix).

10.6.10 EEGenome

An important question in modelling evolution is how to abstract genomic complexity into a computable model. As simple copies of the full sequence of eg. the human genome would fill about 3 GBytes per individual, without abstraction, any simulation of evolution becomes nearly impossible due to computing demands. Therefore, genomic models have to capture the essentials of a particular question to facilitate computation, understanding and perhaps even analytical tractability. Unfortunately, there is an overwhelmingly large number of potential abstractions. Thus ε will provide enough flexibility to build new abstractions and offer a number of classes that allow easy implementation of recurrent models. Some standard approaches to build genomic models have been implemented by the Genetic Simulation Library[107]:

105. Caswell & John (1992) "From the individual to the population in demographic models", pp. 36-61 in: DeAngelis & Gross (eds) Individual-based models and approaches in ecology, New York, Chapman & Hall.
106. Scheffer et al. (1995) "Super-Individuals a Simple Solution for Modeling Large Populations on an Individual Basis", Ecol. Model. 80:161-170.
107. Conery & Lynch (2002) "Genetic Simulation Library" http://www.csi.uoregon.edu/projects/genetics/GSL/ - Conery & Lynch (1999) "Genetic Simulation Library", Bioinformatics 15:85-86.

- The **virtual genome** counts only the number of mutations per genome and is therefore the easiest, smallest and computationally fastest genome model. All new mutations are assumed to hit at previously unmutated sites only, which is reasonable if the overall number of mutations is small compared to the number of basepairs in the genome. Such a genome might be used for asexual populations, where the identity of mutations is not of interest. If an individual's genetic fitness (the phenotype) is updated as new mutations occur, then even variable mutation effects can be modelled. However, this model does not allow for recombination, as it does not record the identity of mutations. For illustration, compare offspring of two parents that are homozygous for the same 20 mutations to offspring of parents with the same number of mutations scattered throughout different sites. While the former will produce identical offspring, recombination is likely to produce quite different offspring in the latter, and a virtual genome model could not tell the difference.
- The **infinite genome** avoids this problem by recording only mutated sites, adding a new site with each new mutation. To save memory, each mutation is represented by a single bit. Then fitness can be computed by counting mutations in a genome, assuming that they have all the same effect and are completely independent of each other (multiplicative fitness model). Recombination between different mutations is essentially free. While this facilitates insight into general effects of recombination, it does not allow for varying degrees of linkage and variable mutation effects, two important factors in evolution. Furthermore, length of evolution is restricted by the assumption that each new mutation hits a new site (memory eventually overflows).
- The **sparse genome** stores each mutation as a separate structure and thus remembers its position in the genome and its effect as well. However, computational complexity is significantly higher than with the infinite genome. This further limits the length of evolution that might be observed.

Over time, more detailed genomic models will be constructed and will be incorporated in to ε. The corresponding classes will use `EEGenome` as abstract base class to facilitate investigation of different genome models in a particular simulator. The following classes might be used to construct more detailed models of genomes.

10.6.11 EEChromosome

Each individual has only one genome, but this might be viewed as a collection of chromosomes that might be autosomes, sex-chromosomes or extranuclear genomes like mitochondrial DNA or plasmids. Each chromosomal type has a particular way of inheritance and consists of a collection of genes where physical distance and genetical distance (frequency of recombination / degree of linkage) are correlated according to the recombinational landscape[108]. Although the latter relation can be assumed to be roughly linear for most modelling efforts, classes derived from `EEChromosome` can facilitate implementation of subtle inhomogeneities.

10.6.12 EEGene

As for genomes, a similar variety of different models can be conceived for genes. From simple counting of mutations up to storing whole sequences, it depends on the particular simulator to decide what details to devote computing capacity to. While it might be reasonable to neglect intragenic recombination in some cases, it should be noted that some questions will not allow that. Such questions might be addressed by inhomogeneous genome models that facilitate detailed observation of a few genes on a simulated, realistically large background of hidden details. To model details of genes like exon-intron structure or relations between protein sequence, structure and effect, even more classes besides `EESequence` and `EEDistributionMutationalEffects` might be necessary (eg. `EEProtein`, `EEIntron`, `EEExon`). These should be added as needed. Besides genetic information, `EEGene` can store properties that describe the corresponding phenotype to facilitate fast computation.

10.6.13 EESequence

If a simulator focuses on a few genes only, the easiest approach might simulate full raw sequences. `EESequence` encapsulates all methods needed to handle nucleotide sequence data. It is worth investing in a careful implementation, as each nucleotide contains only 2 bit of information, so 4 basepairs might be stored in 1Byte, where as a textual representation needs 4 times as much memory. Furthermore, these two bits can be organised to facilitate application of transitions and transversions (change 1 bit only). Fi-

108.Eg.Lander et al. (2001) "Initial sequencing and analysis of the human genome",Nature 409:860-921.

nally, `EESequence` provides a number of methods to help with various evaluations.

10.6.14 EEDistributionMutationalEffects

Sometimes the plain sequence is less interesting than the number and type of mutations it contains. To store large numbers of mutations of very different effects, one can use `EEDistributionMutationalEffects`. It builds a block of linked mutations that are binned according to their effect. As effects vary over many orders of magnitude, the various bins cover a log scale of adjustable granularity. Each bin counts only the number of mutations and, thus, introduces small inaccuracies by forgetting the exact effect of the mutation. However, the overall picture of which bin of mutations plays the most important role is easy to recover. This might help investigate one of the great mysteries in biology: the distribution of mutational effects.

One note of caution should be added for systems with recombination. A careful biological discussion has to evaluate model assumptions concerning intragenic recombination, as a rough calculation quickly shows that recombination events per basepair[109] can easily reach frequencies similar to mutation rates and thus might significantly alter conclusions.

10.6.15 EEMemoryManager

In C++ memory management has to be implemented by the programmer, in contrast to languages like Java, where a garbage collection removes no longer used objects from memory. While implementation of memory management is extra work, it confers the advantage of potentially higher execution speed. In principle, two approaches are possible: Use of `new/delete` operators and reuse of objects. Generally, the `new` operator calls the constructor method(s) of a class to initialise a new object. It has to be matched by exactly one call of the `delete` operator that calls the destructor method(s), when the object is no longer needed. This works well for objects with long lifetimes. However, individuals in populations live for extremely short time spans from an evolutionary perspective and thus many `new/delete` calls confer significant run-time costs. Therefore, reuse of objects is the faster alternative, as has been confirmed in preliminary tests.

109. Lander et al. (2001) "Initial sequencing and analysis of the human genome", Nature 409:860-921 reports 1 or more recombination event per meiosis in 100 Mbp. This equals to 10^{-8} recombinational events per base pair, a figure comparable to the mutation rate per meiosis.

In ε, reuse of `EEBioObjects` that are nodes of a list is mediated through the singleton `eMemory`, a global object of type `EEMemoryManager`. It manages an `EEHeap` for each class of reused objects. Whenever an object is no longer used, it is moved from its current list to the heap, an operation that is very cheap in doubly linked lists. There it waits until a new object is needed, as the corresponding heap is checked for reusable objects, before new ones are generated. This type of memory management is very flexible and allows one to devote memory to those objects that actually need it. All objects in `eMemory` can be deleted upon request. Furthermore, `eMemory` helps simulators to keep memory requirements within given limits by providing methods that compute the current amount of memory used. These are to be called whenever an object is allocated or deleted.

Another important goal of memory management is to avoid the use of virtual memory on modern operating systems. This is necessary, because simulators typically crawl through their whole memory for *each* simulated timestep. This leads to excessive paging and can effectively block even modern pre-emptive multi-tasking operating systems, as paging runs on a high priority, even if the process that asks for it runs on a low priority. Thus simulations whose memory requirements exceed physical RAM on the system that is about to run them should never be started. As there is no ANSI C++ way to determine physical RAM, the participant is asked to provide the maximal amount of RAM he is willing to commit to the simulation.

10.6.16 EEController

After every detail of evolution is implemented and the worldhistory can be run just by calling a single method of EEWorld, a significant amount of administrative tasks remain. Besides starting new simulations and handling user requests, some general parameters and commands need to be implemented. Furthermore, commands for automated access to all interpreters and their parameters are needed eg. to produce a complete list of parameters. User preferences and a graphical user interface need to be managed besides other general simulator functionality. The professional release of the simulator requires implementation of even more commands that are included by conditional compilation.

Such functionality is accessible via `eCEO` (ε central execution officer), a global object of class `EEController`. It helps observe computing statistics, personalises results, contains the main loop of the simulator, and implements similar functionality that is simulator-independent and important to

general simulator functionality (see "The ideal evolution@home simulator" on page 116).

10.6.17 RND generators

No stochastic simulation environment is complete without reasonable random number generators. As random numbers are heavily used in ε and most simulators need only one good general solution, each random number generator is implemented in one global function that is optimised for speed. While one could also use object-oriented technologies to encapsulate various recurrently used functionalities, the corresponding overhead usually does not pay off in terms of code readability or ease of development.

Generation of random numbers is not as easy as one might expect. While it is generally impossible to get true random numbers from a deterministic computer, several methods exist that give sequences of numbers that are very similar to random numbers. Main requirements are a uniform distribution over their range and a period that is much longer than the number of random derivates needed for a simulation. Furthermore, each random sequence needs to be initialised by a unique seed, as identical seeds lead to the same pseudo-random sequence on the same system.

Currently, an adapted version of the long period random number generator of L'Ecuyer with Bays-Durham shuffle is employed[110]. It returns random derivates at 64 bit floating point precision between 0 and 1 and repeats itself only after more than $2x10^{18}$ calls. Assuming that it is called about a million times per second, it would still need more than about 500 000 years of computing time, before repeating itself. In case a super-astronomical period ($>10^{5000}$) is necessary, one might refer to the Mersenne twister algorithm[111].

Seeds for RND generators have to be derived automatically to ensure that each simulation is different. Currently, absolute date and time is transformed into a 32 bit integer that changes with every second. This assumes that simulations of an identical parameter combination are never started in the same second. Validity of this assumption depends on the number of participants and on the way run-files are distributed. While rare repeats by dif-

110. See ran2 on page 282 of Press et al. (1992) "Numerical recipies in C". 2nd, Cambridge, Cambridge University Press. - Original reference: L'Ecuyer (1988) "[Long period random number generator]", Communications of the ACM 31:742-774.

111. Matsumoto & Nishimura (1998) "Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator", ACM Transact. Modeling and Computer Simulation 8:3-30.

ferent participants from different parts of the world should not be regarded as a problem, this mechanism can be dangerous when large clusters of identical nodes are used to compute similar parameter combinations; currently they must not start identical simulations in the same second, an annoying circumstance. Alternative mechanisms include a complex algorithm that changes the current time seed in another random way (eg. using machine name, uptime, etc.) or dynamic distribution of unique seeds by the ε server suite with distribution of run-files.

10.7 Types of Data and their analysis

An instance that significantly complicates simulations of evolution is the enormous diversity of types of data that are being generated. Without a sophisticated system for handling the various types of data, seemingly endless manual searches in piles of results effectively block meaningful analyses. Thus, ε provides an infrastructure for the most common types of parameters, while it still allows for definition of new types of parameter.

10.7.1 Definitions of types of parameters

If one thinks about parameters needed for or produced by the simulation of evolution, the following classes appear:

- **Administrative** parameters store user preferences and other details.
- **Input** parameters define the course of evolution.
- **Output** parameters record results of the simulation
- **Synthetic** parameters are computed after the simulation by simple formulae that combine any input and output parameter. They will facilitate analysis of results and knowledge discovery.

While input parameters and synthetic parameters can be contained by a single 64bit float without loss of generality, some administrative parameters need at least the possibility of storing texts (eg. the name of the participant). Output parameters are the most complex group, as they have to describe any potentially interesting data that can be observed during evolution. Considering all kinds of potential results leads to the following ontology of output parameter types:

- **Single numbers** need only one number to sum up the whole simulation, like eg. extinction time.
- **Time series of single numbers**. Some properties of interest produce one number per timestep or happen even rarer (eg. clicks of

Muller's ratchet). Such time series should be analysed automatically, to extract interesting features like trends, oscillations, mean, standard deviation, etc.

- **Time series of distributions**. Some properties of interest produce a distribution for every timestep (eg. distribution of fitness in a population). Again, such time series should be analysed automatically.
- **Snapshots** in various forms. Sometimes detailed detailed distributions at a particular point in time are interesting. This might include anything from a histogram of the fitness values in a population to detailed spatial data needed for high-end 3D visualisations.

A single ε simulation has no distribution as output

It is remarkable that a single simulation of evolution never generates results of a simple distribution type. It might observe many distributions, but since these are observed for each moment in time, a single distribution is either a snapshot (eg. at the end of the simulation) or a time series of distributions (eg. average over the whole simulation). Thus, ε does not need a single-run-parameter type for distributions, but it should handle snapshots and time series to avoid potentially dangerous artificial reductions of information.

Automated access

One important task of the class `EEParameter` is to provide easy automated access to such diverse types of data for general implementations of save, load and analyse methods. It does so by providing an interface to `EEParameterDataComplex`, an abstract base class that allows for collections of values to be stored efficiently, where each value might have a special name for easy access (eg. mean, stdev, etc.)

Initialisation

The other important task of `EEParameter` is to encapsulate all information about a particular parameter. This information is loaded at run-time from constant arrays of `EEParameterInitializationData` that have been compiled into the code and are pivotal for defining a simulator (each `EEInterpreter` has a corresponding array). As this information will not only be used during simulation, but also in later analysis, quite a number of details need to be provided. Some of these can be changed by scripts at a later stage (eg. plotlimits), while others can not (eg. IDs for console, file and GUI, or some limits). `EEParameter` includes documentation, limits for valid values, factors that facilitate automated time series analysis, all information to make readable output and nice plots, and a number of special flags that further facilitate automation.

10.7.2 How to automate analysis of time series

As indicated in the previous section, the largest part of interesting results from simulations of evolution comes in the form of time series. If each detail of these time series were kept, the resulting amount of data would be too large to handle. Ideally, the required reduction of data should not lose information that might be needed for final analysis of the model. Thus, a sophisticated time series analysis[112] should be undertaken by the simulators themselves, before the results of this analysis are submitted to the server suite for further investigation. Such an analysis might comprise

- recording of minimum, maximum and last value
- parameters of an inferred overall normal distribution
- a reasonable number of quantiles
- a small histogram with appropriate bins
- a linear regression over time to record trends
- exponential, logarithmic and Box-Cox[113] transformations before a corresponding linear regression
- observation of repeated fluctuations, preferably without performing a complete Fast Fourier Analysis.

The art of designing such time series analysis modules is to record enough, but not too much, while keeping computing time short and still present results in such a general way that they can be analysed by automated tools. This will require considerable development effort.

Single-pass versus multi-pass statistics

There are two important areas in automated analysis. First, **single-pass statistics** comprises all those cases where the observed number is only seen once by the analysis method; all relevant calculations have to be performed on the spot (eg. remember the minimum). Second, **multi-pass statistics** can be applied to cases where data is stored and can be screened several times to summarise it (eg. as required for linear regression). Many interesting data in simulations of evolution are only available for single-pass analyses, as storing them can easily become prohibitive (eg. fitness-values of a very large population). For some occasions, however, it is worth remembering values for multi-pass statistics that can be performed by the simulator (as it has the resources to store such data), but not by any central server (as it

112. E.g. pp. 137-217 in Hütt (2001) "Datenanalyse in der Biologie - Eine Einführung in Methoden der nichtlinearen Dynamik, fraktalen Geometrie und Informationstheorie", Berlin, Springer.
113. See pp. 423-426. Sokal & Rohlf (1981) "Biometry: The principles and practice of statistics in biological research. Second Edition", San Francisco, California, W.H. Freeman & Co. - Original reference: Box & Cox (1964) "An analysis of transformations", J. Roy. Stat. Soc. Ser. B 26:211-243.

would be overwhelmed with data to analyse and another distributed computing project would be needed just for analysis). Thus, it is worth investing in quality of automated analysis modules, as their work greatly facilitates overall analysis of a simulator and would never be possible on a central analysis computer after completing all simulations.

Another important area is automated spatial pattern recognition and analysis of spatial results, once simulators become spatially explicit. What has been said about time series can be applied to spatial patterns, only that their analysis is more complex. While it is easy to transform a spatial pattern into a nice picture, automated reduction of analysis results to a few numbers is difficult.

10.7.3 Computation of Multi-run-results

All results discussed above belong to the category of single-run-results, as they are produced by a single stochastic simulation of evolution. However, all single runs have to be repeated to obtain estimates on variability between different runs, ie. real distributions. Such computation of multi-run-results is done by the `SingleRunResultsAnalyzer`. It produces a distribution for each type of value from the single-run-results of a simulator that had the same input parameters. In most cases, this large file of multi-run-results will be the ultimate target of analysis, from automated screens to manually selected plots. As this dataset has a high number of dimensions, in the long term a number of advanced techniques will be necessary to cope with the growing flood of data. This includes efficient access to high-dimensional databases[114], corresponding data mining and knowledge discovery methods[115] and visualisation of high-dimensional data[116].

One approach to the latter would be automatic generation of 2D plots that contain advanced symbols encoding information on up to 55 dimensions using a combination of lines and other sub-symbols whose size, thickness and colour give a rough indication of the value for the corresponding dimensions. While such plots might be largely unreadable where the cloud

114. Kriegel (2001) "Access Methods for High-Dimensional Data Spaces" http://www.dbs.informatik.uni-muenchen.de/Forschung/Index
115. Klösgen & Zytkow, (eds, 2001) "Handbook of data mining and knowledge discovery", Oxford, Oxford University Press. - Kriegel (2001) "Knowledge Discovery in Databases" http://www.dbs.informatik.uni-muenchen.de/Forschung/KKD
116. This might involve approaches like humanIT (1998) "InfoZoom: Der Knowledge Browser" http://www.humanIT.de/ or use of the high-end open-source visualisation tool kit VTK besides all other visualisers discussed in the server suite.

of points has the highest density, they can do a great job of identifying outliers and the mechanistic causes that produce them, if one gets accustomed to reading such plots. Of course an interactive analysis system will also be available for exploration of results (see also "Visualisers and Datamining" on page 108).

10.7.4 Types of simulation projects

Up to now different types of data and results have been discussed, but there is still another level at which simulations might be distinguished: different types of simulation projects define different ways to search parameter space. This deeply affects the way new parameter combinations (and single runs) are scheduled on a global computing system:

- **Gridsearches** basically generate a regular grid of points in parameterspace and simulate each point until the desired accuracy has been reached. Priority setting helps to get trends faster, so that new projects can be scheduled more efficiently.
- **Repeated optimisations** or inverse modelling are fundamentally different. Their goal is to find a particular parameter combination that satisfies a particular set of conditions and lies in a certain area of parameterspace. To archive this, the relevant area is scanned superficially, to find that sub-area that is closest to the goal. This is scanned again and the process is repeated, until desired accuracy is achieved or no further progress can be observed. While this type of project employs grid searches too, it automatically chooses grid properties, as the details of the next simulations depend on the results of the previous round of optimisation. Besides specifically adapted modules, such projects require a very high degree of automation in the server suite.
- **Critical parameter analyses** are a special form of repeated optimisation. It systematically changes input parameter(s) until a desired value is obtained in a particular output parameter. It is well suited for questions of the type "How many advantageous mutations are needed to offset a particular deleterious mutation rate?"
- **Population viability analysis** is a special form of critical parameter analysis, as it is only concerned about survival probability of a population for a given time[117]. Future modules in the ε server suite will facilitate easy application of population viability analysis to any given simulator.

While the current design of the server suite is best suited for grid searches, all types of repeated optimisation projects can be implemented as soon as a higher level of automation has been reached. They only require special versions of a number of modules that need to work closer together than grid searches require.

10.8 First experiences with evolution@home and the Participant's Free Choice System[118]

The global computing system described above, evolution@home, is the first of its kind that allows participants to choose the complexity of work units. Initially, it was not clear how participants would respond to this flexibility. Thus, an extremely simple, version of the server-suite was set up to produce static web-pages ordered according to computational complexity. The following experiences were made with this extremely simple setup.

General features

The evolution@home system started on 3rd April 2001 with Simulator005 Project 1, its first semi-automated global computing project (S005P1). Although there has been no advertising other than an email to global computing review sites[119], about 150 non-anonymous participants had contributed about 13 000 public simulation results with more than 10 years of computing time in 37 weeks for analysis in this section[120]. For the latest statistics, see the evolution@home website[121]. Such a level of participation underscores the importance of advertising, press releases, public participant statistics and a fully automated worker software. However, it is striking that the largest global computing project SETI@home (3 million PCs,

117. Boyce (1992) "Population viability analysis", Ann. Rev. Ecol. Syst. 23:481-506. - Brook et al. (2000) "Predictive accuracy of population viability analysis in conservation biology", Nature 404:385-387. - Coulson et al. (2001) "The use and abuse of population viability analysis", Trends Ecol. Evol. 16:219-221.
118. A slighly modified version of this section appeared in Loewe (2002) "evolution@home: Experiences with work units that span more than 7 orders of magnitude in computational complexity", 425-431. 2nd International Workshop on Global and Peer-to-Peer Computing on Large Scale Distributed Systems at the 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid2002), 21-24 May, Berlin, Germany, IEEE Computer Society. (see http://www.evolutionary-research.net/ for download)
119. Eg. Pearson (2001) "Internet-based Distributed Computing Projects", http://www.aspenleaf.com/distributed/.
120. Final plots in this section were completed in mid December 2001 using Excel.
121. Loewe (2002) "The evolution@home website" http://www.evolutionary-research.net

500000 CPU years in ca. 1.5 years, see http://www.ud.com) has a similar ratio of one-time to continuous participants: about 10%.

Actual work unit sizes

While other global computing projects have relatively uniform work unit sizes, problem domain inherent constraints make work units of evolution@home span more than 7 orders of magnitude, as a work unit is defined as a single-run, the most natural choice for individual- based simulations. To continue extremely long runs on other computers would lead to prohibitive communication and other server side costs, as the longest runs are usually those that generate huge amounts of temporary data. Therefore, participants should choose their commitment in terms of physical RAM and computing time to increase the probability that a given run would be completed. To this end RAM and computing time complexity is predicted for each parameter combination. In the current project S005P1 population size RAM requirements range from 1 KB to 900 MB with predicted computing times between 0.01 sec and 180 years. While evolutionary questions would still ask for more, practical considerations excluded all simulations with more than 30 days predicted computing time from publication. Simulations with less than 15 min were run on a non-public computer. While one can debate their exact location, lower and upper boundaries indisputably exist for tasks whose global computation is feasible.

10.8.1 Lower computing time boundary

The lower boundary T_{CPUlow} can be derived, when communication costs are considered. If transfer of necessary data takes longer than local computation, then distribution is not feasible:

$$T_{CPUlow} = Transfersize \,/\, Transferspeed \tag{13}$$

where ***Transfersize*** is the sum of all Bytes of all communication needed to distribute the task and collect the result over a network with ***Transferspeed*** as real-world bandwidth in Bytes/sec. Thus it makes no sense to distribute work units of less than 25 sec expected computing time over a 28K modem line if they generate 100 KB of traffic. While this figure shrinks to 1 msec with Gigabit Ethernet, it rises considerably for semi-automated projects, where occasional manual interaction is needed.

10.8.2 Upper computing time boundary

The upper boundary $T_{CPUhigh}$ can be derived by considering MOORE's Law: If computing time doubles every T_{Moore} eg. 2 years, then for extremely long computations, it is worth waiting. The break even point is reached at

$$T_{CPUhigh} = 2 \bullet T_{Moore} \tag{14}$$

as a 4 year simulation started today will need only 2 years when started in 2 years on a on a computer that is twice as fast. While practical considerations usually reduce feasible computing times significantly below this theoretical limit, a global computing framework for evolution needs to take it into consideration.

Incomplete runs

Incomplete runs can be used. Computers might crash, participants can stop a run, or they change their commitments. Any of these reasons can lead to incomplete simulations. While many distributed computing projects cannot use such data, evolution@home can. As searching parameter space is one of the most important objectives and computing time predictions are often very crude, incomplete runs can be used to refine predictions, because first glimpses at results usually allow much easier extrapolation, than *ab initio* estimations. Intermediate results have a special parameter value that identifies them. They can be included in many analyses, as they have the same trend indicators as all other results.

10.8.3 The Participant's-Free-Choice System

For this analysis, evolution@home allowed participants to choose their commitment by downloading one of about 50 run-files from the web with up to about 60 simulation tasks in each. To generate these, parameter combinations of one RAM complexity class were ordered according to predicted computing time and then sequentially packed into one or more run-files. Thus one manual interaction (download run-file and submit results by email) allows comfortable scheduling of many simulations. Simulation start-codes contain complexity predictions for participants who want to compose their own special run-files (documentation supplied). While this was occasionally done, many participants made extensive use of their right to choose. Email feedback confirmed that participants like this flexibility, although rather in a fully automated system. Several issues come up when participants can choose the complexity of their contributions:

- Biased participant choices
- Faithfulness of participants regarding their commitment
- Benchmarking computer systems and
- Accuracy of biological foundations of predictions.

In the remainder of this section these issues will be discussed in the light of experiences with evolution@home.

10.8.4 Biases in participant choices

It was initially unclear whether participant bias would be so strong that only a small part of the complexity spectrum would actually be sampled. If one considers complexity of runs, this is not the case. Figure 16A shows a histogram of the number of simulations with observed computing times that approximately double from one category to the next. The left part of the figure includes the more than 9000 non-public simulations of very small complexity for comparison. These represent a more or less unbiased spectrum of the complexities scheduled. When this variation is taken into account, participant choice was more or less uniform for all simulations up to about 16 hours (Figure 16B). Longer computing times were submitted less often with increasing complexity. This is easy to understand, as participants picking complex runs will not submit as many single-run-results as participants picking small runs. Therefore, commitments appear to be quite uniform when total computing times invested in the corresponding categories are considered (Figure 16C). Thus a simulation scheduling system needs considerable fine-tuning capabilities for complex runs, while shorter tasks should be handled largely automatically.

participant choice has little bias

If the distribution of simulations per RAM complexity class is considered, the same largely uniform picture appears, except for very large simulations. There are considerable fewer simulations in the 450 MB class and none in the 900 MB class at that time– a clear reflection of current PC memory sizes (Figure 16D). To find out how this more or less uniform pattern is distributed in time, predicted computing times and RAM complexities were plotted over the date when simulations were started (Figure 16E + F). While this confirms overall rough uniformity, changes in the run-files distributed over the web can be tracked in such plots with a delay of about 3 weeks – an indicator of current scheduling flexibility.

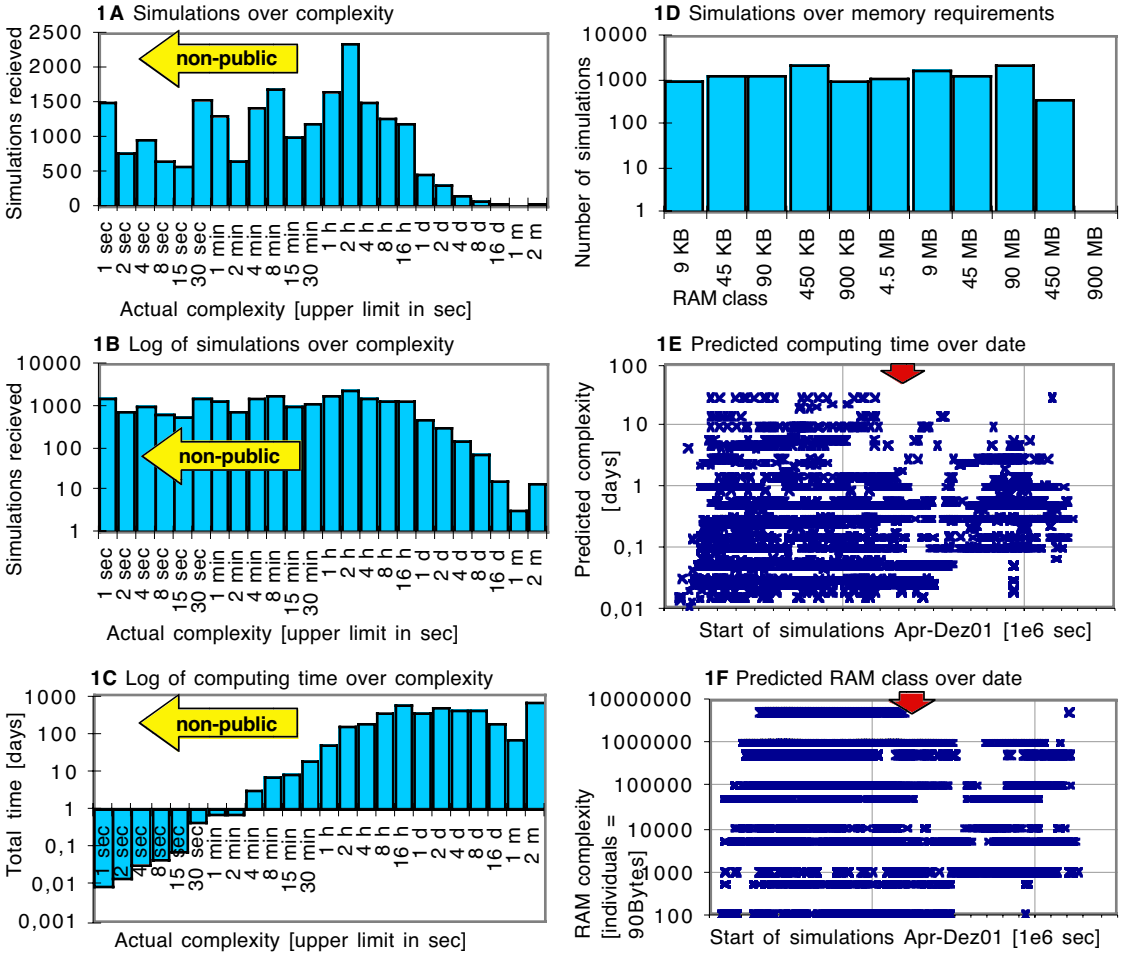


Figure 16 Distribution of participant choices. Of about 22000 total simulations, 9000 non public simulations have been excluded from **D**-**F** and marked in **A**-**C**. The arrow in **E**-**F** denotes the first removal of completed run-files from the web (1 x-unit=11.574 days).

10.8.5 Intermediate results

Nearly 6% of results received from public participants came from interrupted runs. As the current release of the simulator cannot save an intermediate state to disk and continue from there, completed simulations mean that the corresponding computer was running all the time and the simulator application had not been closed. To avoid losing everything in case of a crash, the list of observed parameters was written to file every hour. Given this fact,

it is remarkable that about 94 % of all simulations were able to run uninterrupted, although many of them had long computing times.

Causes for interrupts

Generation of intermediate results might have two causes. First, prediction quality was so bad that the participant quite understandably stopped the apparently never-ending run. After all, a prediction that fails by 5 orders of magnitude is no prediction at all. Second, prediction quality might have been OK, but the participant changed his commitment deliberately or involuntarily (eg. system crash). Figure 17A shows that 23% of all intermediate results might be due to bad predictions, whereas 77% come from real interruptions. (Please note that many completed runs had bad predictions too, see Figure 19C.) Figure 17B shows no trend in the memory complexity of intermediate results. At first glance, it would not harm the project if intermediate results were discarded. In the long term, however, intermediate results permit improvement of computing time predictions. This is especially important for models in evolutionary biology, as their behaviour is often difficult to predict.

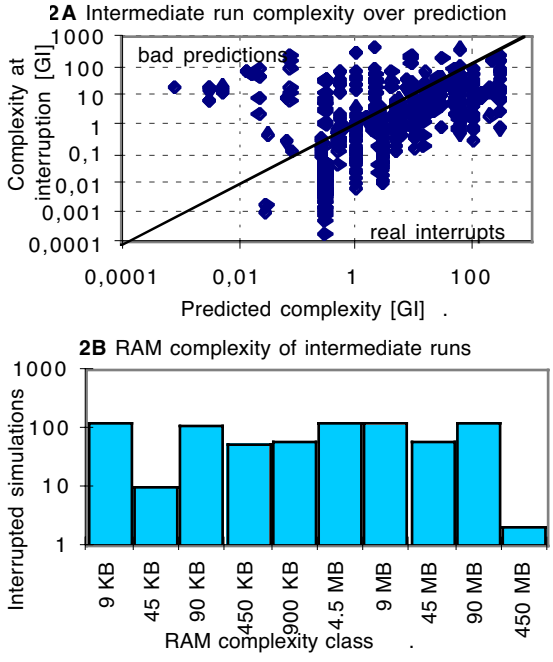


Figure 17 Intermediate Results.
A 23% of all interruptions might be attributed to bad predictions; the rest is due to technical causes or participant decisions; GI = GigaIndividuals, see section on benchmarking.
B RAM complexity of interrupted results shows no trend.

10.8.6 Benchmarking computer systems

Benchmarking is necessary to give the participant a prediction of computing time on his system, and the importance of hitting at least the correct order of magnitude should not be underestimated. Unfortunately, comparison of computing systems is very complex and many of the benchmarks devised concentrate only on a small part of the performance spectrum[122]. Thus some have reached the conclusion that the only valid benchmark is the final application. In global computing, things get worse, as only idle CPU-cycles are used and execution of any other software decreases performance. Simulators of evolution@home have even more problems. As their tasks span nearly 6 orders of magnitude in RAM complexity, cache effects can play an important role.

Currently the problem is solved as follows. When the simulator is started for the first time, it asks for a maximal RAM limit on this computer to avoid use of virtual memory, an effective knockout even for fast pre-emptive multitasking operating systems. Before the first simulation is started, a benchmark population is generated that (i) tests whether allocated memory is really available and (ii) measures an initial performance under current real world conditions. This performance is used for initial computing time predictions. However, once a real simulation is run, the actual performance is monitored and used for an individual prediction whenever the participant requests one. After each simulation, computing time and complexity are recorded in the preferences file to allow computation of a long-term average performance that is used for future performance predictions. Thus the "benchmark" adapts as closely to the real application as possible.

The most natural basic measure of complexity for individual-based models is one individual whose transition from one moment of simulated time to the next is computed. The number of such "individuals" that can be computed in one second is called INDS or MINDS, if a million individuals are used. When future simulators implement models with many different individuals, INDS refers to a well-defined average standard individual that is characteristic of the corresponding simulator and is used in its complexity

122. Gustafson et al. (2001) "HINT Web site" http://www.scl.ameslab.gov/HINT
Gustafson & Snell (1995) "HINT: A new way to measure computer performance". Proceedings of the 28th Hawaii International Conference on System Sciences, January, Wailea, Maui, Hawaii. - Gustafson & Todi (1998) "Conventional benchmarks as a sample of the performance spectrum". Hawaii International Conference on System Sciences, http://www.scl.ameslab.gov/Publications/HICSS98/HICSS98.pdf.

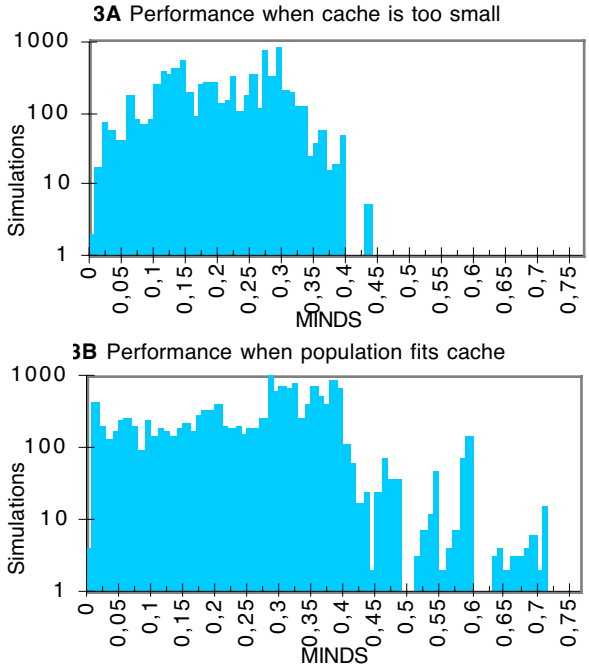


Figure 18 Cache effects on performance distribution. **A** Little or no cache effects can be expected in simulated populations that need > 4 MB RAM. **B** Performance of simulations with 1 KB - 1 MB RAM has a high probability to be influenced by cache effects with the fastest systems being ca. 70 times faster than the slowest.

prediction formulas. One measure of overall computational complexity is GigaIndividuals (GI).

Equipped with this background, the performance spectrum of all single-run-results of evolution@home can be understood (Figure 18 and Figure 19A+B). Performance of computers that contributed results spans nearly two orders of magnitude, if cache-sized simulations are considered. For simulations that are too large for current caches, performance still differs by a factor of 40. Using Moore's Law, this figure can be used to infer that participants of global computing projects use systems that range from brand new to about 10 years old ($40 = 2^{5.3}$; 5.3*2a) – assuming that the slower runs were not due to busy CPUs. Global computing might have to deal with a broader performance spectrum than other applications as people might use it to give their old PCs a new destiny.

Altogether, these results underscore the importance of taking local performance into account when predicting total run time – especially for complex simulations.

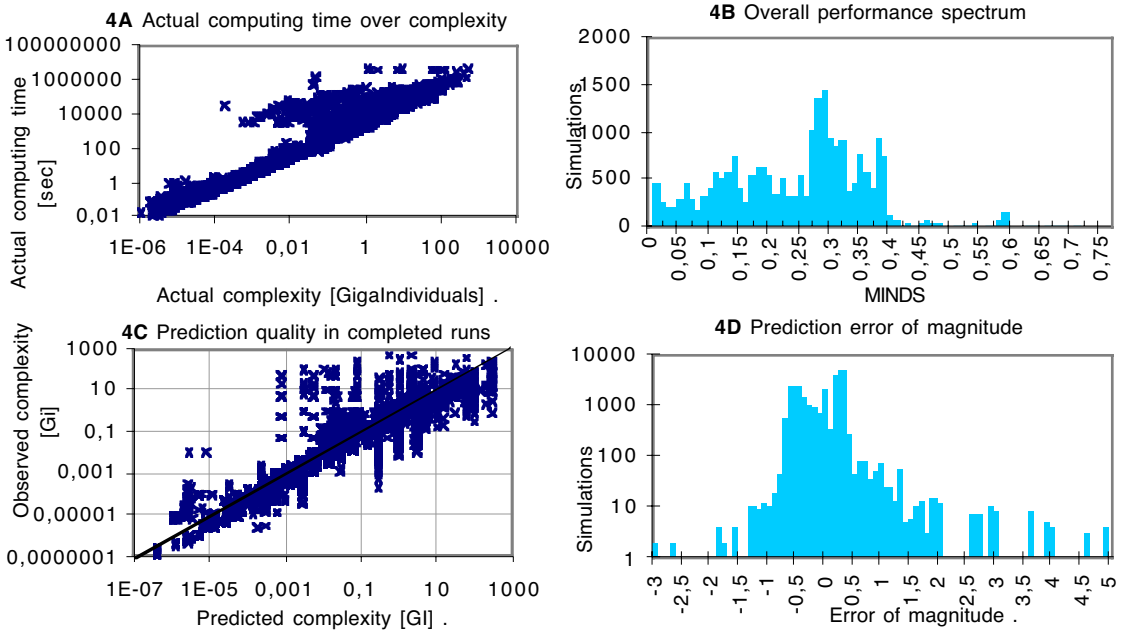


Figure 19 Sources of computing time variability.
A The nice line in the lower part represents maximal performance of the computer used for small non-public simulations. As the simulator uses the lowest possible background priority, other software might cause considerable slowdown. **B** Overall performance spectrum. **C** Actual complexity might considerably deviate from predicted complexity. **D** Histogram of the error of magnitude (see text for definition) of completed runs only. Although most predictions are accurate within half an order of magnitude, outliers can be extreme.

10.8.7 Prediction accuracy

Besides local performance, biological knowledge about the system investigated profoundly influences prediction accuracy. In many simulations that tackle long-term evolutionary questions, models have a general structure that waits for certain events. In most cases, some related parameters indicate arrival of these events. However, to schedule simulations one cannot start them simply to look for the first hints of ultimate complexity (one might need a global computing system just to do this). Thus very simple formulae have to be used to arrive at reasonable predictions. However, systems where such simple formulae give the correct answer are usually not the cutting edge of research. Thus one of the challenges in developing a simulator is the need for such a formula with reasonable prediction quality. This is no small task, as not even a crude order of magnitude can be predicted for many

problems in evolutionary biology, especially not for Muller's ratchet, the topic of Simulator005. Thus, participants and administrators have to live with this inaccuracy. After all, no simulations would have been necessary if the answer had been known already.

To estimate prediction quality of Simulator005, predicted complexities were compared to actually observed complexities for completed runs. As can be seen in Figure 19C, considerable prediction errors exist in both directions.

10.8.8 Definition of the error of magnitude

Unfortunately, one cannot just compute the relative error, as it generates seriously misleading figures due to its inherent asymmetry ([10^{-7} observed – 1 inferred] / 1 inferred value leads to 1 as relative error, although the inferred value is 7 orders of magnitude off and would lead to a relative error of 10^7, if 10^7 had been observed). However, as all relevant values have the same sign, the **error of magnitude**, Err_{mag} (or EoM), can be defined as

$$Err_{mag} = \log 10\left(\frac{C_{observed}}{C_{inferred}}\right) \qquad (15)$$

where **C** denotes the quantity whose error is to be observed (in this case the computational complexity in GigaIndividuals), the suffix "$_{inferred}$" denotes values whose Err_{mag} is being calculated, while "$_{observed}$" denotes values that are used to determine Err_{mag}. This new way of quantifying errors has some nice intuitive features:

Strengths

- It is symmetric around 0, where 0 indicates no error.
- Negative or positive values indicate that the observation is below or above the prediction, respectively.
- An Err_{mag} of 1 or 3 has the intuitive meaning of missing the mark by a factor of 10 or 1000, respectively.

Limitations

However, it has the following limitations:

- It can only be computed if all values have the same sign

To my knowledge, the concept of quantifying errors of magnitude by Equation (15) has not yet been proposed.

Figure 19D shows observed errors of magnitude. Most predictions deviate by Err_{mag} 0.5 or less. This accuracy within a factor of 3 is remarkable, given the fact that we are dealing with more than 7 orders of magnitude here. However, currently observed extremes range up to 3 orders of magnitude below and 5 orders of magnitude above the inferred value.

Intermediate results are important when prediction error is large

In light of such prediction problems, proper handling of intermediate results becomes crucial, especially as this problem is not rare in evolutionary biology. Given the strict rules regarding computing time limits in most super-computer facilities, global computing permits an astonishing flexibility here. Design 5 of ε has a server suite structure that can take this into account. To improve usability of the system, it is proposed that participants enter not only the highest RAM and computing time limits when configuring a simulator, but also an upper limit for the prediction error of magnitude. Thus, runs with poor complexity predictions are deliberately forced into intermediate results that are then used by the framework to improve predictions for the corresponding parameter combination.

10.8.9 Related work

A few aspects of other global computing projects[123] are worth mentioning. Computational complexity of individual work units from global computing projects ranges from fractions of seconds (Photons in Xpulsar@home) or a few minutes (Porivo's peerReview) up to several weeks (some primes in GIMPS) or even more than a year (complex models of the climateprediction.com project). Most projects partition their work into units of several hours up to a few days such as SETI@home or Folding@home. However, no current project has such a diverse work unit complexity as evolution@home.

All such projects encounter the problem of scheduling computation of their large numbers of work units. Thus they share the application-centric view of Application Level Scheduling, an important topic in computational grid research[124]. In contrast to classical resource or job scheduling on supercomputers, Application Level Scheduling optimises timely execution from the perspective of the application by eg. avoiding delays due to batch queues or slow connectivity. Optimisation criteria might be turnaround time, result quality or others. This requires a sophisticated software system that has been specifically adapted to the application and monitors performance relevant features like CPU and Network load. These observations are then used to compute the best schedule under given circumstances. For an example, see "AppLeS"[125]. Such an Application Level Scheduling system might allow

123. Pearson (2002) "Internet-based Distributed Computing Projects", http://www.aspenleaf.com/distributed/.
124. Berman et al. (1996) "Application-Level Scheduling on distributed heterogeneous networks (Technical Paper)". Proceedings of Supercomputing '96.

participants to work (more or less) interactively with applications that are too complex for any single workstation[126]. Recent efforts to support parameter-sweep applications extend Application Level Scheduling to extremely complex problems that are similar to most global computing projects[127].

Global computing projects can be divided into problems with equally interesting work units (eg. cryptography) and problems with unequally interesting work units (eg. evolutionary models). While minimising overall completion time is probably the most important goal of potential scheduling strategies for the former, the latter add a new aspect to scheduling. These problems require fast processing of the most important work units in the first place, while overall completion time is less important. As interesting work units might speed up overall progress of the project or point to interesting regions of parameter space for more intense investigation, they play a key role in steering the whole project. However, the task of finding these important work units is highly specific to a problem domain. Thus, such projects can easily be limited by the need for a sophisticated data management and priority setting system. This has been the case with the current prototypic server suite of evolution@home. The importance of these data handling issues for overall success can not be underestimated; implementation of the Design 5 server suite described in this chapter will be an important step towards the solution of these problems.

For situations, where a sophisticated server suite is not available, the Participant-Free-Choice System presented above has a distinctive feature: simplicity. It neither needs a complicated server suite, nor a specially adapted Scheduler, nor a Network Weather Service, nor a list of known computers, nor standing Internet connections. It can be operated with a simple static web server on the operator's side and occasional dial-in connections on the participant's side. Administrative overheads for scheduling are minimal, as are requirements for participants. If turnaround time is less critical than a low implementation complexity and easy access to more computing power, then the minimalist approach of the Participant-Free-Choice system can be recommended. In other cases more sophisticated Application Level Scheduling or a special server suite will be necessary.

125. Berman (2002) "Application Level Scheduling on the Computational Grid" http://apples.ucsd.edu
126. Eg. Smallen et al. (2000) "Combining Workstations and Supercomputers to Support Grid Applications: The Parallel Tomography Experience". Proceedings of the 9th Heterogenous Computing Workshop, May 2000.
127. Casanova et al. (2000) "Heuristics for scheduling parameter sweep applications in Grid environments", 349-363. Proceedings of the 9th Heterogeneous Computing Workshop (HCW'2000).

10.8.10 Lessons and Conclusions

It was not clear initially whether participants would devote computing time to all complexity classes or rather stick with the most simple tasks. The bandwidth of received results was encouraging and suggests that stochastic parameter-sweep applications might use a similar system, if they need to repeat their runs many times anyway, and if there is no pressing deadline for getting particular results.

A weakness of the current prototype is its dependency on manual analysis of results. This leads to considerable slowness in updating progress and run-files on the web. Ideally, the website would contain only the hottest run-files of the hour, making sure that every simulation with a particular complexity is computed once, before any is computed twice. Such a scheduling mechanism, however, needs something like the server suite that has been described in this chapter. However, its tightly integrated modules that cooperate with the whole evolution@home results analysis workflow cannot be easily imported from outside. Thus the next steps for evolution@home include implementation of such a system.

Another important feature, however, should have been incorporated from the beginning. When participants configure their simulator by providing upper complexity limits, they should have been given the opportunity to specify their tolerance limit for the prediction error of magnitude. This would have helped with the enormous prediction inaccuracies and will be included in future versions.

All in all, evolution@home is the first global computing system that has to deal with work units that span more than 7 orders of magnitude in computational complexity. Initial experiences reported here indicate that global computing has great potential to develop the flexibility needed to cope with such diversity, if frameworks incorporate the corresponding features. For efficient simulations of evolution, tight integration of a global computing framework, an evolutionary model implementation framework and corresponding databases are necessary. Cornerstones of the design of such a framework have been described with Design 5 of ε in this chapter.

11 Perspectives for future developments

Eepslion is build to grow with its applications. Here, current status, next steps and potential for future developments are discussed.

11.1 Current state of development

An overview over the current Design 5 was presented in the last chapter. The current prototypic implementation focused on those features that are most important for Simulator005. To start evolution@home in April 2001, code from Design 3 and 4 was extended to

- include the present range of detail observations of Muller's ratchet
- provide a minimal server suite that can generate grid searches, estimate computational complexity, check integrity of results and compile a large table of the most important results.

The resulting code with well over 20 000 lines C++ is available for MacOS (68K, PPC, Carbon) and Windows (95, 98, ME, NT, 2000, XP). Main development effort since release 1 was improvement of run-time predictions and generation of a Windows-release that truly runs only with idle priority and thus does not interfere with other active programs. This has been achieved with release 5. It was developed with Metrowerks[128] CodeWarrior Professional Release 6 under MacOS. One further maintenance release was made to improve its server suite functionality, quality of predictions and a bug that affected a minor part of the simulations ($U \leq 0.001$).

The current release (S005r6) does not use XML to store results, has no graphical user interface and works only in semi-automated mode (manual download of run-files from the Web and submission of results by email). While some parts of the code are already implemented according to Design 5 standards, most of it still needs to be ported to the new design.

However, the most important feature of the current release is that it is advanced enough to actively run a global computing project that produces meaningful results. See Part V and VI of this work for analysis of results.

128.http://www.metrowerks.com

11.2 The next steps

As the Y2K problem has shown, code of actually used systems lives longer than often suspected. As the enormous task of exploring evolution by computer simulations has only just begun, it is worthwhile using high coding standards for ε. A number of tools will facilitate this and should be put in place, before implementing the major part of the framework:

- **Automated code version control system**. Besides making documentation of code changes easier even for one developer, the use of a version control system is mandatory, once more than one person is working on the project. The easiest time for transition is before the major part of Design 5 with all corresponding libraries is in place.
- **Automated documentation system**. Every major project needs appropriate documentation, if it is ever used by more than one person. However, it can no longer be generated manually. Modern automated documentation systems scan source code to generate HTML or PDF documentation from the structure of classes and the comments that are part of the source code. Such a system will be needed for ε and should be put in place before new code is added.
- **Automated code testing tools**. Complex languages like C++ allow for a large number of dangerous constructs that should be avoided[129]. Moreover, large projects have their own coding standards to facilitate readability of the code (see page 73). Automated tools like CodeWizard[130] can scan source code and point out inconsistencies to significantly reduce errors that only have to be found later.
- **Appropriate development platforms**. As the code produced will have to run on many different platforms, but no compiler supports all platforms (from MacOS to Supercomputers) an appropriate combination of compilers must be selected to commit to (eg. CodeWarrior and Kai C++[131]). Standardised preprocessor switches are then used to handle compiler- or platform-specific parts of the code. However, these should remain few, as one of the important rules in ε development is to stick to ANSI C++ (except for GUI issues).
- **Automated code testing**. A system should be employed that allows code to be tested automatically. Thus while code is being developed,

129. Hyman & Vaddadi (2000) "Effektive C++-Techniken", Bonn, Galileo Press. - Meyers (1995) "Effektiv C++ programmieren". 2. korrigierte, Bonn, Addison-Wesley.
130. ParaSoft (2001) "CodeWizard" http://www.parasoft.com/cplus.htm
131. Kai Software (2001) "Kai C++ is the best multi-platform ISO C++ compiler." http://www.kai.com

the corresponding test methods are implemented as well. The main goal of this system is to check, without further manual interaction, whether the code performs apparently correctly. Although this is no proof of correctness and is no substitute for manual debugging, a large number of errors can usually be found in this way.

Automated analysis

As soon as appropriate solutions have been found for these issues, the server suite should be implemented with these tools and standards. Highest priority will be starting the central database and porting existing results to it. This will facilitate results analysis and scheduling of new runs and thus help govern the next projects of Simulator005.

Automated communications

Then the next big step will be to commit to a cross-platform graphical user interface[132] for C++ to make handling of a simulator easier for participants. It will also facilitate automatic download of tasks and submission of results. Implementation of a graphical user interface and full automation will be major milestones in the development of ε. Soon after this is achieved, the code is ported to different platforms.

Better bio-classes

Then the biological modelling infrastructure shall be reimplemented to meet Design 5 standards. This will take time, as no small number of classes is involved. The first landmark will be implementation of Simulator005 or a similar simple models with these new classes. Then new biological functionality will be added, starting with recombining and other genome models, and followed by spatial structure for populations.

11.3 Long term perspectives

ε will grow

Finally, ε will have reached the state where it implements a reasonable part of the basics needed for simulation of evolution and it can be readily used to build models, start corresponding global computing projects and analyse their results. Then further evolution of the framework will take place by implementation of new details needed for particular models and by implementation of new methods of automated analysis, no matter whether they further advance time series or spatial snapshot analysis in a simulator or improve some knowledge discovery modules in the server suite.

Hierarchical global computing

Once automation in global computing has reached enough maturity and stability, the next major challenge can be addressed: Hierarchical global computing. This is especially important for simulations of multi-level popu-

132.Eg. Smart (2002) "wxWindows Home: Cross-Platform GUI Library" http://www.wxWindows.org

lations that are too complex for any single computer. Such populations can be eg. many parasites in few hosts or many mtDNA molecules in less mitochondria in less cells in few humans. Once aware, it is easy to find many other examples for multi-level populations. When their simulation can be split into an ongoing process with relatively few long-term individuals that are influenced by repeated incidences of short-term individuals building large populations for a short time, they are particularly suitable for hierarchical global computing. Here corresponding simulators with standing Internet connections contain a little server suite each. They hand out simulation tasks of subpopulations and collect the results. These sub-results are then used to compute the next state of evolution of the upper population. Every computer that participates in such a simulation needs a standing Internet connection, and an elaborate system of application level scheduling and server connect information will be necessary to achieve reasonable performance. While it probably depends on the details of the model, whether such code better suits global computing than super-computing, it is expected that at least some problems will be solved better by this approach.

Symmetric multi-processing

Another long-term development will be to provide ε with features that can take advantage of symmetric multi-processing (SMP). Machines with more than one processor per node are not too rare and some complex simulations would significantly benefit from being able to use more CPUs, especially those problems that go to super-computers due to complexity. However, balancing the load between the various CPUs in an automated and general way is no trivial problem for ε, as each simulator might require one to distribute a different level (habitats, locations, populations, or even individuals).

GUI for web databases

As evolution@home results are to be published on the web, over time a number of such databases will accumulate. It will be desirable to build a common graphical user interface for these databases to facilitate retrieval of data for other researchers. The complicated data structures and database sizes make this a reasonable sized project, if the result is to be of actual use for average biologists.

Automated code generation

Finally, in about 20 years, a software suite might be under development that allows automated construction of simulators with the help of a graphical user interface. This might allow biologists to build, run and analyse models although they can not code in C++. Such a system contains enough intelligence to check for errors, produce test results, help improve code, and submit the model for formal registration and distribution. Computing biol-

ogists would use such a system to help with routine tasks, while they still add special code by hand, similar to the process of building web-pages in Dreamwaver. If this state is ever reached, ε will have become a standard tool for evolutionary research.

12 Design, description and tests of Muller's ratchet "Simulator005"

This is the scientific part of the documentation of the first simulator released as part of evolution@home. It is called Simulator005 (=S005) and simulates two Muller's ratchet - processes, one in the foreground ("ratchet") and one in the "background". The mutational parameters of these processes can be chosen freely and quite some details are observed about the foreground ratchet process.

12.1 Why was this simulator built?

Muller's ratchet describes the accumulation of slightly deleterious mutations (SDMs) in asexual genomes. As the overwhelming majority of mutations is believed to be (slightly) deleterious and asexual genetic systems have a surprisingly wide distribution (eg. mitochondrial DNA in mammals), one can suspect that Muller's ratchet might play a role in extinction of species. This problem is especially interesting in light of recently observed high mutation rates in mitochondria ("Mutation rates in conflict" on page 13). Finally, Muller's ratchet is interesting from a theoretical population genetics point of view for the question of whether sex evolved to escape the consequences of deleterious mutations or for other reasons.

However, theoretical extinction times can be very difficult to predict analytically ("Predicting the rate of Muller's ratchet" on page 41). Simulator005 was built to help with predictions and add a few details that make simulations more realistic than the analytical model that has been understood best up to now[133]. It complements and extends predictions from mathematical models by allowing (i) two sets of mutation rate and effect parameters to operate at the same time, ie. a foreground ratchet and a background ratchet, (ii) advantageous mutations, (iii) various distributions of mutational effects, and (iv) observation of new details like multiple clicks, phases of a click, and others under the given circumstances. Results help to

133. Haigh (1978) "The accumulation of deleterious genes in a population - Muller's Ratchet", Theor. Popul. Biol. 14:251-267. - Gessler (1995) "The constraints of finite size in asexual populations and the rate of the ratchet", Genet. Res. 66:241-253. - Stephan et al. (1993) "The advance of Muller's ratchet in a haploid asexual population: Approximate solutions based on diffusion theory", Genet. Res. 61:225-231.

Abbreviation	ParameterID	Explanation
K	1 *	Carrying capacity of the population
Ri	2 *	Initial reproductive capacity
Ub	3 e	Genomic mutation rate for background mutations
Sb	4 e	Selection coefficient for background mutations
dmeb	5 e	Distribution of mutational effects (background)
Ur	6 *	Genomic mutation rate for the Ratchet observed
Sr	7 *	Selection coefficient for the Ratchet observed
dmer	8 e	Distribution of mutational effects (Ratchet)
Cend	9	Clicks to observe before worldhistory can stop
Tend	10	Last generation in standard worldhistory
rfcID	11	ID of project this simulation belongs to
rndseed	12	First seed used to initialize global RND generator

Table 11 Input parameters of Simulator005. The star behind the ParameterID denotes parameters that are included in analytic theory, "e" denotes extensions. Other parameters serve administrative purposes.

Computing Statistics		
err	14	Errors since start of this world history
intm	15	Result is intermediate (=1) or complete (=0)
timev	16	Total seconds needed for last evolve&observe-loop
timpr	17	Total amount of work (GigaIndividuals) predicted
timwh	18	Total seconds computing time for this single-run
Glo	19	Work units in this run (in gigaindividuals)
minds	20	Performance during current single run in MINDS
Unique events in simulated worldhistory		
Teq	21	Time from perfect start to exp(-U) fitness
Tclfd1	22	Time from perfect start to 1.FitnessDownClick
Tclfi1	23	Time from perfect start to 1.FitnessIncrClick
Tclb1	24	Time from perfect start to 1.background click
Tclr1	25	Time from perfect start to 1. Ratchet click
mmpr	26	Date where meltdown is predicted to start
Tmm	27	Date where meltdown started
Tex	28	Extinction time (single run)
wtim	29	Current world time in current simulation
Timeseries summaries: background information		
aclfd	30	Absolute click time for fitness decreases
eclfd	31	Effective click time for fitness decreases
clsfd	32	Click size of fitness decreases
aclfi	33	Absolute click time for fitness increases
eclfi	34	Effective click time for fitness increases
clsfi	35	Click size of fitness increases
atcb	36	Click time (absolute) of background mutations
etcb	37	Click time (effective) of background mutations
tcbpr	38	Predicted click time of background mutations
clsb	39	Click size of background mutations
Timeseries summaries: Ratchet observations		
atcr	40	Click time (absolute) of Ratchet mutations
etcr	41	Click time (effective) of Ratchet mutations
tcrpr	42	Predicted click time of Ratchet mutations
clsr	43	Click size of Ratchet mutations
Multiple clicks		
cl1	44	Absolute time for clicks that fix 1 mutation
cl2	45	Absolute time for clicks that fix 2 mutations
cl3	46	Absolute time for clicks that fix 3 mutations
cl4	47	Absolute time for clicks that fix 4 mutations
cl5	48	Absolute time for clicks that fix 5 mutations
Duration of phases of a click		
T1a	49	Duration of phase 1a for Ratchet (n1 to 1.6*no)
T1b	50	Duration of phase 1b for Ratchet (1.6*no to no)
T2	51	Duration of phase 2 for Ratchet (no to 0)
Sizes of various population subgroups		
Nt	52	Average total population size over world history
Nod	53	Individuals with highest fitness (best class)
Noi	54	Individuals with lowest fitness (worst class)
irNo	55	Analytic best class size for Ratchet
Nor	56	Individuals with least Ratchet mutations
ibNo	57	Analytic best class size, background mutations
Nob	58	Individuals with least background mutations
Snapshot of features of the current generation		
Df	59	Distribution of total genetic fitness in population
Dr	60	Number of Ratchet mutations per genome
Db	61	Number of background mutations per individual

Table 12 Output parameters of Simulator005 arranged according to their larger topic. Abbreviations and ParameterID's are given besides a short explanation.

understand particular species and improve theoretical models of Muller's ratchet.

12.2 Input and output parameters

The input parameters of Simulator005 can be found in Table 11 and the output parameters in Table 12. To reduce the absolute number of output parameters the two ratchets modelled were defined as foreground ratchet (named 'ratchet', short 'r') and as background ratchet (named 'background', short 'b'), although they can equally contribute to evolution of fitness. The ratchet in the foreground is only observed in more detail by a number of additional parameters.

The sum of all parameters with their meaning in the model investigated defines a simulator. New releases of a simulator are not allowed to add new parameters, or omit previously defined ones. However, they might *extend* the range of values that a parameter can assume. This is used for `dmeb` and `dmer`, the distributions of mutational effects for "background" and foreground (= "ratchet") ratchets. Up to the current release 6 of Simulator005 they assume the value 0 that stands for equal mutational effects in current and future simulations. Later releases might add other values that make simulations assume particular other distributions of mutational effects.

Simple distributions

Many output parameters belong to the time series (30-37, 39-41, 43-54,56,58) or snapshot type (59-61). As corresponding complex data types are not yet available in ε, they are treated as Normal distributed data and the following numbers are recorded during simulation (or inferred from recorded values upon request):

- Mean (***m***) of the inferred Normal distribution, upon request.
- Standard deviation (***sd***), upon request.
- Coefficient of variation (**CV = *sd*/*m***), upon request.
- Sample size (n), recorded.
- Absolute minimum (min) of all values, recorded.
- Absolute maximum (max) of all values, recorded.
- Current value (curr = last value observed), recorded.
- Sum of all values (sum), recorded.
- Sum of squared values (ssq), recorded.

Details about some parameters

As meaning of most parameters in Table 11 and Table 12 is self explanatory, here only some details need to be added. If the default value 0 is used as

`rndseed`, a new seed is automatically generated from current universal time. Other values are used directly.

Starting conditions are always important for simulations. In case of the ratchet, the population should have approximately reached mutation-selection equilibrium, before the rate of the ratchet is observed. S005 records the first click of all types of mutations separately, before starting observations of clickrates, to ensure equilibrium. In addition, it records the time from start until fitness has decayed to its expected theoretical equilibrium value e^{-U}, where U is the sum of all deleterious mutation rates[134]. Comparison of these results will facilitate simulation analysis.

There are 4 data streams that record ratchet clicks in detail:

- clicks that decrease fitness
- clicks that increase fitness
- background mutation clicks
- ratchet (= foreground) mutation clicks

Each click in these streams is characterised by

- absolute click time (= generations since last click)
- effective click time (= absolute click time / number of mutations that were fixed during this click)
- click size (= overall selection coefficient for fitness clicks or number of mutations for background and ratchet clicks)

S005 records a distribution for each of these properties for each of the 4 streams.

Only for the "ratchet" mutations, additional properties are recorded. To investigate the role of multiple clicks, distributions of the absolute clicktime are recorded separately for clicks that fix 1, 2, 3, 4 or 5 mutations. This allows comparisons to the overview statistics of the ratchet click stream and contains more details. To observe the various phases of a click that are described by the analytical diffusion theory approach, a distribution of the duration of each of the following times is recorded:

134. Lynch et al. (1995) "Mutation accumulation and the extinction of small populations", Am. Nat. 146:489-518 point out that the mean fitness of founding individuals of a new population will be approximately e^{-U}, if effective size of the ancestral population is greater than 5/s. See also Haldane (1937) "The effect of variation on fitness", Am. Nat. 71:337-349. - Kimura et al. (1963) "The mutation load in small populations", Genetics 48:1303-1312. - Bürger & Hofbauer (1994) "Mutation load and mutation-selection-balance in quantitative genetic traits", J. math. Biol. 32:193-218. - Lynch et al. (1995) "Mutational meltdowns in sexual populations", Evolution 49:1067-1080.

- o `T1a` = Deterministic part of the decay of the frequency of a newly generated best class from its old frequency $\mathbf{N_1}$ (frequency of second best class) to $1.6*\mathbf{N_0}$.
- o `T1b` = Stochastic part of the decay of the frequency of a newly generated best class (from $1.6*\mathbf{N_0}$ to equilibrium frequency $\mathbf{N_0}$).
- o `T2` = Waiting for stochastic loss of the best class (decay from $\mathbf{N_0}$ to 0).

Comparisons might help to improve analytical theory. The same might be true for comparisons between analytical sizes of the best class $\mathbf{N_0}$ and actual observations of the number of individuals in this class for individuals with highest fitness, lowest fitness, least ratchet mutations and least background mutations. Observations are completed with snapshots of the distribution of total genetic fitness and distributions of the number of ratchet and background mutations per genome in the current (ie. last) population size.

12.3 Simulation Model

Simulated world history

Each simulation run follows the same general worldhistory:

- o Initialise RND generators
- o Create a population with `N = K` individuals in a habitat with a carrying capacity of `K`
- o Simulate evolution with the given input parameters while observing the output parameters, until either `Cend` clicks of the ratchet have been observed or `Tend` generations have been reached or the population is extinct.
- o Store an intermediate result, every hour.

The use of two termination criteria (besides extinction of the population) will facilitate flexible allocation of run-time to the most interesting simulations. In project P1, 500 clicks of the ratchet were assumed to be enough for a reasonably precise observation. As smaller populations could be observed for longer simulated time periods, simulations were ended at `Tend` = 10^5, 10^6, 10^7 or 10^8 generations, with one run for each length (expected runtimes over 30 days were sorted out). Other projects might define new limits.

Individuals model

Each individual is described by one genetic fitness variable of 64bit float accuracy, one 32bit counter of ratchet mutations, one 32bit counter of background mutations and pointers to the previous and next individual in the population. The new state of all these properties in the next moment of simulated time (= 1 generation) is computed from the old state that is stored separately. The new state is computed by applying the following steps to

each individual of the parent generation: Compute number of offspring, produce offspring with new mutations and die. Thus the life history of each individual is as follows:

- **Mutation.** Background mutations, ratchet mutations and their combined effect on fitness are inherited from the parent, while new mutations are added between development of the parent and fertilization. New mutations are assumed to be Poisson distributed with the corresponding mutation rate (ratchet or background) as mean. Thus, all new mutations occur before development of the new individual starts and all subsequent mutations in the germline only affect the next generation. This is reasonable for multi-cellular individuals, as the phenotype is derived from cells that stem from the fertilized egg, whereas new mutations in an individual's germline are not likely to influence development of other parts of the body. It is assumed that single-cellular individuals do not behave completely differently in evolutionary terms, although their body and germline are identical; however, this remains to be shown by another simulator.
  Mutational effects are assumed to be completely independent of each other, so that fitness can be computed multiplicatively. Mutational effects are drawn from a specified distribution of mutational effects, where 0 stands for equal effects and other numbers encode distributions that are going to be implemented in the future.
- **Birth and juvenile survival** until this individual starts to consume resources of an adult is 100% in Simulator005. Here is period **Va** for viability selection.
- **Generation time changes** here; ie. the individual was considered 'offspring' before this point and is considered 'parent' after this point.
- **Pre-reproduction survival** is 100% in Simulator005. Here is period **Vb** for viability selection.
- **Reproduction** consists of two parts: Compute the genetically and environmentally determined expectation of the number of descendants, draw a random derivate from the corresponding Poisson distribution (assumes that family sizes are Poisson-distributed) and produce the corresponding number of offspring.
  - **Compute the number of surviving offspring** that are going to be produced in this generation (see Figure 20). Simulator005 reduces effective fertility according to the num-

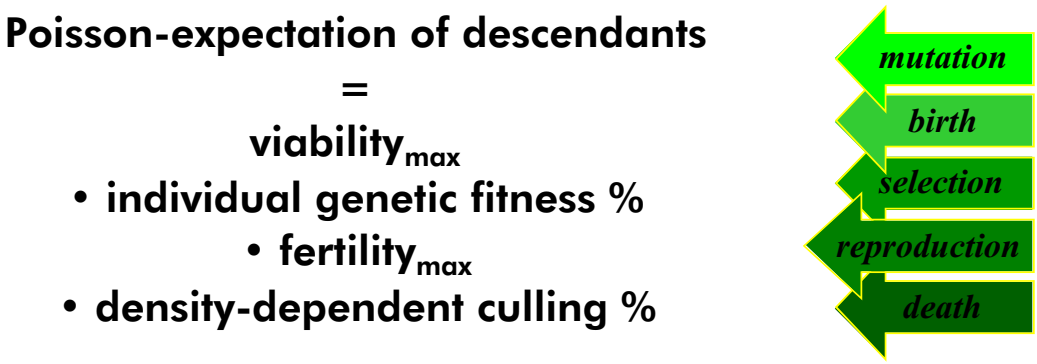


Figure 20 Factors that influence the number of offspring and sequence of events in life history.

ber of mutations in an individual. This should be equivalent to the approach of previous ratchet simulations who reduce juvenile viability, as maximal fertility and viability are nothing but commutable factors in the equation in Figure 20. The red

Model	Zygotes N(t)	Viability V_{max}	Genetic Fitness %	Fertility R_{max}	Density-Culling %
Wright-Fisher	constant	1	individual F	unlimited	implicit pick exactly N zygotes
Mutational Meltdown	some function	1	individual F	fixed	Implicit; pick exactly N zygotes, if possible
Lag Culling	determined by individuals' behaviour and environment	1	individual F	fixed	depends on mean fitness of last generation $K/(N*(\sum(F\%*R_{max})/N))$ maximum = 100%

Table 13 Population density regulation in various simulation models of Muller's ratchet.
The Wright-Fisher model allows observations of the ratchet, but can not lead to extinction[a], while the limited fertility of a mutational meltdown model can drive a population to extinction, if enough deleterious mutations have been accumulated[b]. The red bar denotes factors with a close link in the implementation of the model. Simulator005 uses LagCulling, which allows for extinctions and implements a density culling mechanism that is more natural for some situations.

a. Widely used. For examples see Gessler (1995) "The constraints of finite size in asexual populations and the rate of the ratchet", Genet. Res. 66:241-253. - Charlesworth & Charlesworth (1997) "Rapid fixation of deleterious alleles can be caused by Muller's Ratchet", Genet. Res. 70:63-73.
b. Lynch & Gabriel (1990) "Mutation load and the survival of small populations", Evolution 44:1725-1737. - Lynch et al. (1995) "Mutational meltdowns in sexual populations", Evolution 49:1067-1080.

bars in Table 13 help to illustrate that this is not a fundamental difference: In each case two constants are multiplied with the individual genetic fitness that ranges from 100% to 0%. This applies to viability selection at point **Va**, which is the viability selection point other authors use. Viability selection at the other points is expected to lead to more complicated results and should be explored in separate simulations, if necessary. Furthermore, the explicit mechanism of population density regulation differs fundamentally between S005 and the implicit mechanisms of other models.

- **Produce all offspring** for this individual and observe several statistics for the population on the fly (distributions of fitness, ratchet mutations and background mutations; number of individuals with highest and lowest fitness and with the least ratchet or background mutations).

- **Die.** As generations are discrete, each individual is killed with 100% probability at the end of each simulated step in time, viability selection point **Vc**.

It should be noted that, contrary to other simulations of Muller's ratchet, here each individual of the parent generation is visited exactly one time to compute its offspring. Other simulations use a sampling procedure that randomly picks a parent, produces one offspring, determines whether it survives and then picks the next random parent, until either the offspring population is full or reproductive capacity has been exhausted. This is a typical case, where both abstractions of reality seem to be valid and different simulations are needed to determine effects.

Population model

Each population needs to be somehow regulated in size to avoid unlimited growth, something that no habitat nor any computer simulation can entertain. The easiest density regulation can be implemented from a population's perspective, as described above: The population picks as many randomly produced offspring, until it is filled. However, as the individuals of a population are the active agents in nature, an individual-centric approach might be more natural and is at least worth trying. In S005 each individual determines the number of offspring it produces. In this process it is influ-

enced by its own genetics as well as by the environment. The first three factors in Figure 20 describes the maximal number of offspring an individual might produce under optimal environmental conditions like an empty habitat. Most of the time, however, individuals do not live in empty habitats and, thus some density-dependent reduction of the number of surviving offspring occurs. Whether this is due to decreased fertility or decreased juvenile survival does not matter here; all that matters is that fewer offspring reach maturity than in an empty habitat, ie. without intraspecies competition. The density-dependent culling factor ***DDC*** measures this reduction. It can be computed by considering the maximal number of offspring that all individuals in the current population could produce if they were in a habitat of virtually infinite size. Dividing this number by the carrying capacity of the current habitat yields a quotient that measures the density-dependent reduction of offspring-production in the current population. The inverse of it yields the density-dependent culling factor:

$$DDC = 1 / \left(\frac{\mathbf{N} \cdot \mathbf{AMRC}}{\mathbf{K}} \right) \tag{16}$$

where ***K*** is the carrying capacity, **N** is the population size of the current generation and ***AMRC*** is the Average Maximal Reproductive Capacity of its individuals:

$$\mathrm{AMRC} = \frac{\sum_{i=1}^{N} F_i R_{max} V_{max}}{N} \tag{17}$$

where F_i is the genetic fitness of an individual (between 0 and 1) and R_{max} is the maximal reproductive capacity and V_{max} is the maximal survival to reproduction of the species. Unfortunately, the actual size of the currently computed generation is not available during computation and neither is the average maximal reproductive capacity. Therefore the corresponding values of the last generation are used. Thus ***DDC*** can be computed by

$$DDC = K / \left(\sum_{i=1}^{N_{t-1}} F_i R_{max} \right) \tag{18}$$

However, it is necessary to limit ***DDC*** to values of 1 or smaller, if one wants to observe extinctions. The intuitive meaning of this is that the environment does not compensate for unconditionally deleterious mutations. As the use of last generation values for ***DDC*** produces a one generation lag be-

tween a particular request for culling and the actual culling process, this model is called **LagCulling**. It applies to a number of natural populations[135].

Clicks are observed in EEPopulation

While some parameters are observed in the evolve-method of the individuals (IDs 52-54,56,58-61), some are computed analytically at the start of the simulation (26,38,42,55,57), and some are observed by the global evolve framework (computing performance ParameterID's 14-20). The largest fraction of all other output parameters, however, is concerned with clicks of the ratchet. As these can only occur in a population, they are determined in the `EvolveAndObserve()` method of the corresponding `EEPopulation` derived class. This method is called by the global evolve-method, before the evolve-methods of all individuals are called. Thus, in every generation the distribution of mutations (59-61) is computed from the number of mutations in every individual and the number of individuals in the least loaded classes (53,54,56,58) is recorded. If this number passes limits that characterise transitions from one phase of the ratchet to another, the corresponding durations of that phase are recorded. If a least loaded class contains no individuals, its ratchet has clicked.

12.4 Validation of Simulator005

Code of Simulator005 has been successfully used since 1999. Since then it has been checked for plausibility, compared to analytical extremes, compared to simulation data of other authors and compared to analytical theory. The biologically relevant core-functionality has been checked several times and seems to work properly. It leads to results that are plausible, when compared with others.

Besides a number of trivial errors with non-trivial consequences and absolutely non-trivial debug processes (eg. mix up "+" and "-"), one error might be mentioned as an example for the importance of plausibility tests: in a certain test-run, zero-mutation rates lead to mutation accumulation, a clear indication that something was going wrong. Inspection of the code revealed a problem in the early version of the Poisson random number generator: the ANSI generator employed produces only 16 bit different values. If they are

135. Bjornstad & Grenfell (2001) "Noisy clockwork: time series analysis of population fluctuations in animals", Science 293:638-643. - Saether (1997) "Environmental stochasticity and population dynamics of large herbivores: A search for mechanisms", Trends Ecol. Evol. 12:143-149. - May (1976) "Models for single populations", pp. 4-25 in: May (ed) Theoretical ecology: principles and applications, Oxford, Blackwell Scientific Publications.

divided by the largest value possible, the corresponding double float values are still not more diverse than 16 bit. Thus the code that produced the Poisson derivates then leads to the generation of an event every 1/16bit, which showed up as mutations (ca. 10^{-4}). Having found that bug, an *ad hoc* workaround was constructed that gave correct results for expectations of more than 0.001. However, it had never really been tested for lower values at the time of implementation, because it was not needed for lower values (Simulator003, 1999). When the same Poisson generator code was reused to build Simulator005, it was assumed to work fine, so only a new random number generator was employed[136]. However, the old workaround had been forgotten. Then Simulator005 was released (S005r1-S005r5), results came in and finally, the first large thorough analysis session started. One surprising result was that Muller's ratchet appeared to slow down much more than expected at very low mutation rates ($U \leq 0.0005$) in a region that should be best described by the $1/\boldsymbol{U}$ expectation of clicktime. Before thinking too long about potential biological implications, the old workaround was remembered. After inspecting source code, it was found that indeed this had been the reason for the 'surprising biological phenomenon'. Fortunately, only a fraction of all results was affected by this '`PoissonBug`' (all simulations, where mutation rate $\boldsymbol{U}$ was lower or equal to 0.001, with lower values having effective mutation rates that were *much* lower and an unnaturally elevated level of double mutations that led to an increase in multiple clicks). The analyses presented in this work exclude all corresponding results (except two plots that mention the use of practically unbiased $\boldsymbol{U}$=0.001 explicitly in their legend). After finding the `PoissonBug`, release 6 of Simulator005 was prepared with the corresponding correction and a number of other improvements.

This incident shows clearly the importance of a quality framework. Had the current implementation of ε been better, this bug would have never occurred. Had it been worse, there would have been no chance to easily remove problematic runs and get better estimates in the corresponding region of parameter space, as documentation would have been missing. Thus, investment in a high quality framework for simulations of evolution is something that really pays off in practice.

136. The new generator has 53 bit different values (see precision of double floats) and is described as ran2 on page 282 of Press et al. (1992) "Numerical recipies in C". 2nd, Cambridge, Cambridge University Press. - Original reference: L'Ecuyer (1988) "[Long period random number generator]", Communications of the ACM 31:742-774.

12.5 Future of Simulator005

Future releases of Simulator005 will allow the simulation of various distributions of mutational effects. To this end, just other values of the `dmeb` and `dmer` parameter than 0 have to be linked to the corresponding distribution they define. In the long term, Simulator005 should accompany general framework development until the relevant parameter space has been searched sufficiently. This requires several million single simulations that need several thousand CPU years, if some future projects look deeper into what happens when two very different ratchets meet each other in large populations and with different distributions of mutational effects. Until these projects have been completed, hopefully a fully automated release of the simulator with graphical user interface will be available. As such a fully automated release contains an automated simulator download manager, the next simulators can be expected to have a faster start.

The results computed by S005 will be published in a corresponding database on the Internet, once the server suite is advanced enough to perform the necessary transformations automatically. These results will remain on the web, even after Simulator005 has completed all its projects. After S005 has eventually stopped, it might be reactivated at any time, if particular questions can not be answered from the data available and more detailed simulations are needed. The resulting long-term database should help biologists to answer corresponding questions about Muller's ratchet easily without further complex computations.

13 Analytic predictions of the rate of Muller's ratchet in *Mathematica*

Diffusion theory and quantitative genetics methods allow for approximations of the rate of Muller's ratchet. As they involve complex calculations, Mathematica is needed for actual computation. Here an integrated method is presented that automatically picks the correct method and presents the best analytical prediction currently available.

13.1 Diffusion theory

Diffusion theory has been remarkably successful at approximating the rate of Muller's ratchet for parameter sets, where the least loaded class $\mathbf{N_0}$ has a reasonable size ($\mathbf{N_0} > 1$ or better $\mathbf{N_0} >> 1$). The following approach has been developed by STEPHAN[137,138] and has been successfully used and extended by others[139]. It is based on HAIGH's model of the ratchet[140]. Please consult these references for further details. This chapter focuses on how to compute the rate of the ratchet.

Phase 1

The time between two clicks can be divided in two phases. In phase 1, immediately after each click, the previously second best class $\mathbf{N_1}$ has just become the best class $\mathbf{N_0}$. Thus it has a frequency that is related to the equilibrium size for $\mathbf{N_1}$, which is higher than equilibrium for $\mathbf{N_0}$. Therefore, if the new best class is more frequent than about 1.6 $\mathbf{N_0}$, it will decrease to its new equilibrium value. This takes T_{1a} generations. Then decrease slows down, as drift gains more importance. This time can be computed purely deterministically[141] by

$$T_{1a} \approx \frac{1}{s}\left(1 - \frac{1.6s}{U}\right) \tag{19}$$

137. Stephan & Kim (2002) "Recent applications of diffusion theory to population genetics", pp. 72-93 in: Slatkin & Veuille (eds) Modern developments in theoretical population genetics, Oxford, Oxford University Press.
138. Stephan et al. (1993) "The advance of Muller's ratchet in a haploid asexual population: Approximate solutions based on diffusion theory", Genet. Res. 61:225-231.
139. Gordo & Charlesworth (2000) "On the speed of Muller's ratchet", Genetics 156:2137-2140. - Gordo & Charlesworth (2000) "The degeneration of asexual haploid populations and the speed of Muller's ratchet", Genetics 154:1379-1387. - Charlesworth & Charlesworth (1997) "Rapid fixation of deleterious alleles can be caused by Muller's Ratchet", Genet. Res. 70:63-73.
140. Haigh (1978) "The accumulation of deleterious genes in a population - Muller's Ratchet", Theor. Popul. Biol. 14:251-267.
141. Gordo & Charlesworth (2000) "On the speed of Muller's ratchet", Genetics 156:2137-2140.

where U is the genomic deleterious mutation rate per generation that have the effect s (with a positive selection coefficient for deleterious mutations in this chapter, because the original papers do the same). However, drift might oppose this decay to the new equilibrium. Taking this into account, this process might be described as a one-dimensional diffusion process on the interval $[1.6x_0, x_1]$, where x denotes the equilibrium frequency of individuals in the best (0) or second best (1) mutation class. Using the diffusion theory[142], the expected time can be estimated to be[137]

$$T_{1a} \approx \frac{1}{s} \ln\left(\frac{U}{1.6s}\right) \tag{20}$$

As time for decay from $1.6N_0$ to N_0, denoted by T_{1b}, is a slow diffusion process, it is usually subsumed under T_2, the duration of Phase 2. Simulator005 allows separate observations of T_{1a}, T_{1b} and T_2.

Phase 2

Phase 2 describes the time that is needed for the best class to diffuse from its equilibrium frequency to 0. The infinitesimal operators of the corresponding one-dimensional diffusion process are given by[137]:

$$a(x) \approx ksx\left(1 - \frac{x}{x_0}\right) \tag{21}$$

$$b(x) \approx \frac{x}{N}(1 - x) \tag{22}$$

where k is a special value, N is the population size, s and U are the same as above, x_0 is given by $e^{U/-s}$ and $x \in [0, 1]$. Earlier work used $k = 0.5$. However, $k = 0.6$ has been found in recent calculations[143]. The diffusion process is defined on $[0, 1] = [0, x_0]+[x_0, 1]$, as the frequency of the best class might not only be smaller than equilibrium x_0, but also significantly larger[144] (see Figure 5 on page 34). With these formulae and standard diffusion theory[142], the expected time for Phase 2 can be calculated (see also[145]) as

$$T_2 = T_{0, x_0} + T_{x_0, 1} \tag{23}$$

where $T_{0,x0}$ is computed by integrating Ewens equation 4.39 (page 123 in [142]) from 0 to x_0:

142.Ewens (1979) "Mathematical population genetics", Berlin, Springer Verlag.
143.Stephan & Kim (2002), ibid. - Gordo & Charlesworth (2000), ibid., Genetics 154:1379-1387.
144.Charlesworth & Charlesworth (1997), ibid. - Stephan & Kim (2002), ibid.
145.Appendix in Charlesworth & Charlesworth (1997), ibid.

```
ratchetClickTimeStephan[k_, ne_, u_, s_] :=
 (* Implemented by Loewe 2002-03
   References: Stephan & Kim (2002), Ewens (1979)
       k old is 0.5 k new is 0.6, s>0 is deleterious *)
 Module [ {n0, x0},
      n0 := ne * E^-(u/s);
      result = If [n0 ≤ 1, 1 (* out-of-range dummy result *),
    x0 := E^-(u/s);
    upplim := 1 - (1 / ne);
    tphase1[u, s] := (1/s) Log[ u / (1.6 s) ];
    a[x_, u, s, k] := k * s * x * (1 - x/x0);
    b[x_, ne] := (x/ne) * (1 - x);

    (* the following standard diffusion theory formulas
      are from Ewens, 1979, Eq 4.17, 4.39, 4.40, 4.22 *)

    psi[y_, u, s, ne, k] :=
     E^-2*NIntegrate[a[z,u,s,k]/b[z,ne], {z,0,y}];

    t439[x_ , u, s, ne, k] :=
      (2 * NIntegrate [psi[y, u, s, ne, k], {y, 0, x}])
                  / (b[x, ne] * psi[x, u, s, ne, k]);

    t440[x_ , u, s, ne, k] :=
      (2 * NIntegrate[psi[y, u, s, ne, k], {y, 0, x0}])
                  / (b[x, ne] * psi[x, u, s, ne, k]);

    totaltime[ u, s, ne, k] :=
     NIntegrate[t439[x, u, s, ne, k], {x, 0, x0}]
      + NIntegrate[t440[x, u, s, ne, k], {x, x0, upplim}]
      + tphase1[u, s];

    totaltime[ u, s, ne, k]
   ] (* endif *)
] (* end of module *)
```

Figure 21 Implementation of the diffusion approach to predict the clicktime in *Mathematica*.

$$T_{4.39}(x, x_0) = 2\frac{\int_0^x \psi(y)dy}{\{b(x)\psi(x)\}}, \quad 0 \le x \le x_0 \tag{24}$$

and $T_{x0,1}$ is computed by integrating EWENS equation 4.40 (page 123 in [142]) from x_0 to practically 1:

$$T_{4.40}(x, x_0) = 2\frac{\int_0^{x_0} \psi(y)dy}{\{b(x)\psi(x)\}}, \quad x_0 \le x < 1 \tag{25}$$

while

$$\psi(y) = \exp\left[-2\int_0^y \frac{a(z)}{b(z)}dz\right] \tag{26}$$

is EWENS equation 4.17 (page 119 in [142]). As $b(x)$ is part of denominators, it is not practical to integrate numerically exactly up to x =1, else $b(x)$ becomes 0 and computation ends with an error. Therefore, 1-1/**N** is used as upper limit for the integration of Equation (25) without losing any accuracy.

Computation of click time

Here, total click time is obtained by adding Equation (20) and Equation (23) using k = 0.6 in *Mathematica*[146]. The code used is given in Figure 21 and takes some moments for execution. While single click times are obtained with reasonable speed, it should be noted that the large number of calculations necessary for generation of plots can easily lead to hours (2D plot) or even days (3D plot) of computing time (on a PowerMac G3 at 350 MHz). Unfortunately, instead of giving an estimate of computing time, some parameter choices lead to numerical problems resulting in various error messages of *Mathematica* that are hard to understand for newcomers. Part of this might be due to the fact that the approach presented here selects the upper integration limit automatically (1-1/**N**). Plotting values of the integral over various upper limits leads to a normal flat line at the final value in most cases. However, occasionally quite a strange graph occurs. Nevertheless, this automatic upper limit is usually easier to handle than manual determination of the upper limit for each clicktime prediction. A considerable amount of computing time is spent in Equation (25) that had not been used in an earlier paper[147]. Thus, computation can be speeded up by about an

146.Wolfram Research (2002) "Mathematica Homepage" http://www.wolfram.com
147.Stephan et al. (1993), ibid.

order of magnitude, if $T_{x0,1}$ is omitted from Equation (23), albeit with loss of accuracy in many cases. Although the approximation

$$\psi(y) \approx \exp\left[2Ne^{U}y\left(\frac{y}{2} - x_0\right)\right] \tag{27}$$

has been proposed[145], the resulting double integration still needs a sophisticated maths-package, if one wants to arrive at a click time estimate.

13.2 Quantitative genetics theory

A number of quantitative genetics approaches have been used to predict the rate of Muller's ratchet. They use the change in mean and higher moments of the distribution of mutations per individual and treat it like a quantitative trait[148,149]. The method developed by Gessler nicely complements the diffusion theory approach, as it works well for N_0<1.

It is based on the fact that in this region of parameter space the best class cannot ever attain its mutation-selection equilibrium frequency, since individuals can not be divided. Thus there is a quasi-deterministic force that drives the best class to extinction. This circumstance allows development of a probability mass function for the distribution of the number of mutations per individual, which in turn allows one to compute the rate of the ratchet as done in Equation 8 of Gessler. Inversion gives the effective click time T_{cl}:

$$T_{cl} = \frac{1}{sk - st} \tag{28}$$

where again s stands for a constant selection coefficient that is positive for deleterious mutations, and k and t are computed as follows:

$$k = \min[x \,|\, Ne^{-\theta}\theta^{x}/x! \geq 1, x \in 0, 1, 2, \ldots\theta\}] \tag{29}$$

where N is the population size and $\theta = U/s$, again with U as mutation rate;

148. Gessler (1995) "The constraints of finite size in asexual populations and the rate of the ratchet", Genet. Res. 66:241-253. -

149. Pamilo et al. (1987) "Accumulation of mutations in sexual and asexual populations", Genet. Res. 49:135-146. - Gabriel et al. (1993) "Mullers Ratchet and mutational meltdowns", Evolution 47:1744-1757. - Lynch et al. (1993) "The mutational meltdown in asexual populations", J. Hered. 84:339-344. - Higgs & Woodcock (1995) "The accumulation of mutations in asexual populations and the structure of genealogical trees in the presence of selection", J. math. Biol. 33:677-702. - Prügel-Bennett (1997) "Modelling evolving populations", J. theor. Biol. 185:81-95.

$$t = b - k \tag{30}$$

where b is given by Equation 3 of Gessler:

$$b = \min[x \,|\, N f(x, x) \geq 1, x \in \{k, k+1, k+2, ..., \theta\}] \tag{31}$$

where $f(x, x)$ is given by Equation 2 of Gessler:

$$f(b, x) = \begin{cases} \dfrac{\lambda^{(x-k)}}{(x-k)!}\left[\sum_{i=b-k}^{\infty} \dfrac{\lambda^i}{i!}\right]^{-1} & x \geq b \\ 0 & \text{otherwise} \end{cases} \tag{32}$$

where $\lambda = \theta - k$. To actually determine b, $f(b,x)$ is called with $b = x$, until the condition in Equation (31) is satisfied.

Intuitively, these formulae scan all mutational classes to find the first class that is not removed deterministically. The number of mutations carried by this class is the key for successful prediction. For additional explanations, please refer to Gessler[148].

While there are no theoretical problems with this approach, the excessive use of factorials (!) generates problems, even for such a powerful system as *Mathematica* on a modern Computer. A numerical approximation of 10^5!, 10^6! and 10^7! takes 5 sec, 3 min and about 1 hour on a PowerMac G3 at 350 MHz and surpass by far the range of all machine supported numbers with about 10^{456572}, $10^{5565708}$ and $10^{65657059}$, respectively. Such numbers, however, are needed to use this method for eg. U= 0.1 and s = 10^{-7}. Close investigation of these formulae might lead to considerable speedup (eg. if the loop that computes k moves downward from θ instead of upward from 0). However, the numerical complications should not be underestimated (eg. use of inverted k computation could speed up things considerably, but produced error messages for $N > 10^7$ with θ=1000). As no errors were found for normal, individual computations when θ was 500 or less, the range of this function was limited correspondingly. As with the diffusion approach, proper computation of individual values does not imply proper collaboration with the comfortable plot and other capabilities of *Mathematica*. Please find the code used in Figure 22.

```
ratchetClicktimeGessler500[ne_, u_, s_] :=
 (* Implemented by Loewe 2002-03, Reference: Gessler 1995,
  s>0 is deleterious as in formulas by Gessler *)
 Module [ {n0, b, x, k, lambda, theta, r},
  n0 := ne * E^-(u/s); theta := u / s;
  If [n0 ≥ 1, 100 (* dummy OutOfRangeResult *),
   If [theta > 500, 0.1 (* dummy ComputationTooComplexResult *),
    gesslerk[ne, theta_] :=
         Module[ {x = 0, resultk = -1},
          While[(resultk == -1) && (x ≤ theta),
           If[ne * Exp[-theta] * theta^x / x! ≥ 1,
             resultk = x, (* else *) x = x + 1 ]; ];
          resultk  (* returns k *)
          ] (* EndModule gesslerk *);
    k := gesslerk[ne, theta];
    lambda := theta - k;
    gesslerb[ne, theta_, lambda_, k_] :=
         Module[ {x = k, resultb = -1},
          While[(resultb == -1) && (x ≤ theta),
           If[ lambda^(x-k) / ((x - k)! * Sum[(lambda^i / i!), {i, x - k, Infinity}])
              (* SimpleGesslerf with b=x *)
             * ne ≥ 1,    (* EndOfCondition *)
             resultb = x,  (* else *)  x = x + 1 ]; ];
           resultb (* returns b *)
          ] (* EndModule gesslerb *);
    b := gesslerb[ne, theta, lambda, k];
    t := b - k;
    r :=  s * k - s * t;
    effectiveClickTime = 1 / r  (* returns final result *)
   ] (* endif computation too complex*)
  ] (* endif out of range of Gesslers method *)
 ] (* end of module *)
```

Figure 22 Implementation of Gessler's approach to predict the clicktime in *Mathematica*.

Note added in proof: After submitting this work, Christoph Reisinger[150] pointed out to me what would have been obvious for mathematitians, namely that

$$\frac{a^b}{b!} = \frac{a}{1} \cdot \frac{a}{2} \cdot \ldots \cdot \frac{a}{b} = \prod_{i=1}^{b} \frac{a}{i} \tag{33}$$

which can be combined with the well known result that

$$\sum_{b=0}^{\infty} \frac{a^b}{b!} = e^a \tag{34}$$

to substitute the sum in Equation (32):

$$\sum_{i=b-k}^{\infty} \frac{\lambda^i}{i!} = e^{\lambda} - \sum_{i=0}^{b-k-1} \frac{\lambda^i}{i!} = e^{\lambda} - \sum_{i=0}^{b-k-1} \prod_{j=1}^{i} \frac{\lambda}{j} \tag{35}$$

Similarly, all other factorials in Gessler's approach can be removed. Thus one should be able to compute results by simple loops that do not need the hyperastronomical exponents mentioned above.

13.3 Building a comprehensive method

To facilitate state-of-the-art analytical predictions of the rate of Muller's ratchet, both approaches are combined into one function that automatically calls the correct code: the diffusion approach for $N_0 \geq 1$ and the quantitative genetic approach for $N_0 < 1$ and $\theta < 500$. Parameter combinations outside that range are characterised by an overwhelming mutational pressure suggesting a simple, quasi-neutral expectation like $T_{cl} = 1/U$. The corresponding code for *Mathematica* is given in Figure 21 - Figure 23 to help biologists outside the field access these analytical methods. Please check against typographical errors by recomputing some of the values given by Gessler[148] and Stephan & Kim[137], before relying on the functions. The code given here has passed that test successfully.

```
ratchetClickTime[ne_, u_, s_] := If[s ≤ 0, -1 (* dummy *),
  If [ne * E^-(u/s) ≥ 1, ratchetClickTimeStephan[0.6, ne, u, s],
   If[u / s ≤ 500, ratchetClicktimeGessler500[ne, u, s], 1 / u]
  ]]
```

Figure 23 Integration of analytical approaches to predict the clicktime in *Mathematica*.

150. These helpful discussions took place during a research visit in the Technical Simulation group of G. Wittum at the Interdisciplinary Center for Scientific Computing in the University of Heidelberg, Germany, where CR is currently located. Thank you to GW for inviting me and to CR for taking enough time to look at these details.

14 Fitness measurements in bacteria

One potential measure of fitness is the growth rate of a bacterial culture in liquid medium. It can be readily determined by measuring optical density. However, to increase accuracy of estimations of the true growth rate, many replicates are needed. To analyse about 6000 growth curves with a total of more than 1.5 million OD values, a special software module for Excel VBA was developed. It allows fast high precision fitness measurements. Equipment, software, bacterial strains and other experimental details are described in this chapter.

14.1 Experimental setup

Bioscreen C

Today microtiter plate readers are employed quite often to measure large numbers of growth curves. This study used the *Bioscreen* C (Labsystems, Helsinki, Finland), a plate reader that is particularly well suited to bacterial growth measurements, but does not use standard microtiter plates. Key features that improve accuracy include the following (past tense = experimental conditions during this work):

- **Integration.** Thermal control, shaking and measuring are integrated into one device. Thus a culture can be incubated at a precise temperature and there is no need to change device for shaking or measuring.
- **Thermal accuracy** over a wide range of temperatures is achieved by Peltier-elements that heat or cool the culture plates they enclose. Actual temperatures were recorded automatically for each moment of time and deviated much less than 0.1°C from their nominal value at 37°C, which was the temperature of the lower element. The upper element has a higher temperature (+1°C) to avoid condensation of water on the cover.
- **Continuous shaking** was employed to ensure homogeneity of the cultures. The device is robust enough not to break.
- **Sterile measurements** are facilitated by a special 100 well plate (HoneyComb2). It allowed vertical OD 600 measurements without having to remove the cover. Each well has a total volume of 600 µl, but only 245 µl were used to avoid contamination during shaking.
- **Data export** facilities included complete transfer of all details to a DOS program (`Biolink`) that allowed a quick overview as well as

storing and exporting data to Excel. Biolink export and Excel import filters were developed by Volker Textor.

Pipetting was done under a sterile workbench with a special, programmable 12-channel 250µl pipette (Impact^2, Integra Biosciences, Fernwald, Germany). For pipetting schemes, see Figure 29 on page 198 and Figure 32 on page 210. Without such a high degree of automation the existing database of replicates could not have been accumulated. Besides several test plates, 58 plates with 100 wells each were observed for 24 hours at 5 minute intervals, producing 1.67 million OD values.

Sources of variability

High precision growth speed measurements can be influenced by variability in

- temperature (eliminated by high precision device),
- media composition (eliminated by using the same media production batch for each experiment; first batch for experiments VTX002 + VTX007, second batch for experiments VTX005 + VTX006),
- concentration of inoculum (might lead to varying overlap of lag phase and stationary phase slowing down apparent maximal growth rate; was reduced by transferring standard dilutions from stationary phase; variability was further reduced when high dilutions were used),
- spatial population structure within a well (eliminated by continuous shaking) and
- condensed water on the plate cover.

Condensed water

Condensed water turned out to be a surprisingly complex problem. The *Bioscreen* C hardware is designed to minimise the problem by increased heating of the plate cover. However, occasional low room temperatures led to unexpected and unpredictable condensate production. This is probably due to the fact that the *Bioscreen* C was operated at room temperature (17-23 °C) and not in a special heating chamber, as suggested by the manual that specifies that environmental temperature should not be more than 10 °C away from well incubation temperature. In wells where condensate was built in those 30-60 minutes of maximal growth, optical densities decreased faster than normal and suggested increased growth rates. In wells where condensate dissolved during that time, the contrary was observed. At the end of each experiment, the cover was checked for condensed water, but this did not allow detection of transient condensate on the plate cover during the experiment. Only after comparing observed condensate patterns (no condensate on wells at the border) to occasional strange growth rate patterns

(increased or decreased growth rates in wells at the border), the influence of condensed water became clear. As each plate contained at least one control probe (sterile medium), the timing of OD changes due to condensed water could be checked. In the analysis of results presented below, growth rate measurements were removed if they showed any signs of problems due to condensed water.

The need for an evolution-robot

While the *Bioscreen* C is[151] one of the best systems available for measuring large numbers of high-precision growth curves, actual work revealed *much* room for improvement. From the point of view of evolutionary mutation accumulation experiments, it would be desirable to construct a device that integrates all *Bioscreen* features above with a condensate-free light path and a sterile pipetting facility allowing for automated serial transfers whose growth curves could be observed for each transfer. While development of such an evolution-robot is no small effort, it would be necessary if large numbers of high-precision evolution experiments were ever done. Some initial design-drafts for such an evolution-robot were developed, but more details are not be discussed here due to space limitations.

Most of the actual experimental work was performed by Volker Textor.

14.2 Software

In order to analyse more than 1.5 million single OD values, I developed a software design to automate (i) import and management of raw data, (ii) estimation of exact maximal growth rates including visual control, (iii) comparisons of distributions of growth rates between different serial transfers and between different lines and (iv) Bateman-Mukai analysis to allow fast estimation of mutational parameters. This design was implemented in Visual Basic for Applications (VBA) in Excel (Office 2001 on a PowerMac) by Volker Textor[152]. The following overview is no substitute for his detailed description of features.

BioScreenVBA

The two modules `BioScreenVBA.xls` and `BatemanMukaiVBA.xls` were designed around the following workflow. All raw data files were transferred to `ExcelImportFiles` (EIF) via `BioScreenVBA.xls`. Then the

151. It is currently not produced at the moment.

152. Diploma thesis of Textor (2001) "Programmierung eines VBA-Systems zur Auswertung von Mutationsakkumulationsexperimenten mit Mikroorganismen in Flüssigkultur und Bestimmung der Nachweisgrenzen für Fitnessveränderungen", Institut für Mikrobiologie, Forschungszentrum für Milch und Lebensmittel, Technische Universität München.

same module was used to estimate maximal growth rates by a linear regression of the steepest slope in log OD values over linear time. As time on the x axis was the independent variable, regression of y on x was the correct way to do it[153]. The slope was used to compute the Wrightian fitness w in units of offspring per hour. It can be derived from population size N at time t and time t+1h according to the following formula[154]:

$$w = \frac{N_{t+1}}{N_t} \quad \Leftrightarrow \quad N_{t+1} = N_t \cdot w \tag{36}$$

This formula assumes that the bacterial population has a stable age structure. The software also supports transformation to the doubling time d and to the Malthusian parameter m, which is also known as the specific growth rate frequently used in microbiology[155]. Both have 1 hour as corresponding time unit and are given by the following formulas:

$$m = \ln(w) \quad \Leftrightarrow \quad w = e^m \tag{37}$$

$$d = \frac{\ln(2)}{\ln(w)} \quad \Leftrightarrow \quad 2 = w^d \tag{38}$$

While there are numerous other potential definitions of fitness[156], each with different advantages, w, m or d are often used for situations, where speed of growth in an empty habitat (like a well) is considered as a critical component of fitness.

To facilitate selection of the approximate timeframe of the steepest slope, a number of features were developed, including automated analysis of a fixed timeframe in all growth curves of a given date. However, as automated analysis is not very reliable without sophisticated pattern recognition algorithms, a plot provided easy visual inspection and correction by showing

153. See page 136-141 in Harms (1998) "Biomathematik Statistik und Dokumentation. Eine leichtverständliche Einführung". 7, Kiel, Germany, Harms Verlag.
154. See page 5 f. in Crow & Kimura (1970) "An Introduction to Population Genetics Theory", Edina, Burgess International Group Incorporated.
155. Eg. compare formulas in Painter & Marr (1968) "Mathematics of microbial populations", Annu. Rev. Microbiol. 22:519-548. See also transformations by de Ferro et al. (1999) "Physiological aspects and conservation of a Veillonella strain isolated from the oral cavity. Interaction with streptococci", Anaerobe 5:255-259.
156. Brommer (2000) "The evolution of fitness in life-history theory", Biol. Rev. Camb. Philos. Soc. 75:377-404. - Benton & Grant (2000) "Evolutionary fitness in ecology: Comparing measures of fitness in stochastic, density-dependent environments", Evol. Ecol. Res. Oct 2:769-789. - Murray (1997) "Population Dynamics of Evolutionary Change: Demographic Parameters as Indicators of Fitness", Theor. Popul. Biol. 51:180-184.

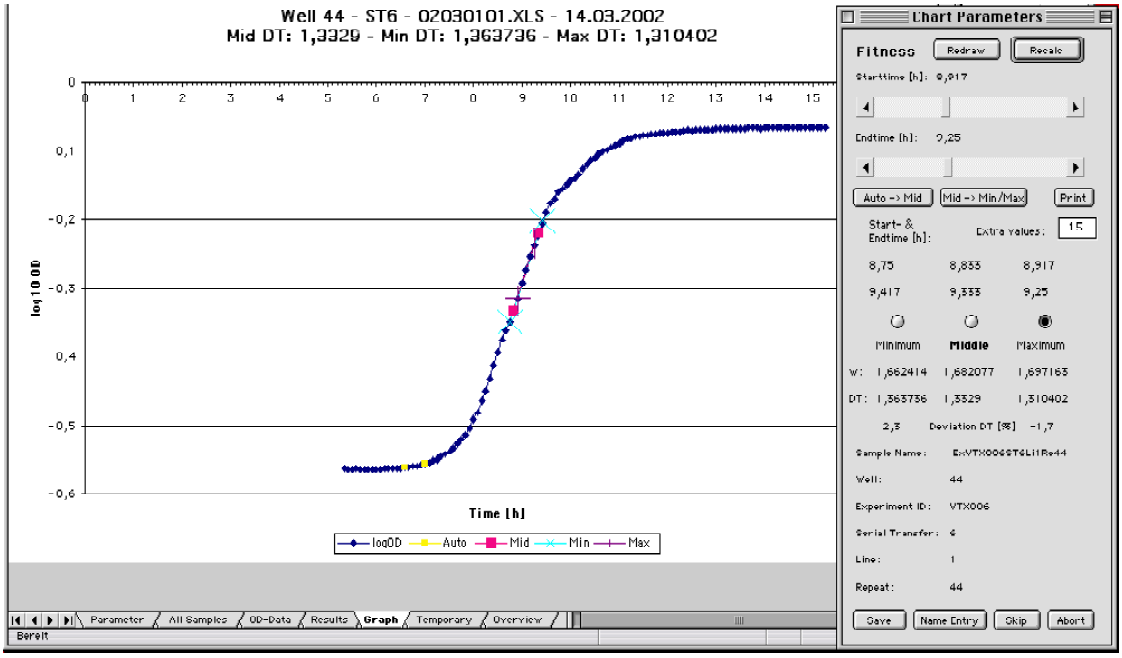


Figure 24 Screenshot of the semi-automated fitness analysis user interface of BioScreenVBA.xls. The OD curve shown is typical in every aspect except the long lag phase, which was selected to produce a nice plot. During normal operation, only 15 values below and above the interesting region are shown. Please note the slight change of growth rate in the middle of the log phase; it is most probably due to a change of substrate. In the present work only the steeper half was used.

raw data and all regression lines used for fitness calculations. The following measures of fitness were computed:

- automatic fitness (only used for fast pre-analysis),
- minimal manual fitness (the smallest fitness, that could still be a true maximum),
- middle manual fitness (the most probable fitness value) and
- maximal manual fitness (the highest fitness value that could be found in the credible region, even if it used the minimal number of raw data points for a regression (=3; correspond to 15 minutes).

The value used in final analyses was `AllManualMean`, the mean of minimal, middle and maximal manually measured fitness values. A screenshot of the graphical user interface with an overview over a typical growth curve can be found in Figure 24. For comfortable analysis (up to two growth curves per minute), zoom features, a preview of fitness values with percentage of deviations and some other features were included.

Complete Data File

After manual analysis, results were exported to the "Complete Data File" that contained all data for one mutation accumulation experiment. Each row contained one analysed growth curve with `ExperimentID`, `SerialTranferID`, `LineID`, `RepeatID`, `AutomaticalFitness`, `ManualMinimalFitness`, `ManualMiddleFitness`, `ManualMaximalFitness`, `OriginalSampleName`, `ExcelImportFileName`, `OriginalWellNumber` and `MeanOfAllManualFitnessValues`. This file was imported by the `BatemanMukaiVBA.xls` module.

BatemanMukaiVBA

This module allowed easy production of plots that compared (i) all replicates of two lines for one specific serial transfer or for all transfers, (ii) all replicates of two serial transfers for one specific line or for all lines and (iii) various overview plots that summarised changes in mean fitness from one serial transfer to the next for a given line. For all comparisons, a t-test[157] was used to compute the probability that observed differences are due to chance.

Furthermore, this module used the BATEMAN-MUKAI method[158] to compute U_{min}, a lower limit for the (haploid) genomic mutation rate per time unit, and s_{max}, an upper limit for the mean selection coefficient (mutation effect is independent of time for a given slope). Assuming that all mutations have the same effect and that non-genetic causes like environmental stochasticity have no influence, U and s are exactly given by

$$U_{min} = \left(p\left(\frac{\Delta M}{t}\right)^2\right) / \left(\frac{\Delta V}{t}\right) \quad (39)$$

$$s_{max} = \frac{\Delta V}{p \Delta M} \quad (40)$$

where ΔM denotes the difference between mean fitness at the beginning of the experiment and mean fitness at the end of the experiment and ΔV denotes the difference between the variance at the beginning and the variance at the end. Diploid organisms have $p = 2$, haploids have $p = 1$ and t is the total length of the experiment in arbitrary time units (eg. generations of the microbe or real-time days). To get meaningful values out of this procedure, the following issues were considered: (i) Build a regression of all mean fitness values and their corresponding variances to increase accuracy. (ii) Al-

157. See p. 73ff. + p. 80ff. in Dytham (1999) "Choosing and using statistics: A biologist's guide", Oxford, Blackwell Science. - Press et al. (1992) "Numerical recipies in C". 2nd, Cambridge, Cambridge University Press.

158. See p. 341-343 in Lynch & Walsh (1998) "Genetics and analysis of quantitative traits", Sunderland, Massachusetts, Sinauer Associates.

low enough time for mutation accumulation, if you want to measure meaningful values. (iii) If environmental conditions change, these formulae become meaningless. (iv) If variance decreases as a trend, you can be sure that environmental conditions obscured real mutational parameters. (v) Be cautious with results that do not make biological sense.

Equations (39) and (40) give exact values for U and s only in the unlikely case that *all* mutations have equal effects. However, it is more likely that most mutations have small effects. In this case $U = U_{min}$ and $s = s_{max}$. Sometimes it is assumed that mutational effects follow a Gamma distribution with the shape parameter β that is even harder to estimate than U_{min} or s_{max}. In that case, the BATEMAN-MUKAI results can be corrected[159] by

$$U_{true} = U_{min} \cdot \left(1 + \frac{1}{\beta}\right) \quad (41)$$

$$s_{true} = s_{max} \Big/ \left(1 + \frac{1}{\beta}\right) \quad (42)$$

However, the true shape of the distribution of mutational effects is a great riddle. There are two other methods for analysis of mutation accumulation experiments, the Maximum-Likelihood method[160] and the Minimum-Distance method[161]. Both allow for variable mutational effects and have considerable computational complexity. Therefore, this study employed only the BATEMAN-MUKAI method. However, one might want to use their estimates of β for the formulas above.

In the experiments analysed here, $p = 1$ for haploid bacteria and t was measured in units of 1 day. The latter convention is rather unusual, but reflects the emphasis of this study on mutation accumulation in the stationary phase.

14.3 Bacterial strains and media

All strains used here had been derived from an *Eschericial coli* B strain that had functional repair genes, could not metabolise L-arabinose, was strictly asexual and had neither functional bacteriophages nor plasmids[162,165]. RI-

159. Keightley (1998) "Inference of genome-wide mutation rates and distributions of mutation effects for fitness traits: A simulation study", Genetics 150:1283-1293.
160. Keightley (1994) "The distribution of mutation effects on viability in Drosophila melanogaster", Genetics 138:1315-1322.
161. Garcia-Dorado (1997) "The rate and effects distribution of viable mutation in Drosophila: Minimum distance estimation", Evolution 51:1130-1139.

Table 14 Strains received from Jeff Blanchard and used in this study.

Strain	Original name	Background	Specific mutations
1	REL606	original	none
2	REL4548	after 10 000 gen. in glucose minimal	none
3	JB1609	original	mut S^- (mutS::Tn5 kanR thi strR ara)
4	JB1618	"	mut $Y68^-$ (mutY68::Tn5 kanR thi strR ara)
5	JB1606	"	mut L^- (mutL::Tn5 kanR thi strR ara)
6	JB1615	"	mut $T1^-$ (mutT1 thi strR ara)
7	JB1621	"	mut $D5^-$ (mutD5 zaf-13::TN10 tetR thi strR ara)
8	JB1624	"	mut M^- (mutM::mini-tet thi strR ara)

CHARD LENSKI had used this strain to start his famous long-term evolution serial transfer experiments[163] and KIBOTA & LYNCH measured deleterious mutation rates in the same strain[164]. Thus there is some background for comparison of results. The actual samples were received from JEFF BLANCHARD, who used transposons to knock out several repair genes to investigate mutator properties[165]. Table 14 reviews the history of the 8 strains used in this study.

Media

Bacteria were grown in 600µl pre-sterilized 'HoneyComb2'-plate wells with a sterile cover that largely blocked exchange of air (Labsystems, Helsinki, Finland). The resulting lack of oxygen is probably the main reason for the reduced growth rate observed (ca. 40 min vs. 20 min under optimal conditions). Each well was filled with 240 µl Luria-Broth medium (LB) and 5 µl inoculum, either from the stationary phase of the corresponding predecessor culture or from a 1:48 dilution of the latter (5 µl stat-phase in 240 µl LB). The resulting dilutions (1:48 or 1:2304) are not radically different from the

162. Main reference: Lenski et al. (1991) "Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2000 generations", Am. Nat. 138:1315-1341. - Other studies that used the same basic strain: Chao et al. (1977) "A complex community in a simple habitat: an experimental study with bacteria and phage." Ecology 58:369-379. - Levin et al. (1977) "Resource limited growth, competition and predation: a model and some experimental studies with bacteria and bacteriophage", Am. Nat. 111:3-24. - Lenski & Levin (1985) "Constraints on the Coevolution of Bacteria and Virulent Phage - a Model, Some Experiments, and Predictions for Natural Communities", Am. Nat. 125:585-602. - Lenski (1988) "Experimental studies of pleiotropy and epistasis in Escherichia coli: I. Variation in competitive fitness among mutants resistant to virus T4", Evolution 42:425-432. - Lenski (1988) "Experimental studies of pleiotropy and epistasis in Escherichia coli: II. compensation for maladaptive effects associated with resistance to virus T4", Evolution 42:433-440.

163. Lenski & Travisano (1994) "Dynamics of adaptation and diversification: A 10,000-generation experiment with bacterial populations", Proc. Natl. Acad. Sci. USA 91:6808-6814.

164. Kibota & Lynch (1996) "Estimate of the genomic mutation rate deleterious to overall fitness in E. coli", Nature 381:694-696.

165. See e.g.: de Visser et al. (1999) "Diminishing returns from mutation supply rate in asexual populations", Science 283:404-406.

dilutions used by KIBOTA & LYNCH for fitness measurements (40µl/4ml) or LENSKI & TRAVISANO for serial transfer experiments (0.1ml/10ml). These dilutions allow for 5.58, 6.64 and 11.17 doublings before reaching the stationary phase again.

To produce an environment that is as homogeneous as possible, two 5 litre batches of LB were produced (50 g Tryptone, 25 g NaCl, 25 g Yeastextract, 4 l distilled water; mix, adjust pH to 7.4 and fill up to 5 l; distribute to many 50 ml or 100 ml flasks and autoclave for 20 min at 121°C). Experiments VTX002 and VTX007 used one batch, VTX005 and VTX006 the other. No antibiotics were used in any media; this might have led to the loss of the mutational repair deficiency mutations, as a potentially important selection pressure was removed and transposons might jump to other locations in the genome without destroying the gene they leave[166].

Before starting the first experiment VTX002, each strain was grown for two consecutive transfers in 100 ml LB at 37°C overnight (the standard procedure used by KIBOTA & LYNCH to minimise freeze-thaw effects on fitness; LENSKI & TRAVISANO used only one day for acclimatisation). The undiluted stationary phases of the 2nd transfer were used to

- inoculate VTX002, serial transfer 1 (ST1) with 5 µl
- produce 6 frozen samples per strain for later experiments (VTX002-VTX007).

Freezing

Freezing of these 1 ml samples was accomplished by mixing 700 µl stationary phase + 300 µl 50% (vol/vol) sterile glycerol in 1.5 ml PE-vials. After mixing and 20 min waiting (cells need to take up glycerol), the samples were moved from room temperature (ca. 20°C) to the freezer (ca. -70°C). Cooling rates were not measured exactly, but are probably in the range of 1°C/min by a factor of about 2. Slower freezing appears to help some cells, at least in the case of *Mycobacterium leprae*[167]. To freeze a plate with grown cultures, 105 µl of glycerol were added to the 245 µl stationary phase culture in each well. All other details were identical to the procedure above, except that cooling rates might slightly vary for different wells (those at the edge cool faster). The glycerol concentration used corresponds to 2 Mol / litre and has been found to confer the smallest damage[168]. The longer frozen samples were stored, the fewer cells could be expected to survive[169]. All in all, freez-

166. Martin Lössner (2001) Personal communication.
167. Colston & Hilson (1979) "The effect of freezing and storage in liquid nitrogen on the viability and growth of Mycobacterium leprae", J Med Microbiol 12:137-142.
168. Fernndez Murga et al. (2000) "Changes in the surface potential of Lactobacillus acidophilus under freeze-thawing stress", Cryobiology 41:10-16.

ing procedures employed in this work were standard microbiological practices.

Mutation accumulation details

To allow long-term mutation accumulation during the stationary phase, plates were wrapped several times in a ca. 30x50 cm piece of Parafilm. This allowed safe storage for many weeks at 37°C (± ca. 1°C) without losing water. Neither regular light nor shaking were employed during that time.

Usually, mutation accumulation experiments start with a single clone that is used to derive a large number of lines, whose subsequent evolution is observed. Many studies define the fitness of the original clone to be exactly 1 with a variance of exactly 0. However, if the first measurements of the distribution of fitness values were made immediately after the lines have been established, inevitable measurement errors would lead to an unchanged mean value with a non-zero variance. This does not influence the slope of the mean regression line, but does decrease the slope of the variance regression line. Thus, the upward biased variance employed usually leads to slight underestimation of U_{min} and to slight overestimation of s_{max}. To avoid this and to catch any potential initial adaptations at the start of the mutation accumulation experiment VTX006, the experiment was started with 4 daily serial transfers that were incubated in the *Bioscreen* C. Thus the variance of the lines at the start could be measured precisely. This more than compensates for the lack of a single-cell bottleneck *immediately* before the start of VTX006; this single-cell bottleneck was generated just a few freeze-thaw events earlier by the researchers who defined the strain.

169. Kim et al. (1998) "Effect of cold shock on protein synthesis and on cryotolerance of cells frozen for long periods in Lactococcus lactis", Cryobiology 37:86-91.

IV. Results Part 1: Experiments

In order to get better estimates of deleterious mutation parameters, a system was developed that allows fast and easy measurements of maximal growth rate in bacteria. This system was used to investigate effects of freezing in glycerol on growth rate and to compare fitness effects of mutations that are neutral from a molecular biology perspective. Finally, a stationary phase mutation accumulation experiment was conducted and analysed with the Bateman-Mukai *technique. The surprisingly high mutation rate is comparable to values known from stationary phase adaptive mutation experiments.*

15 Accuracy of fitness measurements

Accuracy of fitness measurements depends on (i) the number of replicates for a given sample and (ii) the variance within the measurements of the sample. Besides describing the detection limit of this approach, a list of potential problems for high-precision growth rate measurements is given. Finally a number of specific mutations are tested for their neutrality. The result suggests that common techniques in molecular biology have significant impact on fitness from an evolutionary perspective.

15.1 Maximal accuracy

The growth rate is a single number for a given environment. Therefore, the accuracy of a growth rate measurement is determined by the accuracy of the mean of all replicates of the corresponding sample. This accuracy is given by the standard error ***SE***[1]

$$SE = \frac{SD}{\sqrt{N}} \tag{43}$$

where ***SD*** is the standard deviation and **N** is the number of replicates. It says that the true mean will be in the standard error interval around the observed mean in 68% of all cases. For higher confidence levels, 2***SD*** (95.5%) or 3***SD*** (99.7%) should be used instead of ***SD***. Thus, two factors determine the accuracy of a fitness measurement: the number of replicates and the width of the distribution of their values.

In the experiments conducted here, a typical quality distribution usually had standard deviations of about 2%. Thus, 4 replicates result in a standard error of 1%, 10 replicates in a SE of 0.63% and 100 replicates in a SE of 0.2%. Please note that excessive numbers of replicates are needed for moderate increases of accuracy and there will hardly be any situations where one would measure 1000 replicates to archive a SE of 0.063%.

Figure 25 contains the most accurate of all measurements in this study. It shows the doubling time for 99 replicates of *E. coli*, strain 1 measured with the *Bioscreen C* and the Excel analysis software described in this work. Besides the high number of replications, the accuracy is due to the smallest

1. See p. 139 in Sokal & Rohlf (1981) "Biometry: The principles and practice of statistics in biological research. Second Edition", San Francisco, California, W.H. Freeman & Co.

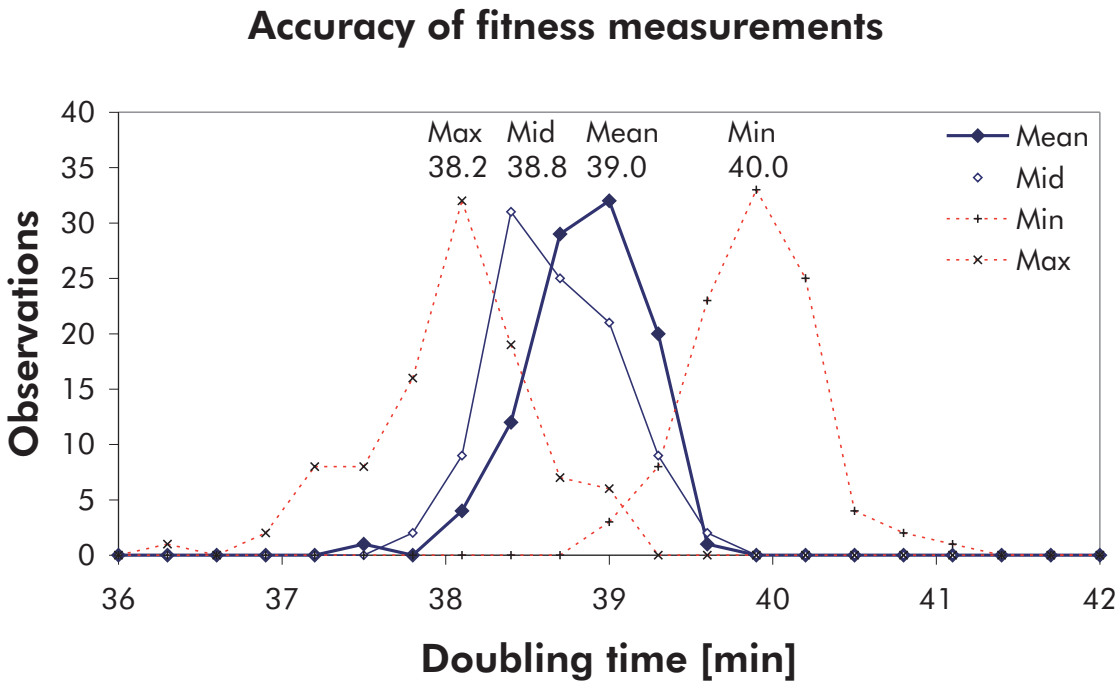


Figure 25 Accuracy of fitness measurements for *E. coli* strain 1 (99 replicates, dilution 1:2304) in experiment VTX006-ST1. The Min/Mid/Max values reflect the raw values as determined by hand in the BioScreen-VBA.xls module. Mean denotes the distribution of means, where one mean refers to the mean of the 3 raw values observed for a particular growth curve. Of all distributions in this figure, "Mean" was used for further analyses. It has μ = 39 min ± σ = 21 sec (0.92%) and a standard error of less than 0.1%.

standard deviation (μ = 39 min ± σ = 21 sec, ie. 0.92%) seen in this study. In this case, the standard error is less than 0.1 %.

Accuracy of other methods

For most measurements of growth rate not much effort is spent to determine their accuracy, so information on detection limits is sparse. Detection limits reported in the literature include 0.92% (1***SE*** for 12 *E. coli* fitness competitions after 15 serial transfers; ***SD***=3.2%)[2], a bit less than 0.8% (1***SE*** for 31 yeast fitness competitions of 'wild type grande' after 36 serial transfers; ***SD***=4.5%)[3] and 0.4% (selection coefficient /generation detection limit under intense competition for growth substrate in chemostat cultures)[4]. A study in *Salmonella typhimurium* reports 0.5% - 1.5% (1 SE for at least 6 rep-

2. Lenski et al. (1991) "Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2000 generations", Am. Nat. 138:1315-1341.
3. Zeyl & DeVisser (2001) "Estimates of the rate and distribution of fitness effects of spontaneous mutation in Saccharomyces cerevisiae", Genetics Jan 157:53-61.

licates of 6 strains, SD = 1.2%-3.7%)[5]. By using cubic and quadratic regressions to derive the steepest slope, others arrived at standard errors of 0.14% - 0.24% (in *E. coli*; SD = 1% - 1.7%)[6], however, without being able to support such accuracy by a few OD values that show actual linear increase on a log scale.

In the light of these values, a remarkable accuracy is achieved when several replicates are measured on the *Bioscreen* C with semi-automated analysis of growth curves. To significantly increase accuracy beyond this level appears to be extremely difficult.

15.2 Potential problems in growth speed measurements

While it is quite simple to determine the approximate growth rate of any given bacterial clone, the following issues can significantly complicate high-precision measurements:

Overview

1. **Different media.** Growth media are an important part of the environment. Different charges might contain slightly different compositions and slight variations of autoclaving conditions might facilitate different chemical reactions. Resulting environmental differences could influence growth rate.
2. **Variable incubation temperature.** Growth rate depends very fundamentally on the temperature of the environment during growth.
3. **Condensed water on plate covers.** If condensate is produced or reduced during the short period, where growth rate is highest, then the values measured lead to biased doubling times. Production of condensate increases apparent growth, while reduction decreases it. Sterile control vessels are needed to exclude such bias.
4. **Dilution of inoculum.** Each growth curve is characterised by the lag-phase (bacteria start to switch to maximal growth speed), the log-phase (maximal growth speed) and the stationary phase (environ-

4. Dean et al. (1988) "Fitness effects of amino acid replacements in the beta- galactosidase of Escherichia coli", Mol. Biol. Evol. 5:469-485. - Hartl (1989) "The physiology of weak selection", Genome 31:183-189. - Hartl & Dykhuizen (1981) "Potential For Selection Among Nearly Neutral Allozymes Of 6- Phosphogluconate Dehydrogenase In Escherichia-Coli", Proc. Natl. Acad. Sci. USA 78:6344-6348.
5. Andersson & Hughes (1996) "Muller's ratchet decreases fitness of a DNA-based microbe", Proc. Natl. Acad. Sci. USA 93:906-907.
6. Kibota & Lynch (1996) "Estimate of the genomic mutation rate deleterious to overall fitness in E. coli", Nature 381:694-696.

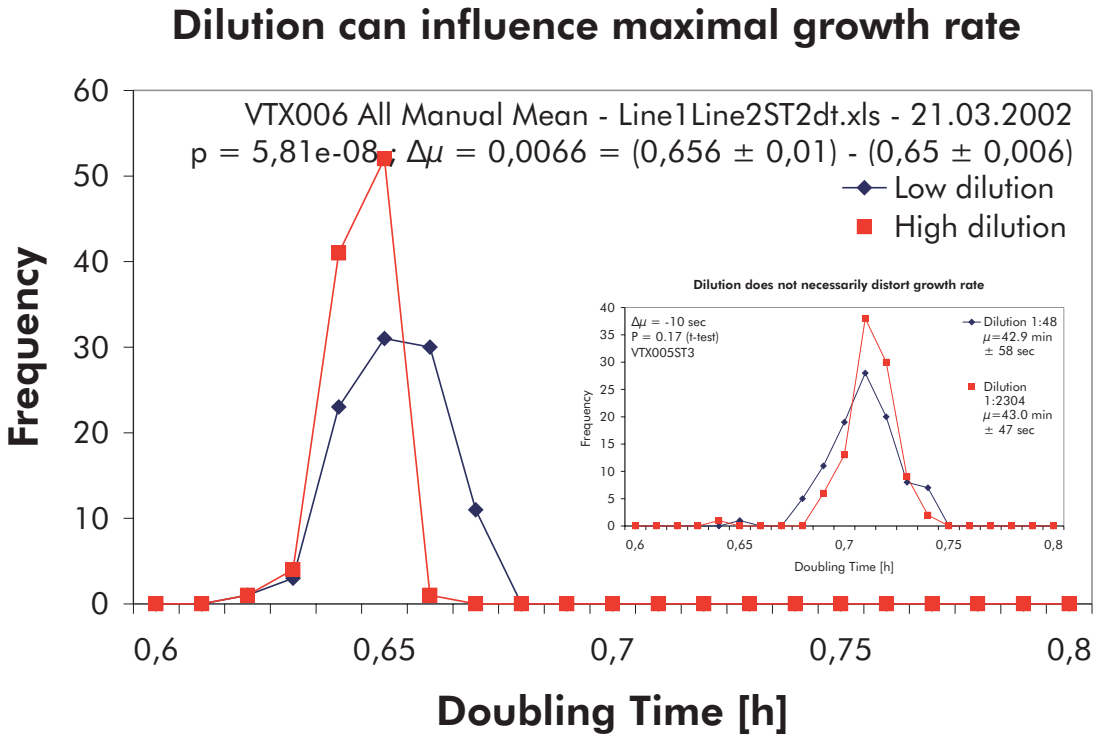


Figure 26 Effect of dilution of inoculum on maximal growth rate. These plots are a direct output of the BatemanMukaiVBA.xls module (except for some reformatting). Low dilution refers to a single 1:48 dilution step ("Line1"), high dilution included 2 such steps (1:2304; "Line2"). Line names reflect numbering in the original experiment to distinguish different dilutions; both such "lines" descended from the original strain 1 (REL606) described in the materials section. $\Delta\mu$ is the difference between the two doubling times (in hours with their standard deviation). The probability that these differences are due to chance has been computed with the t-test. In this case, inoculum dilution differences lead to 1% growth rate differences. The small plot shows, that dilution does not necessarily lead to significant effects on growth rate.

ment becomes overcrowded and growth is slowed down by lack of nutrients and/or increasing concentrations of growth inhibiting waste). Ideally, the log-phase is long enough to allow all cells to switch to maximal growth speed. However, when dilution of inoculum is low, then cells barely start to replicate before they reach the stationary phase. Thus the shallow slope of growth in the lag-phase makes a smooth transition to the shallow slope of growth in the early stages of the stationary phase. Such a growth curve still allows computation of a maximal growth rate, but not of the true one. This might be due to a lack of time (cells need to fully switch to maximal growth speed) or to a high concentration of growth inhibiting waste (transferred

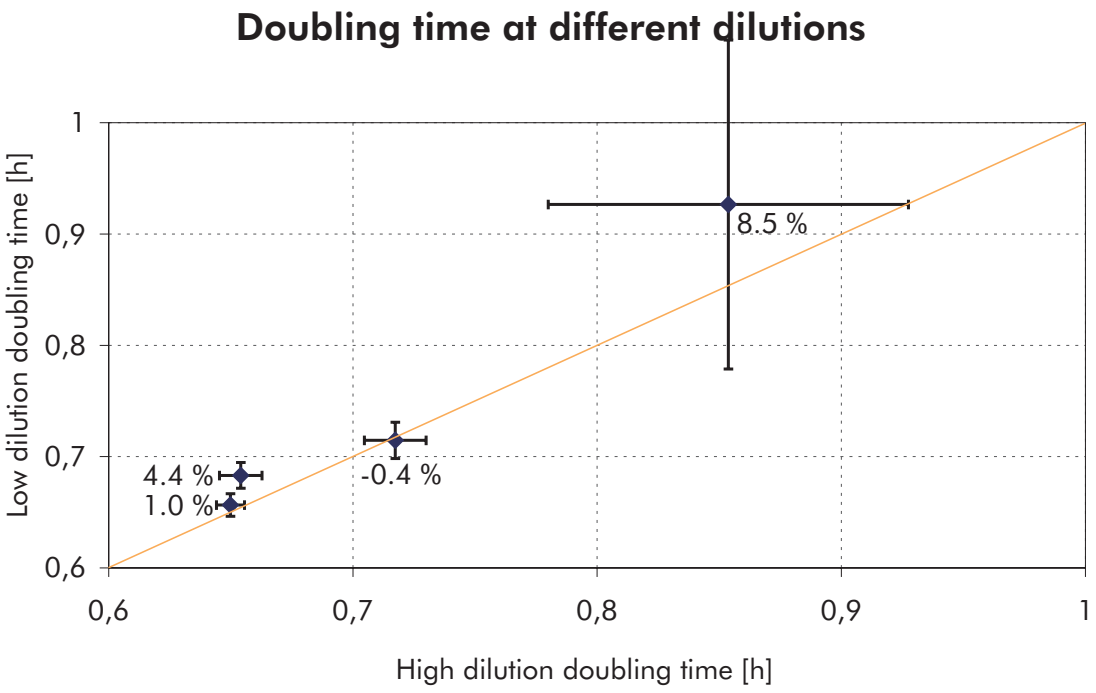


Figure 27 Comparison of the effects of dilution of inoculum for several measurements. Three values shown correspond to the serial transfers ST1 (+1.02%, $p=10^{-7}$), ST2 (+4.42%, $p=10^{-48}$) and ST6 (+8.5%, $p=10^{-5}$) in the stationary phase mutation accumulation experiment VTX006, where ST6 is after 96 days of evolution in the stationary phase. The other value comes from ST4 of VTX005 (-0.40%, p=0.17). Probabilities are computed with the t-test and error bars denote the standard deviation.

with the inoculum from the stationary phase of the parent culture). As can be seen in Figure 26 and Figure 27, a dilution of 1:48 (5.58 generations) and a dilution of 1:2304 (11.17 generations) can indeed lead to slightly different means in a number of cases. While this effect certainly plays no role for dilutions of $1:10^5$ (16.6 generations) as used in some experiments[7], important results in the literature were based on 1:100 dilutions (6.6 generations)[8]. A high concentration of inoculum *per se* does not invalidate growth rate differences observed: If dilution is uniform across wells, the same effect can be expected in

7. Zeyl et al. (2001) "Mutational meltdown in laboratory yeast populations", Evolution 55:909-917.
8. Kibota & Lynch (1996) "Estimate of the genomic mutation rate deleterious to overall fitness in E. coli", Nature 381:694-696. - Lenski & Travisano (1994) "Dynamics of adaptation and diversification: A 10,000-generation experiment with bacterial populations", Proc. Natl. Acad. Sci. USA 91:6808-6814.

each well. However, when dilution of inoculum is low enough to influence growth rate, pipetting problems might become an issue:

Dilution effect

- **Total volume of medium.** Pipetting errors that reduce the growth volume make the stationary phase start earlier.
- **Total volume of inoculum.** Pipetting errors that reduce the volume of inoculum allow for more time to switch to the log-phase leading to higher growth speed. This becomes an issue when inoculum is taken from a long-term stationary phase culture with all kinds of products that change properties relevant to liquid handling.
- **Mixing problems.** If stationary phase cultures are not mixed well enough before inoculum is removed, the number of viable cells transferred can lead to any deviations. This can become a significant problem for spatially structured long-term stationary phase culture that builds clumps of cells that are hard to dissolve.
- **Final concentration in stationary phase.** As inoculum is taken more or less directly from the stationary phase, its actual concentration of cells influences inoculum size. Since the highest concentration of the stationary phase lasts only for a short time, the timing of the transfer of the inoculum can play a role. Deviations in decay of living cell concentration in later stages of the stationary phase lead to variable inoculum sizes.

5. **Temperature of growth medium.** When cells are transferred from a warm stationary phase environment (just taken from the 37°C incubator) to a new well, full of cold medium (just taken from 4°C fridge or 18°C room temperature), they experience a cold shock, which triggers a complex cascade of processes in the cell[9]. This leads to a prolongation of the lag-phase and can decrease the maximal growth rate observed. As cold shocks induce complex changes in the DNA replication complexes, this might even be mutagenic - a topic yet to be explored.
6. **Freeze-thaw effects.** To freeze bacteria for later study of their properties is central to 'experimental palaeontology'[10] and many similar

9. Scherer S & Neuhaus K, (eds, 2002) "Life at low temperatures". The Prokaryotes: An evolving electronic resource for the microbiological community, 3rd Edition. latest update release 3.9 (March 2002), New York.
10. Lenski & Travisano (1994) "Dynamics of adaptation and diversification: A 10,000-generation experiment with bacterial populations", Proc. Natl. Acad. Sci. USA 91:6808-6814.

studies of evolution. Such studies usually assume that freeze-thaw effects have no lasting influence (beyond recovery in 1 or 2 consecutive overnight cultures). However, freezing is a very dramatic event for a bacterial population and only a small fraction survives it[11]. Thus, freezing is a strong selective event, if the frozen population has some variability concerning survival of freezing. It is currently unclear whether freezing can damage DNA *in vivo*. See next chapter.

7. **Adaptive history of the strain.** Bacteria are masters of adaptation. Therefore, changes from one environment to another might easily lead to adaptive evolution, even on timescales as short as 3 days[12]. Thus freezing of bacteria, growing them in 100 ml for two serial transfers to acclimatise them to growth again and finally transfer to a 0.2 ml well of a microtiter plate for the actual measurement can in fact be a complex evolutionary history.
8. **Different microhabitats on a plate.** Each plate has wells at its edges and in its centre, but gas exchange is possible only at the edges. This could lead to different O_2 concentrations that could influence the microenvironment of the various wells. However, such growth patterns could not be detected in this study.
9. **Different cells grow differently.** After all, no cell equals another. Thus some growth rate differences are expected to remain, even if all other details are perfectly equal between measurements.

This complex list shows that high-precision measurements of growth rates are not as straightforward as one might think initially. Please note that this study did not consider measurements that were compromised by one of the problems above, except for those cases where such issues were addressed explicitly (eg. freeze-thaw experiments).

15.3 Neutrality in molecular biology versus neutrality in evolution

Neutrality in molecular biology

A great many studies in molecular biology determine the effects of mutations. However, the methods employed have a rather crude definition of a neutral mutation. Many studies use colony size on specific media to com-

11. Kim et al. (1998) "Effect of cold shock on protein synthesis and on cryotolerance of cells frozen for long periods in Lactococcus lactis", Cryobiology 37:86-91.
12. Rainey & Travisano (1998) "Adaptive radiation in a heteregeneous environment", Nature 394:69-72.

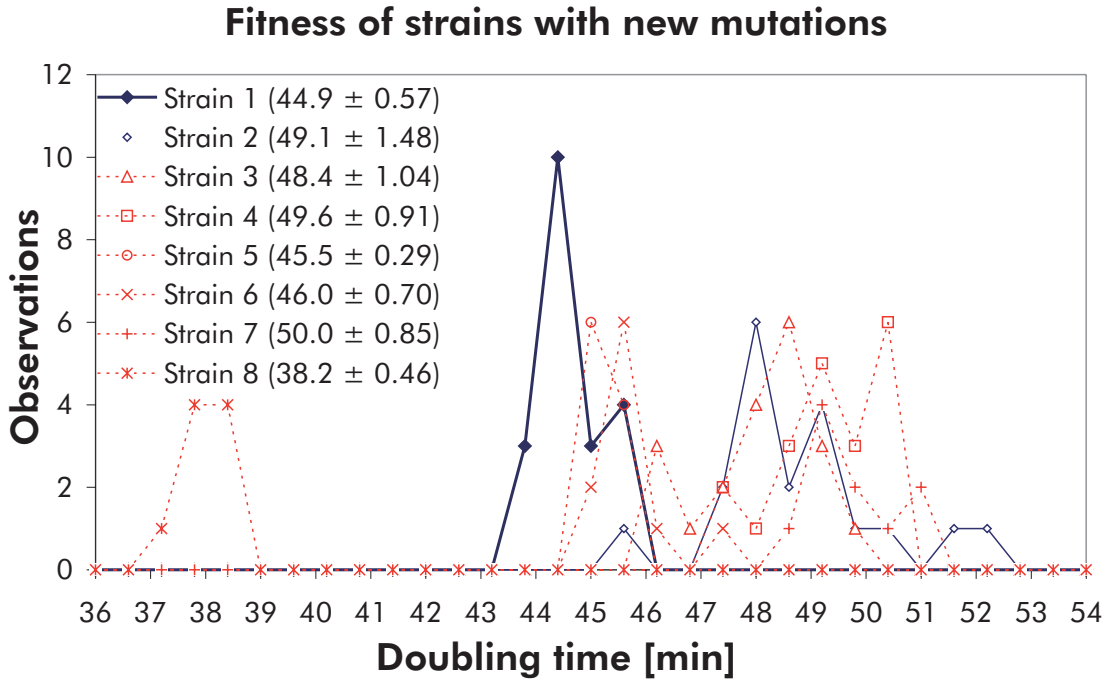


Figure 28 Differences of fitness induced by application of common molecular biological techniques are of considerable interest from an evolutionary perspective.
Doubling time of strains with new mutations in experiment VTX002-ST1. Strain 1 is the original, strain 2 is after 10 000 generation of evolution in glucose minimal and stains 3-8 carry special molecular biological mutations (see Table 14 on page 183). Replicates: 20 (strains 1-4), 10 (strains 5-7) and 9 (strain 8). Comparing strains 2-8 to strain 1 gives -9.4%, -7.8%, -10%, -1,3%, -2,4%, -11% and +15%.

pare growth. Occasionally even growth rates are measured. However, in most cases, more than 80% of the original growth would count as neutral.

Neutrality in evolution

Evolutionary studies are different. Here, a selection coefficient of -0.2 (corresponds to 80% of original growth) is usually considered as seriously deleterious and most interesting mutations have smaller effects. Effective neutrality in evolution means that selection coefficients have to be below $1/N_e$, the effective population size, eg. 99.9999999% of original growth in some cases and *absolute* neutrality implies absolutely no effects. Although such small differences cannot be observed in the laboratory, it is interesting to measure growth rate differences induced by application of common molecular biological techniques. The result can be found in Figure 28, if one is willing to use growth rate as a measure of fitness. It shows that the molecular biological techniques employed in the construction of the strains (antibiotic resistances, transposons, knock out of mutational repair genes, see[13])

have considerable effects on growth rates. This is easy to understand, if one considers that resistances can imply loss of energy due to production of an additional proteins and transposons can hamper non-targeted functions, too. As demonstrated by strain 8, construction of knock-out mutants can also lead to loss of a function that might be important in the wild, but only slows growth under laboratory conditions (assuming that none of the rare gain-of-function mutations had occurred in this case). Thus, before using specific mutants in evolutionary experiments their neutrality should be demonstrated[14].

13. de Visser et al. (1999) "Diminishing returns from mutation supply rate in asexual populations", Science 283:404-406.
14. See test in Lenski (1988) "Experimental studies of pleiotropy and epistasis in Escherichia coli: I. Variation in competitive fitness among mutants resistant to virus T4", Evolution 42:425-432.

16 Freezing can affect growth rate of bacteria

Although it is well known that most individual bacterial cells do not survive freezing in glycerol, it is generally believed that properties of surviving cells are not affected. Here it is shown that freezing in glycerol at -70°C can lead to increase or decrease of doubling time and this can not be avoided just by allowing one or two nights of recovery under assay conditions. Thus, freezing appears to trigger complex evolutionary processes with unpredictable effects. Potential mechanisms are discussed.

Traditionally, it is believed that preservation of cells in glycerol stocks confers no genetic damage[15]. Therefore, this method is widely used to store microbiological samples for later analysis. Its success led to the invention of 'experimental palaeontology', where the course of an evolution experiment can be tracked by the samples frozen at regular intervals[16]. Even mutation-accumulation experiments for measuring mutational parameters have used freezing to preserve their samples for later analysis, although such experiments are very sensitive to any changes in growth rate[17]. Therefore, this study investigated potential effects of the freezing process on maximal growth rate, a frequently used measure of fitness. If freezing has no effect, it would be helpful to know, but if freezing affects growth rate, it becomes crucial to understand the changes induced.

16.1 Experimental design

As can be seen in Figure 29, five freeze-thaw events were studied with the corresponding recovery phases. The first four freeze events were identical (all 1.5 ml vials stored at the same time), except for the duration at -70°C (6, 73, 88 or 94 days). In the fifth event (VTX007 after serial transfer 11), cells

15. Ashwood-Smith (1985) "Genetic damage is not produced by normal cryopreservation procedures involving either glycerol or dimethyl sulfoxide: a cautionary note, however, on possible effects of dimethyl sulfoxide", Cryobiology 22:427-433.
16. Lenski & Travisano (1994) "Dynamics of adaptation and diversification: A 10,000-generation experiment with bacterial populations", Proc. Natl. Acad. Sci. USA 91:6808-6814.
17. Eg. Kibota & Lynch (1996) "Estimate of the genomic mutation rate deleterious to overall fitness in E. coli", Nature 381:694-696. - Zeyl & DeVisser (2001) "Estimates of the rate and distribution of fitness effects of spontaneous mutation in Saccharomyces cerevisiae", Genetics Jan 157:53-61.

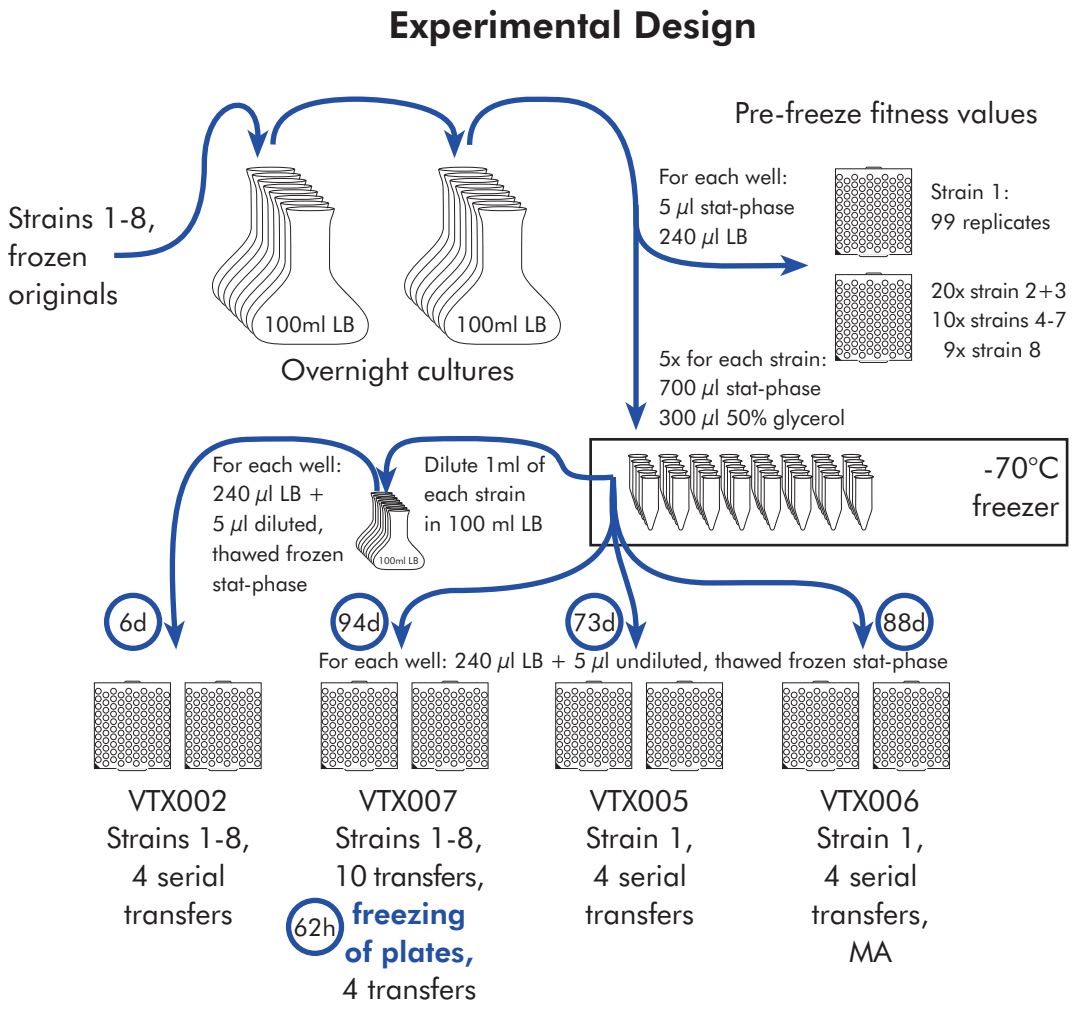


Figure 29 Overview over the experimental design of the various replicated freeze-thaw experiments in this study. Circles denote the time (in days / hours) between freezing to -70°C in 2M glycerol and thawing. The number of serial transfers refers to the number of consecutive one-day growth cycles that allow adaptation of the bacterial population. LB = Luria-Broth growth medium, stat-phase = stationary phase. MA indicates the position where the stationary phase mutation-accumulation experiment was started (see next chapter).

were frozen in the HoneyComb2 plates that had been used to measure growth. Thus the growth statistics reported for this event can be regarded either as nearly 200 freezing events whose effects were measured once or as one freezing event, where many replicates were used to increase measurement accuracy.

Table 15 Freezing can **increase** doubling time on some occasions
Doubling times are given in minutes, ± 2 standard errors with N replicates.

Serial transfer and its meaning	Strain 1 mean±2SE (N)	Strain 2 mean±2SE (N)	Strain 8 mean±2SE (N)
1: Before freezing	45.1 ± 0.17 (99)	49.1 ± 0.68 (19)	38.3 ± 0.31 (9)
2: Directly after freezing	46.6 ± 0.89 (19)	60.0 ± 3.50 (10)	39.1 ± 0.37 (9)
3: 1x overnight recovery	50.9 ± 0.79 (19)	49.7 ± 0.69 (10)	48.3 ± 1.22 (9)
4: 2x overnight recovery	53.2 ± 0.78 (19)	51.1 ± 1.70 (10)	45.0 ± 1.45 (9)
5: 3x overnight recovery	51.8 ± 0.47 (18)	53.2 ± 1.51 (10)	45.6 ± 1.81 (9)

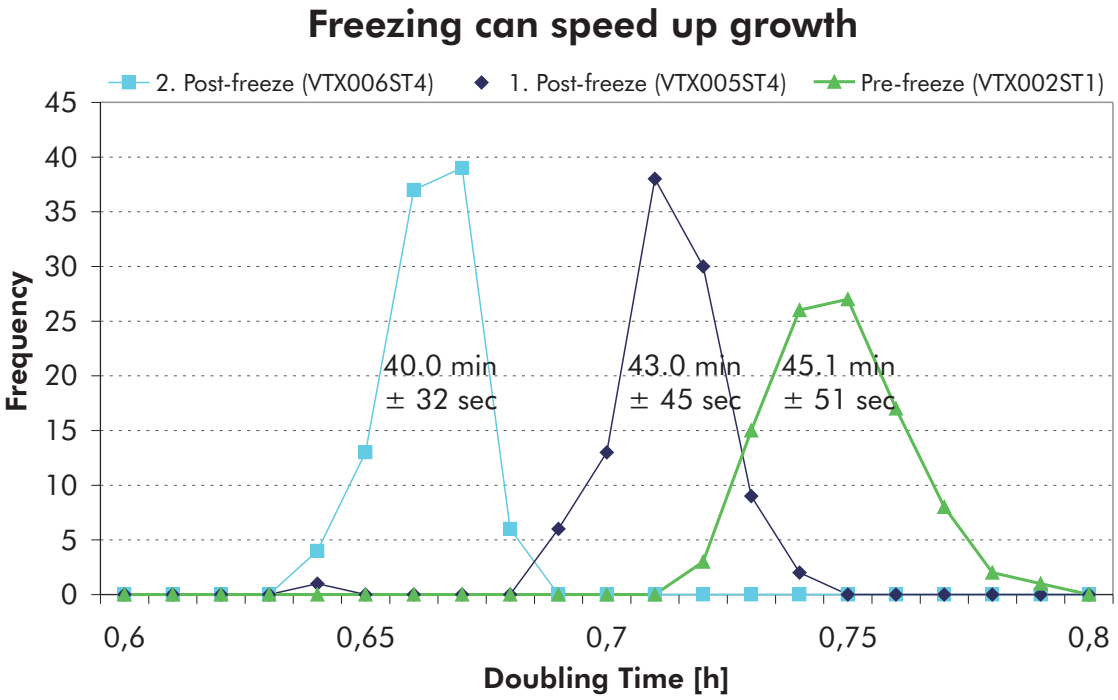


Figure 30 Freezing can **decrease** doubling time to a variable degree on other occasions.
VTX006ST4 and VTX005ST4 measure two independent freeze-thaw events whose original growth rate was measured in VTX002ST1. The differences observed are highly significant and can not be attributed to differences in dilution of inoculum (1:2304 in thawed samples versus 1:48 original), as replicates of VTX005ST4 with 1:48 dilution of inoculum were not significantly different from the higher dilutions. The distributions shown have been measured after recovery for 3 consecutive nights under measurement conditions. However, as VTX006 showed a small decrease in growth rate over that time, freeze effects would have been more pronounced, if measured directly after thawing. The standard deviation is given in seconds.

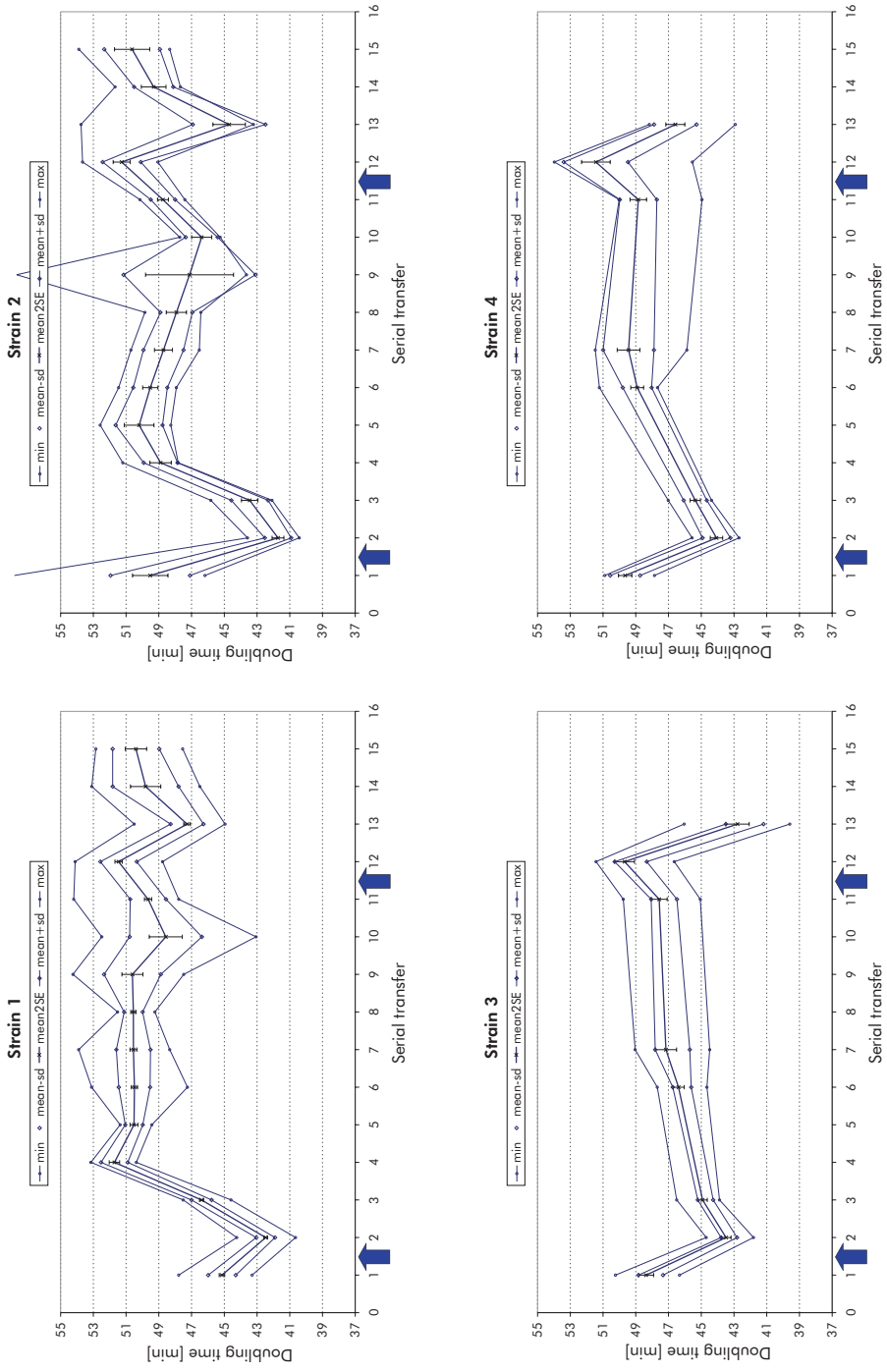


Figure 31a Evolution of doubling time of strain 1-4 in VTX007.

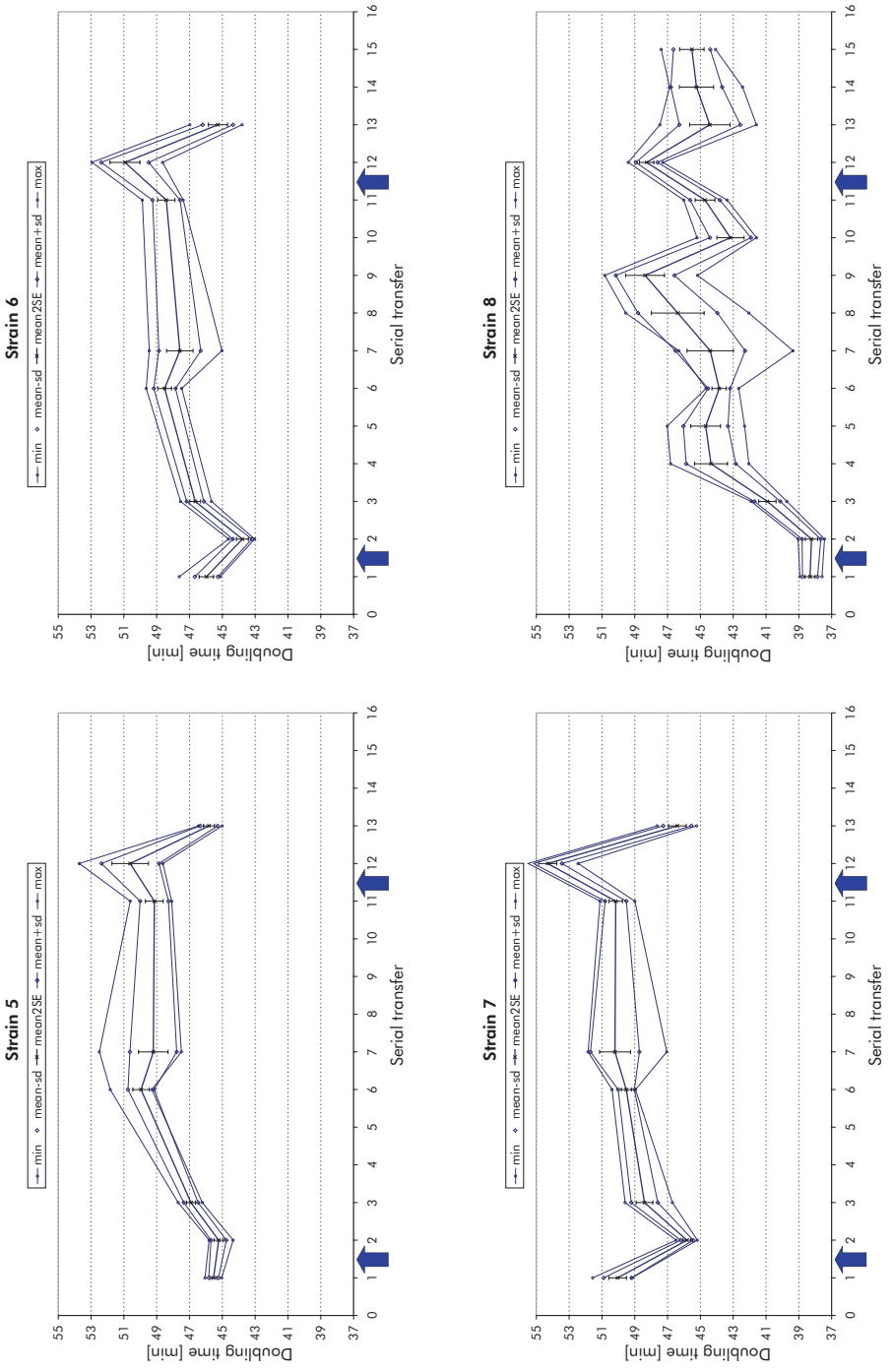


Figure 31b Evolution of doubling time of strain 5-8 in VTX007.

Figure 31 Evolution of doubling time after freezing and thawing is highly unpredictable.
Each plot on the preceding pages shows the evolution of doubling time for a different strain assayed in VTX007. For each serial transfer, the mean (error bars denote 2 standard errors), the standard deviation and corresponding minimal and maximal values are shown. Blue arrows denote freeze-thaw events that lasted for 94 days (between ST1 and ST2) and 62 hours (between ST11 and ST12). Inoculum was diluted 1:48 with fresh medium. Medium was at room temperature during ST1-3+5-6, induced a cold shock ($\Delta T > 20°C$) at the start of ST4 and was preheated to assay temperature in ST7-ST15. Potential effects of cold shock could not be separated from background evolution of doubling time. All values that were blurred by condensed water were not included in these plots.

16.2 Variable effects of freezing can be quite stable

When growth rates before freezing are compared to growth rates after freezing, an amazingly diverse array of behaviours can be observed. In experiment VTX002 a short freezing period appeared to slow down most growth rates, even if recovery for up to 3 consecutive transfers was allowed (Table 15). VTX007 repeated the same experiment with a longer freezing period and led to apparently the opposite result (Figure 31). Even if one sets out to repeat everything with similar durations in the freezer (VTX005 = 73 days vs. VTX006 = 88 days), significantly different growth rates result (Figure 30). While these observations are not incompatible with the hypothesis that short intervals at -70°C lead to increases in doubling time and long intervals in the freezer to decreases in doubling time, they certainly suggest that the individual fate of a bacterial culture during freezing and recovery appears to be highly irregular.

16.3 Potential causes

Several potential causes might explain why changes in growth rate due to freezing and thawing appear to be as poorly predictable as the course of evolution.

16.3.1 Freezing might exhibit strong selective pressure

It is well known that the majority of individual cells in a frozen culture do not survive[18]. Therefore, one would have to assume that survival of freezing is completely independent of any differences between cells, if one wants to

exclude selection. As evidence rather suggests the opposite[19,20], there is a significant probability that the population bottleneck event associated with freezing exerts massive selection pressure.

Selection can influence growth rate in any direction

Generally, there are two ways in which selection can influence growth rate. This depends on what other genes are linked to the selected freeze-survival genes. For example, if survival of freezing depends on fast glycerol import which is accomplished by transporter proteins that can also import nutrients in the log phase, then one might expect that the surviving population would grow faster than before freezing. On the contrary, if survival of freezing depends on higher natural levels of expression of heat shock proteins[19], then growth of most cells after freezing might be slowed down by energy consuming production of heat shock proteins that do not contribute to growth in the log-phase. Similarly, if freeze survival is compromised by an enzyme that degrades the cryoprotectant, a diminished function of this enzyme can lead to better survival[20]. This might result in longer doubling times of survived cells, if this enzyme plays a significant role in growth during the log phase. As cryotolerance and growth rate during the log phase both depend on quite a number of enzymes, many similar pleiotropic interactions might play a role.

Selection needs pre-existing variability

Another issue should be considered here. Selection can only operate on pre-existing variability, because there is no property that could be selected if all bacteria in a population are equal. However, if cells are passed through a single-cell bottleneck, then all descendants are equal apart from the mutations that have been accumulated since that bottleneck. In the present study, the last single-cell bottleneck was some freeze-thaw events ago in the laboratory of those researchers who constructed or defined the corresponding strains. Although the exact prehistory in terms of freezing-events and overnight cultures is unknown, it is hard to imagine that both types combined amount to more than 30 events. Assuming that each of these confers an average mutagenicity like that of 10 standard generations, then the last

18. Fernández Murga et al. (2000) "Changes in the surface potential of Lactobacillus acidophilus under freeze-thawing stress", Cryobiology 41:10-16. - Panoff et al. (2000) "Cryoprotectants lead to phenotypic adaptation to freeze-thaw stress in Lactobacillus delbrueckii ssp. bulgaricus CIP 101027T", Cryobiology 40:264-269. - Colston & Hilson (1979) "The effect of freezing and storage in liquid nitrogen on the viability and growth of Mycobacterium leprae", J Med Microbiol 12:137-142.
19. Chow & Tung (1998) "Overexpression of dnaK/dnaJ and groEL confers freeze tolerance to Escherichia coli", Biochem Biophys Res Commun 253:502-505.
20. Kim et al. (1996) "Disruption of the yeast ATH1 gene confers better survival after dehydration, freezing, and ethanol shock: potential commercial applications", Appl. Environ. Microbiol. 62:1563-1569.

single cell bottleneck is less than 300 generations away. This means that each of the cells in the cultures assayed carries less than one mutation that distinguishes it from its bottleneck ancestor (assuming 1/400 as the genomic mutation rate[21]). This appears to be too little to for the production of the observed variability. However, as discussed below, true mutation rates of freezing might be higher and usual overnight cultures include a stationary phase that might significantly increase mutation rates too (see next chapter). Alternatively, poorly understood cytoplasmatic memory effects might mimic genetic behaviour in the short term; such effects could be produced faster than DNA mutations and might account for much of the variability seen in these experiments.

16.3.2 Freezing induces mutations

As pointed out above, most individual cells do not survive freezing due to the stress associated with this event. It is therefore easy to imagine that freezing profoundly impacts most cells and that this includes their DNA. If part of the individual cell damage is due to DNA mutations, then a considerable number of mutations might impact growth rate after freezing. Mutations in regulatory elements can switch off production of proteins that can be considered as unnecessary baggage for the current environment, leading to an increase of growth rate. Other mutations might impair functionality of vital proteins. The unpredictability of mutational effects might then be reflected in the unpredictability of the effects of freezing. All such mutations disturb evolutionary equilibrium at the optimum for the current environment and therefore trigger evolution in subsequent generations.

Previous studies known to the author are divided in their opinion about the mutagenicity of freezing in glycerol. While some suggest that freezing is mutagentic[22], others show the opposite[23]. A critical issue in all former studies is the lower precision in their measurements of growth rates. Interestingly, freeze-drying was found to induce more mutations than liquid freezing[24], and liquid nitrogen (-196°C) preservation appears to be the least damaging

21. Drake et al. (1998) "Rates of spontaneous mutation", Genetics 148:1667-1686.
22. Calcott & Gargett (1981) "Mutagenicity of Freezing and Thawing", FEMS Microbiology Letters 10:151-155. Calcott & Thomas (1981) "Sensitivity of DNA-Repair Deficient Mutants of Escherichia-Coli to Freezing and Thawing", FEMS Microbiology Letters 12:117-120.
23. Ashwood-Smith (1965) "On the genetic stability of bacteria to freezing and thawing", Cryobiology 2:39-43. - Ashwood-Smith (1985) "Genetic damage is not produced by normal cryopreservation procedures involving either glycerol or dimethyl sulfoxide: a cautionary note, however, on possible effects of dimethyl sulfoxide", Cryobiology 22:427-433.

way of cryopreservation (eg. it did not significantly increase the occurrence of recessive lethal mutations in Drosophila embryos[25]). Since survival of individual cells can depend on the duration of freezing[26], one might conceive that details of the freezing event influence mutagenicity as well. Altogether, mutagenicity of freezing might well contribute to the unpredictability of resulting effects and subsequent evolution.

16.3.3 Stationary phase events

Behaviour of cells in the stationary phase is complex and little understood[29]. However, certain events are known to increase cryotolerance of cells and thus potentially influence the growth rate after thawing by influencing the strength of selection during freezing. Among these are higher expression of heat shock proteins in stress situations and a cold shock prior to freezing[27,26]. In stark contrast to its name the stationary phase of a bacterial culture is not static at all. It rather constitutes an extremely stressful environment where nutrients are low, poison abounds and cells panic to find a way to live. Increased stationary phase mutation rates produce phenomena discussed as 'adaptive mutations'[28]. This phase is best characterised by one mutant after another taking over the majority of the population like one wave after another[29].

Therefore, slight differences might lead to disproportionately large effects on freeze tolerance and subsequently on growth rate. However, in the case of this study the number of variables between different freezing events is quite limited, as all initial freezing events for experiments VTX002 -

24. Ashwood-Smith & Grant (1976) "Mutation induction in bacteria by freeze-drying", Cryobiology 13:206-213. - Tanaka et al. (1979) "Induction of Mutation in Escherichia-Coli by Freeze-Drying", Appl. Environ. Microbiol. 37:369-372.
25. Houle et al. (1997) "The effect of cryopreservation on the lethal mutation rate in Drosophila melanogaster", Genet. Res. 69:209-213.
26. Kim et al. (1998) "Effect of cold shock on protein synthesis and on cryotolerance of cells frozen for long periods in Lactococcus lactis", Cryobiology 37:86-91.
27. Broadbent & Lin (1999) "Effect of heat shock or cold shock treatment on the resistance of lactococcus lactis to freezing and lyophilization", Cryobiology 39:88-102. - Chow & Tung (1998) "Overexpression of dnaK/dnaJ and groEL confers freeze tolerance to Escherichia coli", Biochem Biophys Res Commun 253:502-505. - Baati et al. (2000) "Study of the cryotolerance of Lactobacillus acidophilus: effect of culture and freezing conditions on the viability and cellular protein levels", Int. J. Food. Microbiol. 59:241-247.
28. Foster (1998) "Adaptive mutation: Has the unicorn landed?" Genetics 148:1453-1459. - Rosenberg et al. (1998) "Transient and heritable mutators in adaptive evolution in the lab and in nature", Genetics 148:1559-1566. - Bull et al. (2000) "Evidence that stationary-phase hypermutation in the Escherichia coli chromosome is promoted by recombination", Genetics 154:1427-1437.
29. Finkel et al. (2000) "Long-term survival and evolution in the stationary phase", pp. 231-238 in: Storz & Hengge-Aronis (eds) Bacterial stress responses, Washington, D.C., ASM Press.

VTX007 were started in separated 1.5 ml vials at the same time. While each strain was grown in a different vessel and therefore might be in a different moment of its stationary phase evolution, all vials of a particular strain came from the same stationary phase that had been characterised in VTX002ST1. The only potential differences between these might have arisen in the short time while waiting to be sampled, mixed with glycerol and transferred to -70°C. The resulting slight differences in the exact duration of the cold shock before actual freezing might have generated subcultures that varied in their cryotolerance (and thus in the selection pressure of the freezing event). Potential differences in thawing details are probably much less important[30].

16.3.4 Conclusions

This study shows that freezing of a bacterial culture in glycerol triggers a complex cascade of evolutionary behaviour that is as unpredictable as evolution itself. While evolution is usually associated with extremely long periods of time, this study suggests that evolutionary mechanisms must be considered much more frequently in the daily work of a microbiologist. This conclusion is supported by the work of Rainey & Travisano who observed adaptive radiation in a time period as short as three days[31].

This work is of interest for starter culture production industry[32] and official culture collections that both rely heavily on freezing to maintain identical strains for future reference. However, it should be noted that dynamics of liquid nitrogen freezing and freeze-drying are considerably different from those reported in this study.

Consequences for mutation accumulation experiments that measure mutational parameters are profound, as they assume that fitness of controls and derived generations is not changed by freezing[33]. Thus, future mutation accumulation experimental designs must take this possibility into account and should measure fitness effects directly for each serial transfer without freezing, whenever possible. In analogy to the uncertainty principle of

30. Fonseca et al. (2001) "Operating conditions that affect the resistance of lactic acid bacteria to freezing and frozen storage", Cryobiology 43:189-198.
31. Rainey & Travisano (1998) "Adaptive radiation in a heteregeneous environment", Nature 394:69-72.
32. Grout et al. (1990) "Cryopreservation and the Maintenance of Cell-Lines", Trends Biotechnol. 8:293-297. - Schu & Reith (1995) "Evaluation of Different Preparation Parameters for the Production and Cryopreservation of Seed Cultures with Recombinant Saccharomyces-Cerevisiae", Cryobiology 32:379-388.
33. Eg. Kibota & Lynch (1996) "Estimate of the genomic mutation rate deleterious to overall fitness in E. coli", Nature 381:694-696.

Heisenberg from quantum physics, an **evolutionary uncertainty principle** might state that you cannot observe evolutionary parameters without changing the course of evolution:

Evolutionary uncertainty principle

> *Either you determine evolutionary parameters like mutation rate, selection coefficients etc. with high accuracy at the cost of disturbing the evolutionary process you want to observe, or you do not disturb the original evolutionary process, but have to live with relatively inaccurate estimates of evolutionary parameters.*

17 Deleterious mutation rate and effect in the bacterial stationary phase[34]

Mutation accumulation experiments measure important parameters for prediction of long-term evolution. Here, the first stationary phase mutation accumulation experiment is reported. The deleterious mutation rate (0.07 mutations / genome / day) appears to be surprisingly high, but is comparable to extrapolations from adaptive mutation experiments. The selection coefficient (-3% / mutation) is comparable to that found in other mutation accumulation experiments. Some implications of the findings are discussed.

Mutation rates and selection coefficients are fundamental to any evolutionary theory. Terumi Mukai was the first to determine both mutational parameters for one population[35]. After him, numerous similar experiments have led to considerable insight into spontaneous deleterious mutations[36,37,38]. While most such experiments targeted the fruit fly *Drosophila melanogaster*[39] or the worm *Caenorhabditis elegans*[40], few have studied microorganisms. Besides more general work on RNA-viruses[41], deleterious muta-

34. An improved version of this chapter (including discussion on "How probable are high deleterious stationary phase mutation rates?" on page 320) can be found in L. Loewe, V. Textor & S. Scherer (2003) "High deleterious genomic mutation rate in stationary phase of *Escherichia coli*", Science vol. 302 issue 5650 (28. Nov 2003), pages 1558-1560.
35. Mukai (1964) "The genetic structure of natural populations of Drosophila melanogaster. I. Spontaneous mutation rate of polygenes controlling viability", Genetics 50:1-19.
36. Lynch et al. (1999) "Perspective: Spontaneous deleterious mutation", Evolution 53:645-663.
37. Drake et al. (1998) "Rates of spontaneous mutation", Genetics 148:1667-1686.
38. Keightley & Eyre-Walker (1999) "Terumi Mukai and the riddle of deleterious mutation rates", Genetics Oct 153:515-523.
39. Fry et al. (1999) "New estimates of the rates and effects of mildly deleterious mutation in Drosophila melanogaster", Proc. Natl. Acad. Sci. USA 96:574-579. - Garcia-Dorado (1997) "The rate and effects distribution of viable mutation in Drosophila: Minimum distance estimation", Evolution 51:1130-1139. - Keightley (1994) "The distribution of mutation effects on viability in Drosophila melanogaster", Genetics 138:1315-1322. - Keightley (1996) "Nature of deleterious mutation load in Drosophila", Genetics 144:1993-1999. - Keightley & Ohnishi (1998) "EMS-induced polygenic mutation rates for nine quantitative characters in Drosophila melanogaster", Genetics 148:753-766.
40. Keightley & Caballero (1997) "Genomic mutation rates for lifetime reproductive output with lifespan in Caenorhabditis elegans", Proc. Natl. Acad. Sci. USA 94:3823-3827. - Davies et al. (1999) "High frequency of cryptic deleterious mutations in Caenorhabditis elegans", Science 285:1748-1751. - Keightley & Bataillon (2000) "Multigeneration maximum-likelihood analysis applied to mutation-accumulation experiments in Caenorhabditis elegans", Genetics 154:1193-1201. - Keightley et al. (2000) "Properties of ethylmethane sulfonate-induced mutations affecting life-history traits in Caenorhabditis elegans and inferences about bivariate distributions of mutation effects", Genetics Sep 156:143-154.

tion parameters have been studied in yeast[42], in *Salmonella typhimurium*[43] and in *Escherichia coli*[44]. All these studies have a generation-centric view; they try to estimate mutation rates per generation and neglect the stationary phase. This is appropriate for investigating the log phase.

However, normal bacterial populations spend most of their time under extreme nutritional constraint[45]. As fresh nutrients are consumed much faster than they show up, the situation of most bacteria is dramatic. To evade their potential death by starvation, a subpopulation turns into transient hypermutators and searches the immediate neighbourhood on the adaptive landscape for an adaptive mutation that allows it to start growth again[46]. These dynamics lead to complex evolutionary processes during stationary phase[47,48], where recombination and SOS induced DNA polymerases appear to play a key role[49]. Cells in stationary phase might divide 1 to 2 times a day[50] or less than once every 3 days[51], but to accumulate muta-

41. See eg. Burch & Chao (1999) "Evolution by small steps and rugged landscapes in the RNA virus phi 6", Genetics Mar 151:921-927. - Burch & Chao (2000) "Evolvability of an RNA virus is determined by its mutational neighbourhood", Nature 406:625-628. - Chao (1990) "Fitness of RNA virus decreased by Muller´s ratchet", Nature 348:454-455. - Novella et al. (1999) "Exponential fitness gains of RNA virus populations are limited by bottleneck effects", Virology 73:1668-1671. - Domingo & Holland (1997) "RNA virus mutations and fitness for survival", Annu. Rev. Microbiol. 51:151-178. - Domingo (2000) "Viruses at the edge of adaptation", Virology 270:251-253.
42. Zeyl & DeVisser (2001) "Estimates of the rate and distribution of fitness effects of spontaneous mutation in Saccharomyces cerevisiae", Genetics Jan 157:53-61. - Szafraniec et al. (2001) "Environmental stress and mutational load in diploid strains of the yeast Saccharomyces cerevisiae", Proc. Natl. Acad. Sci. USA 98:1107-1112.
43. Andersson & Hughes (1996) "Muller's ratchet decreases fitness of a DNA-based microbe", Proc. Natl. Acad. Sci. USA 93:906-907.
44. Kibota & Lynch (1996) "Estimate of the genomic mutation rate deleterious to overall fitness in E. coli", Nature 381:694-696. - See also Figure 4 in Cooper & Lenski (2000 "The population genetics of ecological specialization in evolving Escherichia coli populations", Nature 407:736-739), which allows a Bateman-Mukai analysis, although their authors do not discuss that. - Funchain et al. (2000) "The consequences of growth of a mutator strain of Escherichia coli as measured by loss of function among multiple gene targets and loss of fitness", Genetics 154:959-970.
45. Ozawa & Freter (1964) "Ecological mechanism controlling growth of Escherichia coli in continuous flow cultures and in the mouse intestine", J. Infect. Dis. 114:235-242.
46. Bull et al. (2000) "Evidence that stationary-phase hypermutation in the Escherichia coli chromosome is promoted by recombination", Genetics 154:1427-1437. - Rosche & Foster (1999) "The role of transient hypermutators in adaptive mutation in Escherichia coli", Proc. Natl. Acad. Sci. USA 96:6862-6867. - Torkelson et al. (1997) "Genome-wide hypermutation in a subpopulation of stationary-phase cells underlies recombination-dependent adaptive mutation", EMBO J. 16:3303-3311.
47. Finkel et al. (2000) "Long-term survival and evolution in the stationary phase", pp. 231-238 in: Storz & Hengge-Aronis (eds) Bacterial stress responses, Washington, D.C., ASM Press.
48. Finkel & Kolter (1999) "Evolution of microbial diversity during prolonged starvation", Proc. Natl. Acad. Sci. USA 96:4023-4027.
49. Yeiser et al. (2002) "SOS-induced DNA polymerases enhance long-term survival and evolutionary fitness", Proc. Natl. Acad. Sci. USA 99:8737-8741. - For recombination see Bull et al. (2000) ibid.
50. Matin et al. (1989) "Genetic basis of starvation survival in nondifferentiating bacteria", Annu. Rev. Microbiol. 43:293-316. - Finkel et al. (2000) ibid.
51. Ochman et al. (1999) "Calibrating bacterial evolution", Proc. Natl. Acad. Sci. USA 96:12638-12643.

Experimental Design

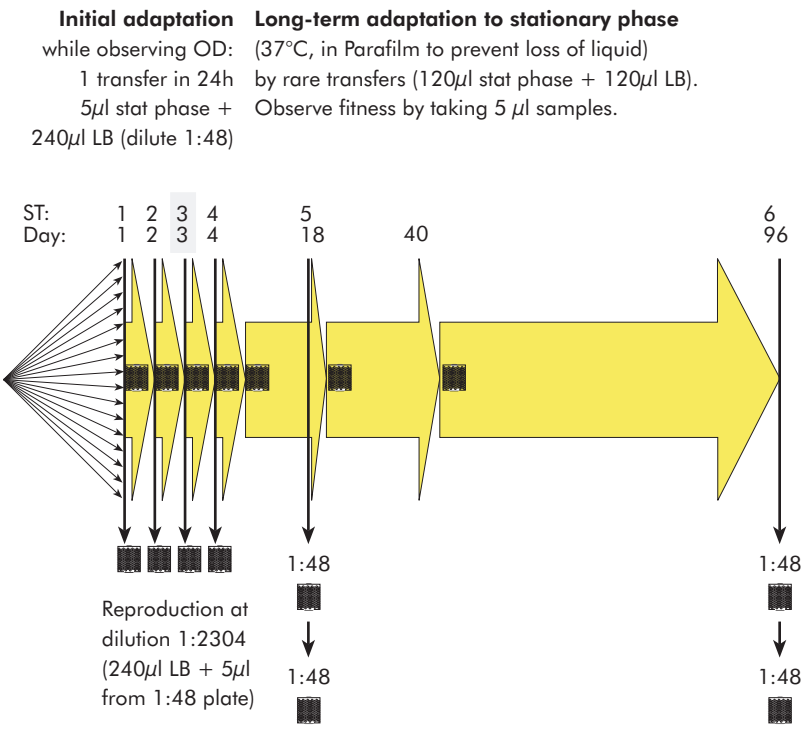


Figure 32 Experimental design of the stationary phase mutation accumulation experiment.
The first 4 serial transfers (ST1-4) were designed to allow regeneration from freezing and to measure the precise initial condition of the 99 lines that were derived from a frozen culture of strain 1. (The same strain had also been used by Richard Lenksi's famous evolution experiments and by Kibota & Lynch's bacterial mutation accumulation experiment). For the longest time after ST4 plates were incubated without shaking (except pipette mixing before the next ST). The blue shade indicates a cold shock (ΔT > 20°C; medium not preheated).

tions they neither need to divide[52] nor do they need to replicate their DNA[53].

Given the importance of stationary phase evolution in nature, it would be interesting, to determine the deleterious mutation rate during the sta-

52. Ryan et al. (1963) "Spontaneous mutation in spheroplasts of Escherichia coli", J Gen Microbiol 30:193-199.
53. Ryan et al. (1961) "Is DNA replication a necessary condition for spontaneous mutation?" Z. Vererbungsl. 92:38-41.

tionary phase. Here, I report the first BATEMAN-MUKAI analysis of a stationary phase mutation accumulation experiment.

17.1 Experimental design

As the stationary phase does not allow one to determine the number of generations with reasonable efforts, the generation-centric view of mutation rates has to be abandoned in favour of a clock-like view. Thus, mutation rates are measured per genome per day. Figure 32 reviews the design of the 96 day mutation accumulation experiment VTX006 that had been started by VOLKER TEXTOR. Fitness is measured in terms of growth rate (Wrightian fitness per hour). To avoid unpredictable skew of growth rates by freezing, all samples were immediately analysed with the *Bioscreen* C system using the same production batch of growth medium for all plates.

17.2 Results

Measurements of the resulting distributions of growth rate showed a decrease of mean and an increase of variance over time (Figure 33), leading to a typical BATEMAN-MUKAI plot (Figure 34). The measurements with most high-quality data (high-dilution series) lead to values of U_{min} = 0.0778 / genome / day and s_{max} = -0.0275 / mutation. If the same calculation is done for those measurements where inoculum had been diluted only 1:48, then U_{min} = 0.060 and s_{max} = -0.046 (after removing measurements that had been blurred by condensed water). Thus, the values to remember are about

$U_{min} \approx 0.07$ / genome / day

$s_{max} \approx -3\%$ / mutation.

The observed quotient of U/s ranges from 2 to 5 and if one wants to estimate long-term consequences of such deleterious mutation rates, one might use simulations from Project 1 of Simulator005 with values of $0.05/-0.01 < U_{min}/s_{max} < 0.1/0.05$.

17.3 Related work

Selection coefficients are as expected

The selection coefficients found here are in general agreement with $s_{max} \approx -1.2\%$ from the *E. coli* log-phase mutation accumulation experiment of KIBOTA & LYNCH[54] and with s_{max} inferred from the long-term serial transfer data reported by COOPER & LENSKI[55]. A similar experiment in *Salmonella* with a much cruder screen for growth rate differences found 5 mutants

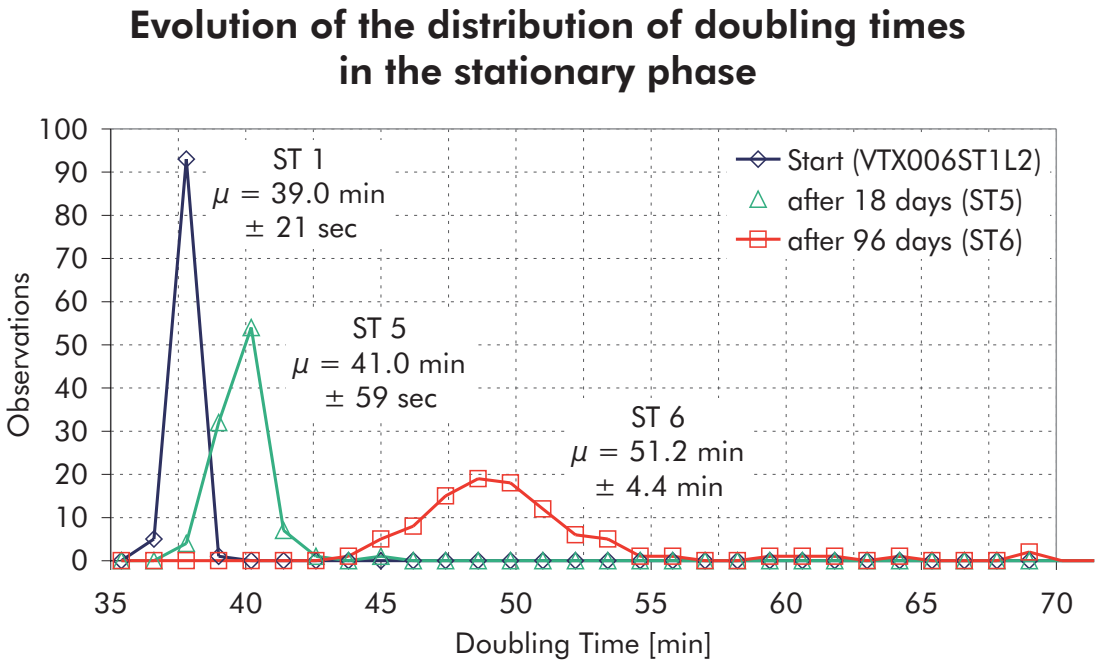


Figure 33 Evolution of the distribution of doubling times in the stationary phase.
Each distribution (mean ± standard deviation) contains one replicate measurement of each of 99 lines for the corresponding serial transfer (ST). Dilution of inoculum was 1:2304. The bins of this histogram are 1.2 min wide and centred around their middle values.

among 444 lines (after 60 growth cycles with 1700 generations) with an average mutant effect of 33%[56]. However, the true average selection coefficient is much smaller, because non-mutated lines have also to be considered and slight mutational effects probably went unnoticed by the experimental design. A hint at the mechanism of the mutations comes from an artificial transposon mutagenesis study that measured random mutation effects[57]. It

54. Kibota & Lynch (1996) "Estimate of the genomic mutation rate deleterious to overall fitness in E. coli", Nature 381:694-696.
55. If a Bateman-Mukai analysis is conducted for values estimated from their Figure 4, then smax is at least -3.3% (if measured between generations 2000 and 10000), more probably -6.1% (generation 0 - 20000) and at most -12% (generation 0-2000), depending on the choice of the timeframe for such an analysis. This assumes a standarddeviation of mean fitness at generation 0 of ± 0.08 (estimated by eye from the same figure). See Cooper & Lenski (2000) "The population genetics of ecological specialization in evolving Escherichia coli populations", Nature 407:736-739.
56. As infered from Andersson & Hughes (1996) "Muller's ratchet decreases fitness of a DNA-based microbe", Proc. Natl. Acad. Sci. USA 93:906-907.
57. Elena & Lenski (1997) "Test of synergistic interactions among deleterious mutations in bacteria", Nature 390:395-398.

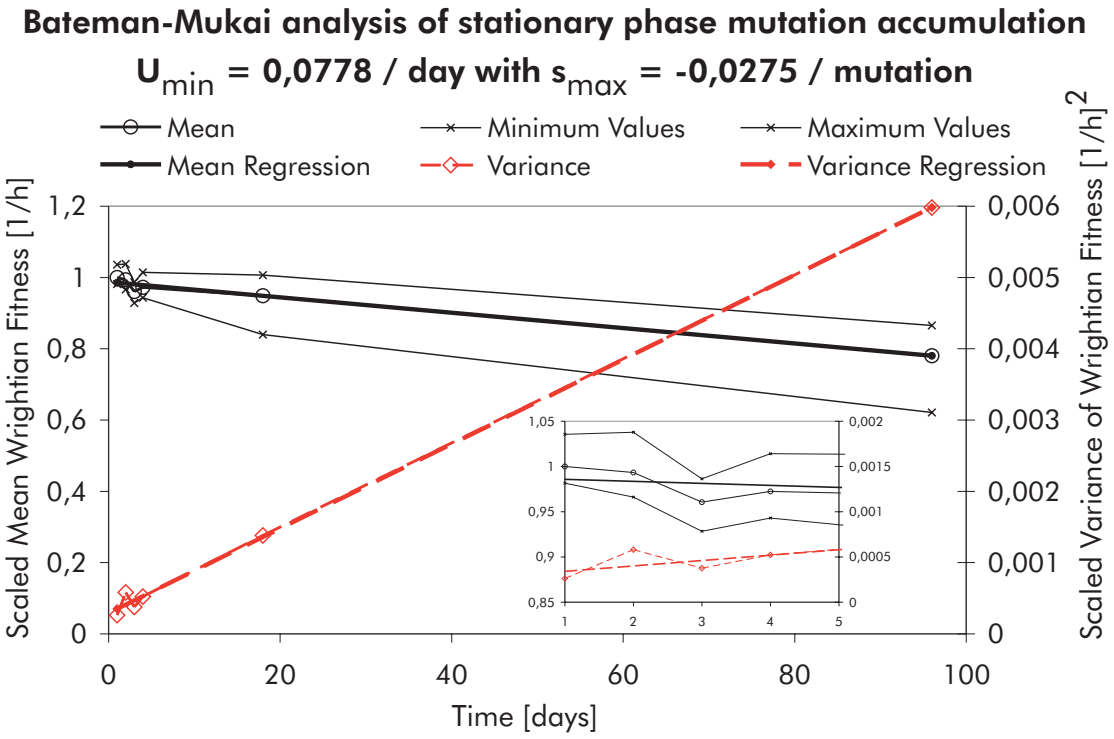


Figure 34 Bateman-Mukai analysis of stationary phase mutation accumulation experiment VTX006. Please note the astonishing overlap between regression lines and individual mean and variance values. The small plot shows a magnified version of the first 4 serial transfers. Fitness is measured in terms of Wrightian fitness (per hour) and actual values are used for the initial distribution of fitness (no artificial mean of 1 and variance of 0). This plot is from the growth measurements made at high dilution.

found average effects of -2.7% per transposon insertion. As such an insertion is likely to disrupt gene function, the mechanism of mutations in stationary phase as observed here probably includes many frameshifting insertions and deletions, as these are much more likely to disrupt gene function than simple point mutations.

Replication based explanations fail

To understand the mutation rate found here is more challenging. Any purely replication-based calculation fails completely: Assuming 2 replications / day in stationary phase[50] and $\mathbf{U}_{total}$ = 0.0025 total mutations / genome / generation[58] results in less than a tenth of the *deleterious* mutation rate found here. The same is true for comparisons with $\mathbf{U}_{min}$ from log-phase based mutation accumulation experiments in *E. coli* ($\mathbf{U}_{min}$ = 0.0002 / ge-

58. Drake et al. (1998) "Rates of spontaneous mutation", Genetics 148:1667-1686.

nome / generation = 0.005 per 25-generation-day)[54]. One would have to assume constant log-phase growth and a mutator phenotype[59] to arrive at a value like the one observed - *per se* not very plausible for a non-mutator strain in stationary phase.

Adaptive mutation experiments help

The riddle is solved when one considers evidence from adaptive mutation experiments that investigate mutations that occur in the stationary phase. It is now well known that a subpopulation of all cells in stationary phase can turn into transient hypermutators[60]. These could produce mutations that confer a Growth Advantage in Stationary Phase (GASP-phenotype)[61]. New GASP mutations repeatedly rise to dominant frequencies in the whole population, only to be overtaken by the next GASP mutation. Therefore, the mutation rates found in the stationary phase of adaptive mutation experiments might apply very well to the experiment conducted here.

Details

Bull and others[62] use the data of Torkelson and others[63] to estimate a mutation rate of 0.007 mutations / 5 mutational targets / hypermutating cell / 4 days starvation at 37°C[64]. This is equivalent to 0.00035 mutations / day / mutational target. If one assumes that the genome contains about 200 hot mutational targets of similar complexity, then an overall deleterious mutation rate of 0.07 / genome / day results. Such a number is easily conceivable. Please note that these mutation rates are not from long-term mutators, but occur in a small fraction of any stationary phase culture that is starved to death. Currently, it is not clear whether this fraction consists of one[65] or several[66] subpopulations. In any case those cells that survive the longest in the experiment discussed here do so because their high mutation rates repeatedly generate genotypes that confer a Growth Advantage in the

59. Sniegowski et al. (1997) "Evolution of high mutation rates in experimental populations of E. coli", Nature 387:703-705. - Funchain et al. (2000) "The consequences of growth of a mutator strain of Escherichia coli as measured by loss of function among multiple gene targets and loss of fitness", Genetics 154:959-970.
60. Rosenberg et al. (1998) "Transient and heritable mutators in adaptive evolution in the lab and in nature", Genetics 148:1559-1566.
61. Finkel et al. (2000) "Long-term survival and evolution in the stationary phase", pp. 231-238 in: Storz & Hengge-Aronis (eds) Bacterial stress responses, Washington, D.C., ASM Press.
62. Bull et al. (2000) "Response to John Cairns: The contribution of transiently hypermutable cells to mutation in stationary phase", Genetics Oct 156:925-926.
63. Torkelson et al. (1997) "Genome-wide hypermutation in a subpopulation of stationary-phase cells underlies recombination-dependent adaptive mutation", EMBO J. 16:3303-3311.
64. The targets of Torkelson et al. (1997) ibid. were multigenic regulons that control the ability to ferment the sugars (1) xylose, (2) maltose and (3) fructose. The other two targets were (4) the *upp* gene, which confers resistance to 5-fluorouracil and (5) a target that controls temperature sensitivity for growth on minimal medium. All these targets except fructose fermentation showed increased mutation rates. This confirms that genomes might contain hot and cold sites as suggested by Rosenberg (1997) "Mutation for survival", Curr. Opin. Genet. Dev. 7:829-834.

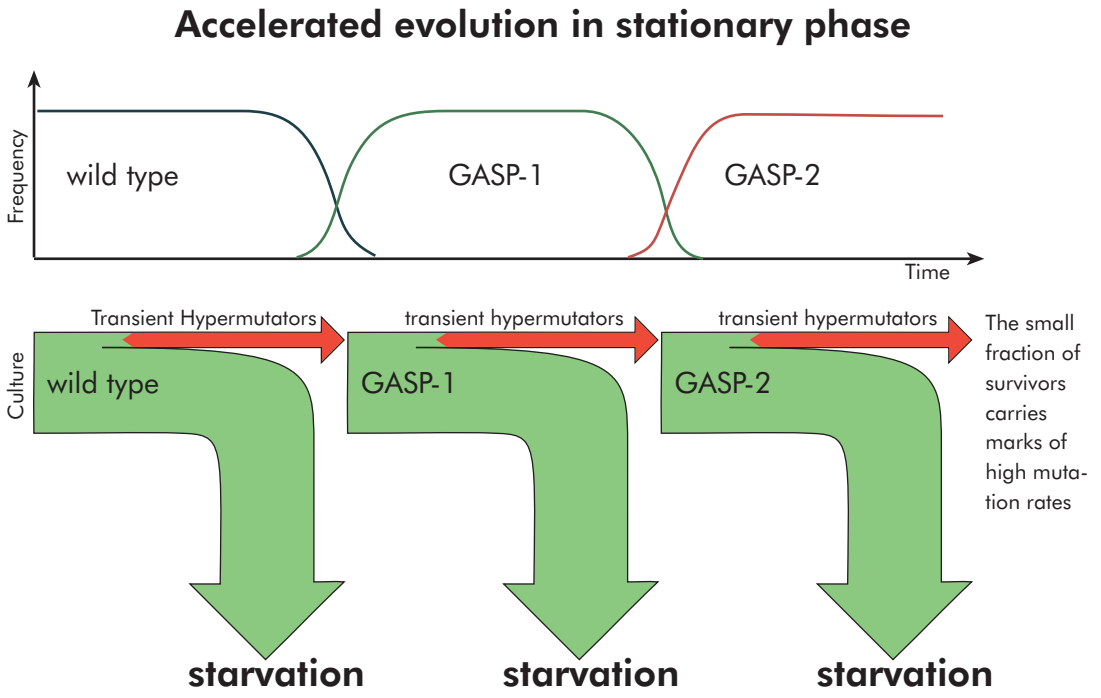


Figure 35 Schematic view of accelerated evolution in the stationary phase.
In each culture a small fraction of transient hypermutators search for new mutations that confer a Growth Advantage in Stationary Phase (GASP). Once such a mutation is found, the resulting genotype spreads over the whole population (see rise in frequency in the upper part) and dominates until GASP-2 is found. These repeated population take-overs lead to a relatively small effective population size and to elevated mutation rates in the surviving line. As stationary phase mutations are known to stem from transient hypermutators, the whole process does not necessarily require permanent mutators.

Stationary Phase (GASP)[67]. As all other cells ultimately starve, effective population size in the stationary phase might well be much smaller than one would think initially. Figure 35 presents an overview over this model of ac-

65. Bull et al. (2000) "Evidence that stationary-phase hypermutation in the Escherichia coli chromosome is promoted by recombination", Genetics Apr 154:1427-1437. - Bull et al. (2000) "Response to John Cairns: The contribution of transiently hypermutable cells to mutation in stationary phase", Genetics Oct 156:925-926. - Torkelson et al. (1997) "Genome-wide hypermutation in a subpopulation of stationary-phase cells underlies recombination-dependent adaptive mutation", EMBO J. 16:3303-3311.
66. Rosche & Foster (1999) "The role of transient hypermutators in adaptive mutation in Escherichia coli", Proc. Natl. Acad. Sci. USA 96:6862-6867. - Cairns (2000) "The contribution of bacterial hypermutators to mutation in stationary phase", Genetics 156:923-923.
67. Finkel et al. (2000) "Long-term survival and evolution in the stationary phase", pp. 231-238 in: Storz & Hengge-Aronis (eds) Bacterial stress responses, Washington, D.C., ASM Press.

celerated evolution in the stationary phase. In the light of this evidence, it appears that more mutations come from highly mutagenic attempts to find new nutrients during starvation in the stationary phase than from replication errors in the log-phase.

How natural are these conditions?

While it is likely that most bacteria in nature are in the stationary phase for most of their time, environmental temperature might have a significant influence on mutation rates. It is easily conceivable that lower temperatures confer lower mutation rates per day, as the overall speed of metabolism is slowed down. If a typical Q_{10} of 2 is assumed[68], then genomic rates of

- 37°C lead to $\boldsymbol{U_{relative}} = 1$ and $\boldsymbol{U_{absolute}} \approx 0.007$/day
- 27°C lead to $\boldsymbol{U_{relative}} \approx 0.5$ and $\boldsymbol{U_{absolute}} \approx 0.0035$/day
- 17°C lead to $\boldsymbol{U_{relative}} \approx 0.25$ and $\boldsymbol{U_{absolute}} \approx 0.00175$/day
- 7°C lead to $\boldsymbol{U_{relative}} \approx 0.125$ and $\boldsymbol{U_{absolute}} \approx 0.0009$/day

are predicted. Still lower temperatures are expected to continue this trend until the cell is frozen (see previous chapter for potential evolutionary processes during freezing). Further experiments will have to test this hypothesis, but the results presented here suggest that mutation rates in nature might be significantly higher than previously thought.

Additional insight might come from comparisons of sequenced genomes of closely related bacteria (*Escherichia coli* K-12[69] and O157:H7[70] *Salmonella enterica* Serovar Typhi[71] and Typhimurium[72]). Such comparisons of genome organisation show a surprising number of smaller differences and reorganisations, while overall 'back-bone' structures retain similarity[73]. However, as the evolutionary history of these sequenced strains is unknown in detail, the ideal experiment would evolve a strain under known stationary phase conditions and then compare genomes of descendants with the original. Unfortunately, high-throughput sequencing technology has to become several orders of magnitude faster before such a study becomes feasible.

68. See p. 378 in Adam et al. (1988) "Physikalische Chemie und Biophysik". Zweite, völlig neu bearbeitete und erweiterte Auflage, Berlin, Springer-Verlag.
69. Blattner et al. (1997) "The complete genome sequence of Escherichia coli K-12", Science 277:1453-1462.
70. Perna et al. (2001) "Genome sequence of enterohaemorrhagic Escherichia coli O157:H7", Nature 409:529-533.
71. Parkhill et al. (2001) "Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18", Nature 413:848-852.
72. McClelland et al. (2001) "Complete genome sequence of Salmonella enterica serovar Typhimurium LT2", Nature 413:852-856.
73. McClelland et al. (2000) "Comparison of the Escherichia coli K-12 genome with sampled genomes of a Klebsiella pneumoniae and three salmonella enterica serovars, Typhimurium, Typhi and Paratyphi", Nucleic Acids Res. 28:4974-4986.

17.4 Mutation rate paradox for bacteria

The mere fact that mutation rates measured here are measured in mutations per genome *per day* suggests comparisons with other known molecular clocks. Ochman and others[74] estimated a substitution rate of 4.5×10^{-9} / site / year for synonymous sites in *E. coli* based on a comparison with *Salmonella* and a divergence date of 100 Myr from host fossil record. As synonymous sites are not selected against, this rate should reflect the total mutation rate. If scaled by genome size (4.7×10^{6} bp), a mutation rate of 5.8×10^{-5} / genome / day results. According to Eyre-Walker 58.4% of these are deleterious (87.8% protein sequences per genome x 95% of amino acid changes are deleterious in nature x 70% of nucleotides change an amino acid)[75]. Thus the predicted deleterious mutation rate from this approach is 3.4×10^{-5} / genome / day. The deleterious mutation rate observed here is about 2000 times higher.

However, the mutation rate of this study is not the first to be in conflict with phylogenetic estimates. The original study of Ochman[74] and others reports a 30-fold difference between the phylogenetic estimate and laboratory experimental replication based mutation rates as reviewed by Drake[76]. This suggests that the mutation rate paradox is more widely spread than one might initially expect (see Chapter 3 for potential solutions and examples in mtDNA).

74. Ochman et al. (1999) "Calibrating bacterial evolution", Proc. Natl. Acad. Sci. USA 96:12638-12643.
75. Keightley & Eyre-Walker (1999) "Terumi Mukai and the riddle of deleterious mutation rates", Genetics 153:515-523.
76. Drake et al. (1998) "Rates of spontaneous mutation", Genetics 148:1667-1686.

V. Results part 2: Extensions of Muller's ratchet theory

This work has led to new observations and insights that extend understanding of traditional models of Muller's ratchet. Its database of currently >28000 simulation results with >16 years CPU time is expected to be fruitful for further research on models of Muller's ratchet.

18 A simple heuristic equation for predicting the rate of Muller's ratchet

Predictions of the rate of Muller's ratchet usually involve mathematical theory and equations beyond the reach of most biologists. The simple equations presented here allow average biologists to estimate orders of magnitude for the rate of the ratchet with their pocket calculator. Thus it becomes easy to check whether the ratchet might play a significant role in a given situation and further investigations have merit. These equations were developed to predict computational complexity of simulations with Simulator005.

Analytical predictions are tricky

Theoretical efforts over more than 3 decades have led to analytical formulae that allow reasonably accurate predictions of the rate of Muller's ratchet, at least in its most simple case. However, papers that investigate biological systems where Muller's ratchet might actually be a problem often address the issue without computing potential extinction times. This might be in part due to the fact that next to nothing is known about the distribution of mutational effects in the critical range and thus any such results have to be treated with caution. However, probably more important are the mathematical skills required to use the sophisticated analytical equations correctly or the computational skills to write special simulation code. While simulations are known to be computationally intense, extensive investigation of parameter space with analytical approaches in *Mathematica* is not free of problems either. One does easily stumble across parameter combinations that generate obscure numerical errors or even apparent internal infinite loops. Thus 'easy' production of 2D or 3D plots like those on the following pages can lead to poorly predictable computing times in the range of hours or days respectively.

The need for a simple equation

Therefore, a rough but simple approximation of the clicktime of Muller's ratchet would be of great help. It could easily be employed to check, which ranges of parameter space merit more detailed investigation of potential extinction times.

Simulator005 needed such a system to predict computing time for simulations, as one of the conditions that would terminate a run is the observation of a given number of clicks of the ratchet (eg. 500 in Project 1). To develop such a system by trial and error, more than 170 intuitively proposed

equations were used to predict clicktimes known from simulations. The following *Equation150* and *Equation172* showed the best results.

18.1 The *Equation150* system

The name '*Equation150*' reflects the fact that this equation is likely to be improved in the future (see *Equation172* below). It is heuristic, since there is no analytic proof for it, although some parts have an analytical meaning (see next section). To compute the rate of the ratchet use

$$T_{cl150} = 1/U, \qquad \text{if } N_0 < 10^{-99} \tag{44}$$

$$T_{cl150} = \frac{e + \pi}{U - U \cdot e^{\left(\frac{N_0}{N_e} - 1\right)/(U \cdot N_0)}} \quad \text{in all other cases} \tag{45}$$

$$T_{cl150} = 10^{15}, \qquad \text{if } T_{cl150} > 100/U \tag{46}$$

where T_{cl150} is the clicktime in generations given by the *Equation150* system, U is the intergenerational genomic deleterious mutation rate, s is the negative constant selection coefficient for all deleterious mutations, N_e is the effective population size and N_0 is the equilibrium size of the best mutational class given by $N_0 = N_e \cdot e^{U/s}$. Please note that in contrast to most ratchet theory papers, s of deleterious mutations is negative here. Equation (44) gives a simple expectation for the case where the ratchet clicks excessively fast. Equation (46) gives a dummy result to indicate that purifying selection removes virtually all mutations in this case. Omitting Equation (46) yields *Equation144*, which was used in releases 1-5 of Simulator005.

Together with the number of clicks needed for mutational meltdown (see Equation (8) on page 39) and a generation time, an approximation for extinction time can be computed with a pocket calculator.

18.2 Meaning of *Equation150* elements

Besides purely heuristic elements (eg. $e + \pi$ in the numerator, 100 as a critical factor for the onset of purifying selection, etc.) some parts of *Equation150* have an actual biological meaning:

Basic frame.

An upper limit for the speed of the ratchet can easily be derived from the intergenerational genomic mutation rate U: There is no reason to expect slightly deleterious mutations to accumulate faster than U mutations per generation, as this is the expected substitution rate for neutral mutations and no special forces should exist that could increase the probability of fixation of a deleterious mutation beyond that of a neutral mutation. However, selection can remove mutations from the population so that they can no longer be fixed. Thus, the rate of the ratchet Cl_r is given by

$$Cl_r = U - U \cdot FractionRemoved \tag{47}$$

Transforming the rate of the ratchet into a clicktime (in generations) gives the basic frame for *Equation150*:

$$T_{cl} = \frac{1}{U - U \cdot FractionRemoved} \tag{48}$$

Processes in the best class

Now the key question is how many mutations are removed by the ratchet. To get a hold on this, consider the average number of new mutations that appear in the N_0 individuals of the best class every generation. As all N_0 individuals are equally likely to catch U mutations, UN_0 new mutations will appear each generation. The average time in generations between the appearances of two new mutations in the best class is therefore given by

$$T_{NewMuta} = 1/(U \cdot N_0) \tag{49}$$

This quantity is found in the negative exponential function of *Equation150*. The fact that *Equation150* produces good results suggests the following: When many new mutations occur in the best class available (low $T_{NewMuta}$), many are selectively removed, whereas when few mutations occur in the best class available, (large $T_{NewMuta}$) few are selectively removed.

It was further found heuristically that for some reason prediction quality is improved by multiplication of the exponent with $1 - N_0/N_e$, the fraction of the population that does not belong to the best class (results not shown).

Basic picture

Thus the following basic picture emerges: The more mutations appear in the best class, the better they can be removed. This makes sense, as under conditions where the ratchet does not operate (very large N_0), most mutations appear in the best class, since the other classes are too small to catch a significant fraction of all new mutations. On the other extreme, under conditions where the ratchet operates deterministically ($N_0 << 1$), the best class is too small to catch a significant fraction of all new mutations.

To turn the number of mutations that occur in the best class into a fraction of mutations that are removed by selection, a function is needed that

- returns not more than 1,
  when new mutations are rare and Equation (49) becomes small
- returns not less than 0,
  when new mutations are frequent and Equation (49) becomes large.

This is achieved by the negative exponential function

$$e^{-\left(\frac{1}{UN_0}\right)\left(1-\frac{N_0}{N_e}\right)} \tag{50}$$

which is identical to the core of *Equation150*.

Considerations from the neutral theory

Another approach to the same basic picture comes from looking at the neutral theory of evolution[1]. Consider the following processes in the most important subgroup of the population N_0 under the simplifying assumption, that N_0 is like a separate population and ratchet mutations are effectively neutral:

- **Fixation time** T_{fix} can be roughly estimated for each of these new mutations by considering T_{fix} for neutral mutations. Standard neutral theory gives a mean of $T_{fix} \approx 4N_e$ for diploid sexual populations. In the case of the ratchet in mitochondrial DNA, this figure has to be divided by 2, because this genetic system is haploid and again by 2, because only the maternal part of the population plays a role. Thus, new mutations need about $T_{fix} \approx N_0$ to get fixed in this class. Two extremes emerge from this:
  When N_0 is large, selection is probably strong and the ratchet is likely to click only rarely.
  When $N_0 < 1$, selection is very weak, but more importantly, fixation

1. This allows application of results from the standard neutral theory of evolution. See Kimura (1983) "The neutral theory of molecular evolution", Cambridge, Cambridge University Press.

coincides with the appearance of a new mutation, as there are no diffusion-fixation processes for such 'population sizes'.

- **Removal time.** New deleterious mutations that occur in the best class have a considerable chance of being removed. To simplify analysis, it is assumed that this happens in $T_{remove} = T_{fix}$ generations. While this is certainly wrong in detail, it might serve as a very crude approximation for this simple estimation. More accurate (and complicated) formulae have been developed by Li[2]. Thus:
  When N_0 is large, selection is probably strong and the ratchet is likely to click only rarely, as times to fixation are long and there are ample opportunities for removal of mutations from the best class.
  When $N_0 < 1$, selection is very weak, and removal is not possible, as usually the appearance of a new mutation coincides with its fixation.
- **Time between fixations.** A rough estimate of the time between fixation events $T_{between}$ can be derived from the neutral theory that states $T_{between} \approx 1/U$, when drift is stronger than selection. When purifying selection is involved, this time is known to be reduced (in combination with a dramatically reduced fixation probability). Then, intuitively speaking, those mutations that get fixed have to do so fast, before they are removed[3].

If $T_{NewMuta}$ is expressed in these terms, then

$$T_{NewMuta} = \frac{1}{UN_0} = \frac{\frac{1}{U}}{N_0} = \frac{T_{between}}{T_{remove}} \qquad (51)$$

If this is applied to the negative exponential function, the following extremes emerge: A negative exponential of Equation (51)

- returns 1 or less,
  when all mutations are removed and (51) becomes very small,
  meaning that $T_{between}$ is much smaller than T_{remove}. While counterintuitive at the first glance ("There is not enough time for a mutation to be removed, before the next one comes"), careful investigation shows that, under these conditions, a large time to removal implies a large N_0 and strong selection. Thus removal is probable.

2. For the neutral case see p. 51 in Li (1997) "Molecular evolution", Sunderland, MA, Sinauer Associates Incorporated. - When selection is involved, see Li & Nei (1977) "Persistence of common alleles in two related populations or species", Genetics 86:901-914.
3. See p. 49 in Li (1997) "Molecular evolution", Sunderland, MA, Sinauer Associates Incorporated.

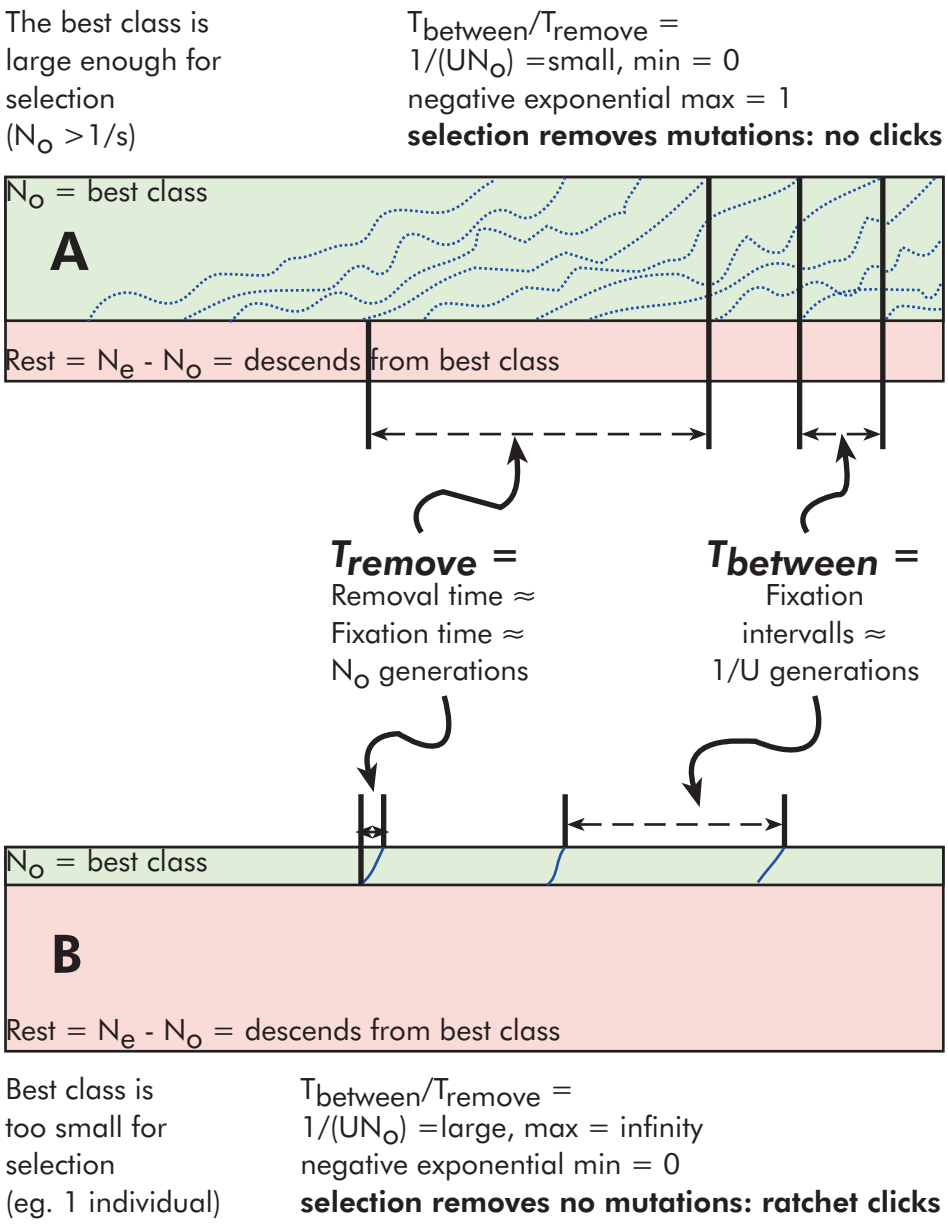


Figure 36 Muller's ratchet as compared to the neutral theory of evolution in two different populations. Populations A and B are shown as graphs with the x-axis as time and the y-axis as frequency in the population. Blue lines denote frequencies of new mutations in the best class (green) over time from origin to fixation under the assumption of effective selective neutrality. As this is wrong in population A (mutations are removed, not fixed in that time), corresponding lines are dotted. The ratchet clicks only in population B. The assumption that new deleterious mutations are removed at least in the same time that neutrals would need for fixation, inspired construction of the simple equations used.

- returns not less than 0,
  when no mutations are removed and (51) becomes large, meaning that $T_{between}$ is much larger than T_{remove}. This is counterintuitive at the first glance, too ("There is enough time for a mutation to be removed before the next one comes"). However, careful investigation shows, that under these conditions, a short time to removal implies a small N_0 and weak to non-existent selection. Especially, if the best class is smaller than 1, appearance of a new mutation and its fixation usually happen at the same time. Thus fixation is probable.

Figure 36 illustrates this logic. After more than 30 years of searching, such analytical considerations strengthen the hope that some day a simple, accurate, analytical equation for the rate of the ratchet might actually be found.

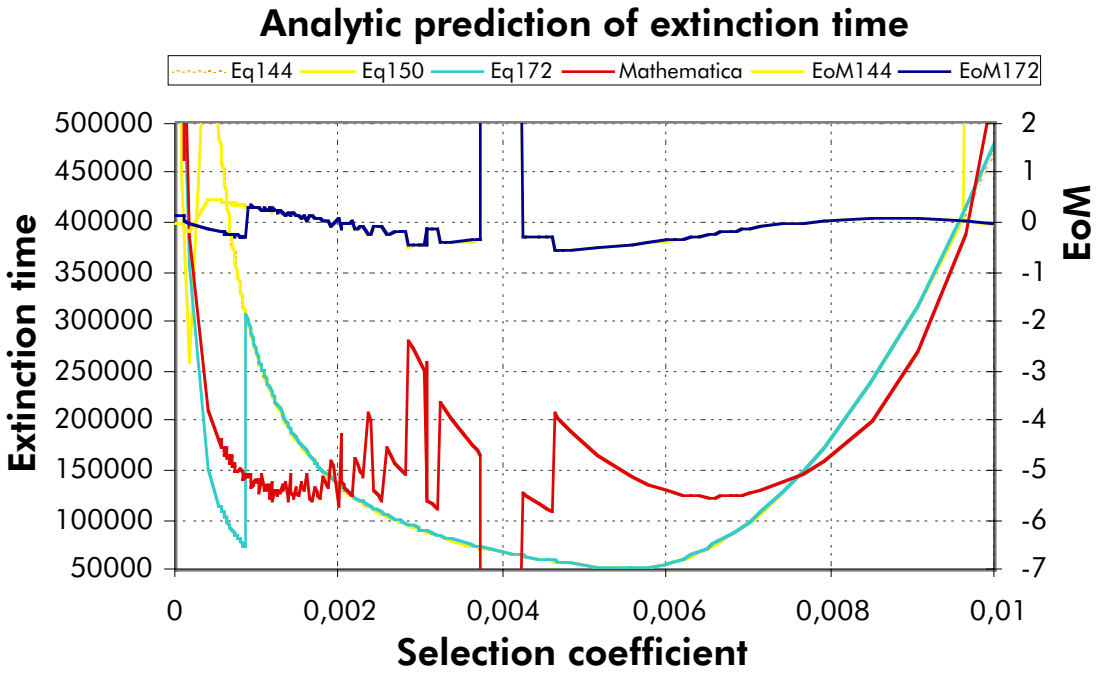


Figure 37 Linear comparison of simple *Equation144,150* and *172* to the best analytical approximations. The population analysed had N_e = 50000, R_{max} = 10 and U = 0.05 under multiplicative fitness. Analytical predictions used in this plot were computed with *Mathematica* according to the analytical methods of Gessler (1995) and Stephan & Kim (2002), as described in chapter 13. Values with T_{ex} below 10^5 are due to unpredictable, complicated numerical errors. Prediction errors of magnitude (EoMs; see text) show, that the simple equations given have a reasonable accuracy over a remarkable range of parameters. Deleterious selection coefficients are given as positive values for easier plotting.

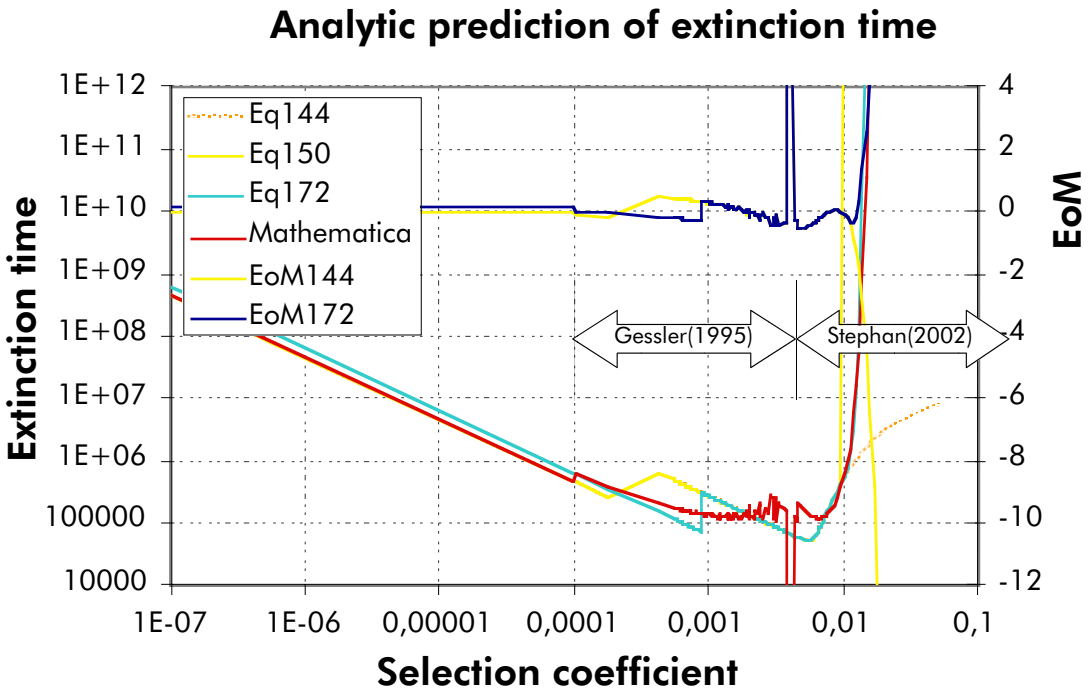


Figure 38 Logarithmic plot comparing simple and analytical approximations as in Figure 37.

18.3 Quality of predictions

Figure 37 compares *Equation150*, its core (*Equation144*, see above) and successor *Equation172* (see below) to the best analytical approximations available on a linear scale. Please note that there are values around $\mathbf{s} = 0.004$ that could not be computed with the current implementation due to complicated numerical issues in *Mathematica*. Such issues occasionally hamper the use of all known analytical approximations that try to predict Muller's ratchet. It is currently unclear whether the oscillatory behaviour of Gessler's approximation is due to numerical limitations during computation or whether it is inherent to its quantitative genetic approach. While the stride of parameter values in the present set of results (S005Project1)is too wide to check for genuinely chaotic behaviour of Muller's ratchet in the corresponding range of parameters, the plots shown below (eg. Figure 39) do not convey the impression of significant chaos.

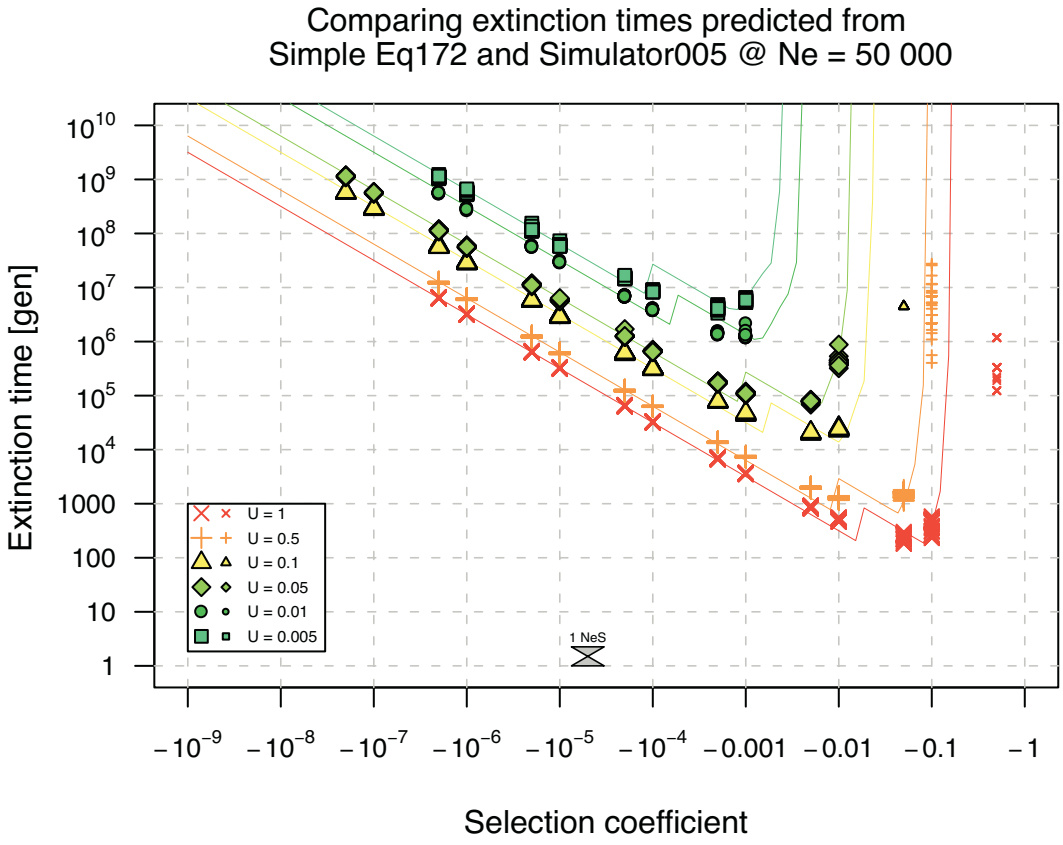


Figure 39 Logarithmic plot comparing *Equation172* and simulation results.
N_e = 50000, R_{max} = 10 and U = 0.05 (yellow triangles) are as in Figure 37. Extinction time is given in generations. The lower mark indicates the border to neutrality (1 N_eS) for the population size used. Thin lines use *Equation172*. Large symbols denote valid extinction time estimates from simulations with at least 2 observed clicks (usually much more). Small symbols denote lower limits for extinction times from simulations without observed clicks, based on the (usually wrong) assumption that the ratchet would have clicked just after stopping the simulation. Each symbol denotes an independent simulation with a different random seed.
When comparing 1845 full observations to predictions from *Equation172*, the following errors of magnitude statistics could be observed: SumOfAbsolute=141, MeanAbsolute=0.0765, Min=-0.28, Max=0.51. This plot contains 1898 simulations with a total of 0.88 years of computing time and was produced by script RS005LL018 at 2002-10-23 22h01m06s.

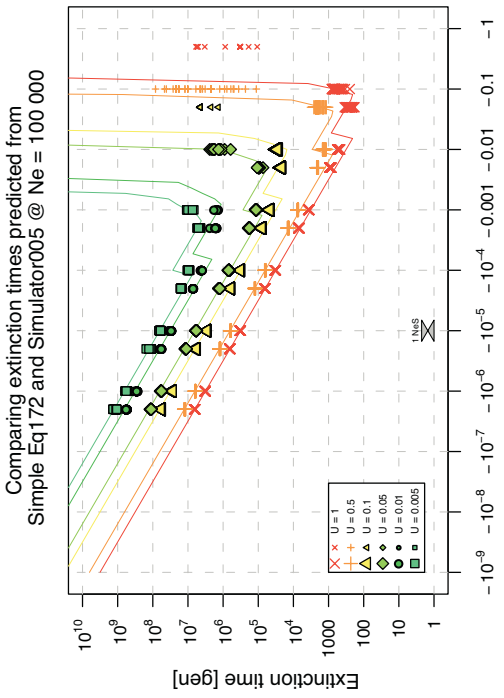
To quantify deviations between simple equations and analytical predictions, the *error of magnitude*[4] is used. It is defined as

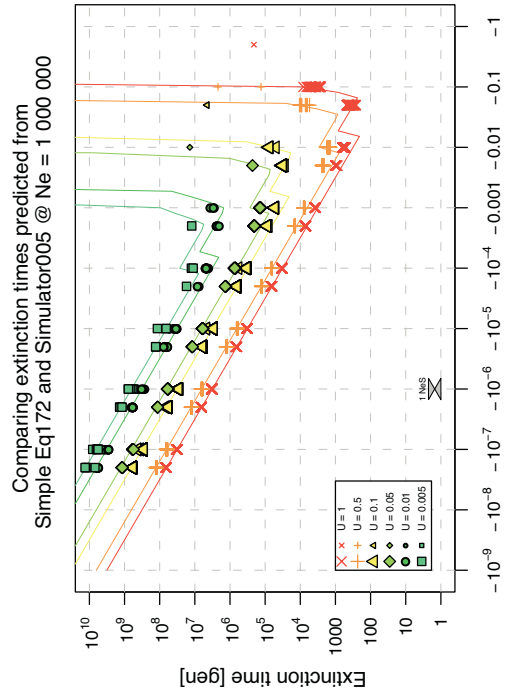
$$Err_{\text{mag}} = \log\left[1 + \frac{Tex_{\text{simple}} - Tex_{\text{analytic}}}{Tex_{\text{analytic}}}\right] \quad (52)$$

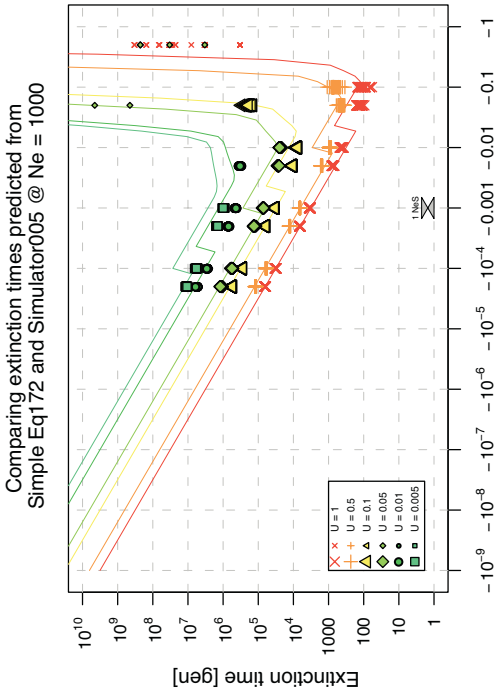
in this context. Figure 37 and its logarithmic counterpart Figure 38 show that prediction accuracy is remarkable for all parameter combinations where the ratchet does click at a reasonable rate. All simple equations predict the location of the 'selection wall' correctly (ie. the selection coefficients that effectively start to stop the ratchet due to purifying selection at a given N_e and U). However, *Equation144* needs an addition to effectively stop the ratchet in the presence of purifying selection, a fact that is also clearly reflected in experiences with the Simulator005 releases 1-5 that used it (see "Sources of computing time variability." on page 146). A comparison of *Equation172* to actual simulation results can be found in Figure 39 using the same parameter combination as in Figure 37. A comparison of a set of parameter combinations for four different population sizes is shown in Figure 40 to convey a feeling for other regions of parameter space. These figures show considerable agreement between *Equation172* and simulation results.

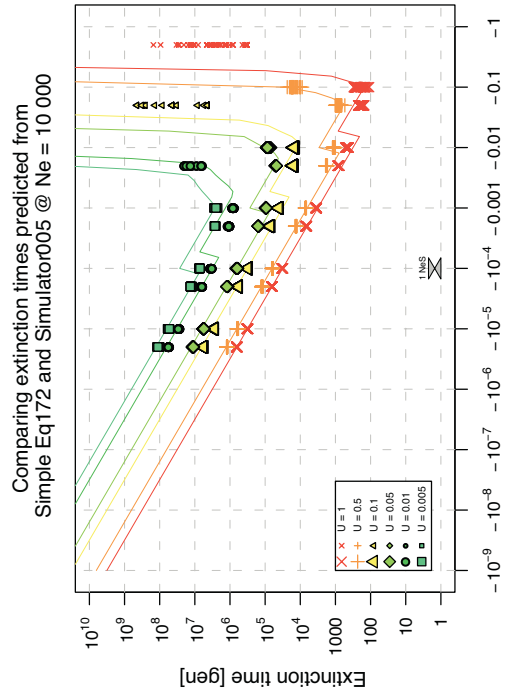
Figure 40 Comparing *Equation172* and simulation results for various population sizes (next page). N_e = 10^3, 10^4, 10^5 or 10^6; R_{max}, U and other details are as in Figure 39.
For N_e = 10^3: 1629 full observations allowed estimation of following errors of magnitude statistics: SumOfAbsolute=290, MeanAbsolute=0.178, Min=-0.57, Max=0.71. (Plot contains 1671 simulations with a total 48 days of computing time and was produced by script RS005LL018 at 2002-10-23 22h05m33s.)
For N_e = 10^4: 1933 full observations allowed estimation of following errors of magnitude statistics: SumOfAbsolute=185, MeanAbsolute=0.096, Min=-0.63, Max=0.58. (Plot contains 2024 simulations with a total 0.54 years of computing time and was produced by script RS005LL018 at 2002-10-23 22h07m15s.)
For N_e = 10^5: 2584 full observations allowed estimation of following errors of magnitude statistics: SumOfAbsolute=250, MeanAbsolute=0.097, Min=-0.33, Max=0.48. (Plot contains 2653 simulations with a total 2.62 years of computing time and was produced by script RS005LL018 at 2002-10-23 22h09m04s.)
For N_e = 10^6: 2326 full observations allowed estimation of following errors of magnitude statistics: SumOfAbsolute=148, MeanAbsolute=0.064, Min=-0.34, Max=0.49. (Plot contains 2521 simulations with a total 2.01 years of computing time and was produced by script RS005LL018 at 2002-10-23 22h10m48s.)

4. See "Definition of the error of magnitude" on page 147. The Error of Magnitude was first defined in: Loewe (2002) "evolution@home: Experiences with work units that span more than 7 orders of magnitude in computational complexity", 425-431. 2nd International Workshop on Global and Peer-to-Peer Computing on Large Scale Distributed Systems at the 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid2002), 21-24 May, Berlin, Germany, IEEE Computer Society.


Comparing extinction times predicted from
Simple Eq172 and Simulator005 @ Ne = 100 000
Extinction time [gen]
Selection coefficient
1 NeS
U = 1
U = 0.5
U = 0.1
U = 0.05
U = 0.01
U = 0.005


Comparing extinction times predicted from
Simple Eq172 and Simulator005 @ Ne = 1 000 000
Extinction time [gen]
Selection coefficient
1 NeS
U = 1
U = 0.5
U = 0.1
U = 0.05
U = 0.01
U = 0.005


Comparing extinction times predicted from
Simple Eq172 and Simulator005 @ Ne = 1000
Extinction time [gen]
Selection coefficient
1 NeS
U = 1
U = 0.5
U = 0.1
U = 0.05
U = 0.01
U = 0.005


Comparing extinction times predicted from
Simple Eq172 and Simulator005 @ Ne = 10 000
Extinction time [gen]
Selection coefficient
1 NeS
U = 1
U = 0.5
U = 0.1
U = 0.05
U = 0.01
U = 0.005

18.4 Complete parameter space overview for Muller's ratchet

While numerical problems prevented computation of complete parameter space overview plots for the analytical approximations known, *Equation150* is particularly well suited for generation of 3D overview plots due to its fixed clicktime in the selection wall area. Thus, for the first time a visual representation of the effects of Muller's ratchet over practically the whole range of biologically interesting parameter combinations can be attempted easily. The results can be seen in Figure 41 for population sizes from 10 to 10^8 and in Figure 42 for population sizes from 10^9 to 10^{30}. The altitude reflects the decadic logarithm of the clicktime T_{cl} in generations as given by *Equation150*. The plain of these triple logarithmic plots varies

- mutation rate U from 1 to 10^{-10} and
- selection coefficient s from -1 to -10^{-10} (again, positive values are used for plotting convenience here).

The upper plateau indicates the area where the ratchet never clicks due to purifying selection. The slope on the mutation rate side reminds one of the fact that a ratchet can click only as fast as deleterious mutations appear. The slight wave in the lower middle is due to parameter combinations where $N_0 = 10^{-99}$, the first critical level in *Equation150*. As soon as the quotient of individuals in the best class over effective population size can no longer be distinguished from 1 because of the limited number of valid digits for double floats (=16), white squares are generated in the selection wall area. This starts for population sizes of $>10^{16}$.

When these plots are compared, the emerging basic picture about the operation of Muller's ratchet shows three distinct areas where the rate of the ratchet is simple to predict:

- **MN (Mutation Neutral) area.** For deleterious mutation combinations of $|U/s| \gg 10$ the ratchet clicks with a speed that is similar to the accumulation of neutral mutations in a population. The remaining area is divided in two subareas:
- **SW (Selection Wall) area.** For deleterious mutation combinations of $|U/s| \ll 1$ and a considerably large population size ($|s| \gg 1/N_e$), the ratchet will not click, because all new mutations are removed by purifying selection.

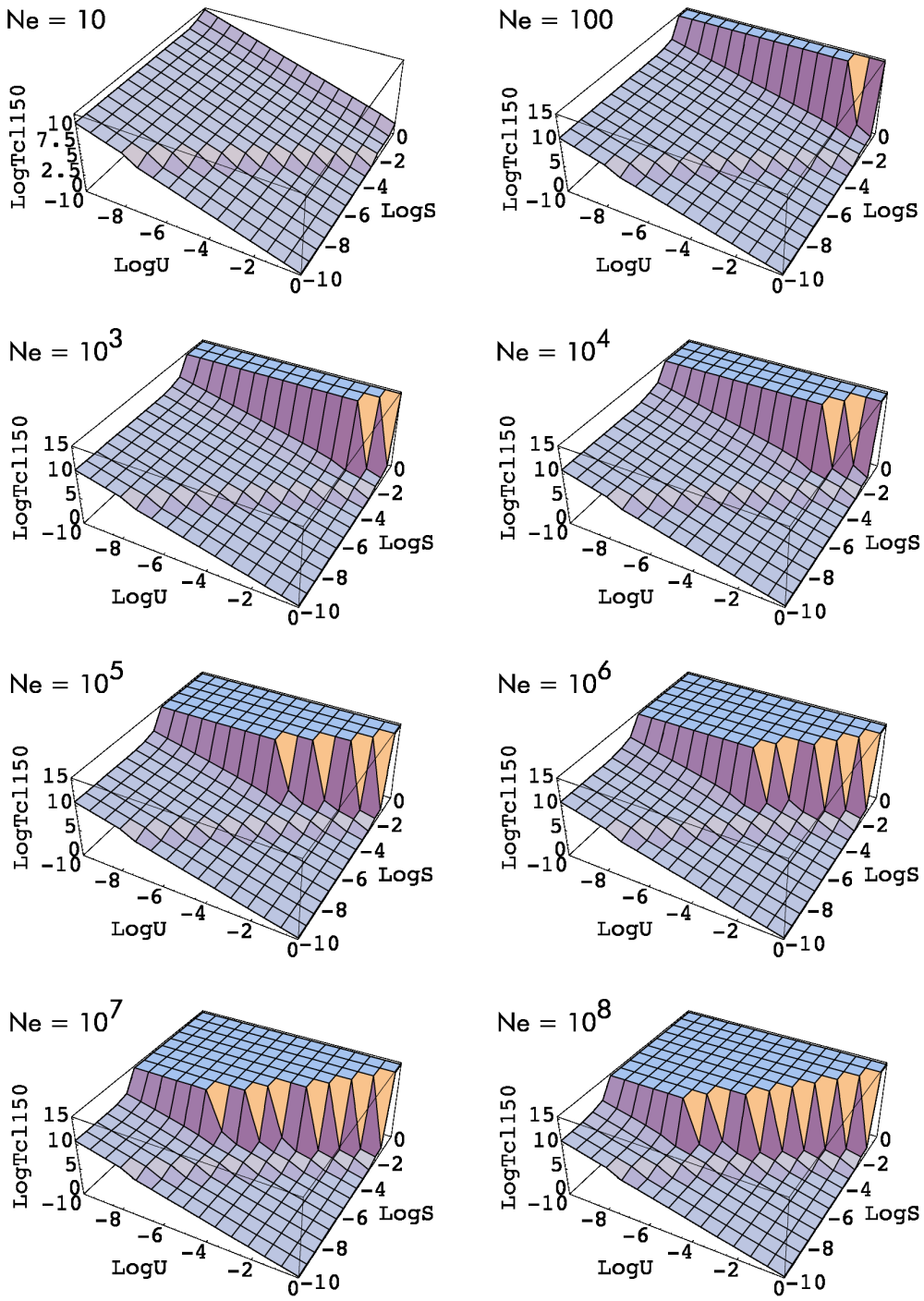


Figure 41 Overview over clicktimes of Muller's ratchet given by *Equation150* for $N_e = 10 - 10^8$

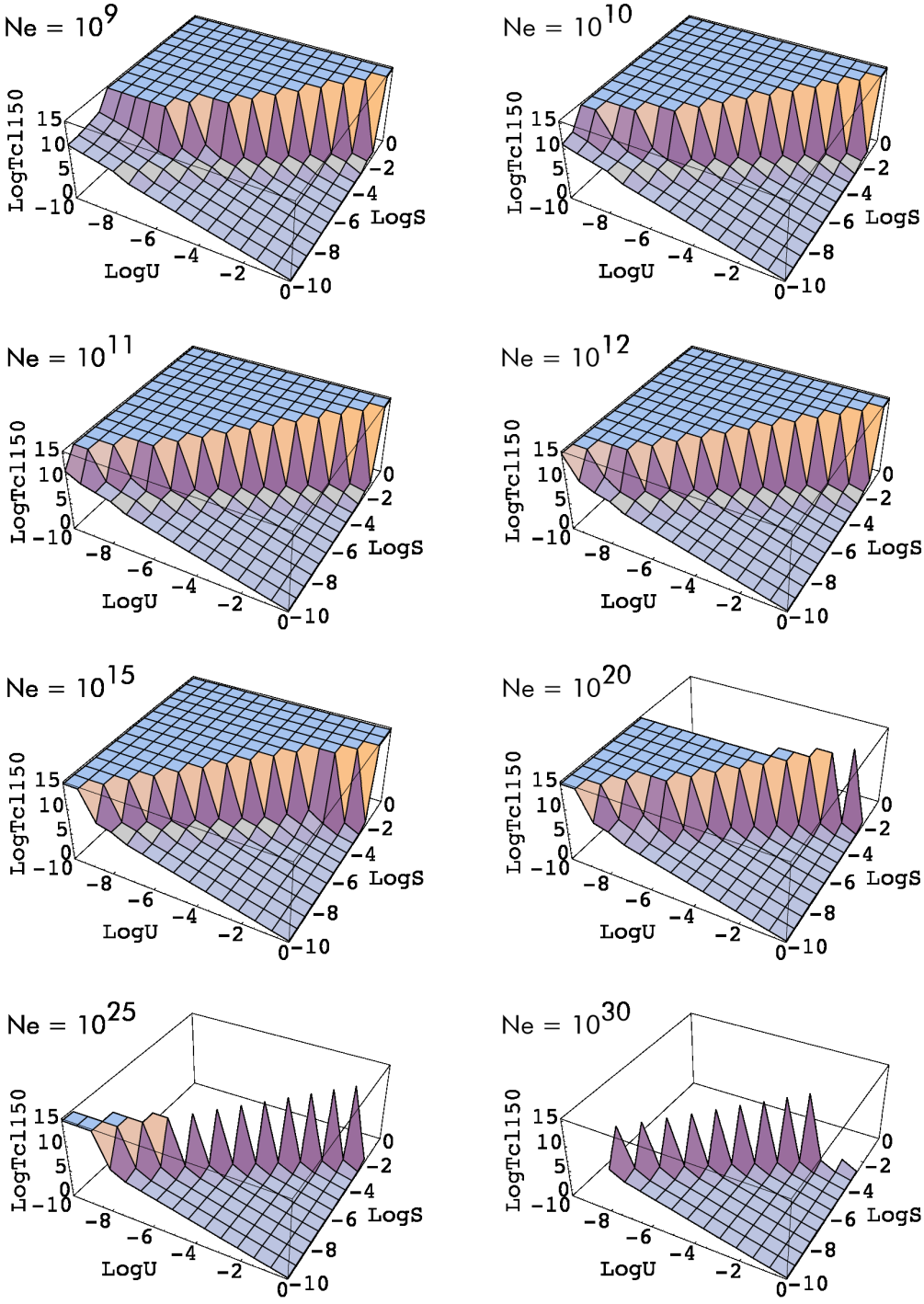


Figure 42 Overview over clicktimes of Muller's ratchet given by *Equation150* for $N_e = 10^9 - 10^{30}$

- **PN (Population Neutral) area.** The ratchet clicks with a speed that is similar to the accumulation of neutral mutations in a population, when $s < 1/N_e$.

Other parameter combinations need more sophisticated approaches. This basic picture can be confirmed, if simulation results are selected that represent two intersections through the 3D plots shown: One through the mutationally-neutral-area and the selection-wall-area (Figure 43) and the other through the mutationally-neutral-area and the populationally-neutral-area (Figure 44).

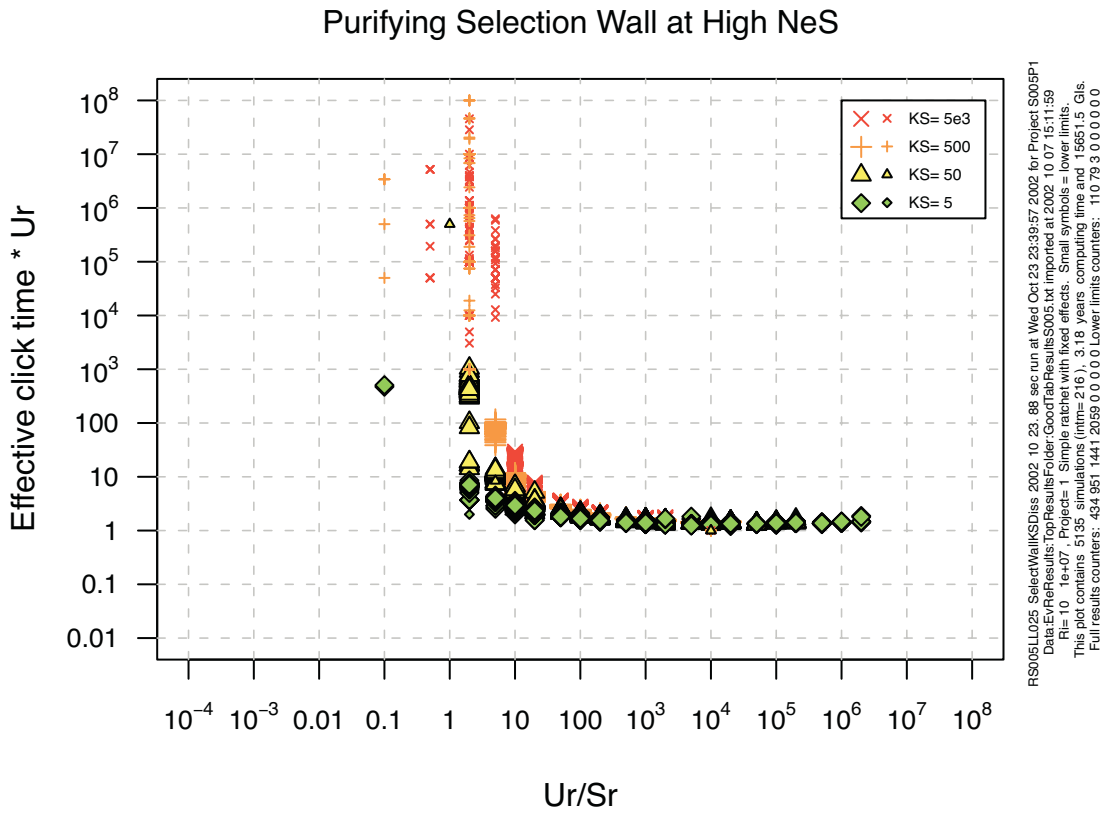


Figure 43 Transition to the purifying selection wall at high $N_e s$.

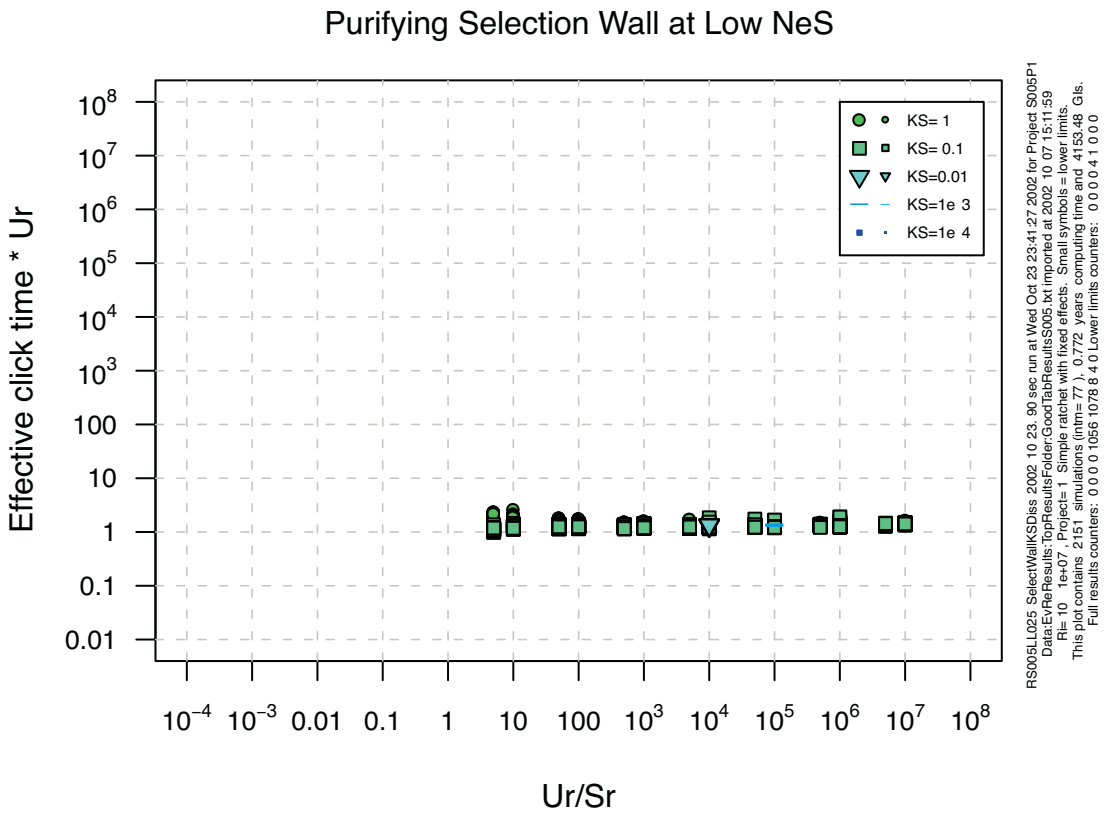


Figure 44 Approximately neutral evolution at low $N_e s$.

18.5 The *Equation172* system

The main limitations of *Equation150* are the highly artificial determination of the selection wall and the abrupt change in clicktime at the transition to neutrality. Considering the selection wall, one improvement comes again from looking at the best class. When the product $N_0 s$ is large (>>1), then there are actual individuals in the best class that can be selectively removed. However, when it is small (<<1), then selection will not see the new mutations in individuals. Thus, the additional waiting time to a click of the ratchet due to the selection wall could be described by

$$\frac{1}{e^{N_0 s}} \qquad (53)$$

for negative **s**. This led to *Equation172* after some further heuristic refinement:

$$T_{cl172} = 1.38/U, \qquad \text{if } N_0 < 10^{-20} \tag{54}$$

The best simple prediction for Muller's ratchet currently available

and for all other cases

$$T_{cl172} = \frac{e + \pi}{U - U \cdot e^{\left(\frac{N_0}{N_e} - 1\right)/(U \cdot N_0)}} + \frac{1}{e^{N_0 s}} \tag{55}$$

This equation represents the current state of the art and was added to the plots above to facilitate comparisons. It is implemented in release 6 of Simulator005. Clearly, more work has to be done before a completely satisfying simple approximation for the rate of Muller's ratchet can be found.

18.6 Future developments

The work above encourages further search for even better simple formulae for predicting the rate of Muller's ratchet. As all such equations will have to be tested against simulation data, it is extremely helpful to have a database of simulation results with which to compare new equations. A potential way to measure the quality of a new prediction formula is computation of the mean absolute error of magnitude for a given set of parameter combinations or simulations, as done here.

In the meantime, wet-lab biologists might use *Equation172* to get a quick, rough overview on what is going on in their systems. The comparisons with analytical and simulation results presented here should help to decide whether a given simple prediction is likely to be very different from analytical or simulation reality or not. Thus it becomes easier to decide whether more detailed analyses of Muller's ratchet are worthwile for a given system.

19 Directions for future work on theoretical models of Muller's ratchet

Simulator005 observes a variety of parameters (some for the first time) that play a role in theoretical models of Muller's ratchet. The resulting database (over 28000 simulations with over 16 years CPU-time) is an excellent opportunity to improve theory by comparing various predictions with actual simulations on a large scale. This chapter suggests a number of theoretical aspects that might benefit from further investigation using the results computed by Simulator005.

19.1 Observations of different phases during a click of Muller's ratchet

If the ratchet clicks slowly and the size of the best class N_0 is well above 1, diffusion theory predicts that the time between two clicks of the ratchet can be divided in two phases:

1. Immediately after a click, the best class contains N_1 individuals (second best class before the click). Under mutation selection equilibrium, this number is expected to be $N_k = N_e \cdot e^{U/s} \cdot (U/(-s))^k / k!$, where k(=1) is the number of mutations with deleterious (= negative) selection coefficients s that appear at the genomic mutation rate U per generation in a population of size N_e[5]. In most cases this frequency will have to fall to its equilibrium value N_0, because $N_1 > N_0$. The time needed for this process is the duration of phase 1.
2. The second phase waits for the frequency of the best class to fall to zero by chance.

Simulator005 can observe the phases

Simulator005 makes the first observations of durations of these phases in actual simulations. The first part of the first phase, `T1a`, counts the generations from a click until the actual frequency of the best class has fallen below 1.6 x N_0 for the first time[6]. This process is usually fast and deterministic.

5. This equation was first found by Kimura & Maruyama (1966 "The mutational load with epistatic gene interactions in fitness", Genetics 54:1337-1351) and plays a central role in analytical approximations for the rate of Muller's ratchet. See Stephan & Kim (2002) "Recent applications of diffusion theory to population genetics", pp. 72-93 in: Slatkin & Veuille (eds) Modern developments in theoretical population genetics, Oxford, Oxford University Press. - Gessler (1995) "The constraints of finite size in asexual populations and the rate of the ratchet", Genet. Res. 66:241-253.

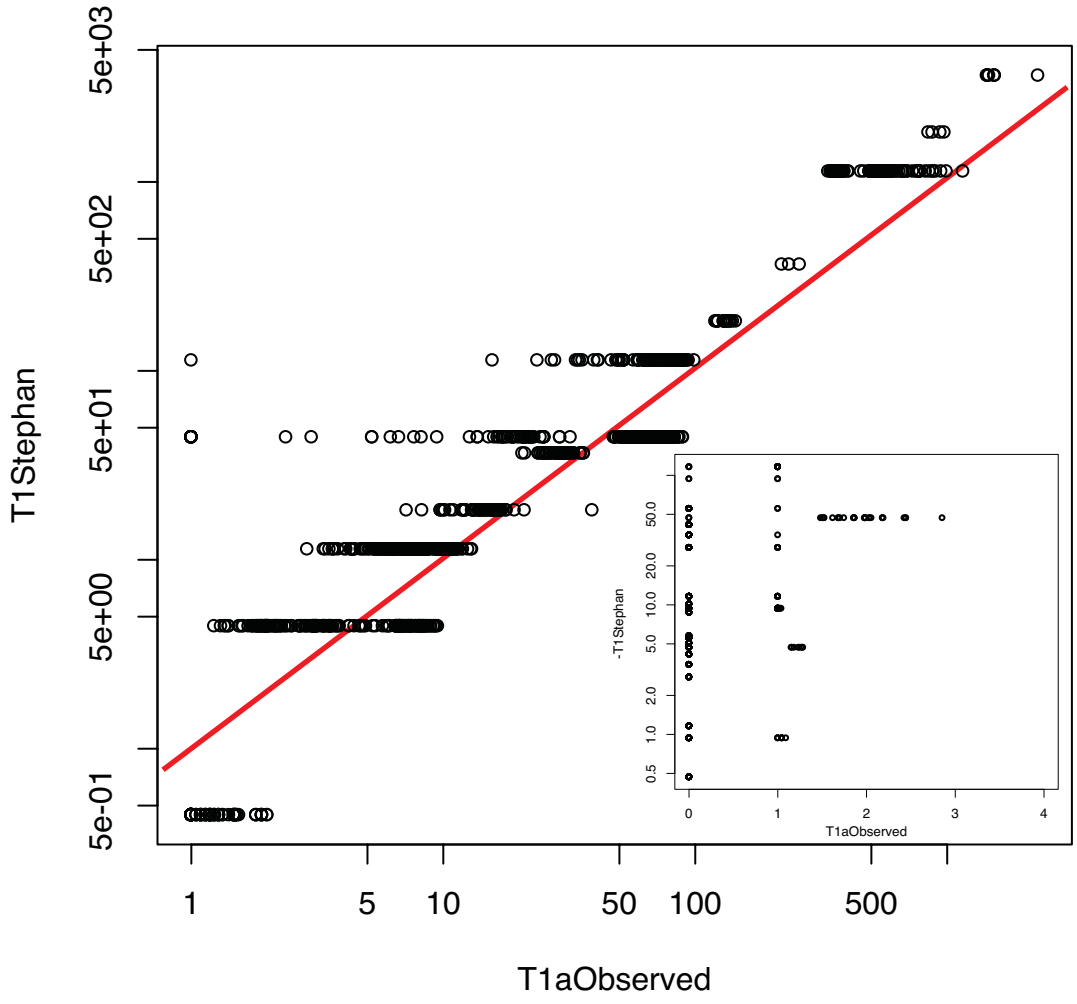


Figure 45 Comparison of the predicted and observed durations of the first phase of a click of the ratchet. The y axis gives the analytical prediction according to Stephan & Kim (2002) for all parameter combinations, whose observed duration of `T1a` is plotted in the x axis. The smaller insert does the same for cases where analytical predictions gave negative values. The red line indicates values where predictions and observations would be identical. The analytical prediction is not valid when selection coefficients are zero. A similar formula by Gordo & Charlesworth[8] does not lead to a significantly different picture. The plots were produced by R on 26 Oct 2002 using all results of S005 from Era1-7s1, where $U>0.001$ and $N_0>5$.

6. Stephan & Kim (2002) "Recent applications of diffusion theory to population genetics", pp. 72-93 in: Slatkin & Veuille (eds) Modern developments in theoretical population genetics, Oxford, Oxford University Press.

The second part of the first phase, `T1b`, counts the generations from the end of `T1a` until the best class falls below its equilibrium expectation $\mathbf{N_0}$ for the first time. This process is usually slow and has a large stochastic component. The second phase, `T2`, counts the rest of generations until the ratchet clicks (stochastic waiting for extinction of best class). Please note that observation of these phases makes sense only if $\mathbf{N_0}$ is significantly larger than 1.

First phase

Figure 45 compares observations of `T1a` to expectations from theory[6]. While general agreement is found in most cases, significant deviations can be observed too. Further improvements of estimation of the first phase are likely to come from investigation of some of these outliers.

Second phase

Unfortunately, computations of analytical predictions of the second phase (and hence the total clicktime) were too complex for comparisons on such scales in this work. However, further investigations of the diffusion theory approximation showed the following interesting detail. Stephan & Kim had suggested that improvements of diffusion theory prediction accuracy might come from a more detailed analysis of their parameter $\boldsymbol{k}$ (not to be confused with the number of mutations in an individual). In computing the length of the second phase, a constant value of $\boldsymbol{k} = 0.5$ was used earlier[7], while later analyses[6,8] suggested $\boldsymbol{k} = 0.6$ (as used elsewhere in this work).

Here, the potential effect of $\boldsymbol{k}$ on the precision of predictions for a given set of parameters was computed by varying $\boldsymbol{k}$ from 0.1 to 0.9 for all 33 examples given by Stephan & Kim[6]. Figure 46 shows 3 typical outcomes. All other examples showed plots that where similar to one of these three types. No clear pattern could be observed in terms of what parameter combination would lead to which curve, but parameter combinations with a low $\mathbf{N_0}$ tended to be like the first plot sample, those with intermediate $\mathbf{N_0}$ like the second and those with the largest $\mathbf{N_0}$ like the third. In most, but not all, of these 33 plots the same clicktime as observed in simulations was predicted by some $\boldsymbol{k}$ between 0.1 and 0.9. In most cases the correct $\boldsymbol{k}$ could not be recognized by some special feature like eg. a global (or local) minimum. Thus if $\boldsymbol{k}$ is the only possible point of improvement in the diffusion theory approximation, it might have to vary over a large range.

7. Stephan et al. (1993) "The advance of Muller's ratchet in a haploid asexual population: Approximate solutions based on diffusion theory", Genet. Res. 61:225-231.
8. Gordo & Charlesworth (2000) "The degeneration of asexual haploid populations and the speed of Muller's ratchet", Genetics 154:1379-1387.

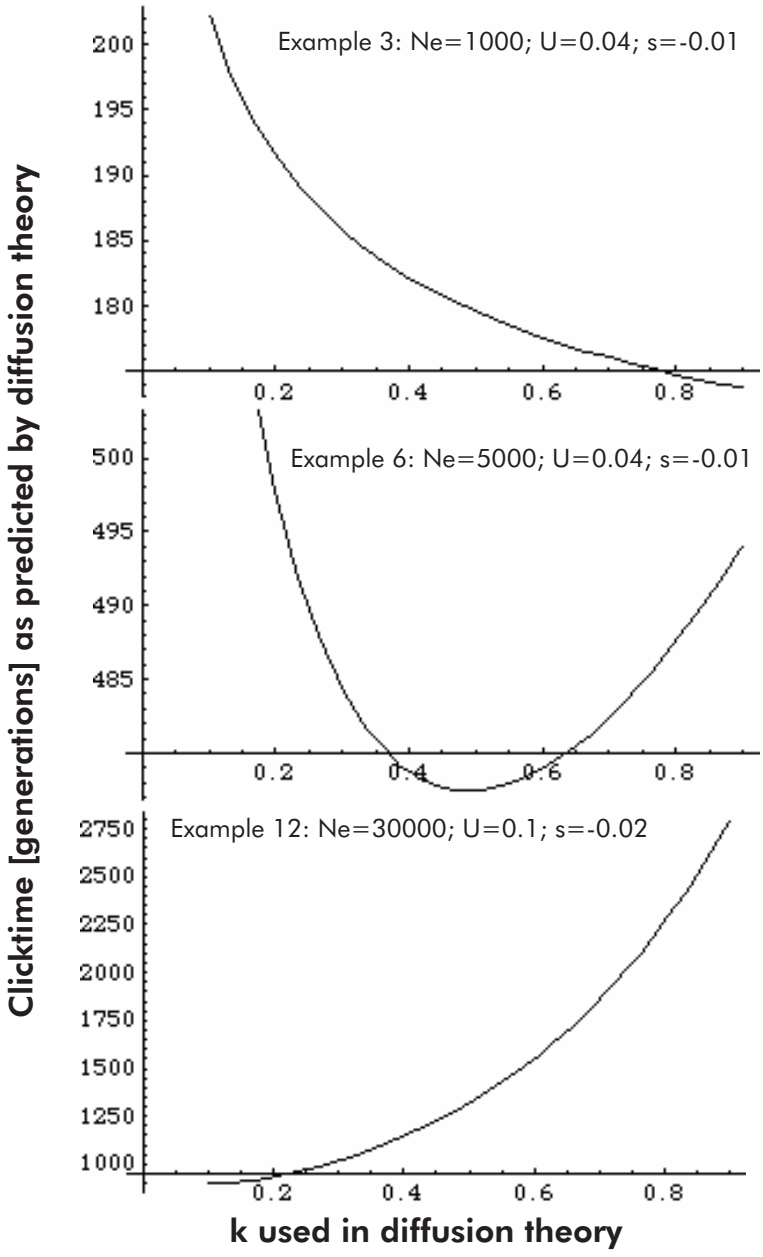


Figure 46 The influence of k on prediction accuracy of the diffusion approximation of the rate of Muller's ratchet.
Ratchet parameters are given for each example (see Stephan & Kim[6]). Gordo & Charlesworth[8] found the following simulation results: Clicktime (±2SE) Ex3=169±39; Ex6=514±99; Ex12=1543±275. Reproduction of these results with S005r0 found the following (assuming `Rmax`=4; `Cend`=100, if no extinction occurred earlier; `Tend` = 10^5): Effective (nearly identical to absolute) clicktime ± StdDev (min-max) Ex3=160±141(12-962); Ex6=478±324(69-1536); Ex12=1356±1187(147-7126). N_o=18, 92, 202.

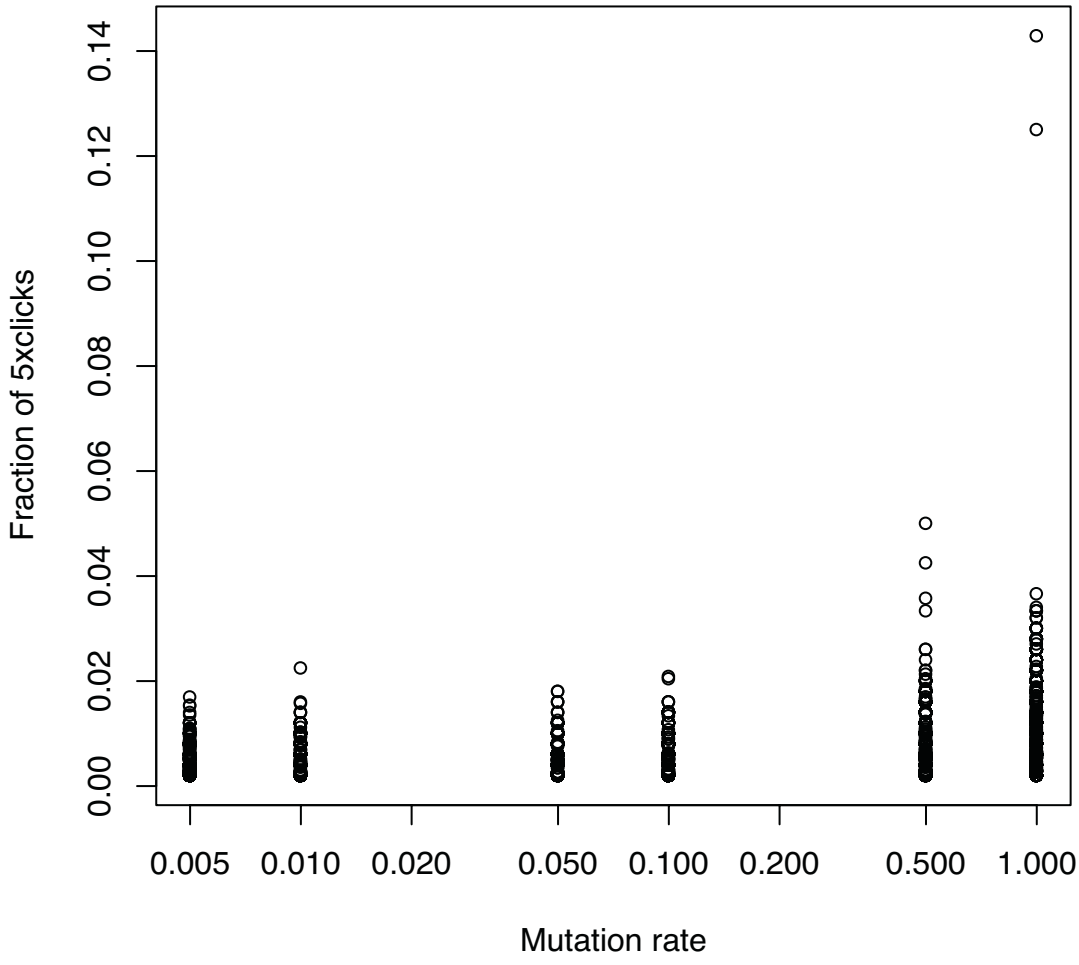


Figure 47 Surprisingly many clicks fix 5 mutations at a time.
The plot was produced by R on 26 Oct 2002 using all results of Simulator005 from Era1-7s1, where $U>0.001$. To reduce the amount of data to be plotted, a little less than half of these simulation results were omitted, as they did not contain any such 5xclicks.

19.2 Observations of multiple clicks of Muller's ratchet

How to observe multiple clicks

If the mutation rate is high enough, the ratchet will be expected to click around several notches in a single generation. Up to now, there have been no quantitative reports on this expectation. Thus simulations conducted with Simulator005 did not only observe clicktimes, but also the number of

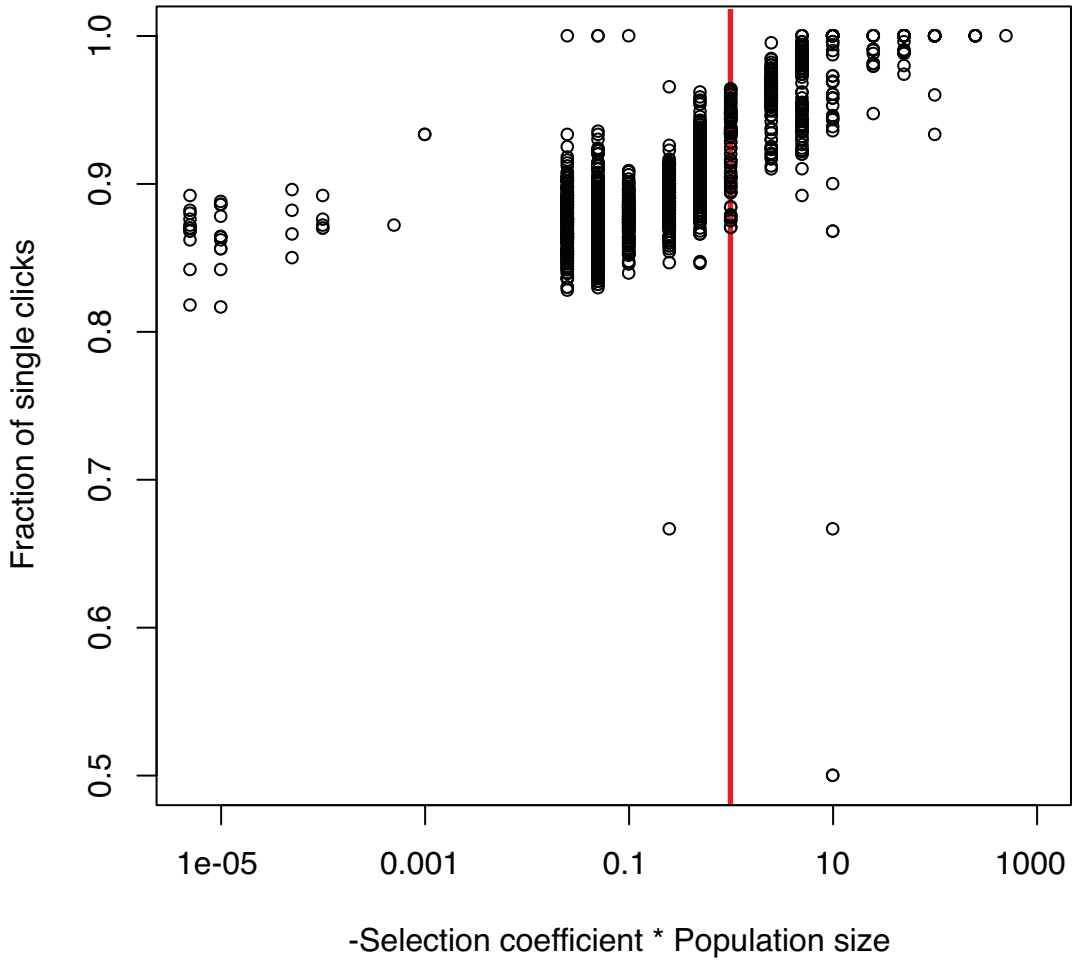


Figure 48 Multiple clicks can still occur when selection is strong.
The plot was produced by R on 26 Oct 2002 using all results of Simulator005 from Era1-7s1, where U=0.005. The red line denotes the approximate border between selection coefficients that are felt in the population and those that are effectively neutral (NeS = 1).

new deleterious mutations that were fixed during a click by comparing the number of mutations in the best class before and after each click. To allow precise quantification of the phenomenon, the following normal distributions (mean, standard deviation, minimum, maximum, etc.) were recorded for each click at the moment it was observed:

- Absolute clicktime (`atcr`) = the absolute time between two clicks (all absolute times are in generations here)
- Clicksize (`clsr`) = difference in the number of mutations of the best class before and after a click
- Effective clicktime (`etcr`) = `atcr/clsr`, computed on the spot
- Absolute clicktime of clicks that fix exactly 1 mutation (`cl1`)
- Absolute clicktime of clicks that fix exactly 2 mutations (`cl2`)
- Absolute clicktime of clicks that fix exactly 3 mutations (`cl3`)
- Absolute clicktime of clicks that fix exactly 4 mutations (`cl4`)
- Absolute clicktime of clicks that fix exactly 5 mutations (`cl5`)

Simple expectation was wrong

Intuitively, one would expect that multiple clicks for Muller's ratchet would occur only when mutation rates are high. It was not clear from the beginning whether it would make any sense at all to look separately for clicks that fix exactly 5 mutations at a time. However, if the fraction of 5fold clicks (`cl5n/etcrn`) is plotted over the mutation rate, one is surprised to see in Figure 47 that the ratchet can still fix 5 mutations at a time in an appreciable number of simulations at a mutation rate as low as 0.005. If the fraction of single clicks (`cl1n/etcrn`) is plotted over the product of selection coefficient and effective population size, it becomes apparent that low-mutation rate multi-clicks can play a role even when selection is no longer negligible (Figure 48). However, as expected, weaker selection further increases the frequency of multiple clicks. Please note that the strength of selection *felt in the population* has to be applied in these plot. Plotting the data from Figure 48 over population size or selection coefficient separately does hardly show significant trends (plots not shown).

Possible biological role of multiple clicks

One might ask why multiple clicks of Muller's ratchet could possibly be interesting. First, a better understanding of such details could be important for future improvements of analytical theory. Second, the existence of multiple clicks indicates that ratchet processes are associated with a high degree of irregularity. Such irregularity could contribute to overdispersion of molecular clocks[9] (see discussion after reporting data on the variability of the rate of Muller's ratchet below). Further research on this issue might turn out to be interesting.

Conclusions

While the exact significance of multiple clicks is still unclear, future reports of simulation results should distinguish between absolute and effective clicktimes, when investigating the rate of Muller's ratchet. It is suggest-

9. Gillespie (1991) "The Causes of Molecular Evolution", New York, Oxford University Press. - Cutler (2000) "Understanding the overdispersed molecular clock", Genetics Mar 154:1403-1417.

ed that in most cases, the effective clicktime is the correct quantity to use, especially when computing extinction times.

19.3 Observation of the variability of the rate of Muller's ratchet

While a number of studies report simulation results for the clicktime of the ratchet[10], only few quantify its variability[11,12,13,14,15] (see Figure 49). This might be partly due to the fact that the *true* standard deviation of the clicktime cannot be estimated easily from a long series of observed clicktimes, because the successive clicktimes are not independent[11].

To investigate the variability of clicktimes of the rate of Muller's ratchet Simulator005 observed the following details:

Recorded data

- Absolute and effective clicktime Gauss distributions were inferred while observing the data, assuming individual values were normally distributed and completely independent from each other. While this assumption is wrong, it gives an easy overview over variability of clicktimes[11].
- Actual individual clicktimes were recorded for the first 500 (or all observable) clicks. However, time constraints did not allow analysis of this large set of data in the present work (ca. 620 MByte from >28000 simulations and >16 years CPU-time). These raw data are available for future analyses that might be able to distinguish between various future hypotheses on variance of clicktime with much more statistical power than would be possible up to now.

Intra-simulation variation

A simple overview over all apparent coefficients of variation (CVs) of effective clicktime in results of Simulator005 can be found in Figure 50. CVs were plotted over U/s, as this appeared to show the clearest trend (s or N_0 gave some trend, too; U or N_e showed no trend): Whenever mutation pressure is much stronger than selection, CVs of 0.7-1 are normal. When selec-

10. See review in Baake & Gabriel (1999) "Biological evolution through mutation, selection and drift: An introductory review", Annual Reviews of Computational Physics 7:203-264.
11. Haigh (1978) "The accumulation of deleterious genes in a population - Muller's Ratchet", Theor. Popul. Biol. 14:251-267.
12. Gessler (1995) "The constraints of finite size in asexual populations and the rate of the ratchet", Genet. Res. 66:241-253.
13. Stephan et al. (1993) "The advance of Muller's ratchet in a haploid asexual population: Approximate solutions based on diffusion theory", Genet. Res. 61:225-231.
14. Charlesworth & Charlesworth (1997) "Rapid fixation of deleterious alleles can be caused by Muller's Ratchet", Genet. Res. 70:63-73.
15. Gordo & Charlesworth (2000) "On the speed of Muller's ratchet", Genetics 156:2137-2140.

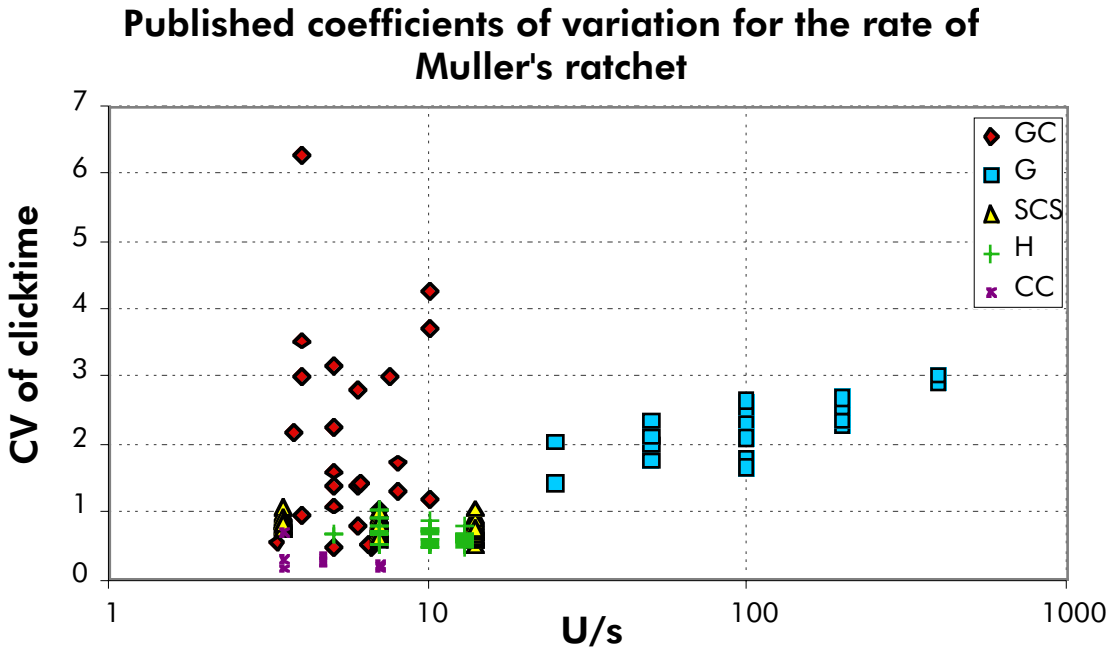


Figure 49 Published coefficients of variation for the rate of Muller's ratchet.
The various data are from H[a], SCS[b], G[c], CC[d] and GC[e]. Please note that the apparent coefficient of variation given here (= apparent StDev/Mean) is not the true one, because individual clicktimes of different clicks in the same population are not completely independent[a]. However, it would be surprising if the true values would lead to a completely different picture. All data except H is inferred from authors SE using technical simulation details in their papers. Thus, accuracy depends on the quality of these details.

a. Haigh (1978) "The accumulation of deleterious genes in a population - Muller's Ratchet", Theor. Popul. Biol. 14:251-267.
b. Stephan et al. (1993) "The advance of Muller's ratchet in a haploid asexual population: Approximate solutions based on diffusion theory", Genet. Res. 61:225-231.
c. Gessler (1995) "The constraints of finite size in asexual populations and the rate of the ratchet", Genet. Res. 66:241-253.
d. Charlesworth & Charlesworth (1997) "Rapid fixation of deleterious alleles can be caused by Muller's Ratchet", Genet. Res. 70:63-73.
e. Gordo & Charlesworth (2000) "On the speed of Muller's ratchet", Genetics 156:2137-2140. Please note, that these authors reported only 2 standard errors, but were not precisely clear on how many clicks they had observed. This plot assumes, that they actually observed as many clicks as they expect fixations in the last columns of their table 1. This is based on a statement in their methods paper that they generally used *about* 5 replicate runs and observed the ratchet for *2 000 - 100 000* generations (Gordo & Charlesworth (2000) "The degeneration of asexual haploid populations and the speed of Muller's ratchet", Genetics 154:1379-1387). Thus the upper CVs could be too high.

tion becomes appreciable, CVs can rise to about 3, before fading into purifying selection. The exact nature of this transition is unclear at the moment, but further simulations can easily target this range of parameters, to see whether CVs shrink before purifying selection becomes active or not. As

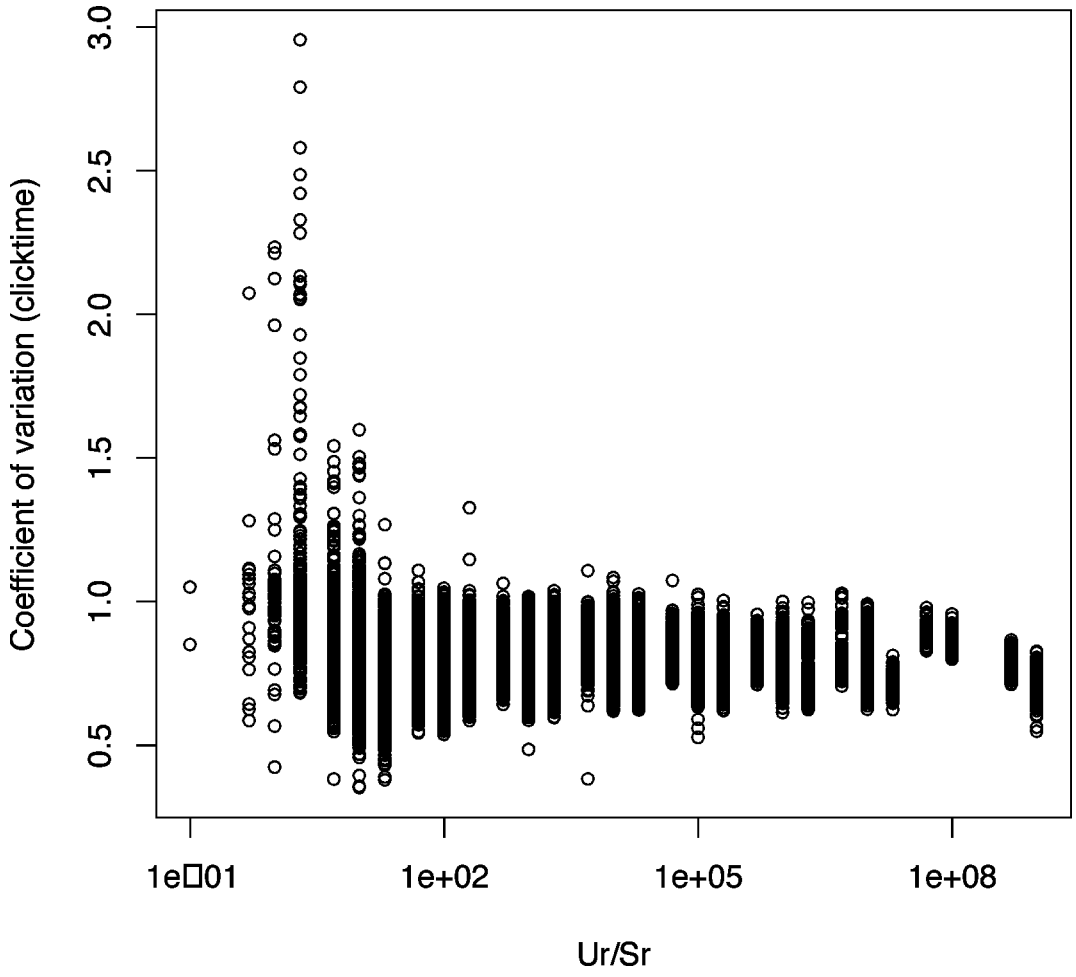


Figure 50 Overview over apparent coefficients of variation of effective clicktime.
The plot was produced by R on 26 Oct 2002 using all results of Simulator005 from Era1-7s1, where $U>0.001$ and more than 5 clicks of the ratchet had been observed. The sign of selection coefficients was inverted.

the bandwidth of CVs is appreciable, further plots were made to see whether CVs would depend on population size for a given combination of U/s. At $U/s = 100$ no striking trend could be observed, although the highest CVs tend not to be found in the largest populations (data not shown). As N_0 and N_e are proportional at a given U/s, this applies to both quantities.

Inter-simulation variance

Then, the whole dataset was partitioned in 4 groups according to N_0 to compute mean μ ± apparent standard deviation and CV for the product of mutation rate and effective clicktime within these groups. The following val-

ues were obtained (by R on 26 Oct 2002 using all results of Simulator005 from `Era1-7s1`, where ***U***>0.001 and more than 5 clicks of the ratchet had been observed):

- $\mathbf{N_0} < 10^{-30}$ (close to neutral): μ = 1.43 ± 0.25; CV = 0.045
- $10^{-30} < \mathbf{N_0} < 1$ (deterministic ratchet): μ = 3.02 ± 2.11 ; CV = 1.47
- $1 < \mathbf{N_0} < 100$ (fast stochastic ratchet): μ = 87 ± 813 ; CV = 7605
- $100 < \mathbf{N_0}$ (slow ratchet): μ = 628 ± 1528 ; CV = 3721

While the first two can be considered as relatively reasonable summaries for a broad range of parameters, the latter two show that the operation of a non-deterministic ratchet depends much more on the actual parameters than on other processes that contribute to variance (compare these extraordinarily high CVs of one big 'inter-simulation analysis' to those significant, but relatively moderate increases of the summary of intra-simulation analyses in Figure 50). While these extraordinarily high variances can be decreased a little bit by decreasing the $\mathbf{N_0}$ analysis window size, the overall phenomenon remains, as the same $\mathbf{N_0}$ does not imply comparable rates of the ratchet[16].

Further work

As Simulator005 observes a large number of parameters at the same time, there is a lot of room for additional analyses (eg. compare CVs of absolute and effective clicktimes; check CVs of multiple clicks; etc...). As simulations from different authors can employ different population density regulations, it will be interesting to investigate potential effects of population density regulation mechanisms on variation in clicktime.

Overdispersed molecular clocks

Muller's ratchet can cause rapid fixation of slightly deleterious mutations[14]. In the long term, such fixations will show up as substitutions in an evolutionary line. Thus variation in the rate of such fixations could contribute to the overdispersion seen in molecular clocks[17,18]. The unexpectedly high frequency of multiple clicks of Muller's ratchet reported above might further contribute to such variation. While lineage-specific effects (eg. generation length[19]) that increase the variance of the molecular clock can be singled out by enough data, the core of the enigma of the overdispersed clock lies in the residual effects ***R*** that remain after removing lineage-specific effects. ***R*** is defined as the ratio of the variance in the number

16. This was found earlier by Gordo & Charlesworth (2000) "The degeneration of asexual haploid populations and the speed of Muller's ratchet", Genetics 154:1379-1387.
17. Gillespie (1991) "The Causes of Molecular Evolution", New York, Oxford University Press.
18. Cutler (2000) "Understanding the overdispersed molecular clock", Genetics 154:1403-1417.
19. Martin & Palumbi (1993) "Body size, metabolic rate, generation time, and the molecular clock", Proc. Natl. Acad. Sci. USA 90:4087-4091. - Ohta (1993) "An examination of the generation-time effect on molecular evolution", Proc. Natl. Acad. Sci. USA 90:10676-10680.

of substitutions in a lineage over the mean number of substitutions. It is expected to be 1 in a simple (Poisson-) model of neutral evolution[17]. If – as sometimes suggested – selection acts only as a selective sieve that removes a fraction of non-silent mutations in proteins, there is no simple reason to expect a higher variance. However, observed values for R can be 5 and more[18]. This is not the place to review the large number of theoretical models that have been invented to explain overdispersion of the molecular clock[17,18,19,20,21].

The Muller's-ratchet-contributes-to-overdispersed-clock hypothesis

This work rather adds a new hypothesis that is waiting to be tested, namely that Muller's ratchet can contribute significantly to the overdispersion observed in molecular clocks. According to this hypothesis, 3 classes of mutations exist:

1. Effectively neutral mutations that drive the ratchet so fast that it will not be significantly different from neutral expectations.
2. Dangerous mutations that drive the ratchet slowly (selection coefficients are too large to be neutral or harmless, but too small to be completely removed by purifying selection)
3. 'Normal' deleterious mutations are removed by purifying selection and do not appear on the long term (comparable to selective sieve).

Classes 1 and 3 do not contribute to overdispersion by Muller's ratchet (although they might do by other mechanisms, like eg. background selection[22]). However, class 2 contains mutations that have been shown to be on the border of purifying selection and are expected to accumulate with a larger variance by itself. Moreover, the present work has shown that the border between purifying selection and accumulation of the most dangerous mutations is almost switchlike compared to the vast range of possible mutation effects. This means that minor changes in environmental conditions could easily switch on or off accumulation of such mutations, leading to behaviour that can be close to chaotic. Thus, although class 2 mutations do not cover the largest part of the mutational spectrum, they can easily make significant contributions to erratic behaviour of the molecular clock.

20. Huai & Woodruff (1997) "Clusters of identical new mutations can account for the "overdispersed" molecular clock", Genetics 147:339-348. - Huai & Woodruff (1998) "With the correct concept of mutation rate, cluster mutations can explain the overdispersed molecular clock", Genetics 149:467-469.
21. Lynch & Jarrell (1993) "A method for calibrating molecular clocks and its application to animal mitochondrial DNA", Genetics 135:1197-1208.
22. Charlesworth et al. (1993) "The effect of deleterious mutations on neutral molecular variation", Genetics 134:1289-1303.

The first principal objection to such an hypothesis is that overdispersion is also observed in genomes that recombine, while Muller's ratchet largely depends on absence of recombination. However, not all genomes have regular sex and even those that do, have at least some degree of linkage equilibrium that prohibits completely free recombination. Thus small local episodes where Muller's ratchet operates could contribute variance of the molecular clock. However, a careful quantification will be necessary before this hypothesis can be accepted or refuted. Probably, this involves a significant number of simulations. Should Muller's ratchet indeed turn out to contribute significantly to overdispersion of the molecular clock, then one might conceive that approaches could be constructed to narrow down the range of selection coefficients that contributes to the observed variance, by employing simulations of the ratchet. That slightly deleterious mutations have repeatedly been discussed[23] as a factor contributing to overdispersion encourages such work.

19.4 The 'neutral slowdown' of Muller's ratchet

In most practical applications of the molecular clock hypothesis, divergence time of two species is calculated by dividing half of the divergence between descendants by the substitution rate. The latter is assumed to be equal to the intergenerational mutation rate for neutral mutations[24]. As discussed in Chapter 3, this approach led to the discovery of the mutation rate paradox in mitochondrial DNA.

Complications of molecular clocks

Several arguments suggest that application of a molecular clock in detail might not be so easy. First, substitution rates are likely to be distributions, and the ratio of (the distribution of) divergence over such rates yields a different distribution, usually much wider than under the assumption of constant substitution rates[25]. Second, observation of simulated pseudogene evolution *in vitro* has illustrated effects of mutational hotspots that slow

23. Ohta (1987) "Very slightly deleterious mutations and the molecular clock", J. Mol. Evol. 26:1-6. - Ohta (1992) "The nearly neutral theory of molecular evolution", Ann. Rev. Ecol. Syst. 23:263-286. - Ohta & Gillespie (1996) "Development of neutral and nearly neutral theories", Theor. Popul. Biol. 49:128-142. - Ohta (1998) "Evolution by nearly-neutral mutations", Genetica 103:83-90.
24. Kimura (1983) "The neutral theory of molecular evolution", Cambridge, Cambridge University Press. - Parsons et al. (1997) "A high observed substitution rate in the human mitochondrial DNA control region", Nat. Genet. 15:363-368. - Loewe & Scherer (1997) "Mitochondrial Eve: the plot thickens", Trends Ecol. Evol. 12:422-423.
25. Haubold & Wiehe (2001) "Statistics of divergence times", Mol. Biol. Evol. 18:1157-1160.

down long-term substitution rates as compared to short-term mutation rates[26]. Third, the frequency of new mutations in a population appears to play a role, if one wants to consider the relationship between short-term mutation rate and long-term substitution rate in detail[27].

Observations

Assuming that each click of the ratchet results in a fixation of a new mutation[28] might add a new point to the story. When selection coefficients are so small that the ratchet runs fully deterministically, then one might envision a transition to complete neutrality. This work observed simulations that qualified in that sense, because the deterministic size of their best class was below $\mathbf{N_0} < 10^{-200}$. Figure 51 shows the corresponding results plotted over a measure for the largest possible strength of selection felt in the population. It is easy to see from that plot that, in many cases, Muller's ratchet does slow down accumulation of very slightly deleterious mutations.

Theory

Intuitively, one might expect the opposite effect when considering the simple expectation of $1/\mathbf{U}$ as clicktime, because the neutral theory predicts fixation of $1/\mathbf{U}$ mutations per generation and Muller's ratchet fixes mutations by its sampling effects, too. However, such additivity does not appear to take place. Rather the following scenario appears to be probable:

- Probability of reproduction is less than 1 for all individuals, because a Poisson distribution for the number of offspring is assumed. (From a Poisson long-term expectation of about 1 offspring per parent, a probability of about 27% for non-reproduction can be deduced).
- As long as all individuals have one mutation only, no bias occurs. However, even after a short time some distribution function describes the number of individuals in the mutational classes of a population.
- Individuals from either end of this distribution have the same probability of reproduction, because all mutations combined together have no effect or nearly no effect.
- As not all individuals reproduce, all mutational classes have the same percentage of individuals that does not reproduce.

26. Vartanian & Henry & Wain-Hobson (2001) PNAS - For a discussion of hot-spots see Chapter 3 and 26.
27. Donnelly (1991) "Comment on the growth and stabilization of populations", Stat. Sci. 6:277-279. - Siguroardottir et al. (2000) "The mutation rate in the human mtDNA control region", Am. J. Hum. Genet. 66:1599-1609.
28. Charlesworth & Charlesworth (1997) "Rapid fixation of deleterious alleles can be caused by Muller's Ratchet", Genet. Res. 70:63-73.

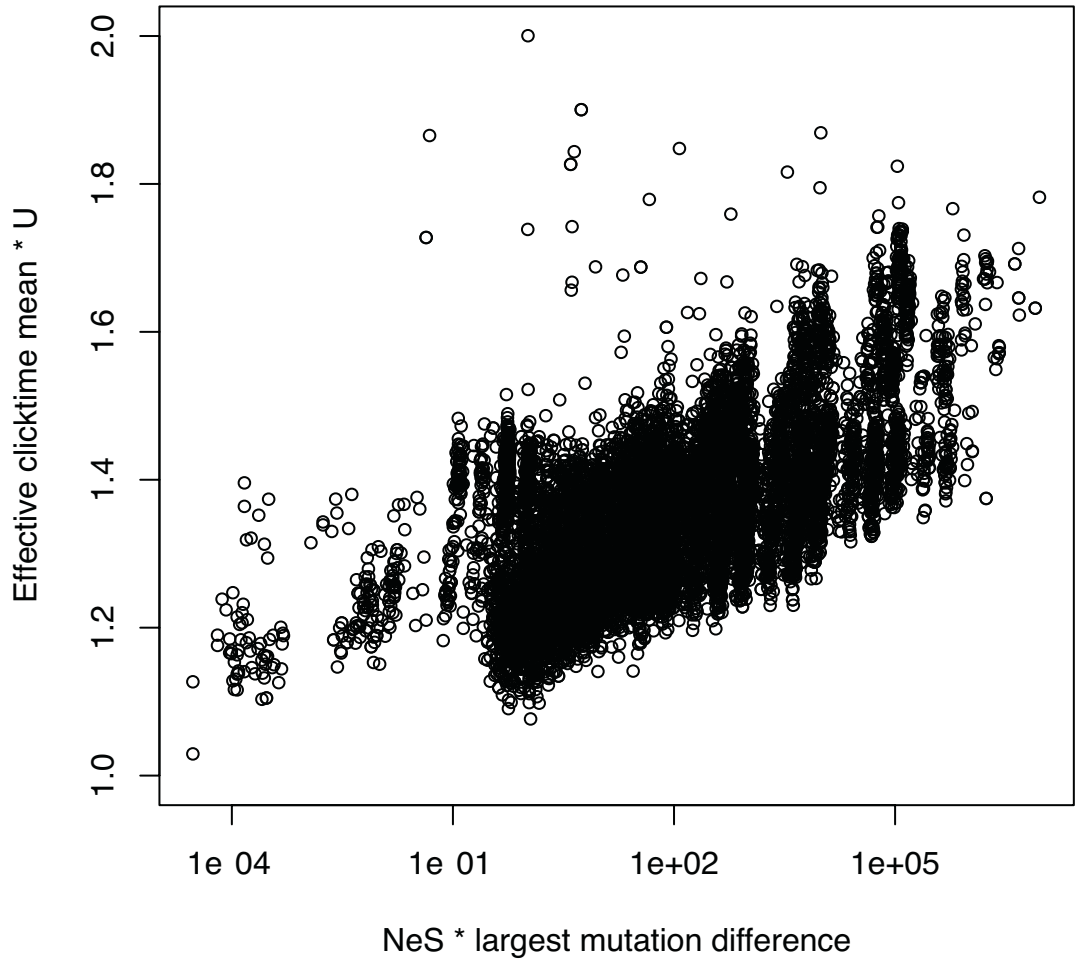


Figure 51 Neutral slowdown over maximal selection strength in population.
To quantify neutral slowdown, the effective clicktime (ie. the assumed average time between accumulation of 2 mutations in an evolutionary lineage) is scaled by the corresponding mutation rate U. The felt amount of selection in a population (N_eS) is scaled by the number differences between the highest and lowest number of very slightly deleterious mutations observed in individuals of the population in the last generation of the simulation. This is done because selection might operate between the fittest and least fit individual.
The mean slowdown of all simulations belonging to this plot is 1.36, standard deviation is 0.11, the smallest value 1.03, the largest (not shown) 2.75. The plot was produced by R on 26 Oct 2002 using all results of Simulator005 from Era1-7s1, where $U>0.001$, $N_0 < 10^{-200}$ and more than 5 clicks of the ratchet had been observed. The sign of selection coefficients was inverted.

- However, the different classes have different numbers of (neutral) mutations that they carry. Thus removal with equal probability removes few mutations from the little loaded classes, whereas many

mutations are removed with individuals from the other end of the distributions. As there are more mutational classes that carry more mutations than little loaded classes, mutation clusters tend to be removed more efficiently.

All in all, individuals with more mutations than those in the best class have a higher probability of removal by drift. Thus, drift usually removes more mutations that it would remove if the whole population only consisted of the best class. This decreases fixation probability to such an extent, that neutral mutations accumulate about 1.4 times slower than they would be expected to do with completely free recombination. Moreover, Figure 51 shows that even a very weak strength of selection can increase neutral slowdown to a more pronounced nearly neutral slowdown.

While these findings are not only of theoretical interest, they might be very important for the mutation rate paradox in (non-recombining) mitochondria. Although they certainly cannot solve the paradox, neutral slowdown can definitively contribute something to a combined solution.

19.5 Other projects

Simulator005 carries a number of parameters that have not been used in the simulations of Project 1, which was analysed in this work. These will facilitate investigation of

- a ratchet of advantageous mutations,
- the minimal advantageous mutation rate needed to stop a given deleterious ratchet (without recombination),
- 'two ratchets'[29], ie. two combinations of mutation rate and selection coefficient (once it has become clear what mutations drive the main ratchet, this might be extended to predict the behaviour of various distributions of mutational effects) and
- different distributions of mutational effects.

As all simulation results are collected in one big database, eventually analysis of these complex questions will be significantly facilitated. Towards this end a significant amount of biological work as well as IT infrastructure-development will be necessary.

29. Similar to Gordo & Charlesworth (2001) "The speed of Muller's ratchet with background selection, and the degeneration of Y chromosomes", Genet. Res. 78:149-161.

20 Compensatory mutations and Muller's ratchet

While back-mutations do indeed play no role in an infinite sites genome, they can become important when the ratchet operates in a finite sized genome over long periods of time. Then the probability increases so that an already mutated site is hit by a compensatory back-mutation. The resulting compensatory mutation rate is further increased by the fact that many deleterious mutations can be repaired by more than one specific mutation on a molecular level. Relevant biological data is discussed.

Review of beneficial mutations in Muller's ratchet

Many typical treatments of Muller's ratchet concentrate on slightly deleterious mutations (SDMs) according to the simple assumption by HAIGH[30] that the effective slightly deleterious genomic mutation rate is obtained by correcting for slightly advantageous mutations (SAMs) according to the simple equation

$$U_{sdmEffective} = U_{sdmReal} - U_{sam} \tag{56}$$

As slightly advantageous mutations are rare, such an assumption has only a minor impact on conclusions. According to the widely read introduction to Muller's ratchet by MAYNARD SMITH[31], back-mutations can be neglected, because the specific back-mutation rate is negligible, compared to the forward mutation rate for new deleterious mutations (see below for the conditions where that is true).

A number of other observations are of general interest here. Variance in mutational effects enhances longevity of asexual populations[32] and compensatory mutations in quantitative traits could effectively stop Muller's ratchet[33]. Compensatory mutations can lead to selection without adaptation[34] and models of compensatory evolution have been employed to ex-

30. Haigh (1978) "The accumulation of deleterious genes in a population - Muller's Ratchet", Theor. Popul. Biol. 14:251-267.
31. Maynard Smith (1978) "Some consequences of sex and recombination - II. Muller's ratchet", pp. 33-36 in: Maynard Smith (ed) The Evolution of Sex, New York, Cambridge University Press.
32. Lynch & Gabriel (1990) "Mutation load and the survival of small populations", Evolution 44:1725-1737.
33. Wagner & Gabriel (1990) "Quantitative variation in finite parthenogenetic populations: What stops Muller's ratchet in the absence of recombination?" Evolution 44:715-731.

plain sequence patterns in RNA[35]. Simulations have shown that only a few slightly advantageous mutations are needed to stop Muller's ratchet in a random mating population, whereas many more are needed under obligate selfing[36]. However, in asexual lineages clonal interference can slow down fixation of advantageous mutations[37]. Slightly advantageous mutations can delay extinctions in a model of alternative fixations of deleterious and advantageous mutations[38]. Finally, an unbounded fitness increase might result if a new mutation changes the effects of all previous mutations in its gene by changing their context[39].

Types of beneficial mutations

One might distinguish 3 types of beneficial mutations:

- Unconditionally **advantageous mutations**. These do not need a preceding deleterious mutation to unfold their effect.
- **CompMOLs**, Compensatory mutations on a molecular level. A deleterious point-mutation can be repaired by these compensatory mutations. If there was no preceding deleterious mutation, the potentially compensatory mutation would be deleterious itself (eg. in RNA stems, where pairing bases are OK while not pairing bases are deleterious).
- **CompQTLs**, Compensatory mutations in quantitative trait loci. These mutations compensate for deleterious effects without actually repairing the structure that has been damaged. While these can stop apparent genomic decay for quite a while[33], they cannot prevent the compensatory potential in the corresponding quantitative traits being reduced. Once this compensatory potential is used up, extinction can no longer be prevented. As the actual compensatory potential of organisms in this sense is currently unknown, no one can tell whether this could delay extinctions beyond the any critical limit.

34. Hartl & Taubes (1996) "Compensatory nearly neutral mutations: Selection without adaptation", J. theor. Biol. 182:303-309.
35. Stephan (1996) "The rate of compensatory evolution", Genetics 144:419-426. - Higgs (1998) "Compensatory neutral mutations and the evolution of RNA", Genetica 103:91-101. - Innan & Stephan (2001) "Selection Intensity Against Deleterious Mutations in RNA Secondary Structures and Rate of Compensatory Nucleotide Substitutions", Genetics 159:389-399.
36. Schultz & Lynch (1997) "Mutation and extinction: The role of variable mutational effects, synergistic epistasis, beneficial mutations, and the degree of outcrossing", Evolution 51:1363-1371.
37. Gerrish & Lenski (1998) "The fate of competing beneficial mutations in an asexual population", Genetica 103:127-144. - de Visser et al. (1999) "Diminishing returns from mutation supply rate in asexual populations", Science 283:404-406.
38. Lande (1998) "Risk of population extinction from fixation of deleterious and reverse mutations", Genetica 103:21-27.
39. Gessler & Xu (1998) "An embarrassment of riches: the stochastic generation of beneficial mutations", Genetica 103:145-155.

Clearly, these three groups exist, although it can be difficult to attribute a specific beneficial mutations to exactly one of the three. However, such a classification is only necessary when the genomic decay stopping potential of all three is assessed in one model. The following treatment concentrates on CompMOLs only.

20.1 A simple model for compensatory mutations on a molecular level in asexual genomes

Where the old theory is right

The cited assumptions that back-mutations can be neglected in treatments of Muller's ratchet[31] is true only for some very special cases:

- **Genome site is infinite.** Thus each new SDM hits a new site and the probability of a specific back-mutation is zero (1/genome size).
- **Only the last SDM can revert.** If for some reason only the last SDM can revert, old SDMs in a finite genome might be hit again, but are never repaired this way, allowing back-mutations to be neglected.
- **The ratchet has started only recently.** When Muller's ratchet starts to click in a perfect, finite genome, all new mutations are likely to hit a new site.

Limitations of the old theory

However, in the majority of cases where a consideration of Muller's ratchet is of interest, a finite genome has been accumulating SDMs for a very long time. This leads to an increase of the probability that the site of an old SDM is hit again. A second hit of a mutation does not necessarily imply repair of a former SDM, but if 1 of 4 bases is correct at a site and the first mutation changed to an incorrect base, then the next mutation at this site has a 1:3 probability of reverting to the original state (assuming only point mutations with equal transition/transversion probabilities). Furthermore, let us assume that this site is part of a stem in an RNA gene and its only function is to pair with some other site. Such a situation allows for 2 ways to repair a SDM at either site: a hit at the original SDM site has a probability of 1:3 of getting a simple back-mutation and a hit at the other side of the pair has a probability of 1:3 of getting a compensatory mutation. Now imagine that this occurs at many sites in the genome, because a considerable number of SDMs have been accumulated. In such a situation it might be conceivable that the compensatory rate of mutation could become so large that it effectively stops Muller's ratchet.

Model

The following models should help to quantify this process. Assume a finite genome with **G** sites each of which can either harbour an original (perfect) base, a SDM or a compensatory mutation. This genome accumulates slightly deleterious mutations at a rate of U_{SDM} per generation. Now assume that each SDM can be repaired in one of several molecular ways. This allows the definition of

$$Repairways = \frac{U_{SpecificComp}}{U_{SpecificSDM}} \tag{57}$$

where $U_{SpecificComp}$ denotes the total sum of mutation rates for all specific mutations that can repair the specific deleterious mutation that had occurred at the rate $U_{SpecificSDM}$. In the case of the RNA example mentioned above, one specific SDM would occur at a given rate $U_{SpecificSDM}$ and could be repaired by

$$U_{SpecificComp} = P_{AnyMuta} \cdot P_{MutaRepairs} \tag{58}$$

where $P_{AnyMuta}$ in the example equals to 2 x $U_{SpecificSDM}$ (hit at either site of the RNA base pair) and $P_{MutaRepairs}$ is 1:3, as only pointmutations without any bias were allowed. Similar equations can be constructed for more complex situations with mutational bias or potentially multiple correction possibilities in proteins. It is easy to see that the genomic deleterious mutation rate U_{SDM} is given by

$$U_{SDM} = G \cdot U_{SpecificSDM} \tag{59}$$

Now, consider a genome that is flooded with SDMs to the degree that nothing of the original sequence is left. It is hardly conceivable that the first SDM that had occurred in that genome is still in its original molecular context that would allow repair as described above. To avoid this, each genome is characterised by a ***Warranty***, that is defined as the fraction of the genome that can change without destroying repairways of the oldest SDM. Thus, the upper limit for the total genomic mutation rate for compensatory mutations on a molecular level, $U_{CompMOL}$, is given by:

$$U_{CompMOL} = U_{SDM} \cdot Warranty \cdot Repairways \tag{60}$$

Thus a simple expectation of equilibrium would predict that, if ***Warranty*** x ***Repairways*** equals one, then the accumulation of ***G*** x ***Warranty*** SDMs will lead to enough compensatory mutations to stop Muller's ratchet.

20.2 Simulation details

To investigate this model in more detail, simulations were conducted with Simulator003 that was never released publicly. Most of its details (individuals' life-history, fitness model, population density regulation, etc.) are the same as in Simulator005 that does not contain the following aspects that allow simulation of the compensatory evolution model described above.

Input parameters

Besides the usual input like environmental carrying capacity ***K***, deleterious genomic mutation rate U_{SDM} and the constant selection coefficient S_{SDM}, this simulator needed ***G***, ***Repairways*** and ***Warranty***. In addition to other data, each individual had a counter for SDM_R (SDMs that could still be repaired, because the total number of SDMs was below ***G*** x ***Warranty***) and a counter for SDM_{UR} (SDMs that were unrepairable, because the genomic context had changed too much already). Once the SDM_R had reached its upper limit, new SDMs were added to SDM_{UR}. The compensatory mutation rate for an individual was computed according to

$$U_{CompMOL} = \frac{1}{G} \cdot U_{SDM} \cdot Repairways \cdot SDM_R \qquad (61)$$

and each compensatory mutation had the corresponding effects on fitness.

Worldhistory

At the start of each simulation, a population with ***K*** individuals was created (as environments were constant, it fluctuated Poisson-like with variance ***K*** around this value, see LagCulling model of Simulator005). Each simulation was run for 500 generations to allow an initial equilibrium to develop. Then the mean fitness of the population F_{start} was measured. After another 500 generations F_{end} was measured. If

$$\Delta F = F_{end} - F_{start} \qquad (62)$$

was negative, this indicated clicking of the ratchet, whereas positive values indicate that compensatory mutations could stop the ratchet.

Measurements of critical repairways

The critical number of repairways needed to stop Muller's ratchet was determined under the assumption that ***G*** = 16000 (like in mtDNA), ***Warranty*** = 0.2 and maximal reproductive capacity = 4. Population sizes of 10,

100, 1000 or 10000 were employed for the following mutational parameters:

- 'High Ratchet': U_{SDM} = 0.4; S_{SDM} = -0.01
- 'Mid Ratchet': U_{SDM} = 0.1; S_{SDM} = -0.001
- 'Low Ratchet': U_{SDM} = 0.02; S_{SDM} = -0.0001
- 'HighULowS Ratchet': U_{SDM} = 0.4; S_{SDM} = -0.0001

'LongLow Ratchet' was identical to 'Low Ratchet' except that fitness values were compared between generations 4500 and 5000. For each of these 20 parameter combinations a list of 9 possible values for ***Repairways*** (doubling from one to the next) was simulated 10 times, observing mean and standard deviation of ΔF. If these fitness changes are plotted over the 9 repairways tested, one could estimate by eye a minimum, mean and maximum of the number of repairways that barely stopped the ratchet (ie. ΔF = 0, plots not shown). This critical number of repairways is discussed below.

20.3 Simulation results

The results of simulations analysed this way can be found in Figure 52. The following observations can be made from the simulations.

Confirmation of expected equilibrium

The theoretically expected equilibrium could be easily found to evolve in the simulations (see Mid, HighULowS and LongLow Ratchet). It evolves when selection is so weak that population size has no influence on the rate of Muller's ratchet and the process has been running long enough for equilibrium to evolve.

Selection reduces critical repairways

When the strength of selection becomes appreciable, the critical number of repairways needed to stop Muller's ratchet is reduced with increasing population size. This is expected, as increasing population size increases the effectiveness of selection and thus fewer compensatory mutations need to occur to counter all SDMs.

Compensatory delay effect

The initial result (observing generations 500 and 1000) of the 'Low Ratchet' parameter combination was hard to understand. How could increasing population size ever lead to a reduced efficiency of compensatory evolution in stopping the ratchet? As no bug in the code could be found to explain this, observing the same runs between generations 4500 and 5000 helped to solve the puzzle. These runs showed the same normal equilibrium expectation seen in other runs with weak selection. Thus the effect appeared to be transient only. It was termed compensatory delay effect and is defined as follows:

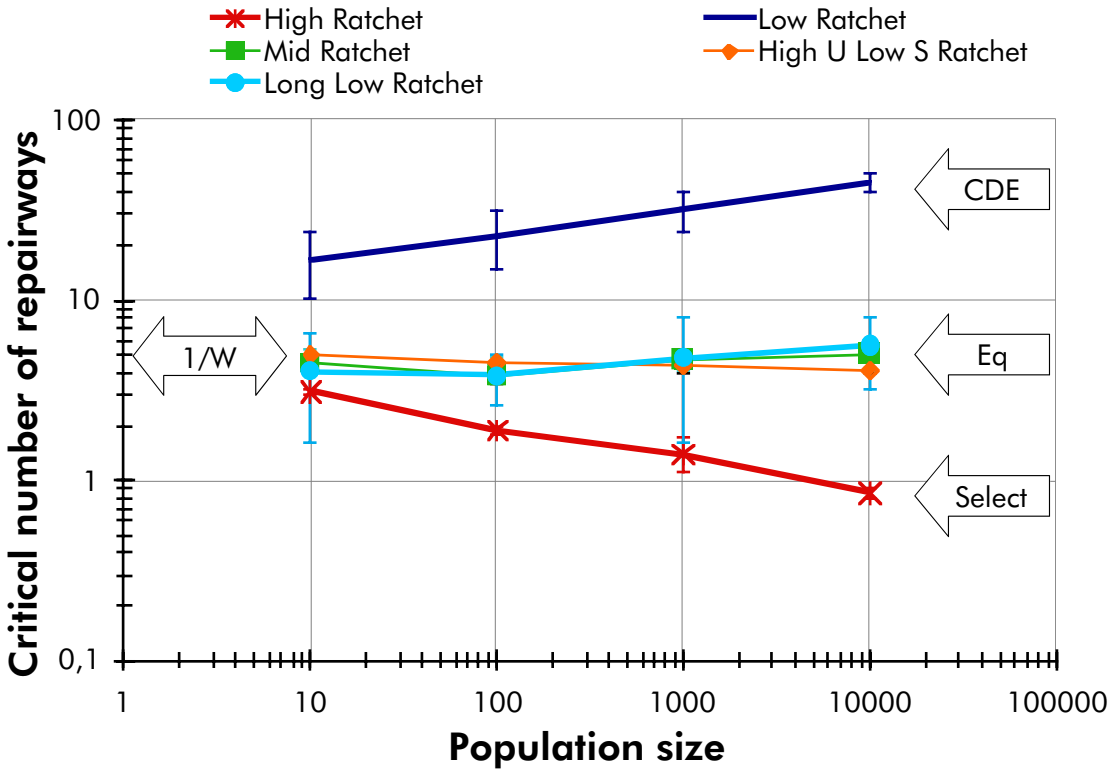


Figure 52 Critical number of repairways needed to stop Muller's ratchet by compensatory evolution. **1/W** marks the expected equilibrium (**Eq**). **Select** marks the influence of selection that increases with population sizes. **CDF** denotes the compensatory delay effect. See text for details on parameter combinations.

Definition CDE

The compensatory delay effect describes the situation where a larger population size demands a larger number of molecular repairways to compensate initial loss of fitness due to Muller's ratchet by compensatory evolution at a molecular level, if selection is very weak and recombination is absent.

The population genetic reasons for its appearance are as follows:

- Fixation time T_{fix} for nearly neutral CompMOLs increases with effective population size N_e ($T_{fix} \approx N_e$ in maternally inherited haploids).
- This increases the probability of CompMOL neutralisation by SDMs, because $T_{fix} * U_{SDM}$ mutations can be expected in each descending line.

- This increase in probability can be stopped by a higher number of repairways (as seen in the initial simulations) or by more individuals with repairable SDMs (see longer runs).

Recombination and stronger selection can be expected to reduce the compensatory delay effect to insignificance.

20.4 What are realistic values for *Repairways* and *Warranty* in biological genomes?

Compensatory evolution can be observed in a virus

It will neither be easy to estimate the average number of ***Repairways*** for a mutation in a genome nor will it be easy to know the ***Warranty*** of a genome. However, the definitions above might allow for eventual success in this search. Such hope is fuelled by Burch & Chao's work[40] on the rugged fitness-landscape of the RNA virus $\phi 6$. The study investigated the recovery from a deleterious mutation by compensatory mutations at different population sizes. After 25 generations of mutation accumulation (5 serial single phage transfers) a sharp decline in fitness (ca. 90%) had occurred, most probably due to a single deleterious mutation. This corresponds to a deleterious genomic mutation rate of ca. 4% and about 400 sites that might have similar effects, if the general mutation rate for RNA viruses of 1/genome/generation is applied to the $\phi 6$ genome size of $G = 10^4$ bp. The large decline in fitness allowed for a compensatory fitness increase of ca. 900%; thus there should be ample opportunity to observe compensatory evolution.

After 15 more generations (8 single phage bottlenecks since start of experiment), 7 different populations for recovery were founded and propagated, each with a different bottleneck size (10, 33, 100, 333, 1000, 2500 or 10000). These populations were maintained for 100 generations or until fitness had recovered (as measured in a single particle picked from the population to avoid measurement of intermediate mean fitness increases due to polymorphism). The following general scenarios could be observed:

General recovery patterns

- **Recovery in a single step**. For bottleneck sizes of 1000 and 10 000 apparently a single mutation had occurred after 25 and 15 generations, respectively.
- **Recovery in one big step and small follow-ups**. For bottleneck sizes of 100 and 2500 the first big, but not complete recovery muta-

40. Burch & Chao (1999) "Evolution by small steps and rugged landscapes in the RNA virus phi 6", Genetics 151:921-927.

tion had occurred at generation 25 and 20, respectively. It was followed by at least one other small mutation.

- **Recovery in many steps**. For bottleneck sizes of 33 and 333 a more or less gradual recovery pattern was observed, where smaller mutations made the start and at least 1-3 other small or large mutations followed.
- **No recovery in 100 generations**. For the bottleneck size of 10, probably a small step of compensatory increase in fitness could be observed, but recovery from the deleterious mutation was not even close to complete during the 100 generations of observation (10 generations per bottleneck).

Nature of the mutations observed

There is some room for debate whether the multi-step recoveries were due to CompMOLs in the sense described here, as the molecular nature of these mutations is not known. Thus unconditionally adaptive mutations or mechanistically completely unrelated compensatory mutations (some viral QTL substitute) could have generated the patterns observed. However, in those cases where a single step led to complete recovery, one is compelled to assume a direct compensatory mutation at the molecular level. Cases in which one big initial step made up for the largest part of the recovery might indicate a compensatory mutation that used a different repairway (not as direct and efficient as complete full repair, but still very close to CompMOLs as defined in this chapter).

Estimation of molecular repairways

While these data do not allow estimation of the ***Warranty*** parameter, they allow calculation of upper limits for the ***Repairways*** in this virus. To estimate a very rough lower limit of the number of particles that could have contributed such a mutation, one can multiply bottleneck size by the number of generations before appearance of the compensatory mutation. This value carries a significant downward bias, because the bottleneck occurred only every 5^{th} generation (for population sizes of 33 and larger) and evolution in that time was neglected. Viruses that were generated immediately after the bottleneck probably still had a significant chance of contributing a compensatory mutation because of selection, while viruses that generated the compensatory mutation only shortly before reaching the final population size of 8×10^9 particles per plaque had a significant probability of being lost by drift.) If the lower limit of the number of particles is estimated this way, the following values can be obtained:

- o *BS=100, GC=25, NC=2500, CMR=1/2500 , R=4
- o BS=1000, GC=25, NC=25000, CMR=4x10^{-5} , R=0.4
- o *BS=2500, GC=20, NC=50000, CMR=2x10^{-5} , R=0.2
- o (*)BS=10000, GC=15, NC=150000, CMR=7x10^{-6} , R=0.07

where * marks observations with incomplete (but large) recovery, **BS** is bottleneck size, **GC** is the first generation after the large compensatory mutation had occurred, **NC** is the lower limit of the number of viruses that could have contributed the compensatory mutation, **CMR** is the apparent total compensatory mutation rate and **R** is the number of ***Repairways*** estimated from dividing **CMR** by an assumed deleterious mutation rate of 10^{-4} per base per generation.

Conclusions

Although these values can be regarded as very rough estimates only, most of them show significant similarity to what would be expected from a simple back-mutation. Although the highest number of repairways found in these calculations stems from a mutation that led to incomplete recovery, it is probably close to the ***Repairways*** needed to stop Muller's ratchet by compensatory evolution alone. However, most of the estimates suggest that compensatory evolution at the molecular level is probably not strong enough to effectively stop Muller's ratchet in the long-term, except perhaps in combination with selection or other potential solutions[41]. Further experiments of this type might lead to more robust estimates.

41. See "The fitness-balance and the genomic decay paradox" on page 335

VI. Results Part 3: Biological Consequences of Muller's Ratchet

Although recombination is almost universal, some genetic systems on earth seem to come along without it. The simulations presented here predict extinction times for various non-recombining evolutionary lines due to Muller's ratchet. For some of these systems, recombination appears to be more important than previously thought; for others there is no simple solution for the resulting genomic decay paradox.

21 Muller's ratchet in mtDNA might cause extinctions in mammals

Mitochondrial DNA is a very common asexual genetic system in otherwise sexual species. The observation of relatively high mitochondrial mutation rates in human pedigrees has led to the question of how such a genetic system could survive the resulting onslaught of Muller's ratchet. Here, the threat from Muller's ratchet is quantified in detail for the first time. It is concluded that for many mammals and humans a range of biologically realistic parameter combinations should have led to extinctions within 20 million years. Some implications for endangered species and evolution of the human line are discussed.

Mitochondria play a key role in supplying cells with energy and are a subject of intense research[1]. They have their own, maternally inherited, tightly packed 16 Kbp genomes that encode for proteins that are essential for the function of mitochondria[2]. Medical genetics has shown clear relationships between some mutations on mitochondrial DNA (mtDNA) and well-

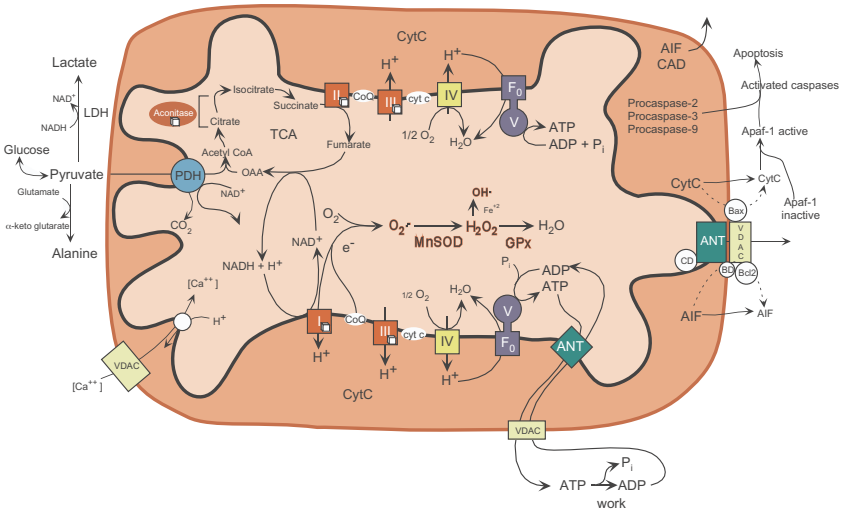


Figure 53 Overview over key metabolic functions of mitochondria. From MITOMAP[2] 1999.

1. Anonymous (1999) "Mitochondria", Science 283:1435-1438+1475-1498.
2. MITOMAP (1999) "MITOMAP Mitochondrial Genomics Database" http://www.gen.emory.edu/mitomap.html

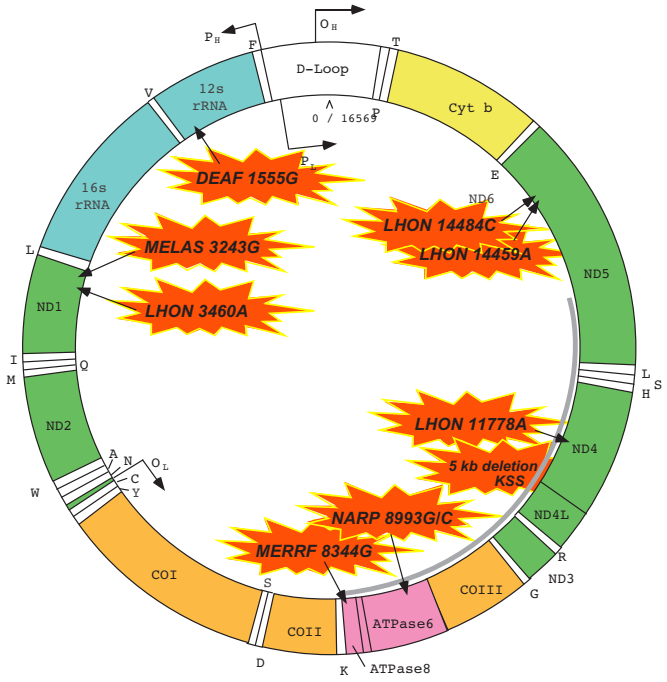


Figure 54 Overview of the mitochondrial genome and known mutations.
The arrows with the red bursts mark some known diseases. The control region (D-loop) has two hypervariable regions, which are often used for evolutionary studies because they contain a large fraction of (supposedly) neutral mutations. Thus, they are over-represented in the statistics of observed mutations. Source of data: manual evaluation of MITOMAP[2] as of 31 Oct 2000.

known diseases[3], underscoring the large potential for deleterious mutations in mtDNA.

Inheritance

Genetics of mtDNA appears simple at a distance but is extremely complicated in detail. From afar, mtDNA looks like a haploid, maternally inherited genome that is selectively neutral. Many phylogenetic studies use it as this, especially since mtDNA is easier to handle experimentally than nuclear DNA. As reality is very complex, each model of mtDNA evolution should be explicit about its simplifying assumptions on

- **Multi-level population genetics**. Many mtDNA molecules replicate in one mitochondrion, many mitochondria replicate in one cell and many cells replicate in one sexual generation. Selection and drift operate at each replicator level and different levels can even be in conflict with each other[4]. These multi-level population genetic aspects of

3. Wallace (1999) "Mitochondrial diseases in man and mouse", Science 283:1482-1488.

mtDNA evolution are being studied[5], but are not considered here. The fact that the mutation rates applied here are not measured in mtDNA directly, but rather average over all details by comparing molecules separated by few generational events in pedigrees or many in phylogenies.

- **Heteroplasmy**. As a direct consequence of multi-level population genetics, transition of one allele to another can take much longer than one sexual generation. Thus individuals can carry two or more different sequence versions of mtDNA. To simplify analysis, heteroplasmies are omitted from the mutations used to compute pedigree rates. This view is supported by the fact that many transitions from one allele to another are very fast owing to a very small mtDNA bottleneck[6] in the germ line.
- **Hot-spots**. Mutation rates vary for different positions[7]. However, the variation observed does not completely explain the difference between pedigree and phylogeny rates[8]. Therefore, the known hot-spots are omitted from calculations of mutation rate and other minor dif-

4. Szathmáry & Maynard Smith (1995) "The major evolutionary transitions", Nature 374:227-232. - Mayr (1997) "The objects of selection", Proc. Natl. Acad. Sci. USA 94:2091-2094.
5. Birky (1991) "Evolution and population genetics of organelle genes: Mechanisms and models", pp. 112-134 in: Selander et al. (eds) Evolution at the Molecular Level, Sunderland, MA, Sindauer Associates, Inc. - Bergstrom & Pritchard (1998) "Germline bottlenecks and the evolutionary maintenance of mitochondrial genomes", Genetics 149:2135-2146. - Kondrashov (1994) "Mutation load under vegetative reproduction and cytoplasmic inheritance", Genetics 137:311-318. - Otto & Hastings (1998) "Mutation and selection within the individual", Genetica 103:507-524. - Otto & Orive (1995) "Evolutionary consequences of mutation and selection within an individual", Genetics 141:1173-1187. - Hastings (1989) "Potential germline competition in animals and its evolutionary implications", Genetics 123:191-198. - Hastings (1991) "Germline-selection: Population genetic aspects of the sexual/asexual life cycle", Genetics 129:1167-1176. - Hastings (1992) "Population genetic aspects of deleterious cytoplasmic genomes and their effect on the evolution of sexual reproduction", Genet. Res. 59:215-225. - Takahata & Slatkin (1983) "Evolutionary dynamics of extranuclear genes", Genet. Res. 42:257-265. - For non-mtDNA systems with similar dynamics, see Rispe & Moran (2000) "Accumulation of deleterious mutations in endosymbionts: Muller's ratchet with two levels of selection", Am. Nat. 156:425-441. - Paulsson (2002) "Multileveled selection on plasmid replication", Genetics 161:1373-1384.
6. Jenuth et al. (1996) "Random genetic drift in the female germline explains the rapid segregation of mammalian mitochondrial DNA", Nat. Genet. 14:146-151. - Koehler et al. (1991) "Replacement of bovine mitochondrial DNA by a sequence variant within one generation", Genetics 129:247-256.
7. Pääbo (1996) "Mutational hot spots in the mitochondrial microcosm", Am. J. Hum. Genet. 59:493-496. - Wakeley (1993) "Substitution Rate Variation Among Sites in Hypervariable Region 1 of Human Mitochondrial DNA", J. Mol. Evol. 37:613-623. - Wakeley (1996) "The excess of transitions among nucleotide substitutions: New methods of estimating transition bias underscore its significance", Trends Ecol. Evol. 11:158-163. - Meyer et al. (1999) "Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA", Genetics 152:1103-1110.
8. Parsons et al. (1997) "A high observed substitution rate in the human mitochondrial DNA control region", Nat. Genet. 15:363-368.

ferences in rates are ignored. This results in only a slight reduction of observed mutation rates.

The model used here further assumes no paternal leakage, no recombination on any level and uniform mutation rates with equal mutation effects.

This study is not the first to suspect that mitochondria might be a dangerous genetic system[9]. One of the early papers on mutational meltdown had a section that asked how mtDNA in general might cope with the threat from Muller's ratchet[10]. The first study that reported the surprisingly high mutation rates in pedigrees asked how mitochondria could possibly survive the resulting onslaught of Muller's ratchet[11]. A comparison of the binding stability and other features of tRNAs from mtDNA versus nuclear DNA showed that mitochondria accumulate more slightly deleterious mutations than the nucleus[12]. While these studies strongly suggest that Muller's ratchet should be in operation in mtDNA and even present the general layout of the U-shaped function of extinction time over selection coefficients[10], it is extremely difficult, if not impossible, to quantify the threat of the ratchet in mtDNA in detail from the literature alone.

The aim of this chapter is to use biologically realistic parameter combinations to quantitatively estimate the threat from Muller's ratchet in mammal and human mtDNA.

21.1 Measures for quantifying the threat of Muller's ratchet in principle

U-shaped plot

In order to allow detailed assessment of the threat from Muller's ratchet in a particular biological system with a given effective population size $\mathbf{N_e}$ and a given genomic deleterious mutation rate per generation $\boldsymbol{U}$, consider the U-shaped plot of extinction time over selection coefficient (see Figure 10 on page 46). Since the least is known about selection coefficients, these are used together with resulting extinction times to quantify the threat by using the following measures:

9. Hastings (1992 "Population genetic aspects of deleterious cytoplasmic genomes and their effect on the evolution of sexual reproduction", Genet. Res. 59:215-225) investigated potential conflicts of selection.
10. Gabriel et al. (1993) "Mullers Ratchet and mutational meltdowns", Evolution 47:1744-1757.
11. Howell et al. (1996) "How rapidly does the human mitochondrial genome evolve?" Am. J. Hum. Genet. 59:501-509.
12. Lynch (1996) "Mutation accumulation in transfer RNAs: Molecular evidence for Muller's ratchet in mitochondrial genomes", Mol. Biol. Evol. 13:209-220. - Lynch (1997) "Mutation accumulation in nuclear, organelle, and prokaryotic transfer RNA Genes", Mol. Biol. Evol. 14:914-925.

MDRS

- $MDRS@T_{ex,min}$. The **M**ost **D**angerous **R**ange of **S**election coefficients denotes those mutational effects that have the shortest extinction times $T_{ex,min}$ within a factor of 2 ('the width of the bottom of the U').

DRS

- $DRS@T_{AgeOfLine}$. The **D**angerous **R**ange of **S**election coefficients denotes those mutational effects that lead to extinction within the estimated age of the evolutionary line considered. This is not necessarily the age of the species, as the number of deleterious mutations is not reset to zero when the border to a new species is crossed in the eyes of a modern taxonomist. While exact values for $T_{AgeOfLine}$ are often hard to get, it will be easy to see that Muller's ratchet will pose no threat when predicted extinction times are a multiple of the age of the earth. In particular the border with purifying selection is such that there is quite a drastic switch from the most dangerous selection coefficients to evolutionarily harmless purifying selection.

In any real biological system investigated today a large degree of uncertainty also exists regarding the other parameters needed to compute extinction times:

- Effective population size N_e,
- Generation time T_{gen},
- Maximal reproductive capacity R_{max},
- Deleterious genomic mutation rate U,

Therefore any detailed assessment of the threat of Muller's ratchet needs to take the upper and lower limits for these values into account. This was done for the following plots. They use the mean effective clicktime of ratchet mutations of Simulator005 together with the means of T_{gen} and R_{max} to estimate T_{ex} for a given simulation (fixed U and N_e). Simulations for various N_e are all plotted with the same symbol that codes for a specific mutation rate. In order to provide a feeling for the errors in T_{ex} due to errors in N_e, T_{gen} and R_{max} simple *Equation172* was used to draw upper and lower limits (thick lines) for the mean (thin lines) extinction times.

J-shaped plot

Sometimes the range of biologically realistic effective population sizes was larger than the range of population sizes that could be computed by Simulator005 due to RAM or CPU time limitations. This is no problem for regions where either drift or purifying selection completely dominate the picture, as population size does not really change the speed of the ratchet there. However, as the most dangerous range of selection coefficients is very close to where purifying selection completely blocks any operation of the ratchet, it is sometimes interesting to see whether larger population sizes

would have an influence on extinction times for a given combination of U and selection coefficient s. Such an overview is given by the J-shaped plots of extinction time over N_e. From them one can easily see whether a thousandfold increase in N_e would effectively stop the ratchet or not. The J-shaped plots allow computation of a safe minimal population size defined as follows:

Critical population size

> *The* **critical population size** $N_{e,crit}$ *is defined as the smallest possible effective population size that does prevent extinction within the assumed age of the evolutionary line due to Muller's ratchet for a given combination of* U *and* s .

Three qualitative behaviours are found in the J-shaped plot. For drift-dominated combinations, no possible increase of N_e can stop the ratchet. For selection-dominated combinations an extremely small N_e effectively stops the ratchet. For the transition between both, one will have to check in detail whether a computed $N_{e,crit}$ is still realistic for a given species.

21.2 Biological parameter values

Mutation rates

This chapter focuses on mtDNA in mammals in general and in humans in particular. It is assumed that the mutation rates found in the D-loop of human pedigrees are within the same order of magnitude as mutation rates for the rest of the mitochondrial genome and that these rates are not highly specific to humans, but rather apply to a broader range of species as suggested in Chapter 3. This is based on two assumptions:

- Human mtDNA biology is nothing exceptional compared to other mammal mtDNAs.
- The mutation rate that is relevant for Muller's ratchet type studies is the intergenerational rate at which complete transitions to new mutations keep appearing, given that they are not known hot-spots.

For comparison, results based on phylogenetic mutation rates are given as well. When calculating deleterious mutation rates for use in ratchet theory, one has to make assumptions concerning the distribution of mutational effects. One cannot assume that all mutations have the same deleterious effect in reality, so instead of applying an arbitrary distribution here, it is equally arbitrarily assumed that either

Distributions of mutational effects

- about 90 % of all mutations have the same deleterious effect and all other mutations can be ignored or that

Table 16 The range of biologically realistic mutation rates in mtDNA.

Substitution rate[a]	Meaning	Deleterious mutation rate[b]	Transformation comment[c]
2.5	First pedigree estimate for coding region[d]	0.08 (0.72)	= x20a/1Myr*16Kbp* 10% (90%) deleterious
1.07	Most complete pedigree estimate for D-loop[e]	0.034 (0.31)	= x20a/1Myr*16Kbp* 10% (90%) deleterious
0.32	Lowest pedigree estimate for D-loop[f]	0.01 (0.092)	= x20a/1Myr*16Kbp* 10% (90%) deleterious
0.26	Highest phylogeny estimate for D-loop[g]	0.0083 (0.075)	= x20a/1Myr*16Kbp* 10% (90%) deleterious
0.1	Typical phylogenetic estimate for D-loop	0.0032 (0.029)	= x20a/1Myr*16Kbp* 10% (90%) deleterious
0.025	Lowest phylogeny estimate for D-loop[g]	0.0008 (0.0072)	= x20a/1Myr*16Kbp* 10% (90%) deleterious
0.017	Phylogenetic estimate of coding region[h]	0.00054 (0.0049)	= x20a/1Myr*16Kbp* 10% (90%) deleterious

Meaning of values used for simulations

Upper limit for deleterious mutation rate based on pedigrees	0.5
Most probable deleterious mutation rate based on pedigrees	0.1
Upper limit for deleterious mutation rate based on phylogenies	0.05
Lower limit for deleterious mutation rate based on pedigrees	0.01
Most probable deleterious mutation rate based on phylogenies	0.005
Lower limit for deleterious mutation rate based on phylogenies	0.001

a. per basepair per million years
b. per whole mitochondrial genome per generation assuming that 10 % of all mutations are deleterious within a given order of magnitude from 10^{-1} to 10^{-9} (values in brackets assume 90 % to be deleterious).
c. Formula = Substitution rate * Generation time in years (Tgen) / 1 million years * mtDNA genome size * fraction of deleterious mutations. The 20 years generation time are from the original pedigree studies.
d. Howell et al. (1996) "How rapidly does the human mitochondrial genome evolve?" Am. J. Hum. Genet. 59:501-509.
e. See Chapter 3; combines data of Parsons & Holland (1998) "Mitochondrial mutation rate revisited: hot spots and polymorphism - Response", Nat. Genet. 18:110-110 and Siguroardottir et al. (2000) "The mutation rate in the human mtDNA control region", Am. J. Hum. Genet. May 66:1599-1609.
f. Siguroardottir et al. (2000), ibid.
g. as cited by Parsons et al. (1997) "A high observed substitution rate in the human mitochondrial DNA control region", Nat. Genet. 15:363-368.
h. Ingman et al. (2000) "Mitochondrial genome variation and the origin of modern humans", Nature 408:708-713.

- about 10% of all mutations have the same deleterious effect within an order of magnitude and all other mutations are either removed by purifying selection or are effectively neutral or belong to another order of magnitude of selection coefficients.

As next to nothing is known about selection coefficients in mtDNA, an array of different mutational effects are combined with the resulting mutation rate. The clicktime of the ratchet is simulated separately for each combination assuming that these mutations are the only ones that occur. How two different ratchets (different mutation rates and different selection coefficients) combine will have to be investigated elsewhere[13].

The resulting rates of mutation with equal effects are then used as an input parameter for Simulator005 that computes effective clicktimes of the ratchet, which in turn are used to estimate extinction times. Table 16 shows how substitution rates discussed in Chapter 3 translate into ratchet-related intergenerational deleterious genomic mutation rates and vice versa.

Effective population size

Effective population sizes for humans are believed to be around 10 000 for the largest part of evolution of the human line[14]. To be on the safe side, this is used as lower limit with 100 000 as upper limit. These values translate into 5000 - 50 000 individuals in the simulations, as only females need to be simulated. For mammals in general, 10^3 - 10^6 simulated females are assumed to place safe bounds on the published estimate of 10^4 - 10^5 (see[15]).

Generation time

Generation time in humans is generally assumed to be 20 years. However, as we go back in time, it is likely to be lower, if one considers that primates have considerably lower generation times[16]. Therefore, the lower limit assumed for humans is 10 years, with a mean of 15 years. For mammals in general generation times of 1 (min), 3 (mid) and 10 (max) years are assumed.

Maximal reproductive capacity

Maximal reproductive capacity $\boldsymbol{R_{max}}$ in this case is the Poisson *expectation* of the maximal number of females that one female can produce under optimal conditions, ie. half the number of total offspring. For humans, the

13. For a start, see the work of Gordo & Charlesworth (2001) "The speed of Muller's ratchet with background selection, and the degeneration of Y chromosomes", Genet. Res. 78:149-161.
14. Rogers (2001) "Order emerging from chaos in human evolutionary genetics", Proc. Natl. Acad. Sci. USA 98:779-780. - Rogers (1995) "Genetic evidence for a pleistocene population explosion", Evolution 49:608-615. - Takahata (1993) "Allelic genealogy and human evolution", Mol. Biol. Evol. 10:2-22.
15. Kondrashov (1995) "Contamination of the Genome by Very Slightly Deleterious Mutations: Why Have We Not Died 100 Times Over?" J. theor. Biol. 175:583-594.
16. Allman et al. (1998) "Parenting and survival in anthropoid primates: caretakers live longer", Proc. Natl. Acad. Sci. USA 95:6866-6869.

values assumed are 8 (min), 10 (mid) and 12 (max) female offspring per generation, certainly an upper limit. For mammals 10 (min), 20 (mid) and 40 (max) are assumed.

Age of line

It is difficult to fix a point in time where the mammal or human evolutionary line came into existence and was not only named as such. If treated with the proper caution, the general appearance of the related evolutionary lines in the fossil record might be taken as an estimate of the time since which deleterious mutations have been accumulating. In this sense, the assumed ages of mammals are 50 (min), 100 (mid) and 200 (max) million years and 2 (min[17]), 10 (mid) and 20 (max) million years for humans.

21.3 Extinction times

Mammals

If the values given above for mammals are used to compute the U-shaped plot of extinction time over selection coefficient, it becomes apparent that a significant part of biologically realistic parameter combinations leads to extinctions of the corresponding species within the assumed age of their existence (Figure 55). This is not necessarily changed when larger population sizes are assumed as can be seen in Figure 56 and Figure 57. However, these figures show that increased mutation rates can be extremely dangerous.

Quantifying the threat from the ratchet for mammalian mtDNA in detail for $\boldsymbol{U}$ = 0.1 and a mean of all other values yields

- MDSR = -0.01 to -0.005 at ca. 100 Kyr and
- DSR = -0.02 to -10^{-6} at 100 Myr,

where as $\boldsymbol{U}$ = 0.01 under the same conditions makes values drop to

- MDSR = -0.0005 to -0.003 at ca. 5 Myr and
- DSR = -0.005 to -10^{-5} at 100 Myr,

as can be seen from the corresponding thin lines in Figure 55.

These results do not mean that all mammals should have gone extinct due to Muller's ratchet in mitochondrial DNA, but they imply that a significant fraction might have. Although we do not know the exact values for most parameters involved in the calculation, the upper and lower limits employed suggest that Muller's ratchet might contribute significantly to observed background extinction rates[18]. To compute the exact magnitude of this contribution, the corresponding ratchet parameters would have to be known for each (living and extinct) species.

17. Templeton (2002) "Out of Africa again and again", Nature 416:45-51.
18. Regan et al. (2001) "The currency and tempo of extinction", Am. Nat. 157:1-10.

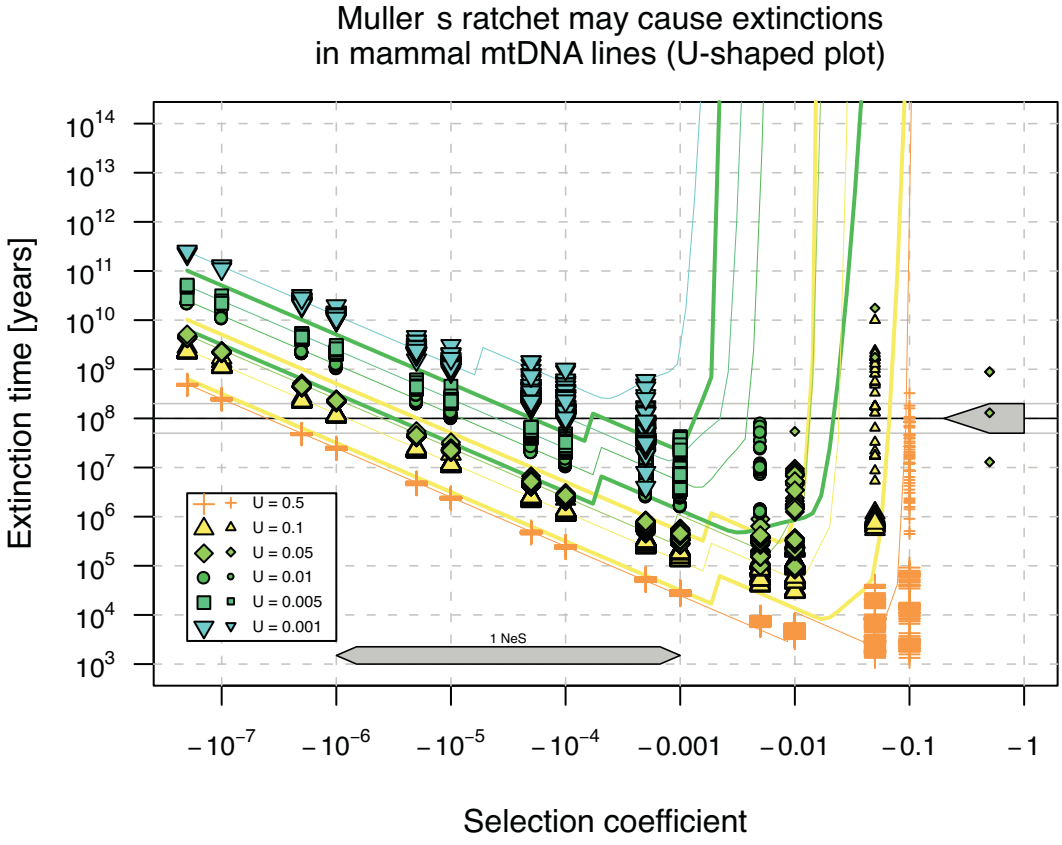


Figure 55 Muller's ratchet might cause extinctions of mammal species due to accumulation of slightly deleterious mutations in mtDNA.

The upper arrow denotes the assumed age of the lines, including limits. The lower arrow marks the border with neutrality for the population sizes used by spanning the selection coefficients from 1 N_eS for the largest (10^6 females) to 1 N_eS for the smallest population (10^3 females). Thin lines use simple *Equation172* (see Chapter 18) with N_e = 31600 and the respective means of T_{gen} and R_{max}. Thick lines for U=0.1 (yellow, higher most probable mutation rate estimate) and for U = 0.01 (mid-green, lower most probable estimate) use the corresponding upper and lower limits of N_e, T_{gen} and R_{max} to provide a feeling for the variability of extinction time estimates. Large symbols denote valid extinction time estimates from simulations with at least 2 observed clicks (usually much more). Small symbols denote lower limits for extinction times from simulations without observed clicks, based on the (usually wrong) assumption that the ratchet would have clicked just after stopping the simulation. Each symbol denotes an independent simulation with a different random seed.

This plot contains 13065 simulations with a total of 9.83 years of computing time and was produced by script RS005LL012 at 2002-10-08 15h36m22s. It is one of two plots in this work that includes results from Simulator005 releases 1-5 with U=0.001, although there was an irregularity with the Poisson RND generator that led to increased variance of effective clicktimes, but still had an effectively correct mutation rate.

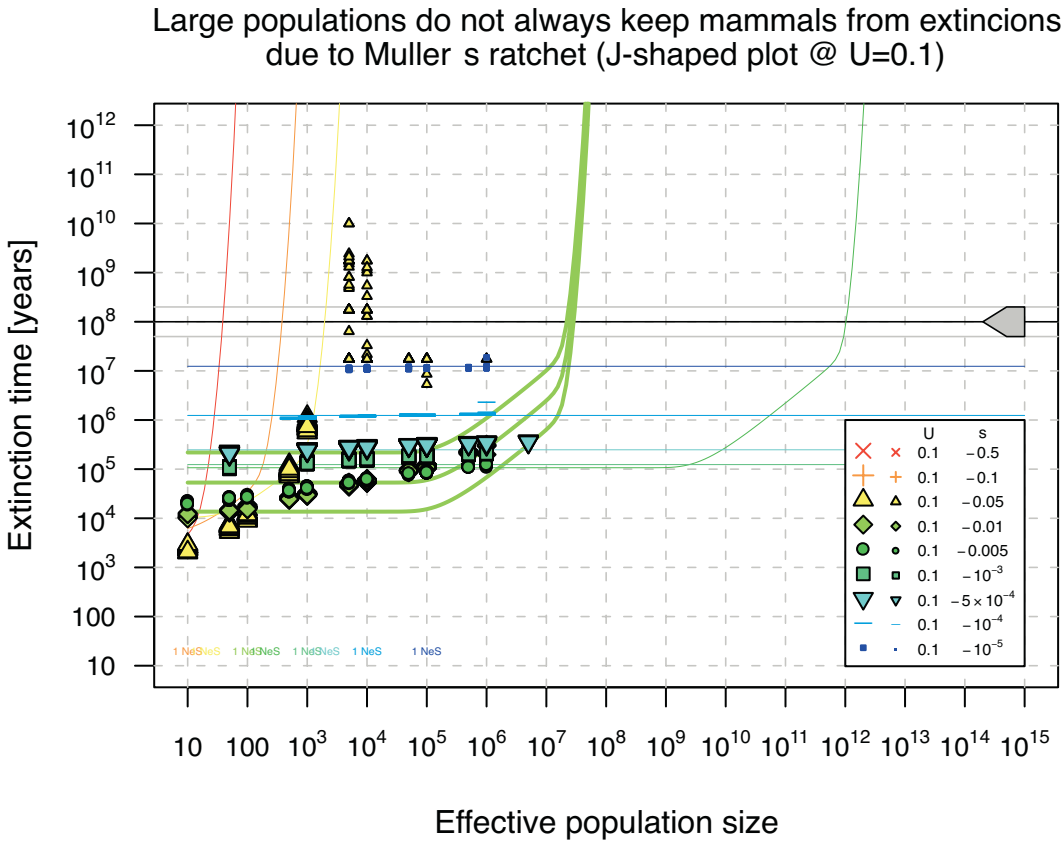


Figure 56 Large populations do not necessarily keep mammals from extinction due to Muller's ratchet in mtDNA (J-shaped plot for U= 0.1).
The arrow denotes the assumed age of the lines, including limits. The small coloured "1 NeS"-labels mark the border with neutrality for the selection coefficients corresponding to their colour. Thin lines use simple *Equation172* (see Chapter 18) with the respective means of T_{gen} and R_{max}. Thick lines mark the most dangerous parameter combination (U=0.1; s = -0.01) using the corresponding upper and lower limits of T_{gen} and R_{max} to provide a feeling for the variability of extinction time estimates. Large symbols denote valid extinction time estimates from simulations with at least 2 observed clicks (usually many more). Small symbols denote lower limits for extinction times from simulations without observed clicks, based on the (usually wrong) assumption that the ratchet would have clicked just after stopping the simulation. Each symbol denotes an independent simulation with a different random seed.
This plot contains 2310 simulations with a total of 1.74 years of computing time and was produced by script RS005LL012 at 2002-10-08 15h40m50s.

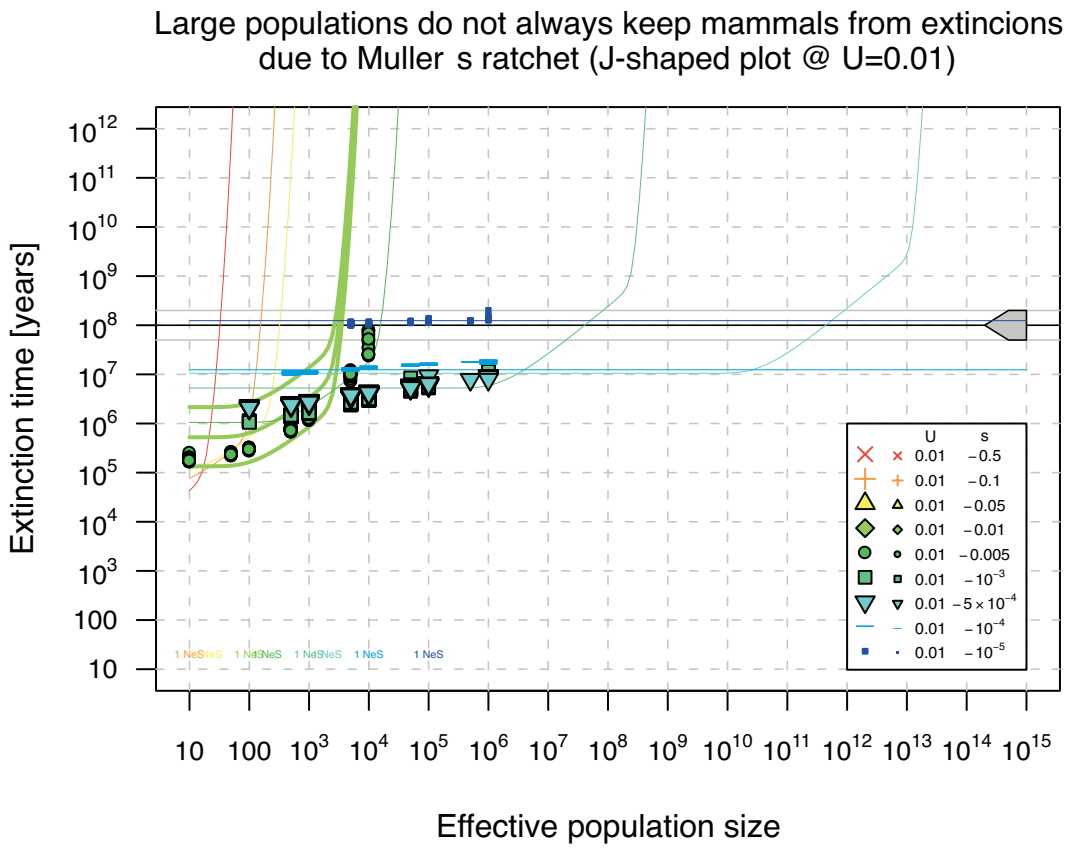


Figure 57 Lower mutation rates do much more to keep mammals from extinction due to Muller's ratchet in mtDNA than larger population sizes (J-shaped plot for U= 0.01).
The arrow denotes the assumed age of the lines, including limits. The small coloured "1 NeS"-labels mark the border with neutrality for the selection coefficients corresponding to their colour. Thin lines use simple *Equation172* (see Chapter 18) with the respective means of T_{gen} and R_{max}. Thick lines mark the same selection coefficient as in Figure 56 (s = -0.01) using the corresponding upper and lower limits of T_{gen} and R_{max} to provide a feeling for the variability of extinction time estimates. Large symbols denote valid extinction time estimates from simulations with at least 2 observed clicks (usually many more). Small symbols denote lower limits for extinction times from simulations without observed clicks, based on the (usually wrong) assumption that the ratchet would have clicked just after stopping the simulation. Each symbol denotes an independent simulation with a different random seed.
This plot contains 1397 simulations with a total of 1.06 years of computing time and was produced by script RS005LL012 at 2002-10-08 15h43m17s.

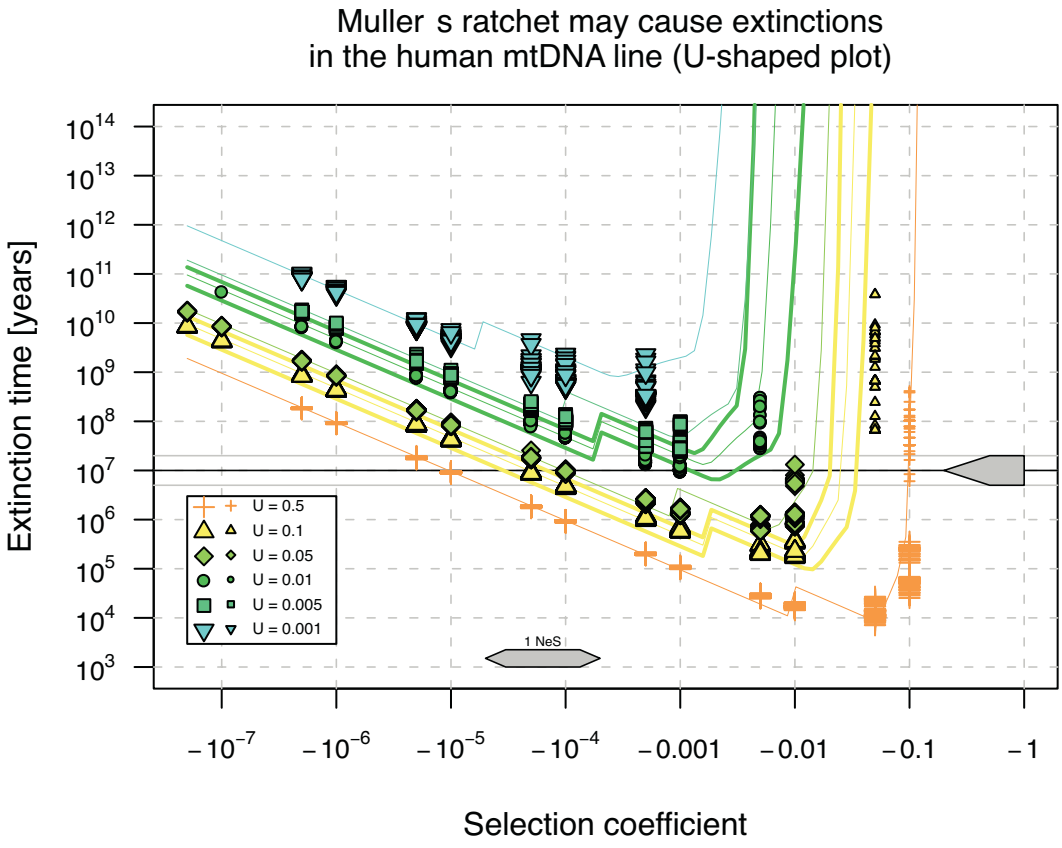


Figure 58 Muller's ratchet might cause extinctions of the human line due to accumulation of slightly deleterious mutations in mtDNA.
The upper arrow denotes the assumed age of the lines, including limits. The lower arrow marks the border with neutrality for the population sizes used by spanning the selection coefficients from 1 N_eS for the largest (50 000 females) to 1 N_eS for the smallest population (5000 females). Thin lines use simple *Equation172* (see Chapter 18) with N_e = 15800 and the respective means of T_{gen} and R_{max}. Thick lines for U=0.1 (yellow, higher most probable mutation rate estimate) and for U = 0.01 (mid-green, lower most probable estimate) use the corresponding upper and lower limits of N_e, T_{gen} and R_{max} to provide a feeling for the variability of extinction time estimates. Large symbols denote valid extinction time estimates from simulations with at least 2 observed clicks (usually many more). Small symbols denote lower limits for extinction times from simulations without observed clicks, based on the (usually wrong) assumption that the ratchet would have clicked just after stopping the simulation. Each symbol denotes an independent simulation with a different random seed.
This plot contains 6042 simulations with a total of 2.29 years of computing time and was produced by script RS005LL012 at 2002-10-08 15h38m21s. It is one of two plots in this work that includes results from Simulator005 releases 1-5 with U=0.001, although there was an irregularity with the Poisson RND generator that led to increased variance of effective clicktimes, but still had an effectively correct mutation rate.

Only with the help of these parameters, distributions of the frequency of predicted extinction times could be plotted[19] and compared with known distributions from the fossil record. But also without such an overview over the lifetime of species in general, the results shown increase interest in genetic causes of extinctions of species.

If the same calculation is done for the evolutionary line leading to humans, a similar result appears (Figure 58). The corresponding J-shaped plots are less interesting, as simulations covered the whole range of potential effective population sizes.

Quantifying the threat for the human line from the ratchet in mtDNA leads to the following values (again for U = 0.1 and a mean of all other values):

- MDSR = -0.01 to -0.005 at ca. 300 Kyr and
- DSR = -0.02 to $-5*10^{-5}$ at 10 Myr,

whereas U = 0.01 under the same conditions makes values drop to

- MDSR = -0.0005 to -0.002 at ca. 10 Myr (from thin line) and
- DSR = -0.001 to -0.001 at 10 Myr (from simulations),

as can be seen from the corresponding thin lines in Figure 58. These results imply that mutation rates as observed in pedigrees might easily lead to extinction within about 20 million years.

21.4 Conclusions

Definition of the genomic decay paradox

The fact that biologically reasonable parameter combinations might lead to extinction of evolutionary lines within their apparent time of existence is the core of the genomic decay paradox:

> *The genomic decay paradox describes situations where the apparent age of an evolutionary line is greater than the currently best predictions of its extinction time.*

Potential solutions

As this definition compares apparent knowledge, room for improvement always leaves the possibility that the paradox will be solved. The simplest solution at the moment would include either much smaller deleterious mutation rates or selection coefficients significantly more deleterious than 1%. The first might increase the minimal extinction times of the most dangerous range of selection coefficients to more than a billion years and make operation of the ratchet irrelevant. The latter would result in strong purifying selection stopping the ratchet. Additional potential solutions are dis-

19. As suggested by Alroy (2002) personal communication, Extinction Thresholds Conf.,Helsinki.

cussed in Chapter 29 and include advantageous mutations, compensatory mutations, other shapes of the distribution of mutational effects, synergistic epistasis and more.

Recombination

Recombination is also a powerful mechanism for stopping genomic decay due to Muller's ratchet. There has been some discussion about whether recombination does occur in mitochondria. Besides some indirect arguments in favour of occasional recombination events[20], no one has ever seen recombination in mammalian mtDNA and an analysis of linkage disequilibrium in 53 complete mtDNA sequences did not show any correlation with genetic distance[21]. Thus it does not appear to be necessary to consider recombination of mtDNA as a way to stop Muller's ratchet.

Similarly, it could be speculated that mitochondrial genes are occasionally repaired by transferring them back from the nucleus. However, while there is good evidence that occasionally mitochondrial genes do end up in the nucleus[22], there is currently no evidence that the process could also work in the other direction. And if it did, it would have to be regular enough to compensate the continuous onslaught of the ratchet.

Where do mutations go?

When looking at the plots given here, one might be tempted to stick to the phylogenetically inferred lower mutation values. However, a note of caution has to be added here: If an intergenerational mutation is observed as done in the pedigree studies, there are not many places it can go:

1. The majority of accidental losses are already accounted for by the drift implicit in the model. Thus no large decrease of long-term substitution rate can be expected. However, a little decrease will be due to processes like neutral slowdown reported on page 249.
2. If the mutation has an effect that decreases the chance of its carrier to have offspring, this is exactly what negative selection coefficients are all about. Such mutations are either
   - so deleterious that they keep the ratchet from clicking (>1% fewer offspring leads to heavy decrease of long-term substitution rate), or

20. Eyre-Walker (2000) "Do mitochondria recombine in humans?" Philos. Trans. R. Soc. Lond. B Biol. Sci. 355:1573-1580. - Awadalla et al. (1999) "Linkage disequilibrium and recombination in hominid mitochondrial DNA", Science 286:2524-2525.
21. Ingman et al. (2000) "Mitochondrial genome variation and the origin of modern humans", Nature 408:708-713.
22. Mourier et al. (2001) "The human genome project reveals a continuous transfer of large mitochondrial fragments to the nucleus", Mol. Biol. Evol. 18:1833-1837.

- they drive a very dangerous ratchet while they accumulate at moderate speeds, or
- they do accumulate basically with the same high speed as neutral mutations, but their combined effect is too small to do any long-term harm to the evolutionary line.

3. If the mutation keeps occurring at the same site flipping it back and forth again and again, then it is one of the few known hot spots.
4. Finally, details of multi-level population genetics could play a role here, but this is unlikely given the fact that pedigree mutation rates have been observed *between* generations and thus should already be the summary of all underlying processes.

Young or dead or what?

Nothing else should be able to stop new mutations from contributing to long-term substitution rate. Thus, when considering human evolution, the most probable alternatives for mitochondrial Eve are either that Eve is young, because pedigree mutations are evolutionarily neutral, or that Eve is dead, because selection coefficients that only *slow down* long-term mutation rate effectively drive a very dangerous ratchet[23]. Before this issue can be comprehensively solved, we need much more data on mutation rates and their effects.

Potential solutions

Currently, it could be speculated that those mutations that do not occur on hot-spots are not very often neutral. Thus they would have deleterious effects of more than a few percent that would lead to complete removal by purifying selection. However, it appears hard to believe that most mutations should lead to more than a few percent fitness decrease, whereas all others are at least effectively neutral and mutations with intermediate effects hardly ever occur. Especially when considering the complexity and robustness of biological mechanisms, one would intuitively expect that there are always many ways to corrupt a function a little less, suggesting that there is a continuum of increasingly smaller mutational effects. The often-used L-shaped distributions of mutational effects[24] have similar implications. Only additional data will be able to resolve these issues and additional models that incorporate the various potential solutions for the genomic decay paradox will have to be computed in order to tell how solid the apparent paradox is. However, for the moment, this study has established details of a mitochon-

23. Loewe (1997) "Mutation Rates, Muller's Ratchet, Genetic Load and Eve: Young or Dead?" First international workshop on human mitochondrial DNA, 25 - 28 October 1997, Washington, D.C.
24. Lynch et al. (1999) "Perspective: Spontaneous deleterious mutation", Evolution 53:645-663.

drial counterpart to the genomic decay paradox predicted for nuclear genomes[25].

Increase of accuracy by global computing

One final word about computing. The plots presented here have regions where further simulations are uninteresting (eg. either drift or selection clearly dominate) and other regions where further simulation looks promising (eg. investigate a given narrow range of selection coefficients derived from future data). One of the strengths of global computing is that after biological analysis a new set of parameter combinations can easily be put to the web again, increasing accuracy as needed without having to wait for time on a super-computer. Finally, automated processing of results via the R statistical system allows easy updating of the plots presented here. Thus, new information about parameters relevant for the ratchet in mtDNA can be easily transformed into better predictions.

25. Kondrashov (1995) "Contamination of the Genome by Very Slightly Deleterious Mutations: Why Have We Not Died 100 Times Over?" J. theor. Biol. 175:583-594. - Crow (1997) "The high spontaneous mutation rate: Is it a health risk?" Proc. Natl. Acad. Sci. USA 94:8380-8386. - Eyre-Walker & Keightley (1999) "High genomic deleterious mutation rates in hominids", Nature 397:344-347.

22 Muller's ratchet might threaten the Amazon Molly and ancient asexuals

There are species that are known to have lived without sex for extended periods of time. These have long been suspected of being threatened by Muller's ratchet. This is the first report that quantifies this threat in detail for the small unisexual fish Poecilia formosa *(Amazon Molly). It is shown that for some biologically realistic mutation rates, Amazon Molly is indeed threatened by Muller's ratchet. However, further data on mutation rates will be needed to finally settle the issue. Similarly, collecting the same data as for Amazon Molly will allow quantifying the threat from Muller's ratchet in other ancient asexuals too.*

22.1 Amazon Molly biology

All the following information on Amazon Molly comes from Dunja Lamatsch[26]. The Amazon Molly (*Poecilia formosa*) is a small fish (3 - 7 cm) that lives in a rather limited range from the Nueces River in South Texas southward to the mouth of the Rio Tuxpan, north of the Sierra del Abra. All these riversystems flow from west to east and have no connection besides the sea[27]. Some mark and recapture experiments could only estimate that the population size has to be larger than 10 000. A rough estimate suggests that there are probably about 10 million individuals. The population is subdivided; subpopulations from South Texas, eg. have no reasonable connection with those in the Rio Purificacion, a fact that significantly affects effective population size $\mathbf{N_e}$ for ratchet considerations[28].

Mode of reproduction

The Amazon Molly reproduces gynogenetically, ie. its eggs contain an unreduced set of chromosomes, but need the sperm of the sister species *Poecilia mexicana* or *Poecilia latipinna* as a mechanical trigger to start development[29]. Usually, only maternal genes are expressed and the paternal genome is expelled.

26. If no other sorce is given, all details are personal communications of Dunja Lamatsch (2002), who discussed some of them with Manfred Schartl.
27. Amazon mollies tolerate marine conditions. Schlupp et al. (2002) "Biogeography of the Amazon molly, Poecilia formosa", Journal of Biogeography 29:1-6.
28. Whitlock & Barton (1997) "The effective size of a subdivided population", Genetics 146:427-441. - Whitlock (2002) "Selection, load and inbreeding depression in a large metapopulation", Genetics 160:1191-1202.
29. Stöck & Lamatsch (2002) "Triploide Wirbeltiere: Wege aus der Unfruchtbarkeit oder Eingeschlechtigkeit", Naturw. Rdsch. 55:349-358.

Paternal leakage effects on rate of ratchet are unclear

Occasional paternal introgression can lead to expression of paternal genes, a process that has been suggested to compensate for Muller's ratchet in Amazon Molly[30]. However, this should not be confused with true recombination, a reason for discussions about the ratchet-stopping potential of paternal introgression[31]. It has been speculated that the paternal genome might also be used as a template for DNA repair[32].

Life history

The Amazon Molly reaches maturity between 4 and 6 months, produces more offspring at larger body sizes and lives for up to 2 years in nature or 3 years under optimal laboratory conditions. They produce 28 (±16) newborn fish every 28 days, depending on water temperature, before they stops about 3 or 6 months before death in nature or in the lab, respectively. This number of newborn fish can be reduced considerably by predation or disease. Thus under optimal conditions in nature the maximal reproductive capacity should be 18 periods times 28 offspring = 504. As a lower limit one might use 50 offspring per lifetime (assuming that many offspring die due to predation or disease and many parents are small and frequently trapped in little ponds without sister species that provide the sperm, needed as a trigger for development). An upper limit of 1560 can be derived from combining the maximal theoretical number of offspring-producing periods in the lab (26) with the maximal conceivable number of offspring per period (60 ≈ Mean + 2 StdDev). This life history results in generation times of 4 month (min), 1 year (mean) and 2.5 years (max).

Age of line

The first Amazon Molly was formed by a hybridisation event between a *Poecilia mexicana* female and a *Poecilia latipinna* male[29,33]. It has reproduced asexually since then. A date for this event can be derived from comparing alleles in Amazon Molly with alleles in their corresponding parent species using a molecular clock. The smallest divergence found was 2 mutations in 1377 bp of various nuclear genes (mostly introns) when comparing mexicana alleles of Amazon Molly to *Poecilia mexicana limantouri*[34]. This was combined with an assumed upper limit of a divergence rate of 2%/Myr (from

30. Schartl et al. (1995) "Incorporation of subgenomic amounts of DNA as compensation for mutational load in a gynogenetic fish", Nature 373:68-71.
31. Beukeboom et al. (1995) "Amazon molly and Muller's Ratchet", Nature 375:111-112. - Beukeboom & van Batenburg (1999) "The effect of paternal leakage on the rate of Muller's Ratchet", handout for poster. 7th Conference of the European Society for Evolutionary Biology, Barcelona, Spain.
32. Beukeboom & Vrijenhoek (1998) "Evolutionary genetics and ecology of sperm-dependent parthenogenesis", J. evol. Biol. 11:755-782.
33. For an overview see Schlupp & et al. (1998) "Dispensable and indispensable genes in an ameiotic fish, the Amazon molly Poecilia formosa", Cytogenet. Cell. Genet. 80:193-198.
34. Schartl et al. (1995) "Evolutionary origin of a parthenoform, the Amazon molly Poecilia formosa, on the basis of a molecular genealogy", Evolution 49:827-835.

mitochondrial clocks) to reach the well-known 100 000 years for the time Amazon Molly has lived without recombination34. Here, 70 Kyr and 130 Kyr are used as lower and upper limits for that age. However, the upper limit might turn out to be too low.

Mutation rate U

Unfortunately, little is known about the mutation rates in Amazon Molly, so some indirect considerations are necessary. The size of a diploid genome is about 1900 Mbp, as derived from DNA content of cells[35]. If this is plainly combined with the mutation rate counterpart of the divergence rate cited above, then about 19 new mutations per diploid genome can be expected in each new generation (1900 Mbp x 0.01 subst/bp/Myr). It can be argued that both copies of the genome should be considered for Muller's ratchet, as both contribute to a functional phenotype in case the other copy is damaged. However, here the more conservative approach is used of considering only a haploid copy. Simulations suggest that further work is needed to clarify this issue in detail[36].

Estimating U by comparison to other species

Another way to estimate intergenerational mutation rates is comparison with known figures for other species[37]. Vertebrates usually have a genome that contains large non-coding parts that are believed to be neutral. Such regions do not contribute to the deleterious mutation rate and should therefore be excluded by calculating an effective genome size that contains only those sequences that affect the phenotype37. The ratio of estimated effective genome size to total genome size is about 1/4 in the worm *Caenorhabditis elegans*, 1/10 in the fly *Drosophila melanogaster*, 1/33 in the mouse and 1/40 in man37. Thus one might assume about 1/20 or more for the vertebrate Amazon Molly, leading to an effective haploid genome size of at least 48 Mbp. If this is combined with 25 germ line cell divisions per adult generation known for female mice37 (similar body size) and 10^{-10} as minimal mutation rate of vertebrates ($1.8*10^{-10}$ in mouse and $0.5*10^{-10}$ in man)37, then at least 0.12 new mutations in the coding region can be expected in each new generation. If one expects at least about half of them to be deleterious, one might assume a mutation rate of at least 0.05 / genome / generation for mutations with effects in the dangerous range.

As this calculation is obviously affected with large errors, an array of other mutation rates is included when quantifying the ratchet.

35. Lamatsch et al. (2000) "Noninvasive determination of genome size and ploidy level in fishes by flow cytometry: Detection of triploid Poecilia formosa", Cytometry 39:91-95.
36. Charlesworth & Charlesworth (1997) "Rapid fixation of deleterious alleles can be caused by Muller's Ratchet", Genet. Res. 70:63-73.
37. Drake et al. (1998) "Rates of spontaneous mutation", Genetics 148:1667-1686.

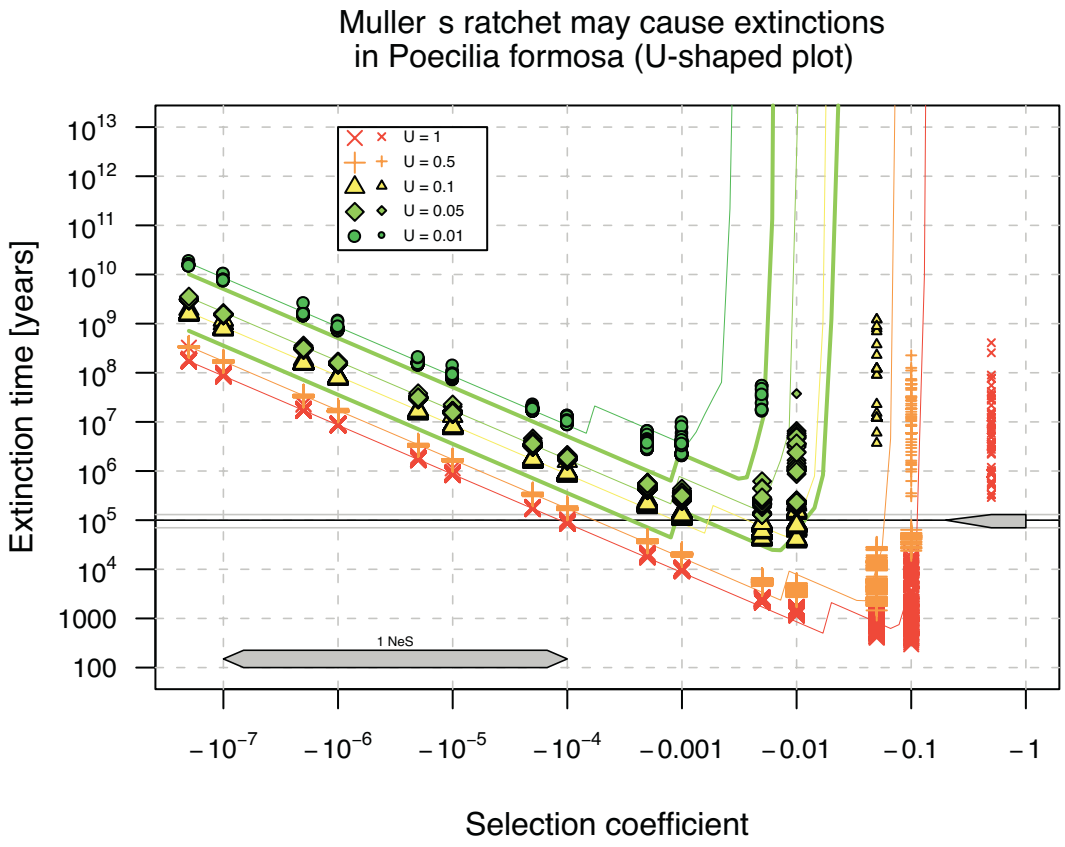


Figure 59 Muller's ratchet might cause extinctions of the ancient asexual Amazon Molly due to accumulation of slightly deleterious mutations.
The upper arrow denotes the assumed age of the line, including limits. The lower arrow marks the border with neutrality for the population sizes used by spanning the selection coefficients from 1 $\mathbf{N_eS}$ for the largest (10^7 females) to 1 $\mathbf{N_eS}$ for the smallest population (10^4 females). Thin lines use simple equation 172 (see Chapter 18) with $\mathbf{N_e}$ = 316000 and the respective means of generation time $\mathbf{T_{gen}}$ and maximal reproductive capacity $\mathbf{R_{max}}$. Thick lines for $\mathbf{U}$=0.05 (first green = lowest most probable mutation rate estimate) denote limits of extinction time with the corresponding upper and lower limits of $\mathbf{N_e}$, $\mathbf{T_{gen}}$ and $\mathbf{R_{max}}$ to provide a feeling for the variability of extinction time estimates due to these factors. Large symbols denote valid extinction time estimates from simulations with at least 2 observed clicks (usually many more). Small symbols denote lower limits for extinction times from simulations without observed clicks, based on the (usually wrong) assumption that the ratchet would have clicked just after stopping the simulation. Each symbol denotes an independent simulation with a different random seed.
This plot contains 10308 simulations with a total of 7.22 years of computing time and was produced by script RS005LL013 at 2002-10-10 10h46m51s.

22.2 Extinction times for the Amazon Molly

Simulation results

If the values given above are used to compute the U-shaped plot of extinction time over selection coefficient for the results computed by Simulator005, it becomes apparent that the higher mutation rates from the set of all biologically feasible values lead to extinction of the Amazon Molly within the assumed age of its existence (Figure 59). This figure includes an estimate of the highest plausible effective population size of the Amazon Molly, so population size will not rescue it from Muller's ratchet here. However, a careful look at the various mutation rates shows that a deleterious mutation rate of 0.05 per generation might be a little too low to cause extinction in 10^5 years. (Details depend on other factors, as a part of the lower extinction time limit is still below the critical age of the line). If the deleterious mutation rate was only 0.01, then an extinction due to Muller's ratchet in the known age of the line can be excluded, even with the current uncertainty in other parameters. However, if deleterious mutation rates are as high as 0.5 or more (which is still conceivable), then the Amazon Molly might have become extinct in a few thousand years.

Quantification of the threat from the ratchet for the Amazon Molly in detail for U = 0.5 (and a mean of all other values) yields

- MDSR[38] = -0.05 to -0.01 at ca. 5 Kyr and
- DSR[39] = -0.06 to -0.0002 at ca. 100 Kyr,

U = 0.1 (and a mean of all other values) yields

- MDSR = -0.01 to -0.005 at ca. 50 Kyr and
- DSR = -0.01 to -0.001 at ca. 100 Kyr,

whereas lower mutation rates under the same conditions lead to minimal extinction times that are just a little bit longer than the assumed age of the line (U = 0.05 leads to extinction in ca. 200 Kyr at a selection coefficient of -0.005 and U = 0.01 to ca. 5 Myr at -0.001). See the thin lines in Figure 59 for a visual overview.

Summary of ratchet in Amazon Molly

This first detailed assessment of the threat from Muller's ratchet in the Amazon Molly is not conclusive, but points in new directions. While higher mutation rates lead to a genomic decay paradox and call for other mechanisms to explain why this fish has escaped extinction thus far, lower estimates for mutation rate do not need additional mechanisms to explain survival. Moreover, at a mutation rate of 0.01 and a selection coefficient of -

38. Most Dangerous Range of Selection coefficients with associated minimal extinction time.
39. Dangerous Range of Selection coefficients that leads to extinction in assumed age of the line.

0.001 the predicted extinction times are between ca. 20 Myr and 100 Myr (depending on population size). In such a case Amazon Molly has lived only through 0.1 % to 0.5 % of its potential lifetime as an asexual. Thus it would carry only few of the most dangerous mutations and one cannot expect to see signs of decreased fitness when comparing it to sexual sister species. This is consistent with the observation that there are no apparent dramatic differences between the maximal reproductive rates for *Poecilia formosa*, *P. mexicana* and *P. latipinna*[26].

22.3 Assessing the threat of Muller's ratchet for other ancient asexuals is easier now

Once rough upper and lower limits are available for the parameters needed to compute extinction times, the system used in this work makes it manageable to quantify the threat from Muller's ratchet for the long list of other ancient asexuals[40]. In a wider sense, this belongs to the process of properly quantifying the well-known fact that asexuals are typically short-lived. While central to theories about the origin of sex, this typical short-livedness still needs to be properly quantified, as pointed out recently by Butlin[41]. Prominent examples for such work include the following organisms:

Darwinula stephensonii

- *Darwinuloideae* are small non-marine ostracods that are believed to have lived for more than 100 Myr without sex (fossil record shows only females)[42]. The species *Darwinula stephensonii* is a member of this group and is thought to exist for more than 20 Myr now42.

Bdelloid rotifers

- The Class *Bdelloidea* of the Phylum *Rotifera* is the largest taxonomic group that has apparently lived completely without sex for at least 40 Myr[43]. The ancient asexuality of these 0.1 to 1 mm long animals is as well established as it can possibly be.
- The clonal, hybrid, gynogenetic mole salamander *Ambystoma* is the oldest known asexual vertebrate with about 5 Myr age[44].

40. Judson & Normark (1996) "Ancient asexual scandals", Trends Ecol. Evol. 11:41-46.
41. Butlin (2002) "Opinion - evolution of sex: The costs and benefits of sex: new insights from old asexual lineages", Nat Rev Genet 3:311-317.
42. See pp. 300-301 in Martens, (ed, 1998) "Sex and parthenogenesis: Evolutionary ecology of reproductive modes in non-marine ostracods." Leiden, Netherlands, Backhuys Publishers. - Butlin et al. (1998) "Asexual reproduction in nonmarine ostracods", Heredity 81:473-480.
43. Welch & Meselson (2000) "Evidence for the evolution of Bdelloid Rotifers without sexual reproduction or genetic exchange", Science 288:1211-1215. - Judson & Normark (2000) "Evolutionary genetics. Sinless originals", Science 288:1185-1186. - Butlin (2000) "Virgin rotifers", Trends Ecol. Evol.15:389-390.

Why quantify the ratchet as often as possible?

In the light of many open questions that can be asked about the general accuracy of the simple standard model of Muller's ratchet used for quantifications presented here, one might ask what the value of such an array of ratchet quantification results would be. First, this approach helps biologists to look for key data needed for *any* quantification of the ratchet. Second, it can be expected that at least some of the ancient asexuals might still be ancient asexuals, but are not an example of the genomic decay paradox, because the known age of their asexuality is smaller than their predicted extinction time. Third, those species with an apparent genomic decay paradox can be subject to a more detailed search for mechanisms that solve the paradox and help them to avoid extinction. Finally, experiences collected with the present system might lead to development of better systems for quantification of genomic decay paradoxes.

44. Spolsky et al. (1992) "Antiquity of clonal salamander lineages revealed by mitochondrial DNA", Nature 356:706-708. - Hedges et al. (1992) "Ancestry of unisexual salamanders", Nature 356:708-710.

23 Muller's ratchet might threaten endosymbionts like *Buchnera*

The enteric bacteria of the Genus Buchnera have lived endosymbiontically in aphids for more than 100 million years. As they are sequestered in the cells of their hosts, they cannot exchange genetic material with bacteria from other hosts. Consequently, it has been suspected that Muller's ratchet should have driven these bacteria to extinction long ago. Molecular signatures of genomic decay have been found in their genomes (loss of genes and excessive accumulation of non-synonymous mutations). However, Buchnera is still alive and nobody knows why. Here the threat from Muller's ratchet to Buchnera is quantified to facilitate further investigation of this intriguing question.

Buchnera

Many bacteria live as endosymbionts in eucaryotic cells and some of them have clinical significance[45], while others are discussed as a model for endosymbiotic evolution in general[46]. The endosymbiontic Genus *Buchnera* is very closely related to *Escherichia coli*. However, the latter can freely exchange genetic material, while the former is trapped in aphid bacteriocyte cells they never leave[47]. This confines any potential recombination events to the population of *Buchnera* cells that live within the same aphid cell or at least within the same aphid individual. As variation in these small populations should be very small, evolutionary repair power of potential recombination events is quite limited and usually neglected. The fossil record of aphids shows that this practical asexuality is very old and suggests that Muller's ratchet might have been operating in this system for a very long time[48].

Aphids

Aphids are major agricultural pest insects and feed on plant sap that contains enough carbohydrates, but lacks eg. some amino acids47. These missing amino acids are synthesised by *Buchnera* that specifically retained the necessary genes[49] and in turn receives nutrients from its host. None of the

45. Horn et al. (2000) "Neochlamydia hartmannellae gen. nov., sp. nov. (Parachlamydiaceae), an endoparasite of the amoeba Hartmannella vermiformis", Microbiology 146:1231-1239. - Horn & Wagner (2001) "Dasein im Verborgenen: Bacterien, die in Acanthamöben leben", Biologie in unserer Zeit 31:160-168.
46. Fukatsu (1994) "Endosymbiosis of aphids with microorganisms: A model case of dynamic endosymbiotic evolution", Plant Species Biology 9:145-154.
47. Baumann et al. (1995) "Genetics, physiology, and evolutionary relationships of the genus Buchnera: intracellular symbionts of aphids", Annu. Rev. Microbiol. 49:55-94.
48. Moran et al. (1993) "A molecular click in endosymbiotic bacteria is calibrated using the insect hosts", Proc. R. Soc. Lond. B Biol. Sci. 253:167-171.

partners of this obligate symbiotic relationship can reproduce without the other47, not the only case where endosymbiontic bacteria completely control reproductive success of their host[50]. Some aphids are believed to be ancient asexuals although some form of cryptic sex might be going on[51].

23.1 Biological ratchet parameters

Multi-level population genetics

Like mitochondria, endosymbionts are part of a complicated population genetic system that has several levels[52,53]. There is a very much CPU-time limited study52 that explicitly addressed Muller's ratchet in *Buchnera* and aphids with two simplified levels of selection by simulating 1-30 bacteria per host in 10 - 90 (occasionally 1500) hosts. While this study helps to understand a number of qualitative features of the system, it did not explicitly try to estimate extinction times for *Buchnera* or its host. Ultimately, a simulator that incorporates the various aspects of endosymbiont multi-level population genetics will have to be built to estimate *Buchnera* extinction times with higher precision. A detailed analysis of multi-level population genetics is outside the scope of this study, but a short overview will help us understand the choice of parameters made for the simple ratchet model used here.

Population sizes

Each *Buchnera* cell has a mean of about 120 copies of the genome[54]. Each adult host contains about 500 000 bacteria that exhibit about 5 replications per host generation[55]. Host populations can burst to enormous sizes

49. Shigenobu et al. (2000) "Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS", Nature 407:81-86.
50. Dedeine et al. (2001) "Removing symbiotic Wolbachia bacteria specifically inhibits oogenesis in a parasitic wasp", Proc. Natl. Acad. Sci. USA 98:6247-6252.
51. Normark (1999) "Evolution in a putatively ancient asexual aphid lineage: Recombination and rapid karyotype change", Evolution 53:1458-1469.
52. Rispe & Moran (2000) "Accumulation of deleterious mutations in endosymbionts: Muller's ratchet with two levels of selection", Am. Nat. 156:425-441.
53. Paulsson (2002) "Multileveled selection on plasmid replication", Genetics 161:1373-1384. - Bergstrom & Pritchard (1998) "Germline bottlenecks and the evolutionary maintenance of mitochondrial genomes", Genetics 149:2135-2146. - Kondrashov (1994) "Mutation load under vegetative reproduction and cytoplasmic inheritance", Genetics 137:311-318. - Otto & Hastings (1998) "Mutation and selection within the individual", Genetica 103:507-524. - Otto & Orive (1995) "Evolutionary consequences of mutation and selection within an individual", Genetics 141:1173-1187. - Hastings (1989) "Potential germline competition in animals and its evolutionary implications", Genetics 123:191-198. - Hastings (1991) "Germline-selection: Population genetic aspects of the sexual/asexual life cycle", Genetics 129:1167-1176. - Birky (1991) "Evolution and population genetics of organelle genes: Mechanisms and models", pp. 112-134 in: Selander et al. (eds) Evolution at the Molecular Level, Sunderland, MA, Sindauer Associates, Inc. - Takahata & Slatkin (1983) "Evolutionary dynamics of extranuclear genes", Genet. Res. 42:257-265.
54. Komaki & Ishikawa (1999) "Intracellular Bacterial Symbionts of Aphids Possess Many Genomic Copies per Bacterium", J. Mol. Evol. 48:717–722.
55. Nancy Moran (2001) Personal communication.

(up to $2*10^9$ aphids /acre47), but polymorphisms are generally low, so effective population size should be small55 (here values between10^4 and 10^6 are assumed for aphids). A study estimated the effective population size of *Buchnera* to be 1 x 10^7 from observed levels of polymorphism[56]. Thus, to omit details of multi-level population genetics, the following approaches are used to estimate extinction times with the simple standard ratchet model of Haigh[57], as implemented in Simulator005:

Simplifications of multi-level population genetics

1. ***Buchnera* bottom-up perspective**. Assume there were only one genetically optimal aphid individual in the world and it would produce one offspring with a probability of exactly 1 as long as it still contains an endosymbiont population. Then the population of endosymbionts in it might be considered as a simple asexual population that will become extinct eventually.
   From the data above, the effective population size can be computed by applying the formula for cyclical changes in census population size[58]. Assuming that 500 000 bacteria live at the end of the cycle, each of the 5 doublings denotes a step in the cycle, and 500 000 /(2^5) bacteria live at the beginning of the cycle, an effective population size of 47 620 would result. Thus, to be on the safe side, the analysis presented here assumes a minimum of 10 000, a mean of 70 000 and a maximum of 500 000 effective endosymbionts in one aphid.
   All other aphids can be considered as replicates of this evolutionary process, since complete lack of exchange of genetic material between them makes them effectively independent.
2. ***Buchnera* top-down perspective**. Assume that aphids were unimportant, because the only thing that counts is the overall effective population size of *Buchnera*. While it is unlikely that an effective population size estimated from polymorphism will be equivalent to an effective population size as needed for analysis of Muller's ratchet, the value given above is nevertheless assumed. Thus, 5 x 10^6, 1 x 10^7 and 2 x 10^7 are assumed in this study.
3. **Aphids top-down perspective**. Assume that aphids have an asexual genetic system called *Buchnera* that is considered at intervals that

56. Funk et al. (2001) "Intraspecific variation in symbiont genomes: Bottlenecks and the aphid-buchnera association", Genetics 157:477-489.
57. Haigh (1978) "The accumulation of deleterious genes in a population - Muller's Ratchet", Theor. Popul. Biol. 14:251-267.
58. See p. 41 and references given there in Kimura (1983) "The neutral theory of molecular evolution", Cambridge, Cambridge University Press.

correspond to aphid generations. Thus the summary of lower-level evolution is entered via ratchet parameters for each aphid generation. This is similar to the approach taken in the case of mitochondria (see Chapter 21), with the important difference that for mitochondria experimental estimates exist that summarise all lower level evolution for intergenerational events. In aphids, however, the only data that exist are the mutation rates at the lower *Buchnera* level, allowing only guesses as to how that translates into intergenerational ratchet parameters for aphids. The effective population sizes for aphids assumed here are 10^4, 10^5 and 10^6.

Generation times

The usual reproductive period of aphids is from about day 8 to day 20 of their maximal 30 day long life47 and the shortest possible generation time observed is 5 days55. Here generation times of 8, 14 and 20 days are assumed for aphids. As *Buchnera* experiences about 5 doublings in that time, its generation time is assumed to be 1, 2.8 and 4 days. Other estimates not considered here assume 3-10 generations per year for aphids and 30-50 generations per year for *Buchnera*[59].

Maximal reproductive capacity

Estimation of the maximal reproductive capacity of bacteria is not straightforward. One could assume a value of 2, as bacteria can produce only 2 descending cells. However, while bacteria cannot produce more offspring they cannot produce fewer either. This raises the question of how Muller's ratchet might affect potentially eternally living bacteria. The answer is simple: It lengthens generation time. Unfortunately the current Simulator005 knows only fixed generation times and one will have to wait for another simulator to be built to model the situation in bacteria in detail. However, the maximal reproductive capacity $\boldsymbol{R_{max}}$ that can be used for analysis of results from Simulator005 might be computed by

$$\boldsymbol{R_{max}} = 2^{\left(\frac{DT_{max} - DT_{min}}{DT_{min}}\right)} \qquad (63)$$

where $\boldsymbol{DT_{max}}$ is the maximal doubling time that does not lead to extinction and $\boldsymbol{DT_{min}}$ is the minimal doubling time of that bacterium before the ratchet started to operate. One can debate whether $\boldsymbol{DT_{min}}$ is always close to the known minimal doubling time for all bacteria (ie. 10-20 min) or is close to the smallest known doubling time of the bacteria considered (ie. 1 day for

59. Clark et al. (1999) "Sequence evolution in bacterial endosymbionts having extreme base compositions", Mol. Biol. Evol. 16:1586-1598.

Buchnera). Here, DT_{min} is assumed to be 30 - 60 minutes, where as DT_{max} is 1-4 days from the fact that a buchnera population that does not manage to double 5 times in 20 days will demographically become extinct from the corresponding aphid. The resulting values for R_{max}, 1.7×10^{7}, 8×10^{28} and 6×10^{57} were used for all analyses here.

Replication based mutation rates

The genome of *Buchnera* has a size of 640 681 bp49 and each bacterial cell carries an average of about 120 copies of it54. The estimated mutation rate for *Buchnera* is $1.4 - 1.9 \times 10^{-10}$ mutations /site /generation[55,59,60]. This results in about 1×10^{-4} mutations/genome/generation (for comparison, in *E. coli* 4.6 Mbp and 5.4×10^{-10} /bp/gen yield 0.0025 mutations/ genome / generation[61]). Assuming that about 10 % of those mutations belong to a given order of magnitude of effectively non-neutral selection coefficients would give a deleterious mutation rate estimate of 1×10^{-5} /genome/generation.

Multiple copies of haploid genomes

The precise effect of the fact that more than 100 copies of the genome reside in each *Buchnera* cell is yet unclear. The simplest assumption would be that additional copies are there to additively compensate for decreased efficiency of existing copies. Thus, all of these copies would be somewhat functional, as each copy contributes in a small way to overall success. Correspondingly, deleterious mutation effects would be only about 1% of what they would be in a haploid single-copy genome. However, the long-term evolutionary cost of such a simple shield from immediate deleterious mutational effects can be devastating, as the overall mutation rate is likely to increase linearly with copy number. Deleterious mutations in each copy are likely to have a small effect, because they are neutral to replication of the genome itself in most cases. Thus, proper scaling for simple ratchet models would use ploidy to scale mutational effects down and mutation rates up. Resulting mutation rates are 0.01 total or 0.001 deleterious mutations/genome/generation. However, more detailed work is needed to arrive at better estimates for the transition form a haploid to a polyploid ratchet[62].

Stationary phase mutation rates

Close relatedness to *Escherichia coli* suggests that similar processes might be at work in *Buchnera*, too. Thus, the upper possible limit for mutation rate used in this study is derived from combining stationary phase mutation rate

60. Ochman et al. (1999) "Calibrating bacterial evolution", Proc. Natl. Acad. Sci. USA 96:12638-12643.
61. Drake et al. (1998) "Rates of spontaneous mutation", Genetics 148:1667-1686.
62. For an analysis for diploids, see Charlesworth & Charlesworth (1997) "Rapid fixation of deleterious alleles can be caused by Muller's Ratchet", Genet. Res. 70:63-73. - Schultz & Lynch (1997) "Mutation and extinction: The role of variable mutational effects, synergistic epistasis, beneficial mutations, and the degree of outcrossing", Evolution 51:1363-1371.

levels observed in this thesis with the multiple genome copy logic described above. The deleterious mutation rate observed at 37°C in E. coli was 0.078 mutations/genome/generation according to the BATEMAN-MUKAI analysis. As this rate probably depends on temperature-dependent enzymatic processes, it most probably has to be scaled down by a factor of 2 for each 10°C (corresponding to Q_{10} of 2)[63]. Furthermore, the differences in genome size (640Mbp/4600Mbp) have to be considered in scaling. Thus at 7°C or 17°C deleterious mutation rates of 0.0014 or 0.0027 /genome/generation would result. If 100 copies of the genome mutate at such a rate, each Buchnera cell could be expected to carry 0.1 new deleterious mutations per generation in its overall genetic material.

Age of line

Finally, endosymbionts in aphids are believed to have lived for 150 (minimum) 200 (mean) or even 250 (maximum) million years in their hosts, based on the fossil record and on matching speciation patterns (between host and endosymbionts)[64].

63. See p. 378 in Adam et al. (1988) "Physikalische Chemie und Biophysik". Zweite, völlig neu bearbeitete und erweiterte Auflage, Berlin, Springer-Verlag.
64. Moran et al. (1993) "A molecular click in endosymbiotic bacteria is calibrated using the insect hosts", Proc. R. Soc. Lond. B Biol. Sci. 253:167-171.

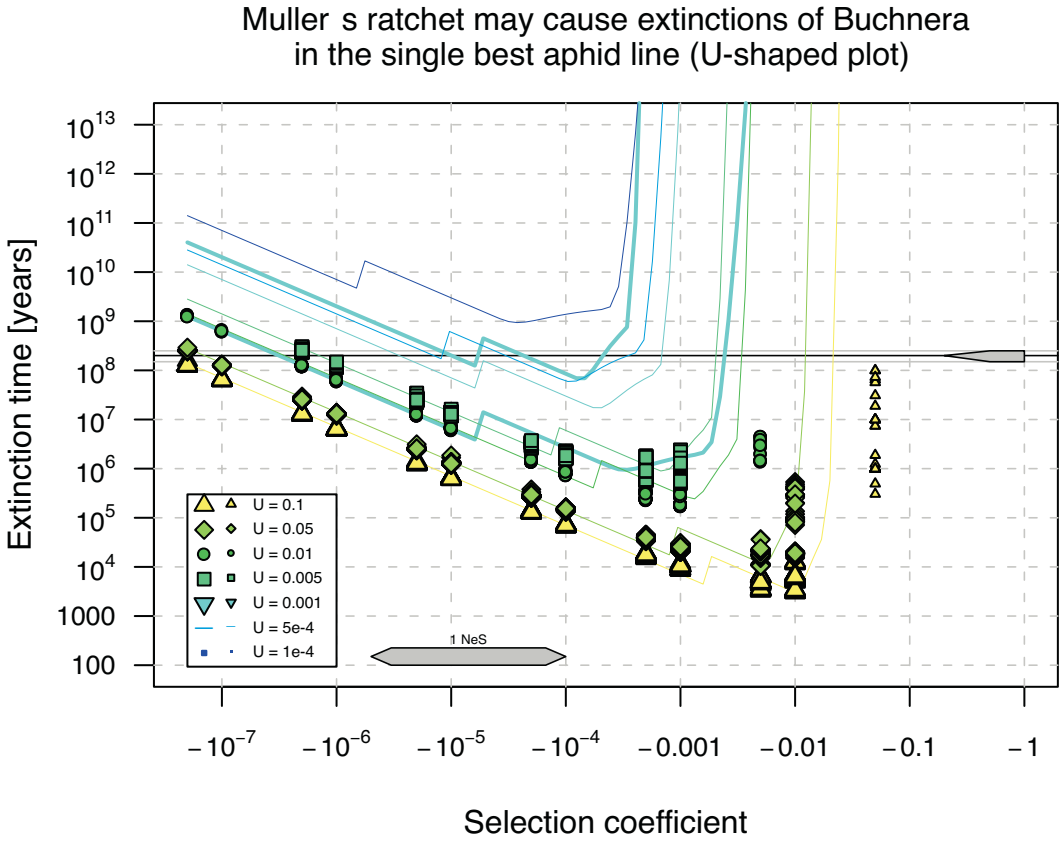


Figure 60 Muller's ratchet might cause extinction of *Buchnera* in the single best line of aphids due to accumulation of slightly deleterious mutations, if no recombination occurs within bacteriocytes or aphids. The upper arrow denotes the assumed age of endosymbionts in aphids, including limits. The lower arrow marks the border with neutrality for the population sizes used by spanning the selection coefficients from 1 N_eS for the largest (500 000 cells) to 1 N_eS for the smallest population (10 000 cells). Thin lines use simple *Equation172* (see Chapter 18) with N_e = 70700 and the respective means of T_{gen} and R_{max}. Thick lines for U=0.001 (first blue = lowest most probable mutation rate estimate) use the corresponding upper and lower limits of N_e, T_{gen} and R_{max} to provide a feeling for the variability of extinction time estimates. Large symbols denote valid extinction time estimates from simulations with at least 2 observed clicks (usually many more). Small symbols denote lower limits for extinction times from simulations without observed clicks, based on the (usually wrong) assumption that the ratchet would have clicked just after stopping the simulation. Each symbol denotes an independent simulation with a different random seed.
This plot contains 4197 simulations with a total of 4.5 years of computing time and was produced by script RS005LL014 at 2002-10-09 16h44m28s.

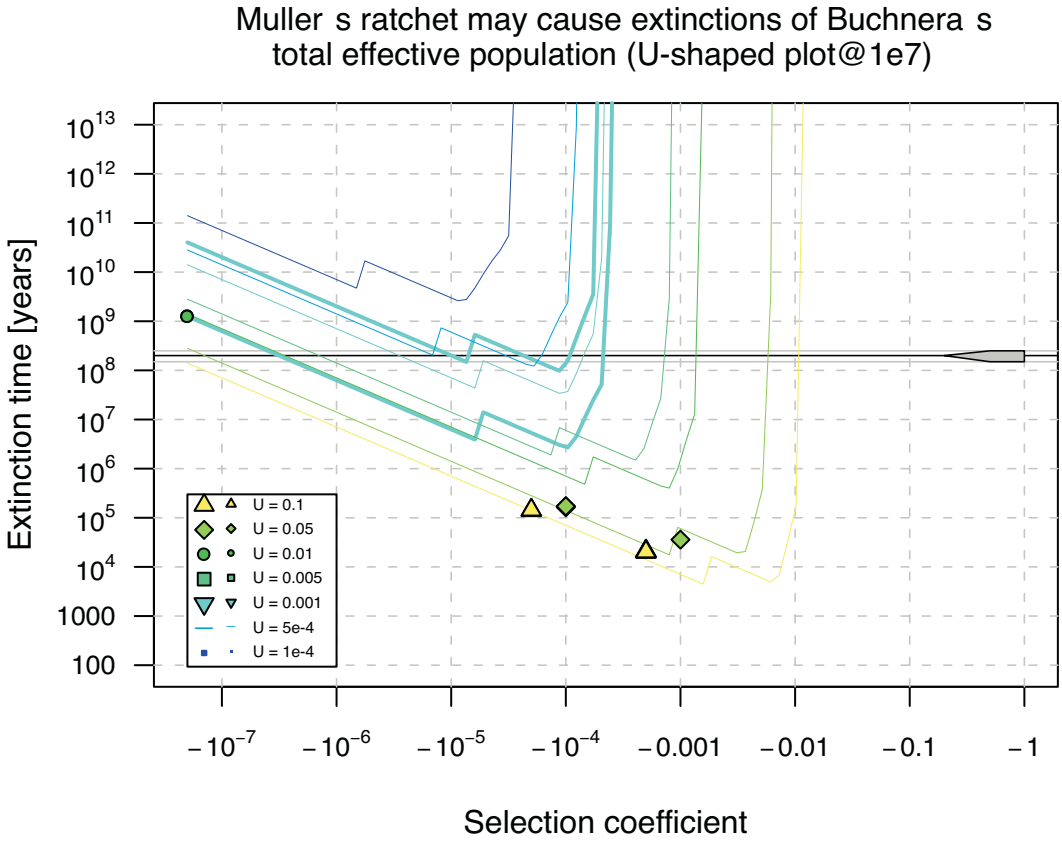


Figure 61 Muller's ratchet might cause extinctions of the total population of *Buchnera* due to accumulation of slightly deleterious mutations, if no recombination occurs within bacteriocytes or aphids.
The upper arrow denotes the assumed age of endosymbionts in aphids, including limits. Population sizes used were 5x10^6 (min) 1x10^7 (mean) and 2x10^7 (max, no simulations at this size). Thin lines use simple *Equation172* (see Chapter 18) with $\mathbf{N_e}$ = 1x10^7 and the respective means of $\mathbf{T_{gen}}$ and $\mathbf{R_{max}}$. Thick lines for $\mathbf{U}$=0.001 (first blue=lowest most probable mutation rate estimate) use the corresponding upper and lower limits of $\mathbf{N_e}$, $\mathbf{T_{gen}}$ and $\mathbf{R_{max}}$ to provide a feeling for the variability of extinction time estimates. Large symbols denote valid extinction time estimates from simulations with at least 2 observed clicks (usually many more). Small symbols denote lower limits for extinction times from simulations without observed clicks, based on the (usually wrong) assumption that the ratchet would have clicked just after stopping the simulation. Each symbol denotes an independent simulation with a different random seed.
This plot contains 15 simulations with a total of 52 days of computing time and was produced by script RS005LL014 at 2002-10-09 16h45m41s. Corresponding simulations need 450 MB RAM (5x10^6 cells) or 900 MB RAM (1x10^7 cells).

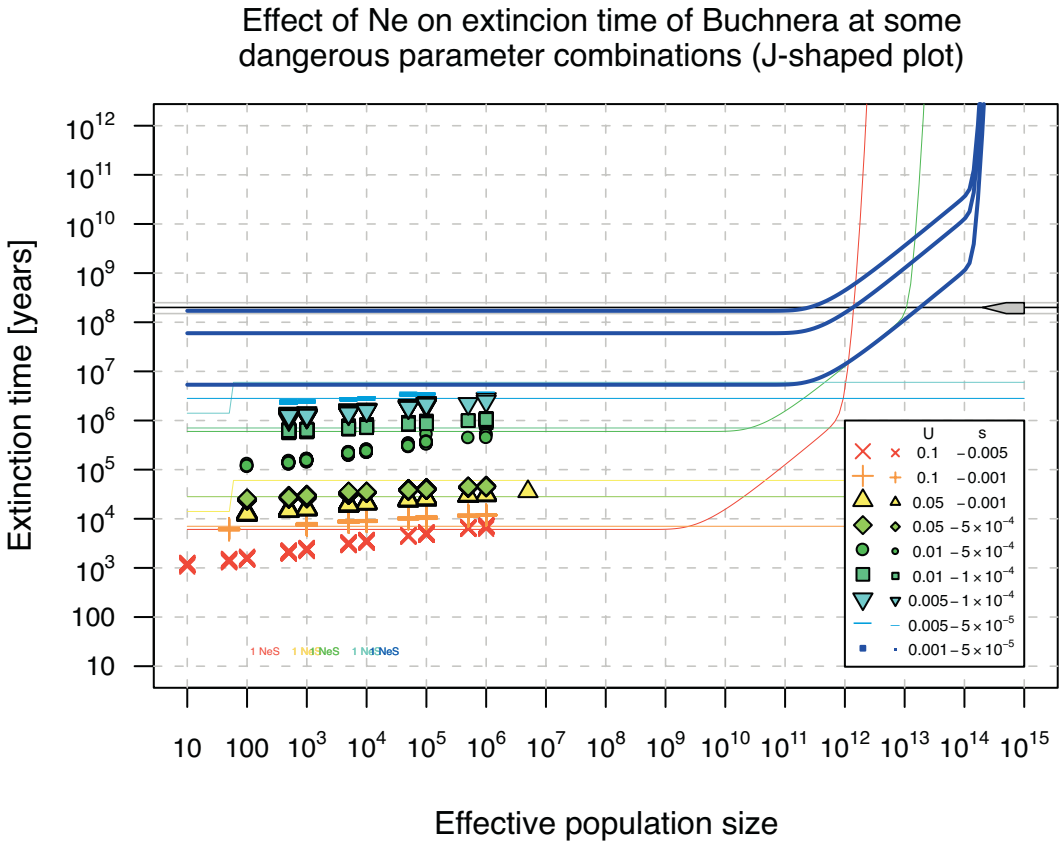


Figure 62 Influence of effective population size (N_e) on extinction time of *Buchnera* at some dangerous parameter combinations (J-shaped plot), excluding potential effects due to recombination within aphids and spatial structuring of the *Buchnera* population in aphids.
It is easy to see from this plot that any biologically feasible population size would not stop Muller's ratchet for the parameter combinations investigated here.
The arrow denotes the assumed age of the lines, including limits. The small coloured "1 NeS"-labels mark the border with neutrality for the selection coefficients corresponding to their colour. Thin lines use simple *Equation172* (see Chapter 18) with the respective means of T_{gen} and R_{max}. Thick lines mark one of the potentially more probable parameter combinations (s = -5x10^{-5}; U=0.001) using the corresponding upper and lower limits of T_{gen} and R_{max} to provide a feeling for the variability of extinction time estimates. Large symbols denote valid extinction time estimates from simulations with at least 2 observed clicks (usually many more). Small symbols denote lower limits for extinction times from simulations without observed clicks, based on the (usually wrong) assumption that the ratchet would have clicked just after stopping the simulation. Each symbol denotes an independent simulation with a different random seed.
This plot contains 2547 simulations with a total of 1.39 years of computing time and was produced by script RS005LL014 at 2002-10-09 16h50m52s.

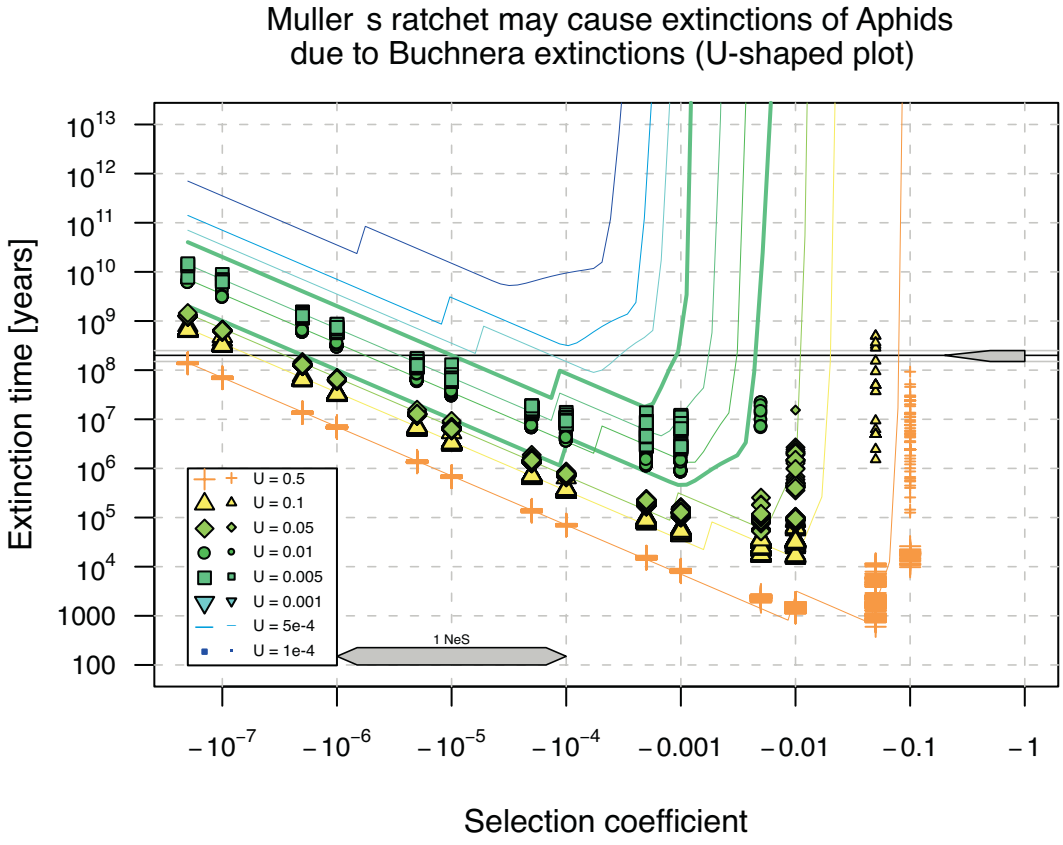


Figure 63 Muller's ratchet might cause extinctions of aphids with *Buchnera* as asexual genetic subsystem accumulating slightly deleterious mutations.
This plot assumes that no recombination occurs within bacteriocytes or within aphids and multi-level population genetics does not stop the ratchet, ie. no selective removal of endosymbionts within aphids
The upper arrow denotes the assumed age of endosymbionts in aphids, including limits. The lower arrow marks the border with neutrality for the population sizes used by spanning the selection coefficients from 1 N_eS for the largest (1 000 000 aphids) to 1 N_eS for the smallest population (10 000 aphids). Thin lines use simple *Equation172* (see Chapter 18) with N_e = 100 000 and the respective means of T_{gen} and R_{max}. Thick lines for U=0.005 (last green=lowest most probable mutation rate estimate) use the corresponding upper and lower limits of N_e, T_{gen} and R_{max} to provide a feeling for the variability of extinction time estimates. Large symbols denote valid extinction time estimates from simulations with at least 2 observed clicks (usually many more). Small symbols denote lower limits for extinction times from simulations without observed clicks, based on the (usually wrong) assumption that the ratchet would have clicked just after stopping the simulation. Each symbol denotes an independent simulation with a different random seed.
This plot contains 7465 simulations with a total of 7.02 years of computing time and was produced by script RS005LL014 at 2002-10-09 16h48m23s.

23.2 Extinction times

The values above were used to conduct simulations of Muller's ratchet with Simulator005 and to analyse corresponding simulation results. Figure 60 shows the U-shaped plot of extinction times over selection coefficients for the evolution of an asexual population of *Buchnera* endosymbionts living in a single otherwise optimal aphid line. Figure 61 shows corresponding extinction times of a single panmictic population of *Buchnera* with an effective size of about 10^7. As few simulation results are available for such large populations, Figure 62 shows the J-shaped plot of extinction time over population size for a series of dangerous combinations of mutation rate and selection coefficient. Finally, Figure 63 shows an U-shaped plot of extinction time over selection coefficient from the perspective of aphids that have an essential, asexual genetic subsystem called *Buchnera*.

These results show that from the endosymbionts perspective, a deleterious mutation rate of 0.001 per generation per genome leads to a genomic decay paradox regardless of what effective population sizes are employed. An exact quantification of the tread of the ratchet to the endosymbiont population in the best aphid line available yields (Figure 60, mean values for U=0.001):

- MDSR[65] = -0.0005 to -0.0001 at ca. 30 Myr and
- DSR[66] = -0.0007 to $-3x10^{-6}$ at ca. 200 Myr.

Similar details for other parameter combinations can be extracted from the thin lines in the figures given. They show that for deleterious mutation rates lower than $1 - 5 \times 10^{-4}$ /genome/generation Muller's ratchet is still operating, but leads to extinctions in time frames that are beyond the assumed age of endosymbionts in aphids.

23.3 Conclusions

This detailed assessment of the genomic decay paradox in *Buchnera* points to two possible directions. That mutation rates are close to the lower limit of biologically feasible values would be the simplest solution of the paradox, but seems unlikely at the moment. Only measurements of actual mutation rates of *Buchnera* in their natural host will be able to tell. If mutation rates turn out to be so large that they lead to a genomic decay paradox (as seems

65. Most Dangerous Range of Selection coefficients with associated minimal extinction time.
66. Dangerous Range of Selection coefficients that leads to extinction in assumed age of the line.

likely now), at least two other potential solutions to the paradox remain. If some processes associated with the complicated multi-level population dynamics do not solve the problem[53,52], one might consider the possibility that recombination of *Buchnera's* within bacteriocytes or even within an aphid might significantly slow down or even stop Muller's ratchet, as in many situations a little bit of sex has effects similar to regular sex[67].

A molecular look at the genome, however, suggests that quite a bit of genomic decay has been going on for quite a while. The genome of *Buchnera*49 is close the known lower limit for genome size[68] and has lost virtually every gene it does not need for replication or for production of the amino acids its host feeds on. Furthermore, the ratio of fixed non-synonymous to synonymous amino acid substitutions is higher in *Buchnera* than in its free-living relatives[69]. This is best explained by fixation of an increased number of slightly deleterious mutations. This evidence suggests that the genomic decay paradox in *Buchnera* might be one of the more complicated examples.

Thus, further research needs to determine more accurate mutational parameters and model multi-level population genetics in more detail to finally solve the puzzle of these scandalous symbionts[70].

67. Green & Noakes (1995) "Is a Little Bit of Sex as Good as a Lot?" J. theor. Biol. 174:87-96. - Peck & Waxman (2000) "What's wrong with a little sex?" J. evol. Biol. 13:63-69.
68. Fraser et al. (1995) "The Minimal Gene Complement of Mycoplasma genitalium", Science 270:397-403.
69. Moran (1996) "Accelerated evolution and Muller's rachet in endosymbiotic bacteria", Proc. Natl. Acad. Sci. USA 93:2873-2878.
70. Hurst & McVean (1996) "Evolutionary Genetics: ... and scandalous symbionts", Nature 381:650-651.

24 Muller's ratchet and microbial diversity

Bacteria can reproduce completely without recombination and can have a significant fraction of mutators in their population. Whenever these mutators become ecologically successful and do not recombine for longer periods of time, Muller's ratchet degrades their genomes. Here, Muller's ratchet is quantified (i) for free-living, asexual bacteria, (ii) for ancient asexual bacteria that have been trapped in sediments and might have grown cryptically and (iii) for RNA viruses. It is concluded that irregular sex as known in modern bacteria and RNA viruses must have played an important role in the evolution of microbes from very early times. However, when bacteria have learned how to survive in the long-term with occasional recombination, an important driving factor for the evolution of regular recombination disappeared. Thus, the mystery of the origin of sex might be bigger than previously thought.

24.1 Ratchet parameters in microbes

Uncultivability instead of extinction in bacteria

As argued in the last chapter, the effect of Muller's ratchet in bacteria cannot be that fewer offspring are produced, as they grow by binary fission only. However, decreased quality and speed of elementary enzymatic processes are expected to lead to an increase in doubling time[71]. An excellent example of this is the leprosy bacillus[72].

Thus the following treatment for bacteria assumes a minimal doubling time $\boldsymbol{DT_{min}}$ of 1 hour , and a maximal doubling time $\boldsymbol{DT_{max}}$ that is 332 times longer (nearly 14 days) in concordance with Equation (63) on page 291. The resulting equivalent of maximal reproductive capacity $\boldsymbol{R_{max}}$ of 10^{100} is used in Simulator005 as a substitute for a special bacterial simulator that models different generation lengths. In reality, quite a bit of variability will be associated with this figure. However, the following treatment does not consider this to avoid numerical problems. Furthermore, the interested reader can easily scale results to a particular situation using Equation (63) on page 291 and the information on "Conditions for extinction" on page 39.

The use of 10^{100} for $\boldsymbol{R_{max}}$ translates into a slowdown of growth due to genetic reasons from a speed that is readily cultured in a laboratory to a

71. See also Koch (1997) "Microbial physiology and ecology of slow growth", Microbiol. Mol. Biol. Rev. 61:305-318.
72. Cole et al. (2001) "Massive gene decay in the leprosy bacillus", Nature 409:1007-1011.

speed that would be considered as "uncultivable" for most approaches. Thus, the following treatment does not consider extinctions of bacteria in a strict sense, but only their transition to a state where they appear as uncultivable for usual attempts.

RNA virus extinction condition

In RNA viruses, conditions for extinction might be determined in a similar way to binary replicating viruses. If hosts are regularly dividing, one might even use host doubling time as a limit for viral doubling times, because a virus that grows slower than its host is not likely to be very dangerous. Alternatively, the following approach might be used for linearly replicating viruses. As these set up a 'virus production machine' that produces many viruses, one might compare them to individuals that have many offspring. All offspring are only one mutational generation away from their ancestor and as soon as this 'virus production machine' no longer substitutes itself, it becomes extinct. To quantify the corresponding parameters in detail can be a significant amount of work, especially as infection probabilities and survival outside of hosts play a role, too. Ultimately, a detailed analysis should use a simulator that is built explicitly for the situation in viruses. For a rough assessment of Muller's ratchet in viruses this study assumes 1000, 3000 and 10 000 as corresponding values for $\boldsymbol{R_{max}}$ for viruses. Again the results can be scaled to other values using Equation (63) on page 291 and the information on "Conditions for extinction" on page 39. However, this does not affect the general conclusions.

Age of lines

Currently, there are no age values that have a special meaning for free-living bacteria or RNA viruses, except the origin of life more than 10^9 years ago. Thus, the following plots mark 1, 10 and 100 million years for convenience. The calculations for ancient bacteria that had been trapped in either amber bees (25 Myr) or salt crystals (250 Myr) are marked with a mean of 100 Myr and corresponding limits.

Population sizes free bacteria

Census population sizes of the sum of all bacterial species are enormous[73] and viruses are even believed to be about an order of magnitude more frequent, at least in the ocean[74]. However, effective population sizes are considerably smaller, as the large numbers of bacteria in habitats that have no future do not contribute genetic material to the next generation. Estimates of effective population sizes for *Escherichia coli* are about 3×10^9

73. Whitman et al. (1998 "Prokaryotes: The unseen majority", Proc.Natl.Acad.Sci. USA 95:6578-6583) estimate that a total of 4-6 x 10^{30} prokaryotic cells belong to 10^5-10^7 procaryotic species.
74. Fuhrman (1999) "Marine viruses and their biogeochemical and ecological effects", Nature 399:541-548.

cells[75], but values of 10^7 cannot be excluded either[76]. Thus, for computing the ratchet in free-living bacteria values between 10^6 and 10^9 are assumed for the U-shaped plot with backup from a J-shaped plot to help estimate influence of population size at dangerous parameter combinations. The extrapolations presented here are based on the simple equation system described in Chapter 18 and probably reflect reality quite well; they should nevertheless be treated with a bit of caution until selected predictions of the simple equation system have been tested for such large population sizes.

Population sizes ancient bacteria

From all the possible ways in which a bacterial culture might survive for millions of years in a salt crystal or in a bee's abdomen in amber, only the possibility of cryptic growth is considered here (see discussion below). This would imply a stationary phase culture (density min 10^5, mid 10^6, max $5x10^6$ per ml according to modern lab experiments[77]) in the cavities of a few µl (min 1, mid 10 max 100 µl), leading to population sizes of 100 (min) 7100 (mean) and 500 000 (max) per closed cavity. This assumes that these cultures stay considerably below the current density record of 10^9 cells/ml for natural aquatic habitats[78] under optimal conditions.

Population sizes viruses

In RNA viruses, the difference between census and effective population size is believed to be even more dramatic than in bacteria, leading to surprisingly small effective population sizes[79]. This study assumes values of 10^4 (min) 10^6 (mean) and 10^8 (max).

Generation times

The exact values for generation time are somewhat difficult to estimate, as Simulator005 would need an 'effective generation time'. However, as generation times influence extinction time linearly, the interested reader can scale the results given to any needed value. For free-living bacteria exactly 1 day generation time was assumed in accordance with estimates from doubling times in stationary phase77. For ancient bacteria 1 (min), 3 (mid) and 10 (max) days were assumed, whereas for RNA viruses 0.1 (min), 1 (mid) and 10 (max) days were used.

75. Whittam (1996) "Genetic variation and evolutionary processes in natural populations of Escherichia coli", pp. 2708-2720 in: Neidhardt et al. (eds) Escherichia coli and Salmonella: cellular and molecular biology. 2, 2, Washington D.C., ASM Press.
76. Maruyama & Kimura (1980) "Genetic variability and effective population size when local extinction and recolonization of subpopulations are frequent", Proc. Natl. Acad. Sci. USA 77:6710-6714.
77. Finkel et al. (2000) "Long-term survival and evolution in the stationary phase", pp. 231-238 in: Storz & Hengge-Aronis (eds) Bacterial stress responses, Washington, D.C., ASM Press.
78. Kirschner et al. (2002) "Extreme productive microbial communities in shallow saline pools respond immediately to changing meteorological conditions", Environ. Microbiol. 4:546-555.
79. Fernando Garcia-Arenal (2002) personal communication. See also Moya et al. (1993) "Genetic structure of natural populations of the plant RNA virus tobacco mild green mosaic virus", Mol. Biol. Evol. 10:449-456.

Non-mutator, replication-based mutation rates

Neither bacteria nor RNA viruses analysed in this chapter were assumed to have multiple copies of their genomes. Thus haploid mutation rates are assumed to apply without change. RNA viruses are known for their high mutation rates of about 1 mutation per replication per genome (eg. 0.8 in Poliovirus, 6.5 in Bacteriophage Qβ, 3.5 in Vesicular Stomatitis Virus and 1 or more in Influenza A)[80]. It is assumed that deleterious mutation rates for a given order of magnitude of mutational effects are 0.1 per genome per generation in RNA viruses. For bacteria like *E. coli* the total genomic mutation rate is believed to be about 0.0025 per replication[81]. If a tenth of these mutations fall into the various orders of magnitude of mutational effects, then 0.00025 deleterious mutations of the most dangerous effect can be expected per replication.

Selection coefficients

This is approximately the deleterious mutation rate observed in a replication-based mutation accumulation experiment[82], albeit at a mutational effect of 1.2 %. Looking for larger mutational effects leads to even smaller deleterious mutation rates[83]. However, selection coefficients of both of these currently known replication based mutation accumulation experiments in bacteria are too large to drive Muller's ratchet (except for mutators). They will be completely removed by purifying selection (see U-shaped plots below). Another estimate of mutation effects arrived at $2x10^{-8}$ for amino acid changing base substitutions and $7x10^{-9}$ for synonymous base substitutions based on a completely different approach[84]. Muller's ratchet would certainly accumulate mutations with such small effects, but their combined damage appears to be too low to cause serious extinctions (see U-shaped plots below).

As there is no reason to assume that mutation effects in nature fall only into those two classes and the methodologies applied would not allow for

80. Drake (1993) "Rates of spontaneous mutation among RNA viruses", Proc. Natl. Acad. Sci. USA 90:4171-4175. - Drake et al. (1998) "Rates of spontaneous mutation", Genetics 148:1667-1686.
81. Drake (1991) "A constant rate of spontaneous mutation in DNA-based microbes", Proc. Natl. Acad. Sci. USA 88:7160-7164. - Drake et al. (1998) "Rates of spontaneous mutation", Genetics 148:1667-1686.
82. Kibota & Lynch (1996) "Estimate of the genomic mutation rate deleterious to overall fitness in E. coli", Nature 381:694-696.
83. Andersson & Hughes (1996 "Muller's ratchet decreases fitness of a DNA-based microbe", Proc. Natl. Acad. Sci. USA 93:906-907) found 5 mutants with an average mutation effect of 33% among 444 lines after 1700 generations suggesting a mutation rate of $7x10^{-6}$ per genome per generation in this mutational effect class. This is consistent with other observations, as the methods used would only detect a small fraction of all mutation effects (large, but not too large mutations).
84. Hartl et al. (1994) "Selection intensity for codon bias", Genetics 138:227-234.

detection of intermediate mutational effects, it is assumed that such intermediate effects appear in bacteria, too.

Importance of mutators in stationary phase

When bacteria lack nutrients in stationary phase, they try every evolutionary trick to find a new source of food, including transient increase of mutation rates in a subpopulation. This phenomenon is well known to be the source of of adaptive mutations[85]. If one of these mutators finds a mutation that allows for growth, the resulting selection coefficient is very large and leads to a growth advantage in stationary phase (GASP) phenotype[86]. If this happens repeatedly, the effective mutation rate of such an evolutionary line will be higher than replication based assays would suggest. Although some adaptive mutations are linked to increased recombination rates[87], the analysis presented here assumes no recombination. If no recombination occurs, all successful mutants carry with them a considerable number of deleterious mutations linked to them[88].

Extrapolation from experiment

The corresponding stationary phase mutation accumulation experiment described in Chapter 17 suggests a mutation rate of 0.078 deleterious mutations per generation, if one generation is assumed to last for 1 day. As many of the processes responsible for stationary phase mutation are probably temperature dependent, the mutation rate observed at 37°C should be scaled using a Q_{10} of 2 to more reasonable temperature of 7°C to 27°C. Q_{10} describes the dependency of the speed of a chemical reaction from temperature in such a way that an increase of 10°C leads to the given factor for the increase in reaction speed[89]. If this experiment and the corresponding scaling apply to natural conditions, deleterious mutation rates could well be in

85. Foster (1999) "Mechanisms of stationary phase mutation: a decade of adaptive mutation", Annu Rev Genet 33:57-88. - Foster & Rosche (1999) "Mechanisms of mutation in nondividing cells. Insights from the study of adaptive mutation in Escherichia coli", Ann N Y Acad Sci 870:133-145. - Rosenberg et al. (1994) "Adaptive mutation by deletions in small mononucleotide repeats", Science 265:405-407. - Rosenberg et al. (1995) "Molecular handles on adaptive mutation", Mol. Microbiol. 18:185-189. - Rosenberg (1997) "Mutation for survival", Curr. Opin. Genet. Dev. 7:829-834. - Rosenberg et al. (1998) "Transient and heritable mutators in adaptive evolution in the lab and in nature", Genetics 148:1559-1566. - Foster (1998) "Adaptive mutation: Has the unicorn landed?" Genetics 148:1453-1459. - Torkelson et al. (1997) "Genome-wide hypermutation in a subpopulation of stationary-phase cells underlies recombination-dependent adaptive mutation", EMBO J. 16:3303-3311. - Finkel & Kolter (1999) "Evolution of microbial diversity during prolonged starvation", Proc. Natl. Acad. Sci. USA 96:4023-4027. - Taddei et al. (1995) "cAMP-dependent SOS induction and mutagenesis in resting bacterial populations", Proc. Natl. Acad. Sci. USA 92:11736-11740.

86. Finkel et al. (2000) "Long-term survival and evolution in the stationary phase", pp. 231-238 in: Storz & Hengge-Aronis (eds) Bacterial stress responses, Washington, D.C., ASM Press.

87. McKenzie & Rosenberg (2001) "Adaptive mutations, mutator DNA polymerases and genetic change strategies of pathogens", Curr. Opin. Microbiol. 4:586-594.

88. Giraud et al. (2001) "Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut", Science 291:2606-2608. - de Visser (2002) "The fate of microbial mutators", Microbiology 148:1247-1252.

the range of 0.05 for tropical climates of 30°C or 0.01 for average temperatures of 7°C. The lines used for determination of these mutation rates were not mutators (at least when the experiment started).

Mutators

Assuming that genetically heritable mutators[90] would have led to correspondingly higher mutation rates would suggest deleterious mutation rates of 0.1 to 1 for mutator lines, depending on the mutator strength. As a fraction of mutators is readily found in natural populations[91], not only among pathogens[92], one might expect that some lines actually accumulate deleterious mutations at such a high speed. Here, this logic is applied to free-living and anciently trapped bacteria alike.

24.2 Muller's ratchet in free bacteria

If the values above are used to compute the U-shaped plot of extinction time over selection coefficient (Figure 64), the following patterns appear for free-living bacteria that live completely without recombination.

89. See p. 378 in Adam et al. (1988) "Physikalische Chemie und Biophysik". Zweite, völlig neu bearbeitete und erweiterte Auflage, Berlin, Springer-Verlag.
90. de Visser (2002) "The fate of microbial mutators", Microbiology 148:1247-1252. - Miller (1996) "Spontaneous mutators in bacteria: Insights into pathways of mutagenesis and repair", Annu. Rev. Microbiol. 50:625-643. - Miller et al. (1999) "Direct selection for mutators in Escherichia coli", J. Bacteriol. 181:1576-1584. - Miller et al. (2002) "Escherichia coli Strains (ndk) Lacking Nucleoside Diphosphate Kinase Are Powerful Mutators for Base Substitutions and Frameshifts in Mismatch-Repair-Deficient Strains", Genetics 162:5-13. - McKenzie & Rosenberg (2001) "Adaptive mutations, mutator DNA polymerases and genetic change strategies of pathogens", Curr. Opin. Microbiol. 4:586-594. - Giraud et al. (2001) "The rise and fall of mutator bacteria", Curr. Opin. Microbiol. 4:582-585.
91. Matic et al. (1997) "Highly variable mutation rates in commensal and pathogenic Escherichia coli", Science 277:1833-1834. - Sniegowski et al. (1997) "Evolution of high mutation rates in experimental populations of E. coli", Nature 387:703-705. - Denamur et al. (2002) "High frequency of mutator strains among human uropathogenic Escherichia coli isolates", J. Bacteriol. 184:605-609. - Mao et al. (1997) "Proliferation of mutators in a cell population", J. Bacteriol. 179:417-422. - Giraud et al. (2001) "Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut", Science 291:2606-2608.
92. LeClerc et al. (1996) "High mutation frequencies among Escherichia coli and Salmonella pathogens", Science 274:1208-1211. - LeClerc & Cebula (1997) "Highly variable mutation rates in commensal and pathogenic Escherichia coli - Reply", Science 277:1834-1834. - Oliver et al. (2000) "High frequency of hypermutable Pseudomonas aeruginosa in cystic fibrosis lung infection", Science 288:1251-1254.

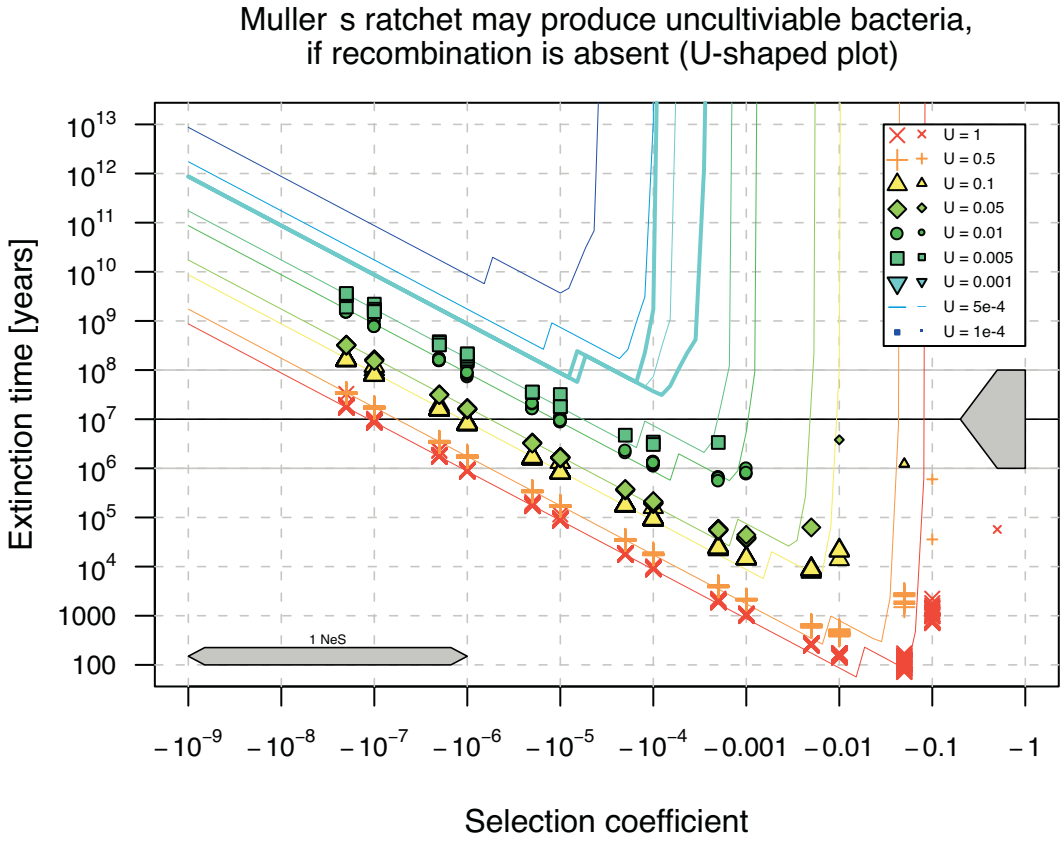


Figure 64 Muller's ratchet might lead to apparent uncultivability of non-recombining, free-living bacteria due to accumulation of slightly deleterious mutations in their genomes.
'Extinction time' denotes the time to loss of cultivability. The upper arrow denotes 1, 10 and 100 million years for comparison. The lower arrow marks the border with neutrality for the population sizes used by spanning the selection coefficients from 1 N_eS for the largest (10^9 cells) to 1 N_eS for the smallest population (10^6 cells). Thin lines use simple *Equation172* (see Chapter 18) with N_e = 31 600 000 and the respective means of T_{gen} and R_{max}. Thick lines for U = 0.001 (highest blue = non-mutator replication based estimate) use the corresponding upper and lower limits of N_e to provide a feeling for the variability of extinction time estimates (limits for T_{gen} and R_{max} are like their mean in this plot). Large symbols denote valid extinction time estimates from simulations with at least 2 observed clicks (usually many more). Small symbols denote lower limits for extinction times from simulations without observed clicks, based on the (usually wrong) assumption that the ratchet would have clicked just after stopping the simulation. Each symbol denotes an independent simulation with a different random seed.
This plot contains 3009 simulations with a total of 2.72 years of computing time and was produced by script RS005LL015 at 2002-10-09 19h33m39s.

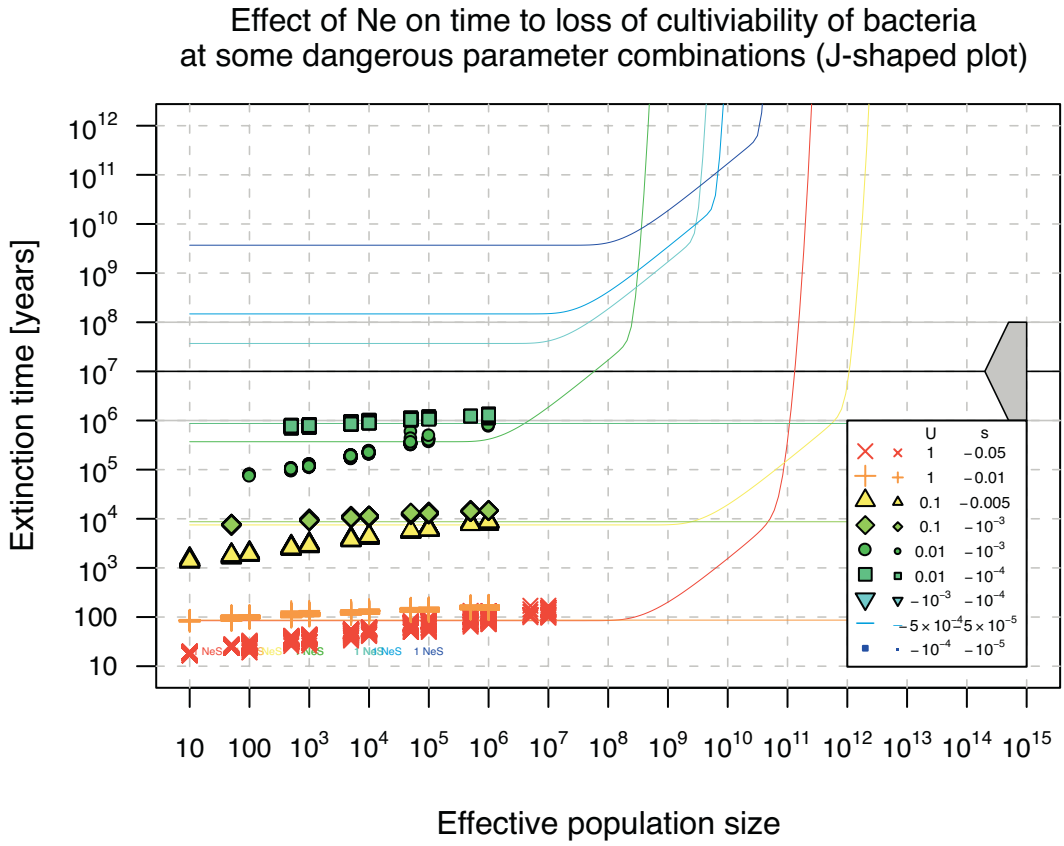


Figure 65 Influence of effective population size (N_e) on time to loss of cultivability of non-recombining, free-living bacteria at some dangerous parameter combinations (J-shaped plot).
It is easy to see from this plot that biologically feasible population sizes would stop Muller's ratchet only for some of the parameter combinations investigated here.
'Extinction time' denotes the time to loss of cultivability. The arrow denotes 1, 10 and 100 million years for comparison. The small coloured "1 NeS"-labels mark the border with neutrality for the selection coefficients corresponding to their colour. Thin lines use simple *Equation172* (see Chapter 18) with the respective means of T_{gen} and R_{max}. Large symbols denote valid extinction time estimates from simulations with at least 2 observed clicks (usually many more). Small symbols denote lower limits for extinction times from simulations without observed clicks, based on the (usually wrong) assumption that the ratchet would have clicked just after stopping the simulation. Each symbol denotes an independent simulation with a different random seed.
This plot contains 2268 simulations with a total of 1.06 years of computing time and was produced by script RS005LL015 at 2002-10-09 19h36m08s.

Extremely low mutation rates survive

The lower, replication-based mutation rates for non-mutators are at the brink of leading to a genomic decay paradox within the assumed age of life on planet Earth. Thus, existence of a completely asexual line of bacteria with such a low mutation rate might not be excluded completely. However, given the fact that most bacteria spend most of their time in the stationary phase with corresponding higher mutation rates, it is unlikely that such a line would exist at all (together with its long-term chemostat environment need to survive as such). Currently, it is unclear whether cells with an anti-mutator phenotype[93] (ie. bacteria with mutation rates below the average) exhibit lower mutation rates in stationary phase or not. If they do, these might play a key role in long-term preservation of bacterial identities.

Stationary phase mutation rates must recombine to survive

For deleterious mutation rates like those observed in Chapter 17 or for replication based mutators, extinction times can be of the order of a few million years or much less. If no other potential solution applies to the resulting genomic decay paradox, such mutation rates suggest that recombination in these bacteria is not 'nice to have', but rather essential for survival. It is unclear whether the levels of recombination that can be associated with stationary phase mutations[94] are enough to ensure survival. Only experimental observations and subsequent simulations will be able to tell.

Large populations can become extinct too

The J-shaped plot of extinction time over effective population size $\mathbf{N_e}$ for various dangerous parameter combinations (Figure 65) shows a number of occasions, where population sizes in the range of 10^8 to 10^{12} could actually change the qualitative conclusion from 'the ratchet is a threat' to 'purifying selection removes deleterious mutations'. However, this occurs only at parameter combinations that are in any case somewhat close to the selective wall. For a number of other dangerous parameter combinations, especially

93. Drake (1993) "General antimutators are improbable", J. Mol. Biol. 229:8-13. - Fijalkowska et al. (1993) "Mutants of Escherichia coli with increased fidelity of DNA replication", Genetics 134:1023-1030. - Fijalkowska & Schaaper (1993) "Antimutator mutations in the alpha subunit of Escherichia coli DNA polymerase III: Identification of the responsible mutations and alignment with other DNA polymerases", Genetics 134:1039-1044. - Goodman & Fygenson (1998) "DNA polymerase fidelity: From genetics toward a biochemical understanding", Genetics 148:1475-1482. - Hadjimarcou et al. (2001) "Identification of a mutant DNA polymerase delta in Saccharomyces cerevisiae with an antimutator phenotype for frameshift mutations", Genetics 158:177-186. - Kunz et al. (1998) "DNA sequence analysis of spontaneous mutagenesis in Saccharomyces cerevisiae", Genetics 148:1491-1505. - Reha-Krantz (1998) "Regulation of DNA polymerase exonucleolytic proofreading activity: Studies of bacteriophage T4 "antimutator" DNA polymerases", Genetics 148:1551-1557. - Schaaper (1993) "The mutational specificity of two Escherichia coli dnaE antimutator alleles as determined from lacI mutation spectra", Genetics 134:1031-1038. - Schaaper (1998) "Antimutator mutants in bacteriophage T4 and Escherichia coli", Genetics 148:1579-1585.

94. McKenzie & Rosenberg (2001) "Adaptive mutations, mutator DNA polymerases and genetic change strategies of pathogens", Curr. Opin. Microbiol. 4:586-594.

where $U/s \geq 100$, population sizes do not appear to play any role. Thus it will be interesting to get more detailed data on actual ratchet parameters in free-living bacteria to see whether large effective population sizes could stop Muller's ratchet for the distributions of mutational effects that actually occur in bacteria. However, as such data is not easily accessible with current methodology, such an analysis is not likely to be possible in the near future.

Non-recombining mutators are degraded to uncultivability

Bacteria that cannot recombine for various reasons (eg. isolation, loss of genes needed for recombination) appear to be doomed to genomic decay. A great deal of uncertainty exists as to how bacterial extinction should be defined exactly[95]. It is proposed here that in light of potential bacterial immortality, the speed of decay of replication rates is the critical parameter in this debate. The decay of replication rate can easily be linked to deleterious mutations, as small impairments in import or processing of nutrients as well as in other enzymatic reactions essential for replication cannot do other than slow down overall speed of growth[96].

For practical purposes this analysis assumes that a bacterial population is extinct, when replication speed has declined to uncultivability because its fastest available generation times are 14 days (see treatment of reproductive capacity above). A real extinction, of course, will happen only if the last (potentially immortal) bacterial cell has ceased to exist. Thus the actual lower limit of reproductive capacity, where a real extinction occurs, depends very much on the (environmentally controlled) practical death rate in that population. Observations in nature will be needed to arrive at more precise estimates.

Conclusion

Such a view suggests that a significant fraction of the large majority of uncultivable bacteria[95,97] might actually be descendants from permanent or transient mutators of known species that have been degraded by Muller's

95. Stahl & Tiedje (2002) "Microbial ecology and genomics: a crossroads of opportunity", (http://www.asmusa.org/acasrc/pdfs/MicroEcoReport.pdf) Washington, D.C., American Academy of Microbiology.
96. See also Koch (1997) "Microbial physiology and ecology of slow growth", Microbiol. Mol. Biol. Rev. 61:305-318.
97. Hugenholtz et al. (1998) "Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity", J. Bacteriol. 180:4765-4774. - Staley et al. (1997) "The Microbial World: Foundation of the biosphere", (http://www.asmusa.org/acasrc/pdfs/Colloquia/microbialworld.pdf) Washington, D.C., American Academy of Microbiology. - Staley & Reysenbach (2001) "Biodiversity of Microbial Life: Foundation of Earth's Biosphere", New York, Wiley. - Nealson & Ghiorse (2001) "Geobiology: Exploring the interface between the biosphere and the geosphere", (http://www.asmusa.org/acasrc/pdfs/Colloquia/geobiology.pdf) Washington, D.C., American Academy of Microbiology. - Amann et al. (1995) "Phylogenetic Identification and In Situ Detection of Individual Microbial Cells without Cultivation", Microbiol. Rev. 59:143-169. - Torsvik et al. (1990) "High diversity of DNA of soil bacteria", Appl. Environ. Microbiol. 56:782-787.

ratchet to the point that their growth rate is below usual detection limits. Alternatively they might have lost genes by mutation that were not needed in their special environments for a long time, but are needed for growth under standard laboratory conditions, a general principle that could be observed on a small scale in the laboratory[98].

24.3 Muller's ratchet in ancient bacteria

The discovery of ancient bacteria[99] like those from the abdomen of bees trapped in 25 Myr old amber[100] or like those from an enclosure in 250 Myr old salt crystals[101] has generated considerable stir. As many of these findings do not appear to be caused by contamination, the main debate is about how biological material in general and DNA in particular could possibly be preserved for such extended periods of time without suffering from razemization, depurination and general degradation[102]. While a number of biochemical mechanisms[103] are being discussed that could potentially stabilise cells to facilitate survival over such long period of time, one of the hypotheses put forth to explain such survival suggests cryptic *in situ* growth[104] to ensure renewal and integrity of the trapped cells. Assuming that supply with nutrients and energy through submicroscopic fissures can actually allow such a population to persist, this section concentrates on the population genetic consequences of such a situation.

98. Cooper & Lenski (2000) "The population genetics of ecological specialization in evolving Escherichia coli populations", Nature 407:736-739. - Giraud et al. (2001) "Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut", Science 291:2606-2608.
99. Kennedy et al. (1994) "Preservation records of micro-organisms: evidence of the tenacity of life", Microbiology 140:2513-2529. - Keilin (1959) "The Leeuwenhoek Lecture: The problem of anabiosis or latent life: history and current concept", Proc. R. Soc. Lond. B Biol. Sci. 150:149-191.
100. Cano & Borucki (1995) "Revival and identification of bacterial spores in 25- to 40-million-year-old Dominican amber", Science 268:1060-1064. - Fischman (1995) "Have 25-Million-year-old bacteria returned to life?" Science 268:977-977. - Cano (1994) "Bacillus DNA in amber: A window to ancient symbiotic relationships?" ASM News 60:129-134. - Yousten & Rippere (1997) "DNA similarity analysis of a putative ancient bacterial isolate obtained from amber", FEMS Microbiol. Lett. 152:345-347. - Greenblatt et al. (1999) "Diversity of microorganisms isolated from amber", Microbial Ecology 38:58-68. - Lambert et al. (1998) "Staphylococcus succinus sp. nov., isolated from Dominican amber", International Journal of Systematic Bacteriology 48:511-518.
101. Vreeland et al. (2000) "Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal", Nature 407:897-900. - Hazen & Roedder (2001) "Biogeology. How old are bacteria from the Permian age?" Nature 411:155-156.
102. Lindahl (1993) "Instability and decay of the primary structure of DNA", Nature 362:709-715. - Herrmann & Hummel, (eds, 1994) "Ancient DNA: Recovery and analysis of genetic material from paleontological, archaeological, museum, medical, and forensic specimens", New York, Springer.

Results

Using general parameter combinations for free-living bacteria and population sizes feasible for such long-term entrapments allow production of the corresponding U-shaped plot of time to uncultivability over selection coefficient (Figure 66). It shows the same general result as the plot for free-living bacteria, except that very small population sizes might lead to significantly more genomic decay. While more detailed information about mutation rates and the distribution of mutational effects will be needed for more specific predictions, it appears that cryptic growth of bacteria would not prevent extinctions due to Muller's ratchet over such extended periods of time. Recombination of individuals from the enclosed population could stop decay, but only for larger population sizes, as mutations with $|s| < 1/N_e$ (ie. very slightly deleterious mutations) accumulate effectively like neutral mutations even in recombining populations[105].

Cryptic growth is threatened by Muller's ratchet

Quantification of Muller's ratchet in ancient bacteria suggests that cryptic growth of bacteria is probably no answer to the question of how bacteria managed to survive that long. In light of the arguments in the previous section about origins of uncultivability in free-living bacteria, it is interesting to note that a reasonable number of bacteria in amber are uncultivable for the techniques applied[106].

103. A number of examples show that specialised proteins can lead to extraordinary stability of DNA against various kinds of damaging influences. However, it is unclear whether such proteins could (or did) stabilise ancient cells long enough to avoid damage increasing with time due to ionising radiation and other dangerous processes. Some preserving processes are described in:
Nicholson et al. (2000) "Resistance of Bacillus endospores to extreme terrestrial and extraterrestrial environments", Microbiol. Mol. Biol. Rev. 64:548-572. - Clegg (2001) "Cryptobiosis-a peculiar state of biological organization", Comp Biochem Physiol B Biochem Mol Biol 128:613-624. - Clegg et al. (1999) "Adaptive significance of a small heat shock/a-crystallin protein (p26) in encysted embryos of the brine shrimp, Artemia franciscana", Am. Zoologist 39:836-847. - Clegg et al. (2000) "Long-term anoxia in encysted embryos of the crustacean, Artemia franciscana: viability, ultrastructure and stress proteins", Cell Tissue Res 301:422-446. - Potts (1994) "Desiccation tolerance of prokaryotes", Microbiol. Rev. 58:755-805. Potts (1999) "Mechanisms of desiccation tolerance in cyanobacteria", Eur J Phycol 34:319-328. - Setlow (1995) "Mechanisms for the prevention of damage to DNA in spores of Bacillus species", Annu. Rev. Microbiol. 49:29-54. - Setlow & Setlow (1995) "Small, acid-soluble proteins bound to DNA protect Bacillus subtilis spores from killing by dry heat", Appl. Environ. Microbiol. 61:2787-2790. - Setlow (1992) "I will survive: Protecting and repairing spore DNA", J. Bacteriol. 174:2737-2741.

104. First proposed in Postgate & Priest (1995) "Putative oligocene spores", Microbiology 141:2763-2764. More details are discussed in Morita (2000) "Is H2 the universal engery source for long-term survival?" Microbial Ecology 38:307-320.

105. Kondrashov (1995) "Contamination of the Genome by Very Slightly Deleterious Mutations: Why Have We Not Died 100 Times Over?" J. theor. Biol. 175:583-594. - Lynch et al. (1995) "Mutational meltdowns in sexual populations", Evolution 49:1067-1080.

106. Greenblatt et al. (1999) "Diversity of microorganisms isolated from amber", Microbial Ecology 38:58-68.

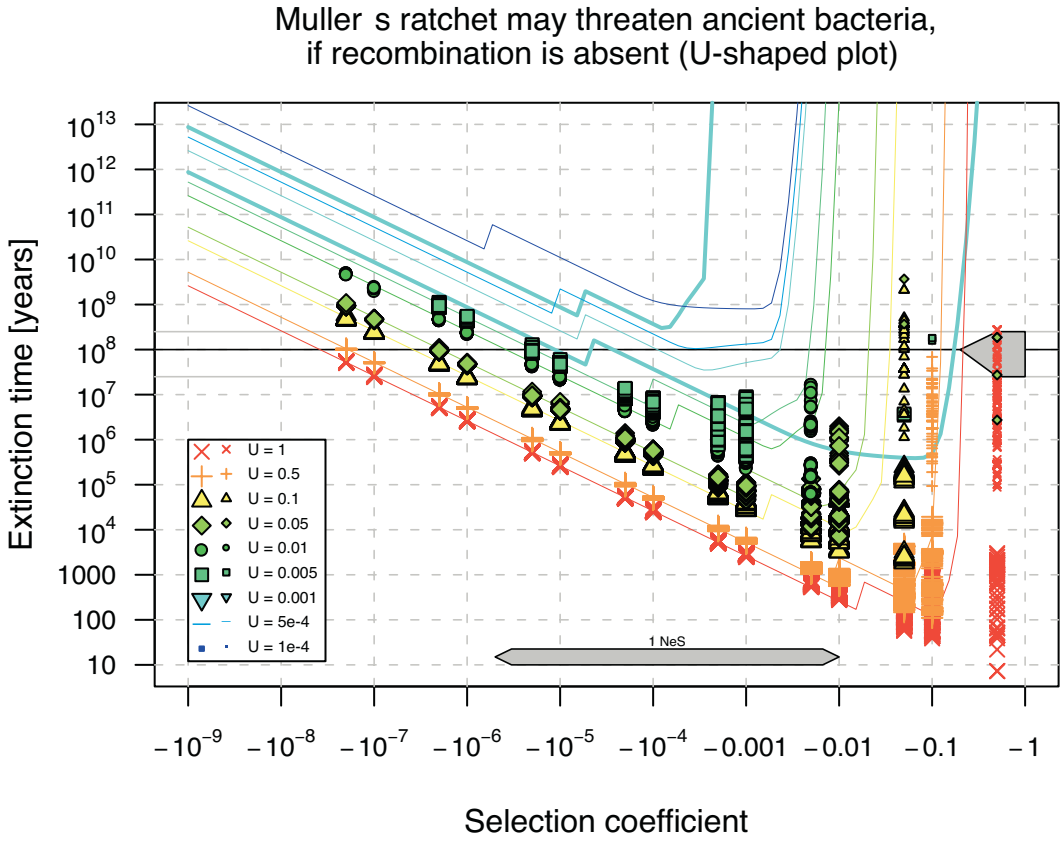


Figure 66 Muller's ratchet might lead to apparent uncultivability of non-recombining, ancient bacteria due to accumulation of slightly deleterious mutations in their genomes.
'Extinction time' denotes the time to loss of cultivability. The upper arrow denotes 25, 100 and 250 million years for comparison. The lower arrow marks the border with neutrality for the population sizes used by spanning the selection coefficients from 1 N_eS for the largest (500 000 cells) to 1 N_eS for the smallest population (100 cells). Thin lines use simple *Equation172* (see Chapter 18) with N_e = 7070 and the respective means of T_{gen} and R_{max}. Thick lines for U = 0.001 (highest blue = non-mutator replication based estimate) use the corresponding upper and lower limits of N_e and T_{gen} to provide a feeling for the variability of extinction time estimates (limits for R_{max} are equal to mean in this plot). Large symbols denote valid extinction time estimates from simulations with at least 2 observed clicks (usually many more). Small symbols denote lower limits for extinction times from simulations without observed clicks, based on the (usually wrong) assumption that the ratchet would have clicked just after stopping the simulation. Each symbol denotes an independent simulation with a different random seed.
This plot contains 15473 simulations with a total of 7.11 years of computing time and was produced by script RS005LL016 at 2002-10-10 11h00m48s.

This might support the view that extinctions of ancient bacteria in amber actually do play a significant role – if cryptic growth of bacteria is possible at all under such circumstances[107].

The sediment-preservation hypothesis

However, if some special preservation proteins preserve ancient bacteria well enough, there is no need to fear Muller's ratchet as genomes are shielded from all mutations induced by biological activity. The only upper limit of survival would then be due to ionising radiation[108]. If cells could survive many millions of years without change in such a state, this could have dramatic consequences for our view on present microbial diversity:

> *For free bacteria that are sooner or later driven to extinction due to Muller's ratchet, extraordinarily well-preserved ancient bacteria trapped in ancient sediments could be a steady source of fresh bacteria that is continually released, as sediments are eroded.*

The repeated finding that geologically ancient bacteria (by a rigorous geological and microbiological assessment) have high molecular similarity to their modern relatives[109] could be interpreted as evidence for this sediment preservation hypothesis[110]. Interestingly, the conflict between slow evolution of bacteria over geological timescales compared to observed modern mutation rates in the same species has the same problem structure as the mutation rate paradox in mtDNA, as described in Chapter 3.

Thus, such a mechanism would add another potential solution to the general genomic decay paradox observed in free-living bacteria: Not only a different distribution of mutational effects or lower mutation rates or recombination, but also reinfection from ancient sediments could explain the persistence of free-living bacteria in the light of the threat from Muller's

107. If protein stabilisation of DNA is the main cause of longevity, uncultivability could also be due to deleterious mutations accumulated by exposure to long-term ionizing radiation. Alternatively, everything might still be OK with the cells, but we just do not know the conditions for cultivation of these specialists.

108. Nicholson et al. (2000) "Resistance of Bacillus endospores to extreme terrestrial and extraterrestrial environments", Microbiol. Mol. Biol. Rev. 64:548-572.

109. Maughan et al. (2002) "The paradox of the "Ancient" bacterium which contains "Modern" protein-coding genes", Mol. Biol. Evol. 19:1637-1639. - Graur & Pupko (2001) "The Permian bacterium that isn't", Mol. Biol. Evol. 18:1143-1146.

110. A precursor of this hypothesis was put forward in a *NewYork Times* interview of October 19, 2002 by one of the co-authors of the paper reporting the finding of *Bacillus permians* (Vreeland et al. (2000) "Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal", Nature 407:897-900.) as a possible explanation for the observed similarity between modern *Salibacillus marismortui* and ancient *Bacillus permians*. However, no connection was made with Muller's ratchet. (Source: Graur & Pupko (2001) "The Permian bacterium that isn't", Mol. Biol. Evol. 18:1143-1146.)

ratchet. Only further research will be able to discriminate between these potential solutions.

24.4 Muller's ratchet in the deep hot biosphere

The last decade has seen an overwhelming number of reports showing that bacteria deep inside the earth contribute a very large part to the existing biomass[111]. For bacteria that live in some rocks, motility can be limited to geological events that move stones, whereas others can move more easily (or be moved eg. by groundwater flow). Depending on the exact circumstances (eg. stationary phase like nutrient limitation in combination with repeated freezing and thawing or high subsurface temperatures), some of these environments are likely to induce a considerable number of mutations. Although exact analyses need to report more details about migration, recombination and mutation rates in the corresponding rocks, and further simulations are needed to integrate such information, a large majority of bacteria from the lithosphere could turn out to be in a similar situation to the ancient bacteria described above. Further investigation of this issue will be interesting, especially in the light of current interest in subsurface bacteria[112].

24.5 Muller's ratchet in RNA viruses

The U-shaped plot of extinction time over selection coefficient for the ratchet parameters specified above for RNA viruses can be found in Figure 67. Although the values estimated for the various parameters are subject to considerable uncertainty, it becomes quite clear that known mutation rates suggest that RNA viruses become extinct in extraordinary short timescales from an evolutionary point of view, if no solution exists for this genomic decay paradox.

111. Fredrickson & Onstott (1996) "Microbes deep inside the earth", Sci. Am. August:42-47 (=Spektrum der Wissenschaft Dezember:66-71). - Gold (1992) "The deep, hot biosphere", Proc. Natl. Acad. Sci. USA 89:6045-6049. - Parkes et al. (2000) "Recent studies on bacterial populations and processes in subseafloor sediments: A review", Hydrogeology Journal 8:11-28. - Krumholz (2000) "Microbial communities in the deep subsurface", Hydrogeology Journal 8:4-10. - Whitman et al. (1998) "Prokaryotes: The unseen majority", Proc. Natl. Acad. Sci. USA 95:6578-6583. - Vorobyova et al. (1997) "The deep cold biosphere: facts and hypothesis", FEMS Microbiology Reviews 20:277-290.

112. Nealson & Ghiorse (2001) "Geobiology: Exploring the interface between the biosphere and the geosphere", (http://www.asmusa.org/acasrc/pdfs/Colloquia/geobiology.pdf) Washington, D.C., American Academy of Microbiology.

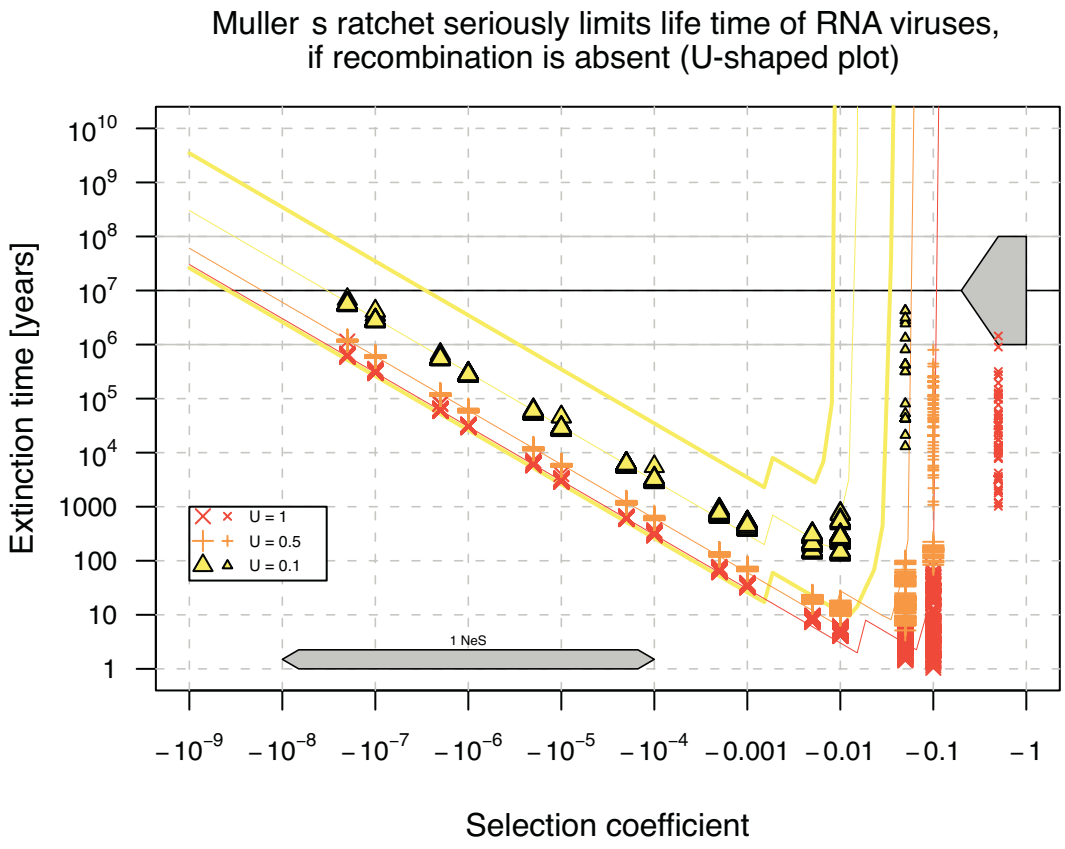


Figure 67 Muller's ratchet might lead to extinction of non-recombining viruses due to accumulation of slightly deleterious mutations in their genomes.
The upper arrow denotes 1, 10 and 100 million years for comparison. The lower arrow marks the border with neutrality for the population sizes used by spanning the selection coefficients from 1 N_eS for the largest (10^8 particles) to 1 N_eS for the smallest population (10^4 particles). Thin lines use simple *Equation172* (see Chapter 18) with $N_e = 10^6$ and the respective means of T_{gen} and R_{max}. Thick lines for U=0.1 (yellow = most probable deleterious mutation rate estimate) use the corresponding upper and lower limits of N_e, T_{gen} and R_{max} to provide a feeling for the variability of extinction time estimates. Large symbols denote valid extinction time estimates from simulations with at least 2 observed clicks (usually many more). Small symbols denote lower limits for extinction times from simulations without observed clicks, based on the (usually wrong) assumption that the ratchet would have clicked just after stopping the simulation. Each symbol denotes an independent simulation with a different random seed.
This plot contains 8194 simulations with a total of 4.29 years of computing time and was produced by script RS005LL017 at 2002-10-10 10h39m33s.

Potential solutions

Genomic decay in some RNA viruses is certainly delayed by recombination[113]; however, small effective population sizes might offset this advantage by increasing the range of very slightly deleterious mutations (ie. effectively neutral mutations, where $|s|<1/N_e$). Proteins in viruses are known to contain only a few positions that can exhibit drastic mutation effects; mutations at all other positions are either very slightly deleterious (ie. effectively neutral) or really neutral (ie. their effects cannot combine to form something dangerous)[114]. Although distinguishing between effectively neutral and really neutral becomes a matter of life and death in this case, current methodologies allow only very limited access to answers. However, as in the case of the other systems investigated, there is currently no reason to assume that most mutations are either strongly deleterious or completely neutral, but do not have selection coefficients from those five to seven dangerous orders of magnitude in between.

Conclusion

The possibility of extinction of RNA viruses is supported by the actual observation of mutational meltdown and extinction of *tobacco mosaic tobamovirus* in an Australian population of *Nicotiana glauca*[115]. The results presented here suggest that some continuous source of RNA viruses must exist to substitute extinct viruses in the long-term. If there is no such source, RNA viruses have either evolved only recently or they exhibit a mutation rate paradox: old viruses mutate very slowly, young viruses mutate very fast. Analyses of phylogenetic trees of retroviruses encounter the same paradox[116]. However, if there is a continuous evolutionary source of RNA viruses, it will be of great interest to find it, because RNA viruses play an important role in ecology[117] and are the cause of a large number of infectious diseases[118]. Much is known about the evolution of RNA viruses[119], but the chances are good that there is even more to discover.

113. Chao (1994) "Evolution of genetic exchange in RNA viruses", pp. 233-250 in: Morse (ed) The evolutionary biology of viruses, New York, Raven Press.
114. Sala & Wain-Hobson (1999) "Drift and conservatism in RNA virus evolution: Are they adapting or merely changing?" pp. 115-140 in: Domingo et al. (eds) Origin and evolution of viruses, San Diego, Academic Press. See also "The mutation effect pyramid." on page 338.
115. Fraile et al. (1997) "A century of tobamovirus evolution in an australian population of Nicotiana glauca", J. Virol. 71:8316-8320.
116. Doolittle et al. (1989) "Origins and evolutionary relationships of retroviruses", Quart. Rev. Biol. 64:1-30.
117. Fuhrman (1999) "Marine viruses and their biogeochemical and ecological effects", Nature 399:541-548.
118. Eg. Krause (1992) "The Origin of Pagues: Old and New", Science 257:1073-1078. - Mitchison (1993) "Will We Survive? As host and pathogen evolve together, will the immune system retain the upper hand?" Sci. Am. September:102-108. - Ewald (1994) "Evolution of mutation rate and virulence among human retroviruses", Philos. Trans. R. Soc. Lond. B Biol. Sci 346:333-343.

24.6 Conclusion

Global computing allows fast new predictions

As pointed out above, considerable uncertainty about the parameters used might lead to slightly different conclusions. However, the probability that more precise parameter values will lead to a completely different picture is quite low in light of the wide error margins employed here. In any case, the global computing system described in this work can easily be used to adjust predictions to more precise (or different) parameter estimates. If new simulations are needed, they can be scheduled to the web within a few hours; the time until new results are received is reasonably fast (days to weeks in the current semi-automated mode of operation; details depend on computational complexity and current participation). The plots presented can be easily generated or updated by corresponding scripts for R, the statistical analysis and programming system[120] that is used for evaluating results. Finally, new simulators can be developed to do similar analyses for more sophisticated biological models, eg. including age structure, recombination and spatial structure.

Although it is still unclear to what degree recombination can stop Muller's ratchet in the natural lifestyle of bacteria and RNA viruses, the analyses above have shown that it is worth taking long-term effects of slightly deleterious mutations seriously when considering evolution in microbes. The resulting overall picture just becomes more robust when a completely independent side is added to the story.

119. Morse, (ed, 1994) "The evolutionary biology of viruses", New York, Raven Press. - Domingo et al., (eds, 1999) "Origin and evolution of viruses", San Diego, Academic Press.
120. R-system (2002) "R: A Language for Data Analysis and Graphics" http://www.r-project.org/ - Ihaka & Gentleman (1996) "R: A Language for Data Analysis and Graphics", Journal of Computational and Graphical Statistics 5:299-314. - Maindonald & Braun (2002) "Data analysis and graphics using R: An example-based approach", Cambridge, Cambridge University Press.

VII. Discussion

By now the reader may ask why bacteria or humans are still alive. Good question. After reviewing the situation in cultivable and uncultivable bacteria, potential solutions for the mutation rate paradox and for the genomic decay paradox are listed. As the simulations in this work consider only a small fraction of all biological mechanisms known to be relevant and as many key parameters are not well known, the solution for these paradoxes requires considerable further work. Finally, some practical consequences are discussed.

25 Origin of the majority of uncultivable microbes

This chapter reviews reasons why a significant part of the uncultivable majority of bacteria is probably generated by genomic decay due to Muller's ratchet and a newly defined 'eco-ratchet'. It might well be that only bacteria that managed to escape this decay by employing some of the specific anti-decay mechanisms discussed have a high probability of being cultured by taxonomists. Despite their potentially decaying genomes, uncultivable bacteria are likely to play important, highly adapted roles in the global ecosystem. If irregular recombination indeed saved many bacteria from mutational meltdown, an important selective force in the evolution of sex has to be reevaluated: why should complicated regular sex evolve, if simple bacterial mechanisms achieve the same result?

The starting point for the considerations here is the high deleterious mutation rate observed in the stationary phase of *Escherichia coli* (Chapter 17). Subsequent analysis of extinction times resulting from such a mutation rate showed that (i) bacteria make the transition to practical uncultivability long before they are likely to become extinct, (ii) assumed normal replication-based deleterious mutation rates are at the border of leading to an genomic decay paradox in free living bacteria and (iii) mutators and stationary phase growth mutation rates almost certainly lead to genomic decay that results in uncultivability within less than the known age of the corresponding lines.

25.1 How probable are high deleterious stationary phase mutation rates?

A key parameter in this whole discussion is the deleterious genomic mutation rate. Previous attempts to measure mutation rates have been largely replication based, ie. bacteria were grown as many generations as possible and observed mutation rates were estimated on a per generation basis. However, the work on adaptive mutations in the last decade suggests a different viewpoint. If stress in the stationary phase leads to exceptionally high transient mutation rates, then the deleterious mutation rate should be measured in the stationary phase and estimated on an absolute time molecular-clock-like basis. As most bacteria are in the stationary phase for most of the time, it is

more likely that the corresponding higher mutation rates apply in nature, even for normal, non-mutator cells.

It even appears that mutation rates measured in replication-based mutation accumulation experiments depend to a degree on the time spent in the stationary phase while waiting for the next serial transfer. Consider the following Bateman-Mukai analyses:

- Kibota & Lynch[1]: 25 generations per day; deleterious mutation rate ≈ 0.00019/genome/generation (bottleneck size 1; mutation-accumulation and assay conditions identical)
- Cooper & Lenski[2]: 6.7 generations per day; deleterious mutation rate ≈ 0.0007/genome/generation (bottleneck size ca. $5x10^6$; mutation accumulation in glucose minimal medium, assay screened for mutations in various targets; data estimated from their Figure 4)
- This work, Chapter 17: stationary phase only (assumed doubling times from 16 hours to 3 days); mutation rate ≈ 0.078/genome/generation (no known bottlenecks; mutation accumulation and assay conditions identical, except for changes due to long-term stationary phase evolution)

The strong correlation of time in stationary phase and deleterious mutation rate is striking (Figure 68) and suggests that stationary phase mutation rates are generally higher than replication-based mutation rates. Further experiments analysing more intermediate values for the plot in Figure 68 are clearly of interest. It would be also interesting to know the usual length of stationary phases in the wild.

It is currently unclear whether mutator cells have only elevated replication-based mutation rates or also elevated stationary phase mutation rates. In any case the higher mutation rates used in theoretical considerations of this work apply to genetically stable mutators that occur at significant frequencies within all natural populations of bacteria. If these mutators do not recombine, they appear to be doomed to uncultivability by Muller's ratchet.

Thus the probability is high that deleterious mutation rates in many free living bacteria are high enough to cause considerable genomic decay due to Muller's ratchet.

1. Kibota & Lynch (1996) "Estimate of the genomic mutation rate deleterious to overall fitness in E. coli", Nature 381:694-696.
2. Cooper & Lenski (2000) "The population genetics of ecological specialization in evolving Escherichia coli populations", Nature 407:736-739.

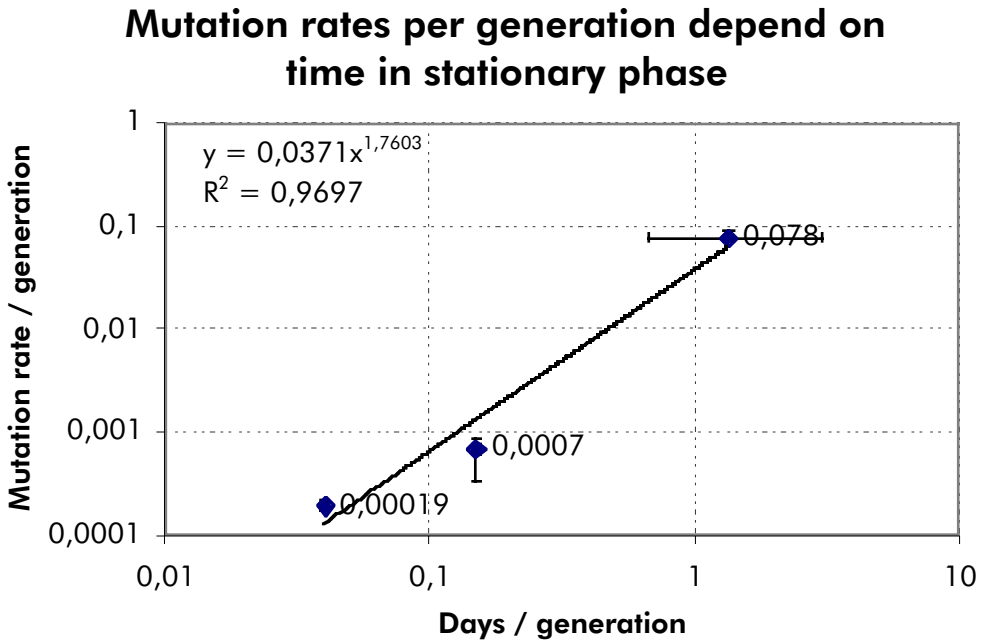


Figure 68 Mutation rates per generation depend on the time spent in the stationary phase.
All estimates are for descendants of the same strain of *Escherichia coli* (Lenski's original) growing at 37°C. Error bars denote probable upper and lower limits. The lowest mutation rate was observed over 7500 generations in 300 days by Kibota & Lynch (1996; "Estimate of the genomic mutation rate deleterious to overall fitness in *E. coli*", Nature 381:694-696), the highest rate in the 96 day stationary phase mutation accumulation experiment in this work (Chapter 17). The data for the point in the middle comes from mutation accumulation in unused genes in Lenski's long-term serial transfer evolution experiment as reported by Cooper & Lenski (2000; "The population genetics of ecological specialisation in evolving *Escherichia coli* populations", Nature 407:736-739; decay of the mean of relative fitness and increase of its variance were estimated by eye from Figure 4 for generation 0 and generation 10 000 to conduct a subsequent manual Bateman-Mukai analysis).

25.2 Adapting to one environment may be bad for life in another

When thinking about the nature of deleterious mutations, one will notice that there are two types of harmful effect:

- **Immediate effects.** Some mutations affect currently needed functionalities by slightly impairing an essential gene. Only slight changes can be fixed this way, as selection would immediately remove any heavily deleterious effects. Since most proteins appear to be quite robust regarding base substitutions[26], it appears that sequence space al-

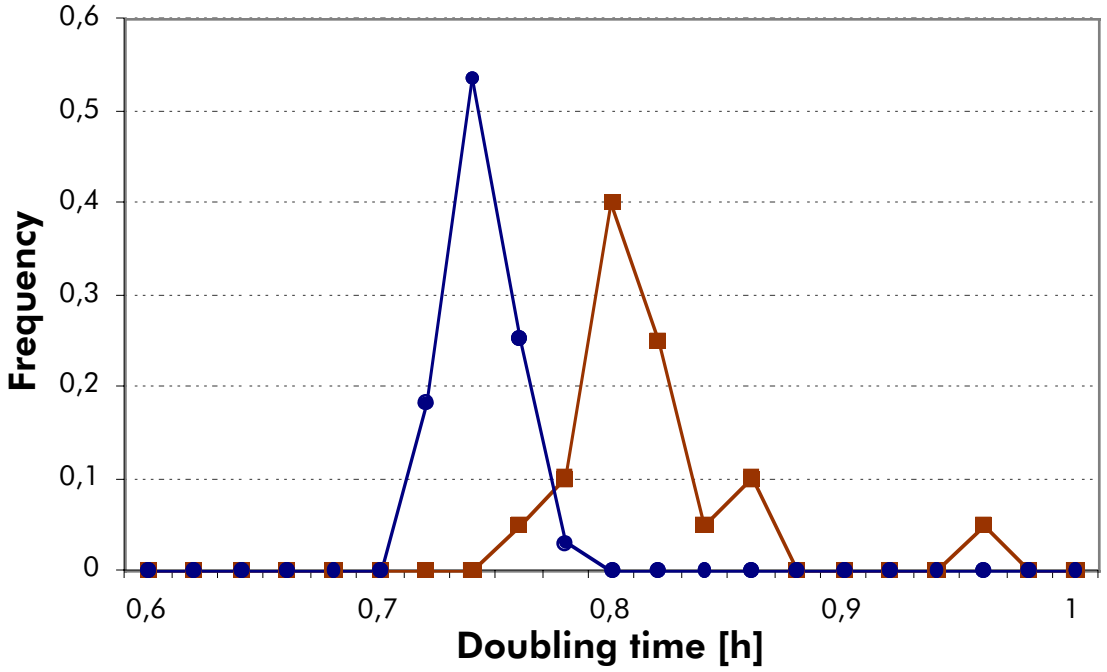


Figure 69 Comparison of fitness in a third environment before and after 10 000 generations evolution in glucose minimal.
The third environment was LB in Honeycomb plates. Growth rate before adaptation to glucose minimal (blue circles; mean doubling time 0,752 ± 0,014 StDev) and after adaptation (red squares; mean doubling time 0,825 ± 0,04 StDev) differ by 4.38 minutes with $P = 10^{-27}$ according to an unpaired t-test (highly significant). Both strains came from Richard Lenski and had been given to me by Jeff Blanchard. For more details, see materials and methods. (Data from: VTX002 All Manual Mean - Line1Line2ST1dt.xls - 21.03.2002; Evolution due to freeze-thaw stress in later transfers of VTX002 and VTX007 obscured this effect, presumably by overlaying with mutations that were more important for the current LB-Honeycomb-well freeze-thaw environment.)

lows for a large number of mutations with very little or no effect. However, measurement errors in usual molecular biological methods do not allow one to determine whether such mutations belong to the dangerous, slightly deleterious class of mutations or to the harmless, really neutral class.

- **Future effects.** When a gene is not needed in the current environment, selection will not protect it. Thus slightly deleterious mutations can easily accumulate in it. However, as small insertions or deletions constitute an appreciable fraction of all mutations, frameshifts are much more likely to destroy the function of the gene before base substitutions can. Since the current environment does not select

for the gene at all, both types of mutation are selectively neutral, despite the fact that frameshifts will have much more pronounced effects in the corresponding future environment. Thus, the probability of losing such genes appears to be considerable, when they are not needed for longer periods of time. While the exact length of such periods will depend on actual mutation rates and needs further investigation, the argument is strong enough to expect long-term evolution in one (eg. benign) environment to have deleterious fitness effects for growth in other (eg. harsh) environments.

Exactly this could be observed when growth rates of strain 1 (LENSKI-original, generation 1) and strain 2 (LENSKI-original after adapting for 10 000 generations to glucose minimal) were compared directly (Figure 69). COOPER & LENSKI found similar results in their large-scale search for ecological specialisation using the same strains[3]. Similar findings have been made frequently in a wide variety of organisms[4]. This suggests the simple rule of thumb that genes that are not essential for life in the current environment are lost very easily within surprisingly few generations due to small insertions or deletions leading to frameshifts.

25.3 The eco-ratchet and the fitness-barrel

Such fast loss of unused genes could be modelled by the fitness-barrel described in Figure 70. With this picture of the fitness-barrel in mind, it becomes easy to define the eco-ratchet:

3. Cooper & Lenski (2000) "The population genetics of ecological specialization in evolving Escherichia coli populations", Nature 407:736-739.
4. Szafraniec et al. (2001) "Environmental stress and mutational load in diploid strains of the yeast Saccharomyces cerevisiae", Proc. Natl. Acad. Sci. USA 98:1107-1112. - Shabalina et al. (1997) "Rapid decline of fitness in panmictic populations of Drosophila melanogaster maintained under relaxed natural selection", Proc. Natl. Acad. Sci. USA 94:13034-13039. - Kondrashov & Houle (1994) "Genotype-environment interactions and the estimation of the genomic mutation rate in Drosophila melanogaster", Proc. R. Soc. Lond. B Biol. Sci. 258:221-227. - Andersson & Hughes (1996) "Muller's ratchet decreases fitness of a DNA-based microbe", Proc. Natl. Acad. Sci. USA 93:906-907. - Peters (1992) "Constraint rather than chance: Regressive and reconstructive evolutionary process in cave fishes progress the same rate", Mitteilungen aus dem Hamburgischen Zoologischen Museum und Institut 89:97-113. - Peters et al. (1993) "Gehirnproportionen und Ausprägungsgrad der Sinnesorgane von Astyanax mexicanus (Pisces, Characinidae) - Ein Vergleich zwischen dem Flußfisch und seinen Höhlenderivaten "Anoptichtys"", Z. zool. Syst. Evolut.-forsch 31:144-159. - Peters (1996) "Über die Ursachen der Gehirnreduktion bei den Haustieren", Verh. naturwiss. Ver. Hamburg NF 35:237-257. - Diamond (1981) "Flightlessness and fear of flying in island species", Nature 293:507-508. - Yamamoto & Jeffery (2000) "Central role for the lens in cave fish eye degeneration", Science 289:631-633. - Sadoglu (1967) "The selective value of eye and pigment loss in Mexican cave fish", Evolution 21:541-549.

The eco-ratchet describes the repeated, irreversible loss of currently unused genes. It clicks when a population inhabits a constant environment for a time that is long enough to allow accumulation of enough frameshifts (or other mutations) for the knockout of a gene that is not used in the current environment and is therefore completely free from purifying selection. Changing to environments with ever decreasing requirements can lead to a highly adapted dead end, where all other environments have higher demands than the genotypes left cold possibly meet. If this last environment is lost, the eco-ratchet will have killed one more population, although habitat destruction will be the formal reason for extinction.

It will be interesting to build an evolution@home simulator that models the fitness-barrel and the resulting eco-ratchet for realistic patterns of environmental change, realistic genome organisations and realistic frameshift mutation rates. The first species where a reasonable definition of such a model will be possible is probably a microbe. Three general types of ecological strategies are likely to be found:

- **Generalists** change their environments too often to allow the eco-ratchet to click.
- **Specialists** lose all other genes and adapt to one stable niche. Their survival depends on the long-term stability of the niche. As their environment does not change, the eco-ratchet cannot click.
- **Losers** are the prey of the eco-ratchet. They live in environments that are stable enough to allow for loss of currently unused genes. However, their environments do change. Thus it would have been advisable not to lose genes for other environments. Repeated redefinitions of the fitness-barrel might lead to evolutionary dead ends that can result in extinction, when the last habitats of these species are destroyed.

Please note that Muller's ratchet can still operate if the eco-ratchet does not. For those occasions where the eco-ratchet operates, it is probably possible to rephrase its mechanics in terms of Muller's ratchet. However, one might also see an inner connection between the eco-ratchet and the Red Queen[5] hypothesis, as changing environments play an important role in both theories.

5. Van Valen (1973) "A new evolutionary law", Evol. Theory 1:1-30. - Maynard Smith (1978) "The Evolution of Sex", New York, Cambridge University Press. - Duarte et al. (1992) "Rapid fitness losses in mammalian RNA virus clones due to Muller's ratchet", Proc. Natl. Acad. Sci. USA 89:6015-6019.

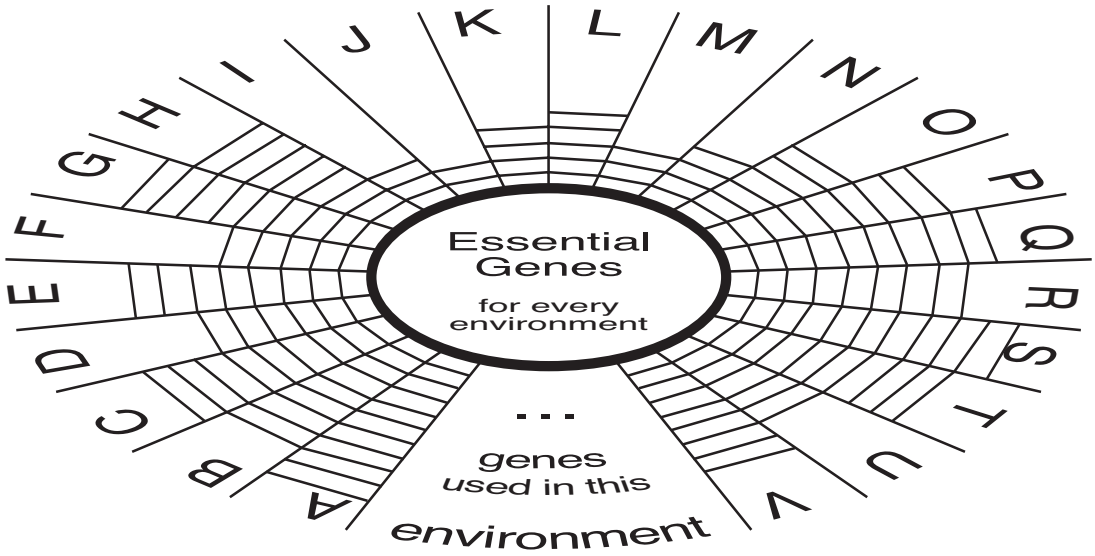


Figure 70 A possible projection of the fitness barrel for a hypothetical organism.
The circle in the core denotes the set of genes that is essential for replication, even under the most benign environment that can be conceived. The letters in the various sections denote different environments, in which the organism could live (environments = the planks that the barrel is made of) . These environments are defined in terms of the extra genes needed for successful reproduction in them. If the organism lives in a certain environment, the upper border of the barrel is defined by the core genes and all other essential genes for that environment. All other environments with their exclusive genes (ie. genes not needed in the current environment) are like planks that protrude over the upper edge of the barrel. Changing to a new environment with new challenges redefines the upper edge of the barrel in terms of a new set of essential genes (that might or might not overlap with the old in more than the core genes). Now imagine that, in order to reproduce, the genome needs to have slightly more gene functions (content of the barrel) than needed to effectively produce exactly 1 offspring for the next generation under optimal conditions in the current environment (height of the upper border of the fitness barrel). As long as the barrel has enough content, offspring is produced. However, while selection guards only content of the barrel that is below its current upper limit (ie. selection guards genes in genome that belong to the essential set for the current environment), all content above the upper limit is easily lost (ie. mutational pressure easily erodes unused genes by frequent frameshift mutations). In this picture, selection of advantageous mutations builds up content of the barrel. Finally, consider the following scenarios with this picture in mind:
The ecological *generalist* frequently changes environments and never spends enough time in any environment to allow mutational loss of genes essential for other environments to become effectively neutral.
The ecological *specialist* picks his favourite environment and stays there for a long time. While this provides the possibility of adapting to its niche, such a strategy removes purifying selection pressure from all genes that are not necessary for that environment. If there is enough time for at least one frameshift in each of these genes, the content of the barrel (ie. the genome) will be degraded to meet the lowest possible standard (ie. the current environment only). After some time this erosion will have affected the whole population. A subsequent change of the environment to a set of genes that had been eroded earlier, results in extinction of the population, because no individual is left that can still meet the current standards (ie. has enough content in its barrel to allow reproduction).

Nevertheless, they are different concepts, because the eco-ratchet focuses on the loss of genetic information instead on the speed of adaptation. Detailed quantification of the eco-ratchet will probably lead to a better understanding of the one big unknown parameter in all models of Muller's ratchet: the distribution of mutational effects.

Until then, quantification of genomic decay in free-living bacteria can be achieved best by models of Muller's ratchet, as described in this work. Combined with the general suspicion that a number of bacterial habitats (eg. soil) do change only slowly (eg. seasons of the year), this work suggests that a significant number of bacterial populations that get trapped in such habitats might actually be forced into the eco-ratchet loser strategy, which most likely leads to uncultivability and eventual extinction.

25.4 What could keep bacteria from genomic decay?

A number of biological issues can keep bacteria from extinction due to Muller's ratchet. It is predicted here that any combination of the following incomplete list of potential solutions for the genomic decay paradox is active in 'typical' bacteria that do not belong to the 'typical uncultivable majority'.

1. **Low mutation rates**. If some cells never become genetically stable mutators nor transient mutators, they could actually have good chances against Muller's ratchet, especially if they are anti-mutators (ie. have mutation rates that are even lower than normal replication-based assays suggest). Only the eco-ratchet could destroy such populations by temporarily switching off selection from some genes.
2. **Dormant states**. Gram-positive bacteria are known to build spores that can bridge hundreds to thousands of years without replication or stationary phase evolution. Even if (unselected!) mutations can occur after extremely long periods of time in a spore, it is easy to see that spores can lead to reinfection of decaying bacterial populations with germs that might even come from a time before the last decay process had started. A single generation of spores might easily bridge thousand years of hot evolution in a wrong direction under starvation stress. This system is comparable to version control systems used for programming large software systems: one of their key purposes is always to allow going back to the last working version of the system.

Ancient bacteria from amber, sediments or salt crystals could have similar 'refugia-like' functions, if they employ a preservation system that does not need cryptic growth for survival.

There are some reports of extraordinary non-spore generation times of supposedly actively replicating bacteria inferred from indirect evidence[6,7,8]. If such estimates (up to 100 000 years![6]) indeed combined with normal replication-based mutation rates as some assume[7], this would lead to effective deleterious mutation rates that hardly ever turn Muller's ratchet.

3. **Recombination**. Even a little bit of occasional genetic exchange could do a lot to save bacteria from decay. This could even repair damage from the eco-ratchet by uptake of lost genes. Recombination in bacteria can occur by a number of mechanisms[9] that allow for successful exchange of genes under a wide array of circumstances. Many known bacteria carry clear signatures of occasional or even frequent recombination[10] and direct comparison of the genomes from two different strains of *E. coli* shows potential molecular effects of these processes[11].
4. **Ideal distributions of mutational effects**. Since selection coefficients are the largest unknown parameter in genomic decay quantifications, one might speculate that some heavily deleterious mutations exist (like those seen in mutation accumulation experiments) and that all other mutations are completely neutral. Thus the broad range of dangerous intermediate selection coefficients (on a log-scale) just would not exist.

---

6. Parkes et al. (2000) "Recent studies on bacterial populations and processes in subseafloor sediments: A review", Hydrogeology Journal 8:11-28.
7. Maughan et al. (2002) "The paradox of the "Ancient" bacterium which contains "Modern" protein-coding genes", Mol. Biol. Evol. 19:1637-1639.
8. Phelps et al. (1994) "Comparison between geochemical and biological estimates of subsurface microbial activity", Microbial Ecology 28:335.
9. See eg. Lim (1998) "Microbiology". 2nd ed., Boston, McGraw-Hill.
10. Whittam (1992) "Population Biology: Sex in the soil", Curr. Biol. 2:676-678. - Lan & Reeves (1996) "Gene transfer is a major factor in bacterial evolution", Mol. Biol. Evol. 13:47-55. - Ochman (1999) "Bacterial evolution: Jittery genomes", Curr. Biol. 9:R485-486. - Ochman et al. (2000) "Lateral gene transfer and the nature of bacterial innovation", Nature 405:299-304. - Koonin et al. (2001) "Horizontal gene transfer in Procaryotes: Quantification and classification", Annu. Rev. Microbiol. 55:709–742. - Feil & Spratt (2001) "Recombination and the population structures of bacterial pathogens", Annu. Rev. Microbiol. 55:561–590.
11. Blattner et al. (1997) "The complete genome sequence of Escherichia coli K-12", Science 277:1453-1462. - Perna et al. (2001) "Genome sequence of enterohaemorrhagic Escherichia coli O157:H7", Nature 409:529-533.

5. **Advantageous and compensatory mutations**. With enough beneficial mutations any genomic decay can be stopped. The key question will be whether enough of these occur. While it is easy to observe some advantageous[12] or compensatory[13] mutations on a short term scale, especially after changing the environment for bacteria, the rates of interest here are the (presumably quite low) sustained, long-term rates of evolutionary innovation.

For other potential solutions, see the next chapter. This work proposes that enough similar anti-decay processes are active in normal cultivable bacteria, whereas many uncultivable bacteria suffer from lack of their help in combination with high mutation rates and extremely rare recombination.

25.5 Conclusions

Sex is more important for bacteria than previously thought

Reviewing the list of potential solutions for the genomic decay paradox in bacteria suggests that recombination plays a key role. While it is not new that bacteria recombine a lot, it is suggested here that they *have to* in order to escape Muller's ratchet. If bacteria do not recombine enough, they slowly slide into that part of the uncultivable majority of bacteria that has been generated by genomic decay.

If bacteria managed to escape from mutational meltdown with irregular sex, why then did regular sex evolve at all?

One of the big puzzles of evolutionary biology is the cause that led to the origin of sex (ie. regular, obligate recombination)[14]. While this complex topic will not be reviewed here, the conclusion reached about the necessity of occasional recombination in free living bacteria has consequences for mutational-meltdown-type theories of the evolution of sex. These assume that sex evolved to save asexual populations from mutational meltdown due to Muller's ratchet. However, if bacteria themselves successfully use occasional recombination in order to survive for geological periods of time, there is no

12. Imhof & Schlotterer (2001) "Fitness effects of advantageous mutations in evolving Escherichia coli populations", Proc. Natl. Acad. Sci. USA 98:1113-1117. - Elena et al. (1996) "Punctuated evolution caused by selection of rare beneficial mutations", Science 272:1802-1804.
13. Burch & Chao (1999) "Evolution by small steps and rugged landscapes in the RNA virus phi 6", Genetics 151:921-927.
14. Bell (1982) "The Masterpiece of Nature: The Evolution and Genetics of Sexuality", Berkeley, University of California Press. - Kondrashov (1988) "Deleterious mutations and the evolution of sexual reproduction", Nature 336:435-440. - Kondrashov (1993) "Classification of hypotheses on the advantage of amphimixis", J. Hered. 84:372-387. - Rice (2002) "Experimental tests of the adaptive significance of sexual recombination", Nat Rev Genet 3:241-251. - Otto & Lenormand (2002) "Resolving the paradox of sex and recombination", Nat Rev Genet 3:252-261. - Butlin (2002) "Opinion - evolution of sex: The costs and benefits of sex: new insights from old asexual lineages", Nat Rev Genet 3:311-317.

need for another recombination system, as everything can be achieved by the simpler bacterial system too. Thus if one is not willing to assume that the more complex regular recombination is older than the simpler bacterial recombination systems or that both systems evolved independently, escape from mutational meltdown cannot be the evolutionary driving force behind the evolution of regular recombination systems.

Why uncultivable bacteria are still interesting

If a large part of uncultivable bacteria is indeed the result of genomic decay due to Muller's ratchet, one might ask why these bacteria should be studied at all. Here are some reasons:

1. First of all, we need experimental estimates of the fraction of uncultivable bacteria that is produced by Muller's ratchet. This can only be obtained by direct experimental observation of such bacteria.
2. Not all bacteria that are uncultivable at the moment can be expected to be the result of a decay process. The probability is high that a significant number of important real bacterial species have not been cultivated yet, even if that number is probably not in the millions. These species need to be described.
3. Some bacteria can be expected to be specialists in the eco-ratchet type sense. In these, genomic decay has destroyed all functions that would be needed in other environments, but within their special constant niche, they are as fit as they can be. These niches and their corresponding species have to be understood for comprehension of global ecosystems.

Even if these points were completely done, the decaying populations themselves are also interesting:

4. Actual decay processes can be studied best by observing their products.
5. The eco-ratchet hypothesis suggests that most uncultivable bacteria are (losing) specialists that try to fill some niche and adapt to a considerable degree to that niche. It will be interesting to make an inventory of such niches and to compare the various roads taken on different occasions to adapt to that niche. Such work can be expected to provide more insight into the large topic of chance and necessity in evolution.
6. The uncultivable majority probably has important biogeochemical functions that need to be understood, if the ecosystems of this earth are ever to be understood.

7. Specific adaptations found in uncultivable microbes could be top engineering ideas for biotechnology, if taken back to an otherwise fitter microbe.
8. Intense investigation of these issues might eventually lead to a bacterial species concept that is more in contact with fundamental biological realities than our current pragmatic phylo-phenetic concept of bacterial species[15]. Whatever could be meant by 'fundamental biological realities' is exactly what will have to be found out in such an effort.

Whatever the outcome of these investigations, one thing seems clear: the uncultivable majority of bacteria will continue to have a significant impact on our views about bacterial evolution[16].

15. Rossello-Mora & Amann (2001) "The species concept for prokaryotes", FEMS Microbiol. Rev. 25:39-67.
16. Hugenholtz et al. (1998) "Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity", J. Bacteriol. 180:4765-4774.

26 Potential solutions for the paradoxes

This chapter discusses most known potential solutions for the mutation rate paradox and for the genomic decay paradox. As the final solution could also be a combination of those listed, an enormous amount of empirical work will be necessary to decide on these issues. However, in most cases some hints indicate how probable it is that a proposed solution contributes to the actual solution.

26.1 Mutation rate paradox

The mutation rate paradox was encountered on various occasions throughout this work: (i) as a problem in understanding the evolution of human mitochondrial DNA, (ii) as a discrepancy between a number of dates inferred from archaeology and molecular clocks, (iii) as a discrepancy between bacterial mutation rates observed in the laboratory and bacterial molecular clocks and (iv) as a discrepancy between the observation of very similar RNA viruses that have been assumed to mutate at a high rate for a long time. It was defined in Chapter 3 as follows:

Definition

The mutation rate paradox consists of two conflicting mutation rates that both should apply to the same evolutionary line: a long-term, low, inferred (phylogenetic) rate and a short-term, high rate supported by independent evidence from recent observations (e.g. sequences in pedigrees or other even phylogenies).

Although details will differ greatly between RNA viruses, bacteria and mitochondria, the following general patterns may contribute to a solution of the mutation rate paradox:

1. **Mutational hot spots**. If most mutations occur only at a few positions that flip back and forth every other generation, observed short-term mutation rates can be extraordinary high while nothing is happening from a long-term perspective.

   This solution is very elegant as it does not carry a mutational load price tag, if the corresponding mutations are neutral. Some positions in the systems investigated clearly fall in the hot spot category, but not all do. Thus in known systems the mutation rate paradox still exists after excluding hot spots from calculation of higher mutation rates.

2. **Selective removal**. If mutation effects are deleterious enough to be removed in the long term, while being small enough to still go unnoticed, then short-term mutation rates are higher because they include mutations that are removed later.
   This process plays with fire. As can be easily seen from the considerations of Muller's ratchet in mammalian mitochondria, only a very small range of selection coefficients leads to such purifying selection without turning on Muller's ratchet in its most dangerous form. When the ratchet clicks, an even a much smaller range of selection coefficients would probably still allow removal of enough mutations to solve the mutation rate paradox, albeit at the cost of getting into the genomic decay paradox. If selection coefficients are small enough to cause no harm during the assumed age of the evolutionary line, then selection will be too weak to remove mutations, as most mutations accumulate in an effectively neutral manner. If recombination is present, the principle is the same, except that purifying selection is several orders of magnitude more efficient and, consequently, the selective removal hypothesis appears to be significantly more credible.
3. **Heterogeneous mutation rates in space or time**. It could be that observations of high mutation rates happened to be made in populations that experienced higher mutation rates than other populations of the same species and that those other populations contribute more to long-term phylogenies than the observed population. Similarly, mutation rates could change over time and observations happened to be made in a period of high mutation rates.
   Only further monitoring of mutation rates in more populations will be able to reveal whether such heterogeneity contributes to a solution of the paradox. It should be noted, though, that observations of high mutation rates are beyond the state where they could be attributed to a statistical fluke. If such heterogeneity exists between different populations and group selection were to remove populations with higher mutation rates (for independent reasons!), then this would also be an elegant solution to the mutation rate paradox.
4. **Drift effects and neutral slowdown**. Under completely free recombination, the neutral theory predicts that intergenerational mutation rates lead to equal long-term substitution rates. However, when mutations are linked together, their fate changes and more mutations are removed on the long term because when drift removes

one individual from the population, this individual carried several mutations on their way to fixation. Simulations conducted in this work show that this actually leads to a small slowdown of long-term substitution rates (see "The 'neutral slowdown' of Muller's ratchet" on page 249) . However, this does not solve the paradox completely. Similarly, yet unexplored details of drift may have similar effects.

5. **Linked loci under selection**. Background selection[17] or hitchhiking[18] can both significantly reduce polymorphism in a group of linked genes. This may affect long-term substitution rates in a way that contributes to a solution of the mutation rate paradox.
6. **Multi-level population genetics**. Perhaps some little understood mechanisms that belong to the multi-level population genetics[19] of mitochondria could explain the paradox, although it is hard to see how at the moment.
7. **Generation time effect**. If longer generation times lead to lower speeds in molecular clocks[20], this might affect the mutation rate paradox too.

Practical consequences

This list is probably incomplete and it appears that a considerable amount of further research will be needed before the mutation rate paradox can be regarded as solved. Certainly, a most complete list of instances where this paradox has been observed will be an important step towards evaluating solutions. In order to collect corresponding data, an email address has been reserved for mails from everybody who wishes to report an instance of the mutation rate paradox. Please send all information with references to:

MutRatParDB@evolutionary-research.net

Eventually this information will be published in a public database on the web under http://www.evolutionary-research.net/.

17. Charlesworth et al. (1993) "The effect of deleterious mutations on neutral molecular variation", Genetics 134:1289-1303.
18. Maynard Smith & Haigh (1974) "The hitchhiking effect of a favorable gene", Genet. Res. 23:23-35.
19. Birky (1991) "Evolution and population genetics of organelle genes: Mechanisms and models", pp. 112-134 in: Selander et al. (eds) Evolution at the Molecular Level, Sunderland, MA, Sindauer Associates, Inc. - Bergstrom & Pritchard (1998) "Germline bottlenecks and the evolutionary maintenance of mitochondrial genomes", Genetics 149:2135-2146. - Kondrashov (1994) "Mutation load under vegetative reproduction and cytoplasmic inheritance", Genetics 137:311-318. - Otto & Hastings (1998) "Mutation and selection within the individual", Genetica 103:507-524. - Otto & Orive (1995) "Evolutionary consequences of mutation and selection within an individual", Genetics 141:1173-1187. - Hastings (1989) "Potential germline competition in animals and its evolutionary implications", Genetics 123:191-198. - Hastings (1991) "Germline-selection: Population genetic aspects of the sexual/asexual life cycle", Genetics 129:1167-1176. - Takahata & Slatkin (1983) "Evolutionary dynamics of extranuclear genes", Genet. Res. 42:257-265.
20. Martin & Palumbi (1993) "Body size, metabolic rate, generation time, and the molecular clock", Proc. Natl. Acad. Sci. USA 90:4087-4091.

26.2 The fitness-balance and the genomic decay paradox

Definition

This work is not the first report of a genomic decay paradox[21]. However, as future research is usually facilitated by clear definitions, it is suggested that we use the following terminology (as introduced in Chapter 21):

> *The genomic decay paradox describes situations where the apparent age of an evolutionary line is greater than the currently best predictions of its extinction time.*

When thinking about potential solutions for the genomic decay paradox, it helps to have the fitness-balance in mind (Figure 71). It weights the various processes in terms of their increase or decrease of the mean fitness of a population. The picture presented here lists only a few prominent processes like selection that increase fitness by fixing advantageous mutations or Muller's ratchet that decreases fitness by fixing slightly deleterious mutations.

Many processes have two faces

Nevertheless, it can be easily seen from the example of genetic drift, that many processes have two faces. Drift can fix very slightly deleterious mutations and thus decrease fitness, but it can also fix very slightly advantageous mutations or pre-adaptive mutations (neutral now, but advantageous in the future). Similarly, recombination does a lot to increase fitness in general albeit at the cost of breaking up favourable allele combinations and migration increases fitness by facilitating escape to better environments at the cost of constantly interrupting local adaptation processes. However, this is not the place to review all the processes that might contribute to the fitness-balance. Potential solutions of the genomic decay paradox influence the balance in such a way that the fitness decreasing side either has less weight than the fitness increasing side or has so little weight that it cannot do any harm within the assumed age of existence of the corresponding evolutionary line. The following processes can be conceived to contribute to a solution of the genomic decay paradox:

1. **Lower mutation rates**. If our current knowledge of mutation rates was collected in populations with higher than normal mutation rates or at times where mutation rates were transiently increased, the true

21. Eg. Muller (1950) "Our load of mutations", Am. J. Hum. Genet. 2:111-176. - Kondrashov (1995) "Contamination of the Genome by Very Slightly Deleterious Mutations: Why Have We Not Died 100 Times Over?" J. theor. Biol. 175:583-594. - Gabriel et al. (1993) "Mullers Ratchet and mutational meltdowns", Evolution 47:1744-1757. - Crow (1993) "Mutation, mean fitness, and genetic load", pp. 3-42 in: Futuyma & Antonovics (eds) Oxford Surveys in Evolutionary Biology. 9, Oxford, Oxford University Press. - Eyre-Walker & Keightley (1999) "High genomic deleterious mutation rates in hominids", Nature 397:344-347.

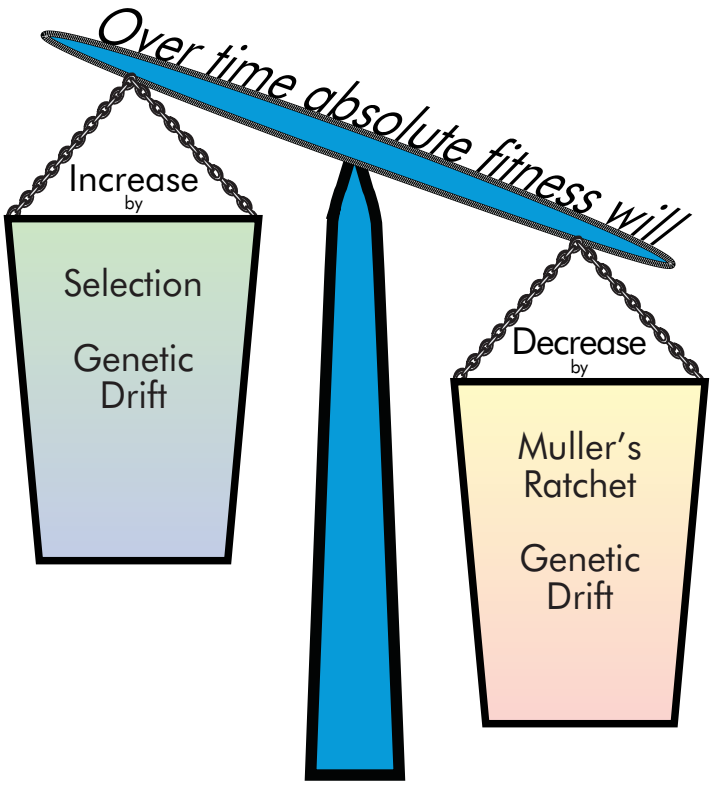


Figure 71 The fitness-balance helps to understand genomic decay.
Most evolutionary processes change the mean fitness of a population. They either increase or decrease it. Some processes appear on both sides, because their contribution can be advantageous on some occasions and harmful on others. Only a few processes are listed in the balance drawn here to facilitate readability.

long-term mutation rate will be lower. If low enough, the paradox might even be solved. However, general anti-mutators are improbable[22].

2. **Hot-spots** further decrease the relevant mutation rate, if frequent.
3. **Small effective genome sizes** translate the total long-term mutation rate to a small dangerous long-term mutation rate, as mutations in completely neutral junk DNA do not contribute to genomic decay[23].
4. **Few dangerous mutation effects**. Very little is known about distributions of deleterious mutation effects[24,25]. If they contain many

22. Drake (1993) "General antimutators are improbable", J. Mol. Biol. 229:8-13.
23. Drake et al. (1998) "Rates of spontaneous mutation", Genetics 148:1667-1686.

considerably deleterious mutations (as those found in typical mutation accumulation experiments) and many completely neutral mutations, but only few really dangerous intermediate mutation effects, this could easily stop genomic decay. Neutrals do no harm and clearly deleterious mutations are easily removed by purifying selection. However, there is no biological reason to suggest that these dangerous intermediate mutational effects should be especially rare. Considering the mutation effect pyramid might rather suggest that dangerous mutation effects could be quite common (Figure 72). This is consistent with the fact that the majority of pointmutations seen in random mutagenesis experiments have no large effect[26]. However, current methods are too insensitive to decide whether these mutations are slightly deleterious and thus dangerous, or absolutely neutral and thus harmless.

5. **Extremely small deleterious mutation effects**. Genomes might be decaying irrevocably most of the time, but decay is too slow to lead to extinction, before the next adaptive episode fixes a further considerable increase in fitness. However, the existence of mutations with clearly measurable effects suggests that such extremely small mutation effects are not the only ones that exist.

---

24. Lynch et al. (1999) "Perspective: Spontaneous deleterious mutation", Evolution 53:645-663.
25. Kingsolver et al. (2001) "The strength of phenotypic selection in natural populations", Am. Nat. Mar 157:245-261.
26. See eg. Loeb et al. (1989) "Complete mutagenesis of the HIV-1 protease", Nature 340:397-400. - Bowie et al. (1990) "Deciphring the message in protein sequences: tolerance to amino acid substitutions", Science 247:1306-1310. - Rennell et al. (1991) "Systematic mutation of bacteriophage T4 lysozyme", J. Mol. Biol. 222:67-88. - Xu et al. (1994) "Random mutagenesis of glutamine synthetase from Escherichia coli: Correlation between structure, activity, and fitness", Journal of Fermentation and Bioengineering 77:252-258. - Markiewicz et al. (1994) "Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence", J. Mol. Biol. 240:421-433. - Kapp et al. (1995) "Alignment of 700 globin sequences: Extent of amino acid substitution rate and its correlation with variation in volume", Protein Science 4:2179-2190. - Suzuki et al. (1996) "Random mutagenesis of Thermus aquaticus DNA polymerase I: Concordance of immutable sites in vivo with the crystal structure", Proc. Natl. Acad. Sci. USA 93:9670-9675. - Wen et al. (1996) "Exploring the allowed sequence space of a membrane protein", Nat. Struct. Biol. 3:141-148. - Martinez et al. (1996) "Exploring the functional robustness of an enzyme by in vitro evolution", EMBO J. 15:1203-1210. - Baumgartner & Hazelbauer (1996) "Mutational analysis of a transmembrane segment in a bacterial chemoreceptor", J. Bacteriol. 178:4651-4660. - Huang et al. (1996) "Amino acid sequence determinants of beta-lactamase structure and activity", J. Mol. Biol. 258:688-703. - deBoer & Glickman (1998) "The lacI gene as a target for mutation in transgenic rodents and Escherichia coli", Genetics 148:1441-1451. - Keefe & Szostak (2001) "Functional proteins from a random-sequence library", Nature 410:715-718. - and many others.

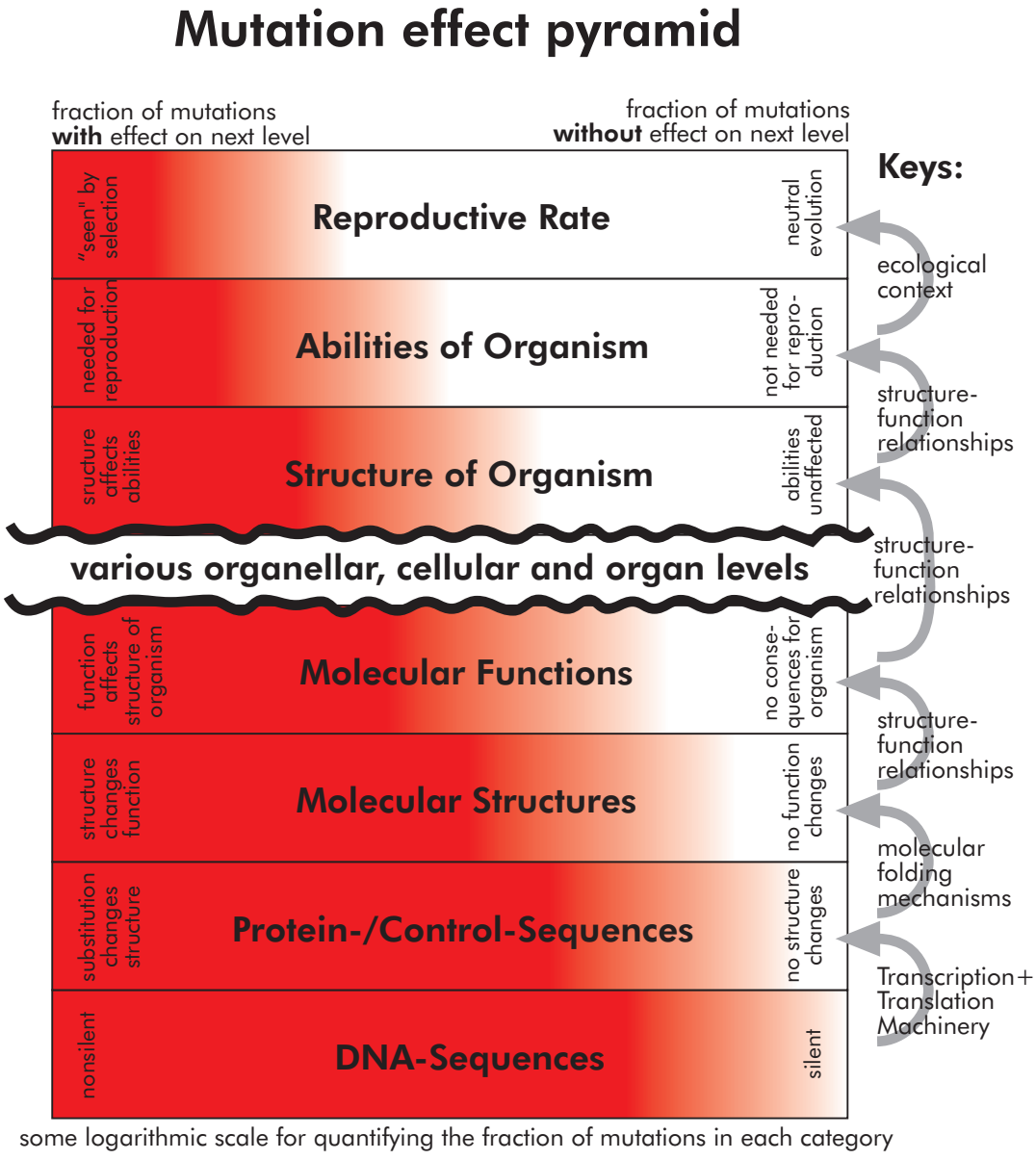


Figure 72 The mutation effect pyramid.
It visualises the principle that always only a fraction of mutations occurring at a lower level will have an effect at a higher level. As mutations always occur at the bottom (DNA), while selection coefficients are measured at the top (reproduction and survival of the fittest), it is easy to understand why only a small fraction of mutations has pronounced deleterious effects. The same logic suggests that mutational effects that can be very pronounced at the bottom (eg. knock-out of a protein), can be attenuated considerably when evaluated at the top. As this is likely to be the case for a considerable fraction of all mutations, there is no reason to suspect that any particular range of selection coefficients on a log scale should be missing.

6. **Advantageous mutations**. Although many examples of advantageous mutations are known and even advantageous mutation rates have been measured in bacteria that adapt to a new environment[27], the long-term sustainable advantageous mutation rate is still completely unknown.
7. **Compensatory mutations**. As described in Chapter 20, the compensatory mutation rate becomes considerable in a genome that carries many slightly deleterious mutations. However, as the number of 'repairways' and the 'warranty period' of modern genomes (see Chapter 20) have never been estimated, it is difficult to predict the role of compensatory mutations in stopping genomic decay.
8. **Quantitative traits compensation**. Quantitative fitness traits could easily compensate for an increased load of slightly deleterious mutations[28]. This is one of the reasons why it is difficult to estimate the maximal theoretically possible reproductive capacities for the multiplicative fitness model employed. A quantitative trait could prevent extinction by masking a considerable amount of deleterious mutations. Ultimately, however, when the reserve of quantitative variation is eventually used up, extinction will occur. Precise quantifications of the compensatory potential of quantitative traits are not available.
9. **Synergistic epistasis leading to quasi-truncating selection**. One of the big puzzles in evolutionary biology is how different mutations combine their effects. If no interaction occurs between different mutations, a multiplicative fitness model like that employed in standard models of Muller's ratchet can be used. However, two mutations can also lead to a combined effect that is larger (synergistic epistasis) or smaller (antagonistic epistasis) than both individual effects. If different mutations combine to produce more pronounced effects under synergistic epistasis, the point could be reached where further mutations have effects that are too large for accumulation under current conditions. Then quasi-truncating selection will remove new deleterious mutations from the population and stop genomic decay (Figure 73). While truncating selection (cull *all* individuals be-

27. Elena et al. (1996) "Punctuated evolution caused by selection of rare beneficial mutations", Science 272:1802-1804. - Imhof & Schlotterer (2001) "Fitness effects of advantageous mutations in evolving Escherichia coli populations", Proc. Natl. Acad. Sci. USA 98:1113-1117.
28. Wagner & Gabriel (1990) "Quantitative variation in finite parthenogenetic populations: What stops Muller's ratchet in the absence of recombination?" Evolution 44:715-731.

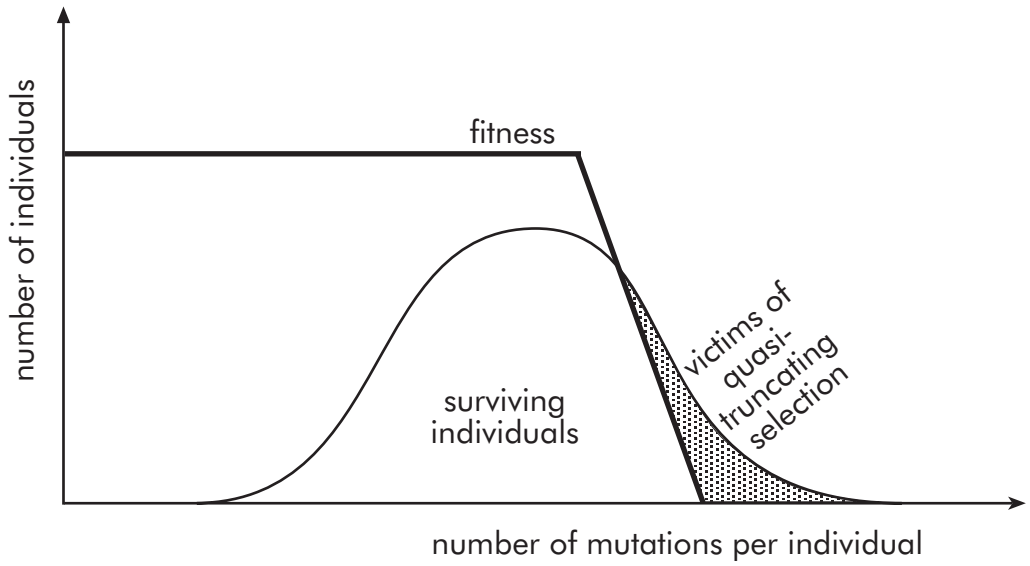


Figure 73 Quasi-truncating selection could stop genomic decay if mutations have equal effects and interact synergistically (ie. the combined effect is larger than independent effects).
The thin line denotes the frequencies of mutation classes in a population. The thick line denotes the fitness of an individual as a function of the number of deleterious mutations it carries. All individuals in the shaded area are removed by quasi-truncating selection, because selection coefficients increase from slightly deleterious to effectively lethal. Truncating selection switches from highest fitness to lethal at an arbitrary number of mutations, something that is rightly suspected to by highly artificial. The shallower the slope of the transition, the more of the raw power of truncating selection is lost. Adapted from Crow (1993) "Mutation, mean fitness, and genetic load", pp. 3-42 in: Futuyma & Antonovics (eds) Oxford Surveys in Evolutionary Biology. 9, Oxford, Oxford University Press.

yond a given threshold) is known as a powerful tool from breeding genetics, its occurrence in nature is highly improbable. The realisation that quasi-truncating selection can be quasi as powerful as truncating selection[29] convinced James Crow[30] and others[31] that this could indeed be a solution for the genomic decay paradox.

However, if mutation effects are not constant, but follow any distri-

29. Crow & Kimura (1979) "Efficiency of truncation selection", Proc. Natl. Acad. Sci. USA 76:396-399.
30. Crow (1993) "Mutation, mean fitness, and genetic load", pp. 3-42 in: Futuyma & Antonovics (eds) Oxford Surveys in Evolutionary Biology. 9, Oxford, Oxford University Press. - Crow (1997) "The high spontaneous mutation rate: Is it a health risk?" Proc. Natl. Acad. Sci. USA 94:8380-8386. - Crow (2000) "The origins, patterns and implications of human spontaneous mutation", Nat Rev Genet 1:40-47.
31. Kondrashov (1994) "Muller's Ratchet under epistatic selection", Genetics 136:1469-1473. - Kondrashov (1995) "Contamination of the Genome by Very Slightly Deleterious Mutations: Why Have We Not Died 100 Times Over?" J. theor. Biol. 175:583-594. - Eyre-Walker & Keightley (1999) "High genomic deleterious mutation rates in hominids", Nature 397:344-347.

bution, the synergistic front against decay breaks down, because in such distributions there will always be a mutation with an effect still small enough to accumulate[32]. Furthermore, the best evidence available on the form of epistasis suggests that on average, mutations are independent of each other; however, when interactions occur, they are synergistic or antagonistic with equal probability[33]. Antagonistic epistasis facilitates mutation accumulation, but may lead to a decrease of overall fitness effects. In light of current evidence it is completely unclear whether epistatic interactions could lead to a slowdown or stop of genomic decay.

10. **Recombination** is one of the most popular answers for stopping genomic decay, as it has been shown[34] to powerfully facilitate removal of deleterious mutations and fixation of advantageous mutations by bringing each of them together in one individual. Thus recombination significantly enhances the power of selection without increasing the genetic death toll. The elegance of this principle has inspired theories about the evolution of sex (i.e. regular recombination)[35].
Although lower levels of recombination can have similar effects to regular recombination, one should not forget that a certain minimal level of repeated recombination is necessary to unfold its power. Furthermore, even a recombining genome can express a large degree of linkage disequlibrium, as several genes are linked together in a region of a chromosome. Finally, recombination can facilitate selection only on mutations that are larger than the inverse of effective population size (denoted as 'slightly' in this work). Mutation effects that are smaller than this limit are freely accumulated even in recombining populations[36]. These limitations should be kept in mind when searching for cryptic sex in apparently asexual genetic systems.

32. Butcher (1995) "Muller's ratchet, epistasis and mutation effects", Genetics 141:431-437.
33. Elena & Lenski (1997) "Test of synergistic interactions among deleterious mutations in bacteria", Nature 390:395-398.
34. Eg. Schultz & Lynch (1997) "Mutation and extinction: The role of variable mutational effects, synergistic epistasis, beneficial mutations, and the degree of outcrossing", Evolution 51:1363-1371.
35. Butlin (2002) "Opinion - evolution of sex: The costs and benefits of sex: new insights from old asexual lineages", Nat Rev Genet 3:311-317. - Otto & Lenormand (2002) "Resolving the paradox of sex and recombination", Nat Rev Genet 3:252-261. - Rice (2002) "Experimental tests of the adaptive significance of sexual recombination", Nat Rev Genet 3:241-251. - Barton & Charlesworth (1998) "Why sex and recombination?" Science 281:1986-1990. - Kondrashov (1993) "Classification of hypotheses on the advantage of amphimixis", J. Hered. 84:372-387. - Muller (1964) "The relation of recombination to mutational advance", Mut. Res. 1:2-9.

11. **Reservoirs of non-decayed genes**. If older, intact, copies of important genetic information are preserved somewhere outside the population of decaying genomes, reintroduction of such considerably fitter genes or individuals could rewind decay history for a considerable amount of time.

    This process is most likely contributing to preservation of bacterial species by constant reinfection with old spores. Other applications conceivable include nuclear copies of mitochondrial genes (that exist but have never been observed to get back into mitochondria) or unknown populations with extraordinarily low mutation rates.

    Occasionally, conservation biologists discuss cryogenetic freezing of endangered species to avoid extinction[37].
12. **Metapopulations**[38]. Genomic decay and extinction of a local population is no problem, if many other populations exist that could recolonise the local patch.

    While this is true from the perspective of the local population, metapopulation structure leads to complicated dynamics that can increase or decrease the power of selection with corresponding effects on genomic decay[39].
13. **Lineage sorting.** If a decaying population among other non-decaying populations exhibits any properties that are negatively selected by group selection, then decay might be stopped by selection of non-decaying groups. All gradual variations in this scenario are conceivable.
14. **Sexual selection of good genes**[40] could have a vital role in maintaining long-term genomic integrity. It could easily magnify otherwise slightly deleterious mutation effects in vital genes, if they are 'beauty-related'. Once we understand genetic blueprints better, we will be able to estimate the true extent of this possibility.

---

36. Kondrashov (1995) "Contamination of the Genome by Very Slightly Deleterious Mutations: Why Have We Not Died 100 Times Over?" J. theor. Biol. 175:583-594. - Kimura (1995) "Limitations of Darwinian selection in a finite population", Proc. Natl. Acad. Sci. USA 92:2343-2344.
37. Schiewe (1991) "The science and significance of embryo cryopreservation", Journal of Zoo and Wildlife Midicine 22:6-22. - Walters et al. (1998) "Refrigeration can save seeds economically", Nature. vol.395 pp.758-758. - Loi et al. (2001) "Genetic rescue of an endangered mammal by cross-species nuclear transfer using post-mortem somatic cells", Nat. Biotechnol. 19:962 - 964.
38. Hanski & Gilpin, (eds, 1997) "Metapopulation biology: Ecology, genetics and evolution", San Diego, Academic Press. - Hanski (1998) "Metapopulation dynamics", Nature 396:41-49.
39. Whitlock & Barton (1997) "The effective size of a subdivided population", Genetics 146:427-441. - Whitlock (2002) "Selection, load and inbreeding depression in a large metapopulation", Genetics 160:1191-1202.
40. Rowe & Houle (1996) "The lek paradox and the capture of genetic variance by condition dependent traits", Proc. R. Soc. Lond. B Biol. Sci. 263:1415-1421.

15. **Harsh intra-organismal selection**. If selection of eggs and sperm is extraordinary harsh, many mutations with deleterious effects in adults could be removed easily and early. The view from a genomic decay perspective suggests that higher taxa possess some sophisticated techniques of checking gamete quality. Investigation of the complicated multi-level population genetics within an individual has only just begun[41].
16. **Larger effective population sizes**. As little is known about effective population sizes relevant for genomic decay, these might be larger than assumed. The main consequence is an increase in the power of selection.
17. **Environments**. If environments change for the better at the right moment, extinction might be avoided by masking deleterious mutations[42]. Thus harsh environments enlarge mutation effects and might keep populations from accumulating dangerous mutations.
18. **Purging of genetic load by inbreeding**. Massive expression of recessive alleles by inbreeding can lead to their selective removal and a population that might have stopped its decay (at least temporarily)[43]. However, inbreeding can also have the opposite effect and accelerate decay by fixation of deleterious mutations.
19. **Molecular solutions** either neutralise (eg. by RNA editing[44]) or decrease (eg. by increasing the amount of chaperones[45]) mutational ef-

41. Paulsson (2002) "Multileveled selection on plasmid replication", Genetics 161:1373-1384. - Rispe & Moran (2000) "Accumulation of deleterious mutations in endosymbionts: Muller's ratchet with two levels of selection", Am. Nat. 156:425-441. - Birky (1991) "Evolution and population genetics of organelle genes: Mechanisms and models", pp. 112-134 in: Selander et al. (eds) Evolution at the Molecular Level, Sunderland, MA, Sindauer Associates, Inc. - Bergstrom & Pritchard (1998) "Germline bottlenecks and the evolutionary maintenance of mitochondrial genomes", Genetics 149:2135-2146. - Kondrashov (1994) "Mutation load under vegetative reproduction and cytoplasmic inheritance", Genetics 137:311-318. - Otto & Hastings (1998) "Mutation and selection within the individual", Genetica 103:507-524. - Otto & Orive (1995) "Evolutionary consequences of mutation and selection within an individual", Genetics 141:1173-1187. - Hastings (1989) "Potential germline competition in animals and its evolutionary implications", Genetics 123:191-198. - Hastings (1991) "Germline-selection: Population genetic aspects of the sexual/asexual life cycle", Genetics 129:1167-1176. - Takahata & Slatkin (1983) "Evolutionary dynamics of extranuclear genes", Genet. Res. 42:257-265.
42. Szafraniec et al. (2001) "Environmental stress and mutational load in diploid strains of the yeast Saccharomyces cerevisiae", Proc. Natl. Acad. Sci. USA 98:1107-1112.
43. Byers & Waller (1999) "Do plant populations purge their genetic load? Effects of population size and mating history on inbreeding depression", Ann. Rev. Ecol. Syst. 30:479-513. - Lacy & Ballou (1998) "Effectiveness of selection in reducing the genetic load in populations of Peromyscus polionotus during generations of inbreeding", Evolution 52:900-909. - Frankham & Ralls (1998) "Conservation biology - Inbreeding leads to extinction", Nature 392:441-442. - Visscher et al. (2001) "A viable herd of genetically uniform cattle", Nature 409:303.

fects or make sure that mutational effects are large enough to be removed[45,46].

20. **Other solutions**. This list is not complete, because constraints on time and space did not allow this work to list all known potential solutions. And even if all known solutions were listed, one can never exclude that an important potential solution for the genomic decay paradox is found in the future.
21. **A bit of everything**. Although not all solutions listed are likely to contribute to the solution of the genomic decay paradox in a given species, it is likely that a combination of several does.

Thus this small excursion through biology suggests that probably some good reasons exist why we have not died 100 times over due to nuclear[47] or mitochondrial[48] genomic decay. Future research will have to show the details.

44. Boerner et al. (1997) "RNA editing in metazoan mitochondria: Staying fit without sex", FEBS Lett. 409:320-324.
45. Moran (1996) "Accelerated evolution and Muller's rachet in endosymbiotic bacteria", Proc. Natl. Acad. Sci. USA 93:2873-2878. - Queitsch et al. (2002) "Hsp90 as a capacitor of phenotypic variation", Nature 417:618-624.
46. Hurst & McVean (1996) "Evolutionary Genetics: ... and scandalous symbionts", Nature 381:650-651. - Gabriel et al. (1993) "Mullers Ratchet and mutational meltdowns", Evolution 47:1744-1757.
47. Kondrashov (1995) "Contamination of the Genome by Very Slightly Deleterious Mutations: Why Have We Not Died 100 Times Over?" J. theor. Biol. 175:583-594. - Eyre-Walker & Keightley (1999) "High genomic deleterious mutation rates in hominids", Nature 397:344-347. - Crow (1997) "The high spontaneous mutation rate: Is it a health risk?" Proc. Natl. Acad. Sci. USA 94:8380-8386. - Muller (1950) "Our load of mutations", Am. J. Hum. Genet. 2:111-176.
48. See this work, Chapter 21.

27 Practical relevance of Muller's ratchet

Discussions about Muller's ratchet and related topics are usually regarded as a highly esoteric topic from theoretical population genetics understood by only very few people. However, these topics can have very practical consequences that are much easier to understand. Some examples are given here.

Our view on selection

In 1893 August Weismann wrote a booklet with a title that can be translated as "The omnipotence of natural selection"[49]. This reflects a popular view about the power of selection and for some students of population genetics it has been a shocking experience to learn from their textbooks that deleterious mutations can be fixed and advantageous mutations can be lost[50]. This work showed further evidence that selection can be very strong on some occasions and extremely weak on others. Much future research will be necessary to estimate accurately the real power of selection in nature. This will help us to understand the evolutionary impact of our decisions concerning environmental and other issues. The following topics are affected.

27.1 Microbiology

The average microbiologist assumes that his carefully constructed and selected strains will not change significantly in the short time in the stationary phase after overnight growth or when stored in glycerol at - 80°C. Usually he is right. However, if this is repeated hundreds and thousands of times, it becomes less clear how closely related the final strain and its original ancestor still are. The work conducted here suggests that considerable adaptive evolution will have occurred, including potential loss of genes.

Best storing practice: Frozen batches

Microbes are best stored by producing a large batch that is frozen under known conditions and serves as a reservoir probably of not identical, but at least very little changed cells. Daily work should proceed with cultures amplified from small, independent samples of this batch. This ensures that all starting conditions are as similar as possible.

49. Weismann (1893) "Die Allmacht der Naturzüchtung. Eine Erwiderung an Herbert Spencer", Jena, Verlag von Gustav Fischer.
50. Eg. see treatment of fixation probabilities in Li (1997) "Molecular evolution", Sunderland, MA, Sinauer Associates Incorporated.

Further work

This work suggests further research to evaluate common microbiological practices in light of short-term evolutionary processes. A recommendation that could well stand at the end of such a work is that for well defined strains it could be important to record durations and temperatures of stationary phases as well as details of freezing events. This is probably the only way to infer the probable extent of evolutionary changes due to handling.

While this is probably unimportant for much of routine microbiological work, it can become very important for large culture collections that serve as *definitive* sources of strains. Most producers of starter cultures already use the batch method mentioned above; now they have another reason to do so.

27.2 Cloning

It has become popular to think of cloning as a technology that will fuel much progress. While this could turn out to be true for some areas, the recent debate often overlooks the vast evolutionary dangers of such a technology, if really employed on a large scale as some appear to suggest. Besides ethical issues about cloning of humans, there are a number of biological and evolutionary considerations that make this prohibitive:

- Cloning is asexual reproduction without recombination. Thus all benefits of recombination in terms of stopping genomic decay are switched off – a major step back in evolution.
- DNA of cloned cells is not taken from the germ-line. Thus the corresponding genome experienced a higher number of divisions per generation than the typical germ line cell that has been optimised to need as few replications as possible.
- If internal germ-line cell quality checks really exist (as suggested in the last chapter as an effective way to slow down genomic decay), cloning would skip them and go on with the corresponding higher load of mutations.
- Moreover, it is probably impossible to construct a DNA based system that checks the quality of cloned cells and has mutation rates as low as the germ line. The main reason is that currently conceivable systems will include extra replications to produce cells that can be checked. Thus the cells checked are not identical with those used. It will be extremely hard to beat the quality of the natural system – if that is possible at all.

- Current technical problems are likely to increase mutation rates even further.

These points make it quite clear that excessive use of cloning is prohibitive from an intelligent long-term perspective of evolution.

Current experiences with cloning support this negative view by a large number of developmental aberrations that are the by-products of one successfully cloned individual (eg. 87 cloned embryos to get one at least apparently normal cat[51]). Thus, when it comes to rescuing an endangered species[52], other means than cloning should be used as long as possible.

27.3 Conservation biology

Many conservation efforts only consider survival over the next 100 years, and it is good to do this, or else nothing will be left to conserve soon. However, if genetic causes do play a significant role in species extinction, it will be important to understand the details more fully. Clearly more research is needed in this area to truly achieve sustainable development[53].

Suggestions

This work suggests that effective population sizes of species are important in the long term, because they determine the border between selectively removed mutations and those accumulating effectively like neutrals despite being deleterious. However, the influence of the mutation rate is probably much more important as can be seen from the U-shaped plots presented in this work (eg. Figure 55 on page 273). Conservation efforts should consider this when regenerating former army or industry sites that may still contain mutagenic substances.

27.4 Sustainability and environmental mutagens

A careful investigation of the U-shaped plots of extinction time over selection coefficients (eg. Figure 55 on page 273) reveals that the genomic mutation rate is probably the most dangerous critical parameter, when it comes

51. Shin et al. (2002) "A cat cloned by nuclear transplantation", Nature 415:859-859. - Matzke & Matzke (2000) "Cloning problems don't surprise plant biologists", Science 288:2318. - Solter (2000) "Mammalian cloning: advances and limitations", Nat Rev Genet 1:199-207.
52. Loi et al. (2001) "Genetic rescue of an endangered mammal by cross-species nuclear transfer using post-mortem somatic cells", Nat. Biotechnol. 19:962 - 964.
53. United Nations Organization (1992) "Rio Declaration: Agenda 21", New York, UNO, Department of Public Information, Project Manager for Sustainable Development.

to genomic decay. While spontaneous mutation rates in organisms cannot be lowered by man, they can definitely be increased by human activity. Modern civilisation has brought forth an enormous array of mutagenic substances and activities which are not be reviewed here. While some of the most dangerous practices (eg. atmospheric tests of nuclear bombs) have been stopped, and a number of mutagenic agricultural chemicals have been forbidden, a large inventory of man-made mutagenicity still remains: from recent findings of mutagenic nitrophen in food[54], smoking, excessive x-raying in medicine and at airports, to the large number of sometimes little understood chemicals that enter the food chain. This list is not to imply that all of these necessarily affect the germ line, but substances or activities that can cause cancer due to mutagenicity are hot candidates for potential impact on the germline too. However, this work shows that it can be very dangerous to take issues of mutation rates lightly. Further research will have to investigate actual impact on the germ line. Where effects are clear, the corresponding dangerous agents need to be banned:

- Remove mutagens from the food chain.
- Remove environmental mutagens as applied in agriculture. (Besides potential effects on the food chain, one should also consider potentially endangered species that come in contact with them).
- Stop M-weapons. While the world is already sensitive to weapons of mass destruction and rightly so, it should become as sensitive about "M-weapons", ie. everything employed in warfare that is mutagenic from uranium ammunition and Agent Orange[55] to so called tactical nuclear weapons.

We borrowed this world from our children

To keep mutation rates at the spontaneous level by removing mutagens from the environment will be pivotal in achieving long-term sustainability[56] of our world, its biodiversity and, last but not least, human health[57].

54. RTECS (The Registry of Toxic Effects of Chemical Substances) (2000) "Nitrophene (Ether, 2,4 - dichlorophenyl p - nitrophenyl) known biological effects" http://www.cdc.gov/niosh/rtecs/kn802c80.html - NTP Chemical Repository (2001) "H&S: Nitrofen 1836-75-5" http://ntp-server.niehs.nih.gov/htdocs/CHEM_H&S/NTP_Chem1/Radian1836-75-5.html - National Toxicology Program (2002) "NTP Homepage" http://ntp-server.niehs.nih.gov/
55. Agent Orange was employed in the Vietnam war to remove the leaves of trees. It is now leading to a serious number of birth defects in the local population. See Dalton (2001) "Bilateral Vietnam study plans to assess war fallout of dioxin", Nature 413:442-442. - Cyranoski (2002) "US and Vietnam join forces to count cost of Agent Orange", Nature 416:252-252.
56. United Nations Organization (1992) "Rio Declaration: Agenda 21", New York, UNO, Department of Public Information, Project Manager for Sustainable Development.
57. Crow (1997) "The high spontaneous mutation rate: Is it a health risk?" Proc. Natl. Acad. Sci. USA 94:8380-8386. - Crow (1999) "The odds of losing at genetic roulette", Nature 397:293-294.

VIII. APPENDIX

Further reference material is listed here.

Introduction to evolutionary bioinformatics : A-2
Mutation rates paradox table for mtDNA : A-32
Related posters presented at international conferences : A-38
Glossary: A-43
Abbreviations: A-45
References: A-48
CV: A-72
Acknowledgements: A-76

Introduction to Evolutionary Bioinformatics[1]

Current bioinformatics mainly compares sequences and predicts their structure to understand their immediate function. Nevertheless, grasping their adaptive evolutionary significance is the key to much deeper understanding. Here evolutionary bioinformatics is defined as all computer-based analyses that use evolutionary models explicitly or implicitly to either understand data or the models themselves. While analytical models of evolution are important for solid foundations, mathematical tractability severely limits the amount of detail they can describe. Individual-based models overcome this barrier at the costs of higher computational complexity and often a more limited analytical understanding. As every model is an arbitrary abstraction of reality, we need to check many models in order to make reasonable predictions. Perspectives and pitfalls of this approach are discussed. At its heart evolutionary bioinformatics does just that: compare many models of evolution in order to understand the real causes that shaped the data we observe today.

Deep in the centre of every living cell on this planet lies a book with many seals and secret codes: the genome. Its information profoundly influences the life of an individual and the life of all descendants that inherit from this finely tuned collection of genes imprinted[2] in various ways. It is truly one of the most fascinating mysteries of biology.

In 1953 the century of biology began

It was only in 1953 that ink and paper of this book were discovered: Watson & Crick uncovered the structure of DNA that encodes everything in its four chemical letters Adenine (A), Thymine (T), Cytosine (C) and Guanine (G). It took about another decade before the (nearly[3]) universal genetic code was cracked. Thus researchers could translate a given DNA sequence into the corresponding chain of amino acids that would make a functional protein, when folded correctly. Today researchers are addressing the next big problem: actual prediction of the three-dimensional structure and function from a linear sequence[4]. All this will be one of many prere-

1. A shorter version of this chapter is under revision for publication in Briefings in Bioinformatics.
2. Hurst & McVean (1998) "Do we understand the evolution of genomic imprinting?", Curr. Opin. Genet. Dev. 8:701-708.
3. Osawa (1995) "Evolution of the genetic code", Oxford, Oxford University Press.
4. Berendsen (1998) "Protein folding - A glimpse of the holy grail?" Science 282:642-643.

quesites to be able to predict effects of mutations[5] and to truly understand evolution.

After the discovery of methods to cut, copy, paste, read and specifically modify DNA, the age of molecular biology[6] took off. Together with technical advances and increasing computing power it ushered in the age of genomics. The sequencing of the complete genome of the bacterium *Haemophilus infuenzae*[7] was only the first in a long row of genome projects[8] that now includes many bacteria and archaebacteria, a yeast[9], a fly[10], a worm[11], a plant[12] and the human genome[13].

0.1 Why do we need bioinformatics?

Databases explode

This flood of data contains information that generations of biologists could only dream of. However, while biological databases explode, the number of researchers is growing only slowly and the danger of drowning in the flood increases. To avoid this, beginning in the 1960s[14] a new discipline emerged that really took off only in the 1990s: computational biology, short bioinformatics[15]. With it came a variety of novel fields and methods that all heavily depend on computational methods like genomics[16], proteomics[17] and micro-array analyses[18]. Today, computer simulations are even used to check statistical relevance of experimental observations[19].

5. Charlesworth (1996) "Open questions: The good fairy godmother of evolutionary genetics", Curr. Biol. 6:220.
6. Sambrook et al. (1989) "Molecular Cloning: A Laboratory Manual. Second Edition". Cold Spring Harbour, Cold Spring Harbor Laboratory Press.
7. Fleischmann et al. (1995) "Whole-Genome Random Sequencing and Assembly of Haemophilus influenzae Rd", Science 269:496-512.
8. For current updates see Kyrpides (2002) "GOLD (TM): Genomes OnLine Database Homepage" http://wit.integratedgenomics.com/GOLD/
9. Goffeau et al. (1996) "Life with 6000 genes", Science 274:546-567.
10. Adams et al. (2000) "The genome sequence of Drosophila melanogaster", Science 287:2185-2195.
11. The C. elegans Sequencing Consortium (1998) "Genome sequence of the nematode Caenorhabditis elegans: A platform for investigating biology", Science 282:2012-2018.
12. The Arabidopsis Genome Initiative (2000) "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana", Nature 408:796-815.
13. Lander et al. (2001) "Initial sequencing and analysis of the human genome", Nature 409:860-921. - Venter et al. (2001) "The sequence of the human genome", Science 291:1304-1351.
14. Hagen (2000) "The origins of bioinformatics", Nat. Rev. Genet. 1:231-236.
15. Both terms are used interchangably here as done by many other authors. They do not imply the usual emphasis (i.e. Bio-*informatics* = *informatics* applied to biology or computational *biology* = *biology* using computers). Go to eg. http://compbiology.org, to find a different use of terms.
16. Brenner (2000) "Genomics. The end of the beginning", Science 287:2173-2174.
17. Pandey & Mann (2000) "Proteomics to study genes and genomes", Nature 405:837-846.

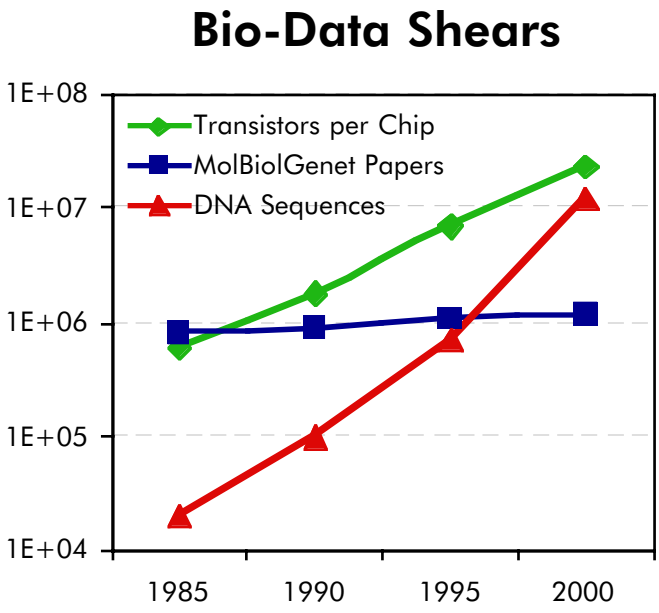


Figure 74 The Bio-Data-Shears[a].
To avoid drowning in the flood of new DNA sequences, a link between exploding computing power and exploding DNA databases has to be established for average biologists. Bioinformatics does just that.
The graph shows approximate DNA Sequences in GenBank, molecular biology and genetics papers in PubMed and Transistors per CPU on Intel Chips.

a. After Boguski (1998) "Bioinformatics - a new era", Trends Guide to Bioinformatics Suppl. 1998:1-3.

Building the *link*

The mission of bioinformatics is to close the bio-data shears with the help of computers (see Figure 74). It lifts biology to a new level by building a *link* between the explosion of biological data and the explosion of computing power. Reading genomic data is not enough, we need to understand it[20]. However, this link is not static. It is not enough to write a set of programs that analyse incoming data while the rest of biology remains unchanged. Bioinformatics rather profoundly changes the face of biology:

- To keep the link up to date, considerable bioinformatics manpower is needed - currently a big problem[21].
- Average biologists have to be trained to use the link in their daily work to take advantage of the latest research.
- Many biologists will have to understand more about models in mathematical language - not the strength of a typical biologist[22].

18. Ideker et al. (2001) "Integrated genomic and proteomic analyses of a systematically perturbed metabolic network", Science 292:929-934.
19. For an example see Stothard (1997) "Phylogenetic inference with RAPDs: Some observations involving computer simulation with viral genomes", J. Hered. 88:222-228.
20. Koonin (2001) "Primer: Computational genomics", Curr. Biol. 11:R155-R158.
21. Marshall (1996) "Hot property: Biologists who compute", Science 272:1730-1732. - Spengler (2000) "Bioinformatics in the information age", Science 287:1221-1222.

Table 17 Attempt at a systematic overview of bioinformatics (Wingender 2001, personal communication).

Approach	descriptive	analytical	synthetic
Sequence	GenBank, EMBL, DDBJ	sequence assembly	primer design hybridisation probe design
Structure	PDB, ProSite, PFAM, CATH	first pass annotation by gene finding tools 1D+3D homology searches	protein design protein structure prediction protein-ligand docking
Function	KEGG, TRANSFAC, TRANSPATH	Regulatory signal characterisation	Simulation of regulatory networks

To get many biologists to use the link, development of widely available, standardised, integrated and user-friendly packages is of paramount importance. A number of such packages and services are being developed by publicly funded groups [23] and by commercial companies[24].

Overview

For an in-depth review of how bioinformatics currently uses computers to analyse biological information please consult textbooks[25], resources on the Internet[26] or human genome analysis papers[13] and references therein. While there are other ways to present a systematic overview[27], Table 17 sets the stage for discussing evolutionary bioinformatics, which addresses issues even more complex than regulatory networks, the most complex entry in the table. In Table 17 complexity increases from descriptive to synthetic and from sequence to functional approaches. While the lower levels of complex-

22. Alberts (1998) "The cell as a collection of protein machines: preparing the next generation of molecular biologists", Cell 92:291-294.
Hillis (1993) "Why physicists like models and why biologists should", Curr. Biol. 3:79-81.
Chicurel (2000) "Mathematical biology. Life is a game of numbers", Nature 408:900-901.
23. Eg. see EMBnet (2002) "EMBOSS Homepage" http://www.uk.embnet.org/software/EMBOSS - DKFZ Heidelberg (2002) "HUSAR" http://genome.dkfz-heidelberg.de - National Center for Biotechnology Information (2002) "Entrez - PubMed" http://www.ncbi.nlm.nih.gov/entrez
24. Eg. see LION Bioscience (2002) "SRS, Bioscout" http://www.lion-bioscience.com/ - Accelrys formerly GCG (2002) "Accelrys - Software for pharmaceutical ... research" http://www.accelrys.com/
25. Baxevanis & Quelette, (eds, 2001) "Bioinformatics: A practical guide to the analysis of genes and proteins". 2nd, New York, John Wiley & Sons. - Brenner & Lewitter, (eds, 1998) "Trends Guide to Bioinformatics". Trends Supplement 1998, Cambridge, Elsevier Trends Journals. - Setubal & Meidanis (1997) "Introduction to computational molecular biology", Boston, PWS Publishing Company. - Peruski Jr. & Peruski (1997) "The Internet and the new biology: Tools for genomic and molecular research", Washington, D.C., American Society for Microbiology. - Benton (1996) "Bioinformatics. Principles and potential of a new multidisciplinary tool", Trends Biotechnol. 14:261-272. - Schulze-Kremer (1996) "Molecular bioinformatics: Algorithms and applications", Berlin, Walter de Gruyter. - Waterman, (ed, 1995) "Introduction to computational biology: Maps, sequences and genomes", London, Chapman & Hall.
26. Eg. Luz & Vingron (2001) "Online Lectures on Bioinformatics" http://www.dkfz-heidelberg.de/tpi/bioinfo/index.html
27. Benton (1996) "Bioinformatics. Principles and potential of a new multidisciplinary tool", Trends in Biotechnology 14:261-272.

ity usually serve as a foundation for higher levels, a lot of parallel development is taking place. Thus recent years have seen remarkable simulations of regulatory networks[28] for systems that are known well enough. Such work is important, as it may help elucidate one of the great mysteries of biology: adaptation[29].

A unifying framework is needed

As bioinformatics progresses, many more tools will be developed to extract information from sequences. This generates a flood of results that follows the flood of sequences. To make sense of these results we need a unifying framework theory. From a pharmaceutical company perspective the answer is simple: every result that may be used in drug design is important. While one can archieve a lot this way, drugs are not everything, so many biologists would keep on searching.

0.2 Evolution as a unifying theory

One does not have to go as far as Dobzhansky[30] to recognise the unifying power of the modern theory of evolution: it is currently the only theory that may unify biology.

Molecular biology

As can be seen in Figure 75, everything from biochemistry to molecular and developmental biology will be needed to predict rates and effects of mutations in an individual. The simulation of regulatory networks is an important step towards a complete and meaningful computation of mutational effects in an individual. This will be the only way to get detailed information on the distribution of mutational effects that is not limited or biased in a special way, as details can be accessed in the computer. However, this only makes sense, when enough biological information is used to feed such simulations. Ultimately, they will provide us with a genotype-phenotype map that might allow us to determine the probability of spontaneous mutations towards a specific phenotype.

Population biology

Then it will take all population biology from ecology to population dynamics and population genetics, as well as a number of other disciplines like palaeontology, climate prediction and environmental modelling to determine the environment that newly mutated individuals encounter. As the environment is the main agent of natural selection, it determines the cur-

28. Chicurel (2000) "Mathematical biology. Life is a game of numbers", Nature 408:900-901. See also footnote [fi] in Table 18.
29. Rose & Lauder, (eds, 1996) "Adaptation", San Diego, Academic Press.
30. Dobzhansky (1973) "Nothing in biology makes sense except in the light of evolution", American Biology Teacher 35:125-129.

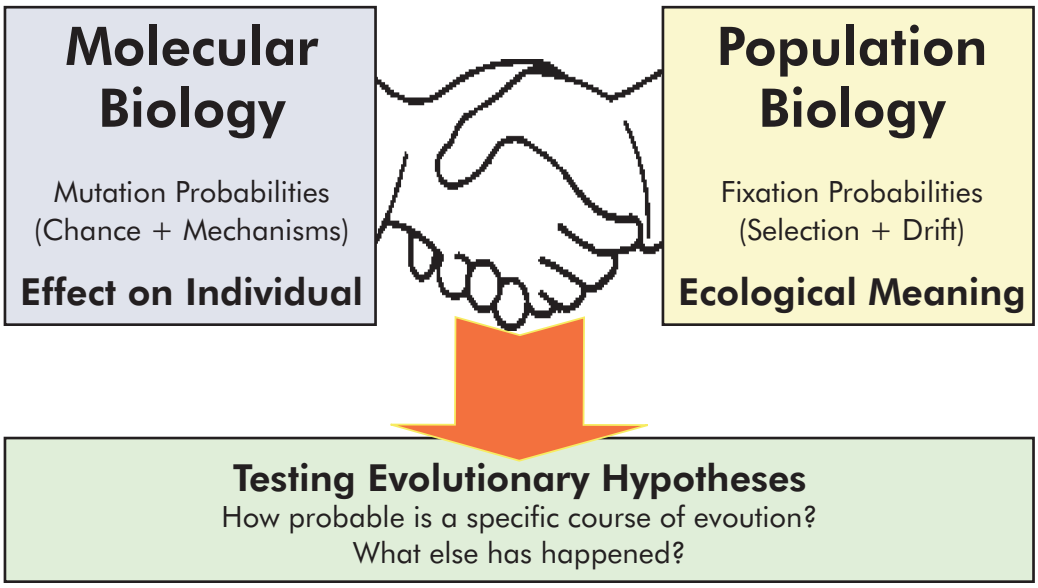


Figure 75 Evolution has a great potential to unify biology.

rent selective value of each genotype. Ultimately, population biology will answer the question, how probable it is that a given genetic feature will fix in the population - once it arises.

Combining taxonomic, comparative morphological and molecular data[31] will then allow us to formulate many specific hypotheses about adaptive evolution comparable to the work of Dean & Golding[32]. When all molecular mutation probabilities towards certain "desired" phenotypes are known together with the fixation probabilities in their respective environment and population structure, then an overall probabilitiy for such a given specific course of evolution can be computed. Likelihood ratio tests[33] can then be used to discriminate between alternative hypotheses for adaptive evolution of a specific functional feature, just as they are used today for

Read here what biology will do in hundred years

31. Huelsenbeck et al. (1996) "Combining data in phylogenetic analysis", Trends Ecol. Evol. 11:152-158.
32. Dean & Golding (1997) "Protein engineering reveals ancient adaptive replacements in isocitrate dehydrogenase", Proc. Natl. Acad. Sci. U.S.A. 94:3104-3109. - Golding & Dean (1998) "The structural basis of molecular adaptation", Mol. Biol. Evol. 15:355-369. - Dean (1998) "The molecular anatomy of an ancient adaptive event", Am. Sci. 86:26-37.
33. Huelsenbeck & Rannala (1997) "Phylogenetic methods come of age: testing hypotheses in an evolutionary context", Science 276:227-232.

Table 18 Examples of some important advances towards a new level of biological understanding

Work	Relevance
Invention of individual-based models[a]	Easy transformation of biological knowledge into software is possible via object-oriented programming languages
Observation of genome wide epistatic interactions[b]	What was done here with artificial digital organisms will one day be possible with more realistic individual-based genome models
Complete functional inventory of genes of an organism[c]	This is the first step towards a detailed understanding and modelling of all functions of a free living organism.
Uncover how DNA governs development of an organism[d]	To truly understand evolution, we must know how genes are turned into shapes. Although important progress in "evo-devo" is being made, a comprehensive theory is still missing.
Description of algorithms for morphological plant patterns[e]	Simple algorithms are capable of generating a great variety of beautifully complex plant morphologies.
Simulation of functions of a gene network in detail[f]	Functions of all proteins involved must be understood. If the model is not robust, some component may be missing[g]. At least this level is needed to predict consequences of mutations.
Uncover details of an ancient adaptive event[h]	When absolute probabilities for hypotheses on adaptive evolution with such a degree of detail can be computed, integration of biology will have reached an important milestone.
Experimental verification of an *in silico* metabolic model[i]	It is generally possible to build complex models *in silico* that make predictions that can be verified *in vivo*.

a. Reviewed in Huston et al. (1988) "New computer models unify ecological theory", Bioscience 38:682-691. - Judson (1994) "The rise of the individual-based model in ecology", Trends Ecol. Evol. 9:9-14. - McGlade, (ed, 1999) "Advanced Ecological Theory", Oxford, Blackwell Science. - DeAngelis & Gross, (eds, 1992) "Individual-based models and approaches in ecology", New York, Chapman & Hall.

b. Lenski et al. (1999) "Genome complexity, robustness and genetic interactions in digital organisms", Nature 400:661-664 - Yedid & Bell (2001) "Microevolution in an electronic microcosm", Am. Nat. 157:465-487. - Wilke et al.(2001)"Evolution of digital organisms at high mutation rates leads to survival of the flattest", Nature 412:331-3.

c. Blattner et al. (1997) "The complete genome sequence of Escherichia coli K-12", Science 277:1453-1462. For a minimal genome see Hutchison et al. (1999) "Global transposon mutagenesis and a minimal Mycoplasma genome", Science 286:2165-2169. For other genomes see footnotes [7-13] . However, it will still take considerable time until the functions of all genes are understood.

d. Goodman & Coughlin (2000) "Introduction. The evolution of evo-devo biology", Proc. Natl. Acad. Sci. USA 97:4424-4425 and the following 23 papers (pp.4426-4540). - Gehring & Ikeo (1999) "Pax 6: mastering eye morphogenesis and eye evolution", Trends Genet. 15:371-377. - Raff (1996) "The shape of life: genes, development, and the evolution of animal form", Chicago, University of Chicago Press.

e. Prusinkiewicz & Lindenmayer (1990) "The Algorithmic Beauty of Plants", New York, Springer-Verlag. - Investigation of molecular basis has only begun, see Parcy et al. (1998) "A genetic framework for floral patterning", Nature 395:561-566.

f. Barkai & Leibler (1997) "Robustness in simple biochemical networks", Nature 387:913-917. - Alon et al. (1999) "Robustness in bacterial chemotaxis", Nature 397:168-171. - Yi et al. (2000) "Robust perfect adaptation in bacterial chemotaxis through integral feedback control", Proc Natl Acad Sci U S A 97:4649-4653. - von Dassow et al. (2000) "The segment polarity network is a robust developmental module", Nature 406:188-192. See also the special issue on "Modelling cell systems", Briefings in Bioinformatics (2001), vol. 2, issue 3, pp. 219-288.

g. see the work of von Dassow et al. 2000, ibid.; when they added 2 little known proteins, their model worked.

h. Dean & Golding (1997) "Protein engineering reveals ancient adaptive replacements in isocitrate dehydrogenase", Proc. Natl. Acad. Sci. U.S.A. 94:3104-3109.
i. Endy et al. (2000) "Computation, prediction, and experimental tests of fitness for bacteriophage T7 mutants with permuted genomes", Proc Natl Acad Sci U S A 97:5375-5380. Edwards et al. (2001) "In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data", Nature Biotechnology 19:125-130.

checking much simpler models of neutral evolution[33]. While it may well take another century before biology routinely operates at such a high level, glimpses of what it would be like might be obtained in systems that are simple enough to be analysed today (Table 18).

Challenges to future biology

To meet these and similar challenges, corresponding software packages and databases need to reach new levels of integration over the coming years. Ultimately, these questions pose enormous computational problems[34]. Such work will help us to understand biodiversity[35], make realistic ecological predictions[36] for solving environmental problems[37], save endangered species by evaluating consequences of management decisions[38], fight evolution of antibiotic resistance in pathogens[39], predict shelf-life of nutrients[40], predict

34. Levin et al. (1997) "Mathematical and computational challenges in population biology and ecosystems science", Science 275:334-343.
35. Bisby (2000) "The quiet revolution: biodiversity informatics and the internet", Science 289:2309-2312. - Edwards et al. (2000) "Interoperability of biodiversity databases: biodiversity information on every desktop", Science 289:2312-2314. - Inchausti & Halley (2001) "Investigating long-term ecological variability using the Global Population Dynamics Database", Science 293:655-657.
36. Carpenter (2002) "Ecological futures: Building an ecology of the long now", Ecology 83:2069–2083. - Clark et al. (2001) "Ecological forecasts: an emerging imperative", Science 293:657-660. - Smith (2000) "Nice work - but is it science? Untestable ecological theory won't help solve environmental problems." Nature 408:293-293.
37. Eg. Hallam et al. (1996) "Modeling effects of chemicals on a population: Application to a wading bird nesting colony", Ecol. Model. 92:155-178.
38. Examples: Jager et al. (1997) "Modelling the linkages between flow management and salmon recruitment in rivers", Ecol. Model. 103:171-191. Many similar studies can be found in the Journal "Ecological Modelling". Besides such specialized individual-based models, population viability analysis is another powerful prediction tool: Shaffer (1997) "Population Viability Analysis: Determining natures's share", pp. 215-217 in: Meffe & Carroll (eds) Principles of conservation biology, 2nd, Sunderland, MA, Sinauer Associates. - Boyce (1992) "Population viability analysis", Ann. Rev. Ecol. Syst. 23:481-506. - Brook et al. (2000) "Predictive accuracy of population viability analysis in conservation biology", Nature 404:385-387. - Coulson et al. (2001) "The use and abuse of population viability analysis", Trends Ecol. Evol. 16:219-221.
39. Baquero & Blazquez (1997) "Evolution of antibiotic resistance", Trends Ecol. Evol. 12:482-487. - Levin et al. (2000) "Compensatory mutations, antibiotic resistance and the population genetics of adaptive evolution in bacteria", Genetics 154:985-997. - Davies & Roberts (2002) "Antimicrobial resistance: An ecological perspective", (http://www.asmusa.org/acasrc/pdfs/Colloquia/Antimicrobialrpt.pdf) Washington, D.C., American Academy of Microbiology.
40. Ross & McMeekin (1994) "Predictive microbiology", Int J Food Microbiol 23:241-264. - Whiting & Buchanan (1997) "Predictive Modeling", pp. 728-739 in: Doyle et al. (eds) Food microbiology, Washington, D.C., ASM Press. - McDonald & Sun (1999) "Predictive food microbiology for the meat industry: a review", Int J Food Microbiol 52:1-27. - Kreft et al. (1998) "BacSim, a simulator for individual-based modelling of bacterial colony growth", Microbiology 144:3275-3287.

evolution[41] and minimise deleterious changes in the course of evolution due to human actions[42].

0.3 Evolutionary bioinformatics: One definition and three viewpoints

Current computational biology often touches evolution explicitly or implicitly and there is a variety of phylogenetic-evolutionary approaches to bioinformatics[43]. Construction of phylogenetic trees[44], homology-based genome annotation and even a simple sequence alignment[45], all imply some evolutionary model (that usually excludes selection and assumes some mutational spectrum to deal with gaps). Coalescent[46] and other simulations from population genetics address evolution more explicitly, whereas ecological studies chose to either model rapid evolution[47] or assume no genetic changes. Finally, besides software for teaching purposes[48], a number of meta-analysis and results databases touch evolution and importance of good databases with useful links should not be underestimated[49]. As average biologists are usually little aware of the models beneath the tools they use, it might be helpful for increasing awareness to explicitly define evolutionary bioinformatics as follows:

Definition

Evolutionary bioinformatics encompasses all computer-based analyses that use evolutionary models explicitly or implicitly.

41. Hall (2001) "Predicting Evolutionary Potential. I. Predicting the Evolution of a Lactose-PTS System in Escherichia coli", Mol Biol Evol 18:1389-1400.
42. Palumbi (2001) "Humans as the world's greatest evolutionary force", Science 293:1786-1790. - Shugart et al. (1992) "The potential for application of individual-based simulation models for assessing the effects of global change", Ann. Rev. Ecol. Syst. 23:15-38. - Liu (2001) "Integrating ecology with human demography, behavior, and socioeconomics: Needs and approaches", Ecol. Model. 140:1-8.
43. Pagel (2000) "Phylogenetic-evolutionary approaches to bioinformatics", Briefings in Bioinformatics 1:117-130.
44. Miyamoto & Cacraft (1991) "Phylogenetic analysis of DNA sequences", New York, Oxford University Press. - Hall (2001) "Phylogenetic trees made easy: A how-to manual for molecular biologists", Sinauer Associates.
45. Mindell (1991) "Aligning DNA sequences: Homology and phylogenetic weighting", pp. 73-89 in: Miyamoto & Cacraft (eds) Phylogenetic analysis of DNA sequences, Oxford University Press.
46. Donnelly & Tavare (1995) "Coalescents and genealogical structure under neutrality", Annu. Rev. Genet. 29:401-421. - Slade (2001) "Simulation of 'hitch-hiking' genealogies", J Math Biol 42:41-70.
47. Thompson (1998) "Rapid evolution as an ecological process", Trends Ecol. Evol. 13:329-332.
48. Meir (1996) "EcoBeaker 1.0 - An ecological simulation program. EcoBeaker Laboratory Guide and the EcoBeaker Program Manual", Sunderland, MA, Sinauer Associates, Inc. - Meir (2002) "Welcome to Ecobeaker 2.0 - Ecology teaching software" http://www.ecobeaker.com/ - Alstad (2002) "POPULUS" http://www.cbs.umn.edu/populus/ - Lemmon (2002) "EvoTutor - Learning through interactive simulation" http://www.evotutor.org/
49. Karp (1996) "Database links are a foundation for interoperability", Trends Biotechnol. 14:273-279.

Table 19 Viewpoints of evolutionary bioinformatics.

Viewpoint	Visibility of evolution models	Goal	Examples
Genotypes analysis	explicit	Assume evolutionary model to analyse seqences (build tree, estimate parameters)	PHYLIP, PAUP, DNA-SP, Arlequin, Mesquite[a]
	implicit	Multiple sequence alignments; genome annotation	ClustalW, MALIGN, HMMER[b]
Model analysis	explicit	Model evolutionary processes to understand them better or to make predictions	Coalescent, individual-based, etc. simulations, evolution@home[c]
	implicit	Model ecological processes to understand them better or to make management decisions for endangered species etc.	Individual-based or metapopulation simulations without genetics[d]
Results analysis	explicit	Provide easy access to often scattered detail results relevant for evolution	Tree of Life, TAED, evolution@home[e]
	implicit	Make annotated genomes and comparative genomics analyses accessible	Human genome browsers[f]

a. Felsenstein (2002) "PHYLIP Homepage" http://evolution.genetics.washington.edu/phylip.html. - Swofford (2002) "PAUP 4.0" http://paup.csit.fsu.edu/. - Excoffier (2002) "Arlequin's home on the web" http://lgb.unige.ch/arlequin. - Maddison & Maddison (2002) "Mesquite - A modular system for evolutionary analysis" http://mesquiteproject.org/

b. European Bioinformatics Institute (2002) "ClustalW" http://www.ebi.ac.uk/clustalw. - Eddy (2002) "Sean Eddy Lab Homepage + HMMER" http://HMMER.wustl.edu/

c. Examples: Eyre-Walker et al. (1998) "Investigation of the bottleneck leading to the domestication of maize", Proc. Natl. Acad. Sci. USA 95:4441-4446. - Kawata (1995) "Effective population size in continuously distributed populations", Evolution 49:1046-1054. - Gavrilets et al. (2000) "Dynamics of speciation and diversification in a metapopulation", Evolution 54:1493-1501. - Gordo & Charlesworth (2000) "The degeneration of asexual haploid populations and the speed of Muller's ratchet", Genetics 154:1379-1387. - evolution@home is dedicated to development of such models, see http://www.evolutionary-research.net.

d. Hanski (1998) "Metapopulation dynamics", Nature 396:41-49. - Hanski & Gilpin, (eds, 1997) "Metapopulation biology: Ecology, genetics and evolution", San Diego, Academic Press. - Heino & Hanski (2001) "Evolution of migration rate in a spatially realistic metapopulation model", Am. Nat. 157:495-511. - Humphries et al. (1996) "An individual-based model of alpine plant distributions", Ecol. Model. 84:99-126.

e. Maddison (2002) "The Tree of Life Homepage" http://phylogeny.arizona.edu/ - Liberles et al. (2001) "The Adaptive Evolution Database (TAED)", Genome Biol 2:RESEARCH0028 (http://www.sbc.su.se/~liberles/TAED.html). - One of the goals of evolution@home is to generate public databases that contain simulation results for well defined models to allow other researchers to easily check, whether a given process applies to their model system or to develop better analytic models for understanding the processes investigated.

f. See http://genome.ucsc.edu/ or http://www.ensembl.org

It can have several viewpoints (Table 19). Such an overview might become a starting point to raise interest about evolutionary issues in the next generation of biologists. The rest of this text will focus on the model-analyis viewpoint in general and on individual-based models in particular.

0.4 Why we need good *in silico* experiments

Most people are only aware of two approaches to understanding biology: experiment and theory. However, a careful look suggests a more differentiated picture. In physics, for example a "gedanken experiment"[50] is a valid way of gathering knowledge and computer simulations are now widely used to explore stochastic processes[51]. Recently, simulations of molecular dynamics have become increasingly fruitful for molecular biology[52]. As processes in ecology and evolution have strong stochastic components when a few individuals play key roles, investigation of models that are beyond extreme simplicity depends more and more upon simulation[53].

Resistance

However, as ROWE[54] points out, not a few researchers "*trained in the classical, analytical approach to mathematics or physical sciences often feel that computer simulations cannot provide rigorous results, or that the 'brute force' approach excludes a more elegant, compact solution to the problem.*" This is probably because good theory in the past *had to be* analytical, as fast computers were lacking. As we now have both opportunities, simulations will emerge as the way to answer inherently complex problems that have no clean, elegant, analytical solution[54].

0.4.1 Ways to understand biology

Biology-Man

Table 20 describes several approaches to biological knowledge with an analogy. Just consider "BIOLOGY-MAN", an imaginary being that symbolises the scientific discipline of biology. As science in the sense of POPPER[55], BIOLOGY-MAN lives from making predictions about nature that actually come to pass, although they could have been proven false. When all ways of understanding are used to complement each other, BIOLOGY-MAN is fine: he makes predictions based on his knowledge and checks them by observations. Now let us examine two extremes:

Extremes

- If there were no *in silico* experiments that incorporate the observed data to build models of reality that are too complex for analytical

50. Eg. Hess & Philipp (2001) "Bell's theorem and the problem of decidability between the views of Einstein and Bohr", Proc. Natl. Acad. Sci. USA 98:14228–14233.
51. Huston et al. (1988) "New computer models unify ecological theory", BioScience 38:682-691.
52. Rapaport (1995) "The art of molecular dynamics simulations", Cambridge, Cambridge University Press. - Goodfellow, (ed, 1995) "Computer modelling in molecular biology", Weinheim, VCH.
53. Wilson (2000) "Simulating ecological and evolutionary systems in C", Cambridge University Press.
54. Rowe (1994) "Theoretical Models in Biology. The Origin of Life, the Immune System, and the Brain", p. vii, Oxford, Clarendon Press.
55. Popper (1963) "Conjectures and Refutations". - Popper (1989) "Logik der Forschung" 9th ed.

Table 20 Ways to understand biology.

latin name	meaning	BIOLOGY-MAN analogy	strength	weakness
in ratio	analytic model	hard, dry bone	well understood, precise predictions	limited to simple models by mathematical tractability
in silico	simulations of more realistic models	flesh	can be very realistic, can use original data, precise predictions	sometimes too hard to understand, computer limited
in vitro	experiment without anything alive	air to breathe	precise molecular observation and manipulation possibilities	expensive, extrapolation to *in vivo* is not always possible
in vivo	laboratory experiment with living cells or organisms	water to drink	controlled environment allows specific manipulations	relevance for natural settings not always clear, limited mechanistic understanding
in natura	observation of organisms in their natural setting	bread to eat	get information on actual natural processes, allows falsification of *in ratio* and *in silico* models	historic, usually limited by ~3 year funding periods, limited mechanistic understanding
in-tuitio	find good questions[a]	spirit with bright ideas	inexpensive and fast	is no scientific proof in itself

a. Krebs (1979) "On asking the right kind of question in biological research". Molecular Mechanisms of Biological Recognition: Proceedings of the Sixth Ahron Katzir-Katchalsky Conference in conjunction with the Minerva Symposia in Biology, Göttingen and Braunlage/Harz, September 24-30, 1978, Elsevier.

maths, then biology would have to link bread, water and air to dry bones. This is a picture of much of biology today, where wet-lab biologists feel rather intimidated by the analytical models of theoreticians. On the other side, theoreticians often dismiss actual observations as "realism that clouds the pure effect under analytical investigation". This may lead to the 'trap of originality'[56].

- If simulations were not complemented by the other approaches, then biology would starve and have no bones - like a pile of rotten flesh. Results would say virtually nothing about the real world and confirm only the prejudices of people who do not like simulations.

56. Smith (2000) "Nice work - but is it science? Untestable ecological theory won't help solve environmental problems." Nature 408:293-293.

All possible ways to understand biology need each other

Only if all approaches work together, new levels of understanding will be reached. James Crow once said: "You can know more than you can prove."[57] This should be kept in mind when striving for analytical understanding, as not all useful models are mathematically tractable and even very simple population models can lead to very complicated behaviour[58]. When it comes to modelling, two approaches are possible[59]: On the *prediction-oriented* extreme, a realist may use a neural network black box to arrive at good predictions for a very specific system without understanding it. On the *knowledge-oriented* extreme a theoretician might prove a general principle that is too abstract to apply to any biological situation on this planet. While each researcher has to find his or her own trade-off between generality, realism and precision[60], individual-based models are inherently knowledge-oriented, but allow either emphasis. Ultimately, however, we need both, as science is about the *knowledge* needed to make good *predictions*.

0.4.2 How to make simulations that are junk

Unfortunately the prejudices against simulations are not completely unfounded. It is as easy to make bad simulations as it is to make bad experiments. However, the fact that lots of measurement artefacts are observed in poor-quality experiments does not invalidate the results from good experiments. The same is true for simulations, besides the fact that current biology had over 10-fold more time to refine experimental approaches compared to simulations. To help raise the standard for good simulations, here is a list of what to do to produce junk simulations[61]:

- Your computer is always right, so do not question it.
- Do not waste your implementation time with coding conventions, speaking variable names, extensive code comments or other documentation. Raw source code is enough documentation.
- Once you have compiled your program, it is free from serious errors. So do not waste your time searching for any.

57. Crow (1999) "You can know more than you can prove." personal communication.
58. May (1976) "Simple mathematical models with very complicated dynamics", Nature 261:459-467.
59. Mooij & Boersma (1996) "An object-oriented simulation framework for individual-based simulations (OSIRIS): Daphnia population dynamics as an example", Ecol. Model. 93:139-153. - Starfield & Bleloch (1986) "Building models for conservation and wildlife management", London, Collier Macmillan.
60. Levins (1966) "The strategy of model building in population biology", Am. Sci. 54:421-431.
61. For additional points see p.116 of Law & Kelton (1991) "Simulation modelling and analysis". 2nd., New York, McGraw-Hill.

- Do not try to understand or explain what is going on in your simulation. The world is too complex as it is.
- Do not waste time checking extremes with analytical theory.
- Do not worry about rounding errors or bad random number generators[62]. They are never a problem.
- It is enough to check only few important parameter combinations.
- A few simulations are enough, as statistics lie anyway.
- Your model is interesting enough in itself. So do not worry about biologically relevant parameters or simplifying assumptions.
- It is not necessary to build different models of the same thing, as your model is the truth already. So do not waste your time comparing models, as they all will give the same true answer.
- After you have completed your work, extrapolate your results to cases where you will never be able to check them. Then call it reality.

While no model is mature when development starts, it is important to strive to remove all junk points with time in knowledge-oriented models that are built to actually capture mechanistic essentials of the true process in nature for better general understanding.

Alternatives

In case of individual-based prediction-oriented models, it is important to have actual evidence for the causal relationships in the model. This is sometimes as difficult to get as reasonable parameter estimates[63]. Thus completely different approaches like regression models, principal component analysis[64], fuzzy logic[65], neuronal networks[66] or heuristic equations may sometimes lead to similar or better prediction quality.

0.4.3 Building good models is an art

Building good models is an art that is best learned by experience. However, some important features are shared by all good knowledge-based modelling efforts:

62. Press et al. (1992) "Numerical recipies in C". 2nd, Cambridge, Cambridge University Press.
63. Rigler & Peters (1995) "Reductionism versus holism: An old problem rejuvenated by the computer", pp. 95-115 in: Rigler & Peters (eds) Science and limnology, Oldendorf/Luhe, Germany, Ecology Institute. - Omlin et al. (2001) "Biogeochemical model of Lake Zurich: sensitivity, identifiability and uncertainty analysis", Ecol. Model. 141:105-123.
64. van Tongeren (1995) "Data-Analysis or Simulation-Model - a Critical-Evaluation of Some Methods", Ecol. Model. 78:51-60.
65. Enea & Salemi (2001) "Fuzzy approach to the environmental impact evaluation", Ecol. Model. 136:131-147.
66. Najjar et al. (1997) "Computational neural networks for predictive microbiology: I. Methodology", Int J Food Microbiol 34:27-49.

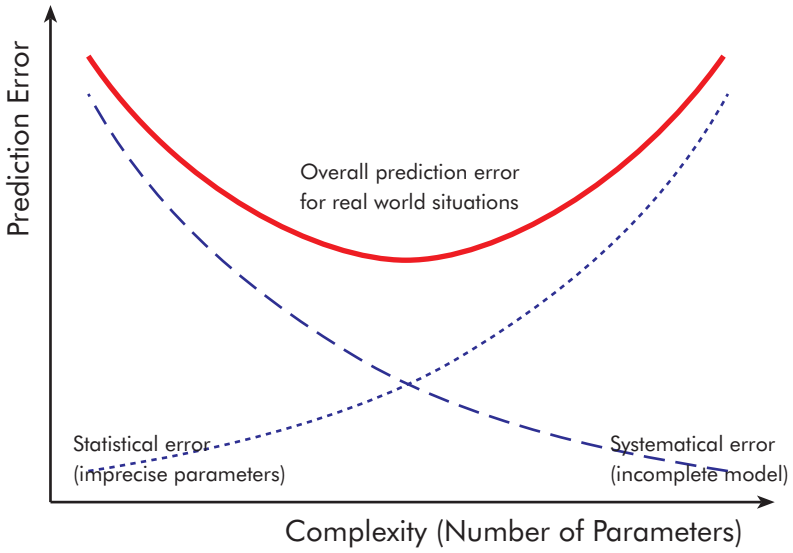


Figure 76 Relationship between systematic and statistical errors.
Including more parameters does not necessarily increase real world prediction accuracy[a].

a. After O'Neill, 1973, "Error analysis of ecological models", Deciduous Forest Biome. Memo Report 71-15, cited after Wissel (1989) "Theoretische Ökologie: Eine Einführung", Berlin, Springer-Verlag.

Have a goal

- **Formulate your question.** You need a goal for your model to make it successful[67]. Do you want to predict actual behaviour of a real species, or do you want to investigate theoretical relationships?

Capture the essentials

- **Find an efficient abstraction.** While simulations allow you to include more reality than analytical models, you are still far from being able to model every detail of your system. So you have to develop concepts that capture the essentials. An excellent example is the idea of fitness. It can reduce all the complexities of an individual's life to just a single number in many models. The degree of details necessary depends on your question. If you want to reach conclusions about real world situations with your model, consider the trade-off between systematic and stochastic prediction error (Figure 76)[68].

Distribution model?

- **Decide whether an i-state distribution model is better.** Individual-based models are usually i-state *configuration* models (ie. each individual has its own state of configuration). However, sometimes it is better to use i-state *distribution* models, also referred to as demograph-

67. Bart (1995) "Acceptance criteria for using individual-based models to make management decisions", Ecolog. Applic. 5:411-420.
68. Snowling & Kramer (2001) "Evaluating modelling uncertainty for model selection", Ecol. Model. 138:17-30.

ic or state-variable models[69]. These models use various properties of individuals (= state variables like age, size, physiological state, ...) to classify a population. Thus the computer does not track each individual, but rather the number of individuals in the corresponding classes. This approach facilitates simulation of large populations[70] and is accessible to a number of mathematical tools that may produce results more powerful and general than is possible for i-state configuration models[69]. However, when (i) the state of individuals becomes too complex, (ii) demographic stochasticity becomes important or (iii) spacial structure plays a role, then it becomes much easier to model each individual separately[69]. To integrate accurate small-scale with comprehensive large-scale models[71], it is sometimes attempted to resolve discrepancies between these approaches[72].

Get a framework

- **Choose a good framework.** Besides the right modelling approach, you need a good software environment that supports it. If you have none, you may spend considerable time developing it, as many others have done before you[73].

Celebrate when you find an error

- **Test your model extensively.** Check your model, your code, your test results in every possible way for implementation errors, logical errors, rounding errors and strange results. Check your results for plausibility and compare them with analytical theory where possible. No (!) piece of non-trivial software on this planet is free of errors, so you have to prove your code gives at least reasonable results. Do not underestimate the time needed for such quality checks.

Invest in input

- **Get good input parameter values.** If you want reliable output, you have to provide quality input. If you cannot get independent values

69. Caswell & John (1992) "From the individual to the population in demographic models", pp. 36-61 in: DeAngelis & Gross (eds) Individual-based models and approaches in ecology, New York, Chapman & Hall.
70. Scheffer et al. (1995) "Super-Individuals a Simple Solution for Modeling Large Populations on an Individual Basis", Ecol. Model. 80:161-170.
71. Tang et al. (2001) "Simultaneous equations, error-in-variable models, and model integration in systems ecology", Ecol. Model. 142:285-294.
72. Wilson (1998) "Resolving discrepancies between deterministic population models and individual-based simulations", Am. Nat. 151:116-134.
73. E.g.: Sequeira et al. (1997) "Implementing generic, object-oriented models in biology", Ecol. Model. 94:17-31. - Mooij & Boersma (1996) "An object-oriented simulation framework for individual-based simulations (OSIRIS): Daphnia population dynamics as an example", Ecol. Model. 93:139-153. - Maley & Caswell (1993) "Implementing I-State Configuration Models for Population- Dynamics - an Object-Oriented Programming Approach", Ecol. Model. 68:75-89. - Gathmann & Williams (1998) "Inter-site: a new tool for the simulation of spatially realistic population dynamics", Ecol. Model. 113:125-139.

for your input parameters, you may have to estimate them by fitting model output to observations of the real system. If you do that, be cautious, as incorrect models can sometimes be made to give the correct answers by parameter fitting[74]. Be prepared for surprises[75].

Check for robustness.

- **Carefully analyse output.** Only if you compare many repeats of simulations over a wide range of input parameter space, you will get a feeling for the robustness of individual results. Try to estimate the uncertainty of your model's predictions[76].
- **Document your model properly.** Make sure that others can understand the strengths and weaknesses of your model before it is used for making decisions[77].

Finally, as there are many ways to describe the reality you are trying to model, comparing different models is pivotal to assessing model quality.

0.4.4 Advantages of good *in silico* experiments

After overcoming the pitfalls, good *in silico* simulation experiments recently emerged as a powerful means of understanding nature:

- They can be much more realistic than analytical maths.
- They are explicit about the model they simulate.
- They allow 'clean' analysis (no measurement artefacts).
- They allow access to details that often cannot be measured, but may be crucial for understanding the dynamics of the system.
- They can often incorporate data in its original format.
- They can simulate long periods of time.
- They are cheap, *if* your software environment is good.

Lack of computing biologists and proper frameworks hinders progress

Good examples for this can be found in Table 18. Why then is *in silico* biology not as widely distributed as it could be? First of all, reasonable levels of computing power have only been available for a few years. One cannot expect the overall approach to biology to change in such a short period of time. Second, average biologists do not understand enough about comput-

74. Hopkins & Leipold (1996) "On the dangers of adjusting the parameters values of mechanism-based mathematical models", J. theor. Biol. 183:417-427.
75. Barlund & Tattari (2001) "Ranking of parameters on the basis of their contribution to model uncertainty", Ecol. Model. 142:11-23.
76. Omlin & Reichert (1999) "A comparison of techniques for the estimation of model prediction uncertainty", Ecol. Model. 115:45-59.
77. Bart (1995) "Acceptance criteria for using individual-based models to make management decisions", Ecolog. Applic. 5:411-420. - Benz & Knorrenschild (1997) "Call for a common model documentation etiquette", Ecol. Model. 97:141-143. - Benz et al. (2001) "ECOBAS - modelling and documentation", Ecol. Model. 138:3-15.

ing and average programmers do not understand enough about biology to build reasonable models. Third, those that do model, often have a hard time getting past the barriers to making good models and thus confirm prejudices against models. Fourth, unrealistic expectations of what models can do lead to frustration. Fifth, lack of good framework software makes most researchers start from scratch. Before we discuss the types of models that can be built, we need to look at general limits of knowledge that apply to all simulations.

0.4.5 General limitations of *in silico* biology

As *in silico* biology is a part of science, so all logical limits of science apply[78]: (i) No model will ever be proven right. It will only be successful, if it can be falsified, but survives attempts to do so[55]. (ii) History of particular (evolutionary) events cannot be proven by reproducible experiments. (iii) Conclusions cannot be drawn from indecisive data and (iv) the tension between reductionism and holism prevails[79,85], as some scientists prefer explanations that contain as little realism as necessary, while others prefer to add as many facts as possible. Other limitations that come to the surface by using the *in silico* approach include the following:

- **Fundamental unpredictabiliy.** It is easy to see why *exact* outcomes cannot be predicted by *stochastic* simulations. However, if there are no stochastic events with unknown outcome, a more fundamental cause for unpredictability shows up: chaos. The literal wing of the butterfly that chances future weather is often encountered in surprisingly simple systems[80]. Chaos leads to situations where changes of input parameters can fundamentally change results, although these changes were smaller than the highest measurement accuracy available. Chaotic systems can exhibit some very strange phenomena[81] making it popular to investigate complex, chaotic systems[82]. However, chaos does not completely prohibit predictions[83] and sometimes it is even possible to distinguish chaos from environmental noise[84].

Biologists have to get used to results like "outcome x has probability p in model y with parameterset z"

78. Casti (1996) "Confronting science's logical limits", Sci. Am. October:78-81.
79. Nurse (1997) "Reductionism: The ends of understanding", Nature 387:657-657.
80. May (1976) "Simple mathematical models with very complicated dynamics", Nature 261:459-467. Huisman & Weissing (2001) "Fundamental unpredictability in multispecies competition", Am. Nat. 157:488-494.
81. Nusse & Yorke (1996) "Basins of attraction", Science 271:1376-1380.
82. Science (1999) "Complex systems", 284:79-109. - Nature (2001) "Complex systems", 410:240-284.
83. Shukla (1998) "Predictability in the midst of chaos: A scientific basis for climate forecasting", Science 282:728-731.

Most chaotic systems have regions in parameter space without chaos and even in the chaotic regions, probabilities for certain specific outcomes can be computed. Such results are not fundamentally different from stochastic models, where a specific outcome can only be predicted with a certain probability too. As most biological systems involve strong stochastic components and pure chaos is mainly a domain of deterministic models it should be natural for biologists to think in probabilities. Physicists were forced to adopt this thinking with the advent of quantum physics. Many biologists, however, still have the deterministic world view of the 19th century, something that should change in the next generation of biologists[22].

Computing power

- **Computing complexity.** To compute probabilities, many similar simulations have to be run. Thus lack of computing power may severely limit investigations of complex models. However, according to Moore's law this problem is halved every 2 years or so. Since microcomputers reached the power of former supercomputers in the late 1990s, a wide range of simulations has become possible. Nevertheless, this is often not enough.

To get all data can be impossible

- **Limited manpower and knowledge.** For many ecological or evolutionary models all the biologists in the world would not be enough to collect all the data that would be needed for "realistic" simulations[85]. While no *in silico* trick can substitute for that, it is a delusion to think that a system is understood when all these data were available[79]. To understand a system, more is needed than a bag of elementary relationships that lead to some poorly understood outcome, when simulated. What exactly this *more* is, however, depends on individual taste[79].

Parameters may mask mechanisms

- **Some mechanisms are hard to confirm.** Adjustment of parameter values in knowledge-oriented models can be dangerous. Sometimes wrong mechanisms can be tuned to give correct answers[86]. Although it may pretend otherwise, such a model may have no more

84. Sugihara & May (1990) "Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series", Nature 344:734-741.
85. Rigler & Peters (1995) "Reductionism versus holism: An old problem rejuvenated by the computer", pp. 95-115 in: Rigler & Peters (eds) Science and limnology, Oldendorf/Luhe, Germany, Ecology Institute.
86. Hopkins & Leipold (1996) "On the dangers of adjusting the parameters values of mechanism-based mathematical models", J. theor. Biol. 183:417-427.

value than a descriptive, prediction-oriented model that hides mechanisms like a neuronal network.

- **Software issues.** Without an appropriate framework, a functional simulation code may be too expensive to develop, as several non-trivial issues have to be addressed.

Development costs may be prohibitive

Equipped with realistic expectations we shall now explore different ways to build individual-based models *in silico*.

0.5 General simulation approaches

As there are a number of excellent textbooks about computer simulation modelling in general[87], we shall focus on the fundamental approaches to implementing dynamic knowledge-oriented models. Advantages and disadvantages are discussed from an evolutionary perspective.

- **Deterministic models** are defined by the absence of any stochastic variation. The big advantage of such models is that you have to run them only once for one paramter combination – if your system is not chaotic. The big problem for such models in evolutionary biology is the discrete nature of individuals. If your formula predicts that 1.56 offspring will be produced, this may be the correct average, but nature will either produce 0 or 1 or 2 ... individuals. While this is no problem in large populations, such effects may lead to completely different results when few or rare individuals play important roles. Then stochastic simulations are needed to support deterministic models[88].
- **Stochastic models** are defined by the inclusion of some random components. They have to be run several times for one parameter combination, to estimate the true mean, variance, etc. of some interesting property. While a stochastic model can be built by adding random deviations to a deterministic model, the full potential is only re-

87. Law & Kelton (1991) "Simulation modelling and analysis". 2nd., New York, McGraw-Hill.
Bossel (1994) "Modeling and Simulation", Wellesley, MA, A. K. Peters.
Mehl (1994) "Methoden verteilter Simulation", Braunschweig, Germany, Vieweg.
Gershenfeld (1999) "The nature of mathematical modeling", Cambridge University Press.
Taylor & Karlin (1998) "An introduction to stochastic modeling". 3rd., San Diego, Academic Press.
Page (1991) "Diskrete Simulationen -Eine Einführung mit Modula 2", Berlin, Springer-Verlag.
Siegert (1991) "Simulation zeitdiskreter Systeme", München, Oldenbourg Verlag.
Mitrani (1982) "Simulation techniques for discrete event systems", Cambridge University Press.

88. Wilson (1998) "Resolving discrepancies between deterministic population models and individual-based simulations", Am. Nat. 151:116-134.

alised when meaningful biological entities (like an individual) are chosen as basic units of the model.

- **Continuous models** have differential equations that describe the changes of the systems variables for *every* moment in time. A prominent example is Lotka-Volterra predator-prey models [89].
- **Discrete models** do not need explicit differential equations. As they simulate only certain discrete time points or events, any algorithm that computes the state of the model for the next point in time can drive it. This results in high degrees of freedom when implementing biological knowledge or data into such a system.

Two types of discrete-time models can be distinguished according to their basic approach to computing simulations:

How to simulate time

- **Event-based models** put all simulation events into an event queue. Simulation time jumps to the time of the most imminent event in the queue. After processing it, corresponding successive events are scheduled. Such an approach is well suited for systems where events occur only rarely. In simulations of evolution, however, individuals are key building blocks and their birth and death are very frequent events. Thus a population requires scheduling of large numbers of events in the same generation and simulation time will never be able to jump over time periods longer than a generation. Thus the overhead needed to schedule events will usually not be worthwile, as there are too many events per time unit in evolution. Only if few events occur per time unit should an event-based model be considered.
- **Time-based models** do not need an event queue and the memory it consumes. Instead they visit every important unit of the system at each moment of simulated time to compute its state in the next moment. The length of such a 'moment' is fixed throughout a simulation and can be considered as the time resolution of a model. It is a very fundamental quantity and simulation code is specifically adapted to its value. As simulations of evolution usually have a high event / individual ratio for a given moment of time, this approach will usually be the most simple for individual-based models.

While general simulation approaches can be further classified[90], it is sufficient to conclude that the most reliable, flexible, and simple way to simulate evolution is to do

89. Begon et al. (1996) "Ecology: Individuals, populations and communities". 3rd, Oxford, Blackwell.
90. See pp.8-12 in Bossel (1994) "Modeling and Simulation", Wellesley, MA, A. K. Peters.

- o stochastic (reliable, does not pretend artificial accuracy)
- o discrete (flexible, easy implementation of biological knowledge)
- o time-based (simple, as many events occur at each moment of time)

simulations, if no analytical solution for the problem can be found.

0.6 Types of evolutionary simulations

Over the last decades three distinct communities discovered that simulations of evolutionary processes in the computer contribute to scientific progress from their perspective: Evolutionary Computation, Artificial Life and Evolutionary Biology and Ecology. Although differences are great despite similar vocabulary, information is starting to flow between these fields[91]. Figure 77 depicts their relationships and how they might benefit each other.

0.6.1 Evolutionary computation

Evolutionary computation is a rapidly evolving field[92] in informatics that aims at using the biological idea of evolutionary adaptation by selection to seach for optimal solutions in complex technical systems from plane wings to protein folds[93]. It differs fundamentally from biological research on evolution, as such work focuses on finding good solutions for various problems that may even come from bioinformatics[94]. Thus parameters like mutation rate, crossover probability, population size and selection strength are chosen

91. Mitchell & Taylor (1999) "Evolutionary computation: An overview", Ann. Rev. Ecol. Syst. 30:593-616. - Foster (2001) "Evolutionary computation", Nat. Rev. Genet. 2:428-436. - Toquenaga & Wade (1996) "Sewall Wright meets Artificial Life: the origin and maintenance of evolutionar novelty", Trends Ecol. Evol. 11:478-482.
92. See also references in Foster (2001) "Evolutionary computation", Nat. Rev. Genet. 2:428-436. - Bäck (1996) "Evolutionary algorithms in theory and practice", New York, Oxford University Press. - Rechenberg (1994) "Evolutionsstrategie ´94", Stuttgart, Fridrich Frommann Verlag. - Holland (1975) "Adaptation in Natural and Artificial Systems. An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence", Ann Arbor, The University of Michigan Press. - Holland (1992) "Genetic Algorithms", Sci. Am. July:44-50. - Heistermann (1994) "Genetische Algorithmen. Theorie und Praxis evolutionärer Optimierung", Stuttgart, B.G.Teubner Verlagsgesellschaft. - Davis, (ed, 1991) "Handbook of genetic algorithms", New York, Van Nostrand Reinhold. -
93. Dandekar & Argos (1996) "Identifying the tertiary fold of small proteins with different topologies from sequence and secondary structure using the genetic algorithm and extended criteria specific for strand regions", J. Mol. Biol. 256:645-660. - Schneider et al. (1998) "Peptide design by artificial neural networks and computer-based evolutionary search", Proc. Natl. Acad. Sci. USA 95:12179-12184. - Rosin et al. (1999) "Coevolutionary analysis of resistance-evading peptidomimetic inhibitors of HIV-1 protease", Proc. Natl. Acad. Sci. U.S.A. 96:1369-1374.
94. Schulze-Kremer (1996) "Evolutionary computation", pp. 211-271 in: Schulze-Kremer (ed) Molecular bioinformatics: Algorithms and applications, Berlin, Walter de Gruyter.

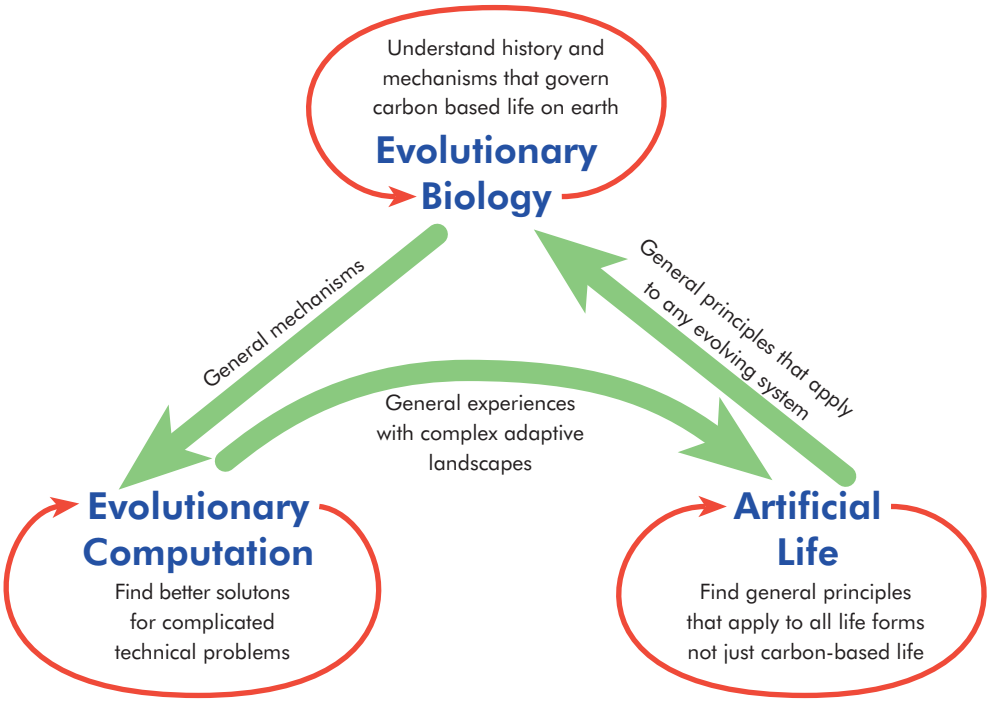


Figure 77 Relationships between fields that simulate evolution in the computer. Besides other journals, each field has a regular publication with the corresponding name.

to facilitate rapid evolution of optimal solutions. Data structures and algorithms are designed to archieve that goal[95] and often only the mere concept of evolution itself is biological.

Do not confuse populations from solution space with biology

Thus these systems do not resemble *anything* biological and are not built with the intention of learning more about biology. If natural populations had parameter combinations like those frequently used in evolutionary computation, they would often not be viable (eg. very high mutation rates in small populations). On the contrary, these technical systems allow something that contemporary biologists could only dream of: computation of fitness by a fitness function that translates genotypes (= representations of problem solutions) into measures of performance (= how good is that solution).

Can evolutionary computation make direct contributions to evolutionary biology? While evolving solutions for technical problems rarely advances

95. Michalewicz (1992) "Genetic Algorithms + Data Structures = Evolution Programs", Berlin, Springer.

evolutionary biology, a major contribution comes from the general experiences gained in exploring complex adaptive landscapes[96]. For example sometimes even a genetic algorithm can become stuck in local optima[97] and evolutionary strategies pioneer Ingo Rechenberg reports a slow-down of evolution with increasing complexity in most cases[98].

0.6.2 Artificial Life

The field '*Alife*'[99] originated from the desire to understand those general principles that apply to all self reproducing automata as defined by John von Neumann[100]. This includes pieces of self-reproducing software in the computer[101], all imaginable forms of life[102], and among others, carbon-based life. A prominent topic are cellular automata[103].

Simulations that target questions from evolutionary biology are usually a very special case of *Alife*, as *Alife* has a much broader perspective. On the contrary, some *Alife* simulations are too remote from models of carbon-based life to contribute to evolutionary biology. However, as some overlap exists, a link between these fields emerged[91]. Today, some spatially structured population models that are implemented as cellular automata contribute to biology[104] and the study of digital genomes gives clues to understanding carbon based genomes[105].

96. Wagner & Altenberg (1996) "Complex adaptations and the evolution of evolvability", Evolution 50:967-976. - Lenski et al. (1999) "Genome complexity, robustness and genetic interactions in digital organisms", Nature 400:661-664. - Wilke et al. (2001) "Evolution of digital organisms at high mutation rates leads to survival of the flattest", Nature 412:331-333.
97. Kemp et al. (1998) "Population partitioning in genetic algorithms", Electron. Lett. 34:1928-1929.
98. Rechenberg (1996) "Vorwort des Herausgebers", pp. 5-7 in: Koch-Schwessinger (ed) Wasserstoffproduktion durch Purpurbakterien, Stuttgart, frommann-holzboog.
99. Adami (1998) "Introduction to Artificial Life", New York, Springer. - Langton, (ed, 1989) "Artificial Life: The proceedings of an interdisciplinary workshop on the synthesis and simulation of living systems held September 1987 in Los Alamos, New Mexico", Redwood City, California, Addison-Wesley. - Prata (1993) "Künstliches Leben: Evolution auf dem PC erleben", München, te-wi Verlag. - See also http://alife.org
100. von Neumann & Burks (1966) "Theory of self-reproducing automata", Urbana, University of Illinois Press.
101. Adami et al. (2002) "Digital Life Laboratory: Software Avida" http://dllab.caltech.edu/avida Ray (2002) "Tierra Homepage" http://www.isd.atr.co.jp/~ray/tierra
102. Lipson & Pollack (2000) "Automatic design and manufacture of robotic lifeforms", Nature 406:974-978.
103. Adamatzky (1994) "Identification of cellular automata", Taylor&Francis.
104. Silvertown et al. (1992) "Cellular automaton models of interspecific competition for space - the effect of pattern on process", Journal of Ecology 80:527-534. - Molofsky (1994) "Population dynamics and pattern formation in theoretical populations", Ecology 75:30-39. - Ellison & Bedford (1995) "Response of a wetland vascular plant community to disturbance: A simulation study", Ecolog. Applic. 5:109-123. - Talia (1998) "Cellular automata thrive on parallel systems", Scientific Computing World October:21-22.

0.6.3 Simulations of biological evolution

Simulations in evolutionary biology and ecology aim at learning more about the biology of particular real systems. Ideally, they take a mechanistic understanding of the processes that govern the system studied, disregard unimportant processes, add biologically feasible parameters, and compute results that offer a simple explanation for some previously not understood phenomenon. A number of books describe how to build such models[106]. The following approaches can be used to do that:

Analytic

- **Top-Down analysis.** Starting from a bird's eye overview, formulae are derived that describe the changes of some overall properties of the system like the frequency of particular alleles. Such an analytical maths approach critically depends on correct identification of those processes that govern the dynamics - no easy task. This classical approach of population genetics often describes deterministic models or i-state distribution models with stochastic components. Today, simulations often check such analytical results.

Genes only

- **Coalescent.** Recent years have seen the rise of the coalescent approach[46]. It is based on the fact that all individuals in a population have one common ancestor, if you go far enough back in time. Coalescent simulations compute genealogies of alleles assuming a particular evolutionary model. Thus patterns observed in sequences can be used to estimate model parameters like ancient effective population size for humans[107]. Simulated time runs backwards from the present to that moment in the past when the last two alleles coalesced in the most recent common ancestor of the sequences investigated.

Individual-based

- **Bottom-Up synthesis.** Starting with the elementary processes that are believed to govern the dynamics of a system, one piece of simple,

105. Lenski et al. (1999) "Genome complexity, robustness and genetic interactions in digital organisms", Nature 400:661-664. - Yedid & Bell (2001) "Microevolution in an electronic microcosm", Am. Nat. 157:465-487. - Wilke et al. (2001) "Evolution of digital organisms at high mutation rates leads to survival of the flattest", Nature 412:331-333.

106. DeAngelis & Gross, (eds, 1992) "Individual-based models and approaches in ecology", New York, Chapman & Hall. - Wissel (1989) "Theoretische Ökologie: Eine Einführung", Berlin, Springer-Verlag. - Wilson (2000) "Simulating ecological and evolutionary systems in C", Cambridge University Press. - McGlade, (ed, 1999) "Advanced Ecological Theory", Oxford, Blackwell Science. - Gillman & Hails (1997) "An introduction to ecological modelling - putting practice into theory", Oxford, Blackwell Science. - Mode (1985) "Stochastic processes in demography and their computer implementation", Berlin, Springer-Verlag. . Ebert (1999) "Plant and animal populations - Methods in demography", San Diego, Academic Press.

107. Sherry et al. (1997) "Alu evolution in human populations: Using the coalescent to estimate effective population size", Genetics 147:1977-1982.

essential, elementary information after another is coded into the computer model until all essential processes have been covered. Then code is added to observe the properties of interest. Results of the simulations are then checked for feasibility, preferably with analytical maths. This approach allows for the greatest freedom in building realistic models that incorporate data as it is observed in nature. The individual is the elementary unit of these individual-based models and they are widely used in ecological and evolutionary research[108]. Although sometimes ridiculed as non-analytical, they are often the only approach to assessing certain biological questions and, when care is taken to perform high quality simulations, a lot can be learned by investigating these models.

The following bottlenecks in the simulation of evolutionary models apply to all three approaches in general, but to knowledge-oriented, individual-based models about biological evolution in particular.

0.7 How to remove bottlenecks in simulating evolutionary models

If modelling is a subjective art that leads to different descriptions of the same complex reality, then it is imperative that we compare different descriptions of nature to reach firm conclusions. This in turn demands that coding of the model itself be as easy as possible to speed up the process and that results be stored in such a format as to facilitate comparisons between output of different models. However, practical computing issues generate severe bottlenecks that make many simulation efforts fail[109].

These bottlenecks could be removed by a framework

- **Time to implement.** Before a meaningful number of simulations can be run, one needs to organise parameter space, overall control loops, good basic containers (for individuals in populations, etc.), script automation, data-file handling, controls for the progress of a simulation, and enough infrastructure to help analysis at the end.

108. Huston et al. (1988) "New computer models unify ecological theory", BioScience 38:682-691. - DeAngelis & Gross, (eds, 1992) "Individual-based models and approaches in ecology", New York, Chapman & Hall. - Judson (1994) "The rise of the individual-based model in ecology", Trends Ecol. Evol. 9:9-14. - McGlade, (ed, 1999) "Advanced Ecological Theory", Oxford, Blackwell Science. See pp.2-5 for background on the development of individual based models.
109. Railsback (2001) "Concepts from complex adaptive systems as a framework for individual-based modelling", Ecol. Model. 139:47-62.

- **Repeated optimisations of the code.** Every non-trivial program evolves for some time, before it reaches a stable version. Fast initial results analyses need to be carried out to check for errors in the code and further refine the model.
- **Computing time.** As large populations are often those that are most interesting, biologists try to simulate the largest populations that can still be computed in reasonable time. Thus without a professional solution, one can either easily occupy a computer for months or distribute the tasks to many computers manually or reset run-priorities all the time to have only currently interesting simulations running. If all this is done manually without a good system, this can take a considerable time that would be better used in biology. Even when individual simulations are not too complex, the fact that many individual-based models have up to 50 or more input parameters generates enormous computing needs, if rigorous model analyses of all meaningful parameter combinations are needed. Currently, biologists often stop when simulations have used enough computing time, rather than stopping when enough parameter combinations have been analysed[110]. It is not by chance that population biology poses enormous computing challenges[111].
- **Time to analyse.** If all simulations have been completed, results files have to be collected and stored in such a way that their widely differing types of data can be found again. Then numerous plots have to be made only to discover that some more simulations are needed. Finally, if results from different models are to be compared, then the struggle with incompatible data formats and other related problems begins.

A work-flow is needed

The severity of these bottlenecks can be greatly reduced by a framework that facilitates implementation of biological models in such an integrated way that distributed computing and final analysis are part of a defined work-flow. Work-flow solutions have been under development for some time[112] and it should be possible to generate such a system for model-building that

110. Bart (1995) "Acceptance criteria for using individual-based models to make management decisions", Ecolog. Applic. 5:411-420. - Eg. Mooij & Boersma (1996) "An object-oriented simulation framework for individual-based simulations (OSIRIS): Daphnia population dynamics as an example", Ecol. Model. 93:139-153.
111. Levin et al. (1997) "Mathematical and computational challenges in population biology and ecosystems science", Science 275:334-343.
112. Peccoud (1995) "Automating molecular biology: A question of communication", Bio/Technology 13:741-745. - Trammell (1996) "Work-flow without fear", Byte April:55-60.

can be used for more than a few special applications. It will consist of a complex problem-solving environment for generation of individual-based models of evolution and of the entire infrastructure needed to distribute simulation tasks to different computers, preferably over the Internet[113]. In the search for such a solution, one encounters different types of software.

Go to another discipline

- Software frameworks from evolutionary computation or artificial life are well suited for their corresponding problem domains, but do not have much of the infrastructure needed for simulations in evolutionary biology.

Use teaching software

- Quite a few people have written easy-to-use simulation programs for teaching evolution. Examples include Selection, PopBio, Populus, EcoBeaker and EvoTutor among many others[114]. While being excellent for simple questions, they lack the flexibility to answer complex research problems.

Code yourself in maths packages

- There are a number of large commercial maths-packages that can be used for simulations of nearly everything. Examples include MathLab, Mathematica, Maple, ModelMaker, Berkeley Madonna, SAAM II, STELLA/ithink, PowerSim, SysQuake, Mesquite and others[115]. Usually lacking biological infrastructure, some biologists use these to implement evolutionary models with their built-in programming capabilities. They provide reliable numeric code, allow easy plotting of results and sometimes support distribution of computations among CPUs where a licensed copy of the corresponding expensive package is installed.

Code yourself from scratch

- Any general purpose programming language can be used to implement simulation models. Biologists have been using C/C++, Java,

113. Internet distributed computing is a powerful way to solve the computing time bottleneck. For sample projects of this type see Pearson (2001) "Internet-based Distributed Computing Projects", http://www.aspenleaf.com/distributed/.

114. Alstad (2002) "POPULUS" http://www.cbs.umn.edu/populus/. - Meir (1996) "EcoBeaker 1.0 - An ecological simulation program. EcoBeaker Laboratory Guide and the EcoBeaker Program Manual", Sunderland, MA, Sinauer Associates, Inc. - Meir (2002) "Welcome to Ecobeaker 2.0 - Ecology teaching software" http://www.ecobeaker.com/. - Lemmon (2002) "EvoTutor - Learning through interactive simulation" http://www.evotutor.org/

115. Maple (2002) "Maple Homepage" http://www.maplesoft.com/ - Wolfram Research (2002) "Mathematica Homepage" http://www.wolfram.com/ - Mathworks (2002) "MathLab Homepage" http://www.mathworks.com - Cognitus (2002) "STELLA / ithink Homepage" http://www.cognitus.co.uk/bus-solution.html - Modelkinetix (2002) "ModelMaker Homepage" http://www.modelkinetix.com/modelmaker - Berkeley Madonna (2002) "Berkeley Madonna Homepage" http://www.berkeleymadonna.com - SAAM (2002) "SAAM II" http://www.saam.com/ - Mesquite Software Inc. (2001) "CSIM18: The simulation engine preferred by systems designers, programmers and educators" http://www.mesquite.com - PowerSim (2002) "PowerSim - Custom Simulator Solutions" http://powersim.com/ -

Fortran or even Basic due to speed, cross-platform availability, tradition, simplicity or price. Libraries with evolutionary or ecological background exist (eg. GSL, SIMEX, OSIRIS, *Inter-site* and others[116]). However, as they are either too specific, no longer under development or not available, most people start from scratch and build their own collection of code, often re-using simulation code from other disciplines. Books have even been written to teach biologists how to do that[117].

Use super-computers

- If computing demands rise considerably, one would want to use a local super-computing centre. There usually some tools are available to distribute simulations among several processors to speed up execution time. However, programming these machines is not a simple matter and their power has to be shared with many other users.

Code yourself with global computing

- Recently, globally distributed computing over the Internet has emerged as a possible alternative for massive parallel applications like most evolutionary simulation problems (see Chapter 2). There are a number of programming frameworks that facilitate this (see[118] for examples). However, none of them facilitates the integrated modelling work flow needed for rapid model building in evolutionary biology, and running a global computing project can be complicated.

Ideally, an integrated framework would combine the strength of all approaches above. However, as such a framework is currently not available, evolutionary biologists do one of two things. They either live with the limitations of some other framework or they start to develop their own highly specialised framework that in turn is rarely re-used by others.

This realisation was the defining moment that led to the development of the framework behind evolution@home which is the first global computing system for evolutionary biology[119]. Although still in its infancy, a very simple first model is operational. The vision behind it is to fill exactly the gap described here.

116. Conery & Lynch (2002) "Genetic Simulation Library" http://www.csi.uoregon.edu/projects/genetics/GSL/ - NMSR (2002) "About SIMEX" http://www.nmsr.labmed.umn.edu/nmsr/simex/ - Mooij & Boersma (1996) "An object-oriented simulation framework for individual-based simulations (OSIRIS): Daphnia population dynamics as an example", Ecol. Model. 93:139-153. - Sequeira et al. (1997) "Implementing generic, object-oriented models in biology", Ecol. Model. 94:17-31. - Gathmann & Williams (1998) "Inter-site: a new tool for the simulation of spatially realistic population dynamics", Ecol. Model. 113:125-139.
117. Wilson (2000) "Simulating ecological and evolutionary systems in C", Cambridge University Press. Ebert (1999) "Plant and animal populations - Methods in demography", San Diego, Academic Press.
118. Pearson (2001) "Internet ... Computing Projects", http://www.aspenleaf.com/distributed/.
119. Loewe (2002) "The evolution@home website" http://www.evolutionary-research.net

0.8 Blessing and curse of interdisciplinary research

Building a software framework for computer models of evolution is a highly interdisciplinary adventure, just like evolutionary bioinformatics. While it is important to understand current research in evolution to build meaningful models, it is also important to follow developments in computer science to be able to make good implementation choices.

Despite the praise that interdisciplinary approaches often get[120], they have a down-side too. In the time one spends programming, one can not do biological research. When one does the latter, one will lack time to improve code or learn some new efficient method from the IT-world. Whenever one tries to learn something really new, one will look foolish, as one has to ask some very elementary questions. As the goal of simulations of evolution is to understand evolution better, much of the informatics work will be 'just' developing applications in a problem-solving environment. All the time-consuming technical implementation details needed to get such a system off the ground are uninteresting to most scientists anyway.

However, once such a system is working, one will be able to investigate evolutionary models with unprecedented accuracy and speed. All the data will be in a format that allows easy comparison of models. One no longer needs to spend time reinventing the simulation basics wheel over and over again, but can concentrate on evolutionary biology at a new level. As evolutionary problems are among the most complicated in the world, such investigations are likely to lead to new approaches for informatics, too.

120.Garwin (1995) "In praise of interdisciplinarity", Nature 376:547.

Mutation rates paradox table for mtDNA

Table 21 Instances where a higher mutation rate in mitochondrial DNA may lead to different conclusions about archaeological issues.

Topic Reference	Phylogenetic rate[a] (%/Myr)			Date from phylogenetic rate[b]			Conclusion[c]	Alternative conclusion	Independent date[b]	Inferred mutation rate (%/Myr)[d]		
	min	mid	max	min	mid	max				min	mid	max
Dog domestication[e] (MRCA[f])	3.7	3.8[g]	3.9	130	133[h]	137	Multiple ancient origins[i]	Single domestication (clade 1)[i]	14[k]	35	36[l]	37
Sheep domestication[m] (MRCA, RFLP)	1	1.5[n]	2	375	563[o]	750	Multiple domestications	Single domestications	11[p]	34	77[q]	136
Cattle domestication[r] (MRCA, MMD[ba])	3.8[s]	6[t]	7.5[u]	524	655[v]	1034	Multiple domestications	Single domestication[x]	10[w]	87	166[x]	197
Goat domestication[y] (MRCA, cytB)	3.8	4.6[z]	5.4	201	242[aa]	282	Multiple domestications	Single domestications	10[ab]	76	111[ac]	152
Goat expansion[y] (MMD[ba])		leads to conflicts[ad]					(before domestication)	In nomadic "kingdoms"[ad]	4.2[ad]	225	267[ad]	303
Horse domestication[ae] (MRCA)	1.8[af]	2.7[ag]	4.8[ah]	320[ai]	481[aj]	630	Multiple domestications[ak]	Single domestication[al]	6[am]	96	216[an]	500[ao]
Pig domestication[ap] (MRCA, cytB)	(1)	1[aq]	(1)	500	665[ar]	785	Multiple domestications	Single domestications	11.5[as]	43	48[at]	68
Post ice age recolonisation of voles[au]	(2)	2[av]	(2)	150	200[aw]	250	Many founder lines	One founder line	11.2[ax]	27	36[ay]	45

Printed at 16 Apr 2003

Laurence Loewe (2002) Diss, TU-München

Table 21 Instances where a higher mutation rate in mitochondrial DNA may lead to different conclusions about archaeological issues.

Topic Reference	Phylogenetic rate[a] (%/Myr)			Date from phylogenetic rate[b]			Conclusion[c]	Alternative conclusion	Independent date[b]	Inferred mutation rate (%/Myr)[d]		
	min	mid	max	min	mid	max				min	mid	max
Noctule bat recolonization of Europe[az] (MMD[ba])	(3.2)	10[bb]	(12.6)	20 21 5.2	39[bc] 43 19	58 71 27	No bottleneck during ice age[az]	Ice age caused bottleneck: Bats confined to small refugia	10[bd]	20 21 5.2	39[be] 43 19	58 71 27
Gene flow in Orang Utans[bf]	1	1.5[n]	2	750	1000	1500	Effective gene flow barriers	Gene flow during ice-age	15	50	100	200
Peopling the[bg] Americas (MMD)	10.3	12.7	15	30	37	43	Early migration	Late migration	12[bh]	28	39	54
Ngobe Amerind bottleneck[bi]	7.5	11	15	6.8	11.4	14	Ethnogenesis based	Conquest based	0.5	102	251	420
Human demographic expansion[bj] (MMD[ba])	-	16.5	-	-	110[bk] 71 57 42	-	Pleistocene expansions	Neolithic expansions, advent of agriculture	10	69	117 94	181

a. This substitution rate (= divergence rate /2) was inferred from paleontological divergence ages with an outgroup. If no such data were available, values were assumed to be similar to the molecular clock of a related species.
b. Dates are given in Kiloyears b.p. (thousands of years, Kyrs)
c. Author's conclusion based on the phylogenetic rate.
d. This substitution rate (= divergence rate /2) was inferred from archeological date and observed sequence diversity using the alternative conclusion.
e. Vilà et al. (1997) "Multiple and ancient origins of the domestic dog", Science 276:1687-1689.
f. MRCA = Most Recent Common Ancestor. Indicates, that the dates computed are coalescence dates.
g. Paleontological calibration based on 1 Myrs divergence time and 7.5% sequence divergence between wolf and coyote.
h. Divergence between the most different genotypes in the most diverse clade of dogs is no more than 1% (1% / 2 / 3.75%/Myr = 0.133Myr)
i. Vilà et al. (1997) speculate that man's initially domesticated dogs still had the morphology of wolves until nomadic hunter-gatherer societies changed to more sedentary agricultural population centers. Resulting new selective regimes may have caused the difference to wolves seen today.
j. If the dog indeed was domesticated several times, then the other mixed dog-wolf clades of Vilà et al. (2001) can be interpreted as evidence for that.
k. or at least 11-12 Kyrs, see Nobis (1979) "Der älteste Haushund lebte vor 14 000 Jahren", Umschau 79:610. - Morey (1994) "The early evolution of the domestic dog", American Scientist 82:336-347. - Mason, (ed, 1984) "Evolution of domesticated animals", London, Longman Group Limited.
l. 3.8% * 133 Kyrs / 14 Kyrs = 36%, applies to D-loop only, primary data were sequences.
m. Hiendleder et al. (1998) "Analysis of mitochondrial DNA indicates that domestic sheep are derived from two different ancestral maternal sources: no evidence for contributions from urial and argali sheep", J Hered 89:113-120.
n. Divergence rate 2-4% from Brown et al. (1979) "Rapid evolution of animal mitochondrial DNA", Proc. Natl. Acad. Sci. U.S.A. 76:1967-1971. This general animal rate applies to the whole molecule, not only to the control region.

o. Min and max values by Hiendleder et al. (1998) from mean sequence divergence of 0.716% (max 1.032%) in mitochondrial RFLP analysis of 243 sheep.

p. Ryder (1984) "Sheep", pp. 63-85 in: Mason (ed) Evolution of domesticated animals, London, Longman Group Limited. - Zeder (1997) "Sheep and goats", pp. 23-25 in: Meyers (ed) The Oxford Encyclopedia of Archaeology in the Near East, New York, Oxford University Press.

q. The usual extrapolation is here applied to RFLP data from the whole mitochondrial DNA: Min:1%*375Kyr/11Kyr = 34% - Mid: 1.5%*563Kyr/11Kyr = 77% - Max: 2%*750Kyr/11Kyr = 136%. Hiendleder et al. (1998) also observed 4.43% divergence in the control region between one European and one Asian sheep. Assuming domestication before 11 Kyrs b.p. leads to a substitution rate of 201 % (4.42% / 2 / 11Kyr).

r. Loftus et al. (1994) "Evidence for two independent domestications of cattle", Proc. Natl. Acad. Sci. U.S.A. 91:2757-2761. - Bradley et al. (1996) "Mitochondrial diversity and the origins of African and European cattle", Proc. Natl. Acad. Sci. U.S.A. 93:5131-5135. - Bailey et al. (1996) "Ancient DNA suggests a recent expansion of European cattle from a diverse wild progenitor species", Proc R Soc Lond B Biol Sci 263:1467-1473. - Troy et al. (2001) "Genetic evidence for Near-Eastern origins of European cattle", Nature 410:1088-1091.

s. To derive a minimum, extend the sequence considered beyond the 240 bp used by Troy et al. (2001). Then more transversions are observed (8 more versus 18 additional transitions in positions 15853-16338 of the reference sequence). This totals to 55 mutations in 725 bp = 7.6% divergence. Rest see[†].

t. Based on the following approach Troy et al. (2001) derived a substitution rate of 38%/Myr for the 240 bp of the D-loop they studied: The transition-transversion ratio for their sequence data is 61/1. Palaeontological evidence indicates 1 Myr for the *Bison-Bos* divergence (see Loftus et al. 1994). As they observe only 3 transversions between *Bison* and *Bos* in the 240 bp they studied, they estimate that these 3 transversions indicate that 183 transitions have occurred between *Bison* and *Bos* (183/240 = 76% divergence / Myr or one substitution in 11 Kyrs). However, there are some problems with their approach.

   If this rate were true, then one would expect that *Bison-Bos* sequences would only have a similarity on the order of (100%-2*38%) = 24% today, unless *very* extreme rate heterogeneity would make most mutations hit only a few sites. Thus *Bison-Bos* sequences should be full of transitions. However, an alignment of 8 *Bison bison* D-loop haplotypes (Genbank accession numbers AF083357-AF083364) with the most probable *Bos taurus* root sequence of Troy et al. (2001) shows 4 transversions, 4 indels (3x 1bp, 1x 7bp) and only 21 transitions in the 240 bp region studied. There were also sequence pairs with 3 transversions (e.g. *Bison* AF083364 - *Bos taurus* AF336737). This results in 88% similarity between *Bison-Bos taurus* and is remarkably higher than the 24% expected from the rate of Troy et al. (2001). Rate heterogeneity would have to be extremely high to achieve this only by hitting the same few transitions again and again. If a standard phylogenetic rate is to be calculated for the system, then it will probably look more like the 12% divergence / Myr than like the high estimates from Troy et al. (2001). Loftus et al. (1994) applied the same reasoning. They use the fact that Indian and Afro-european cattle mtDNA show 74% of the divergence between *Bos* and *Bison* to consider divergence dates of "possibly as much as 1 million years ago". However, as such a conclusion seemed unlikely, Loftus et al. (1994) preferred to use the human 12%/Myr estimate of Stoneking et al. (1992) "New approaches to dating suggest a recent age for human mtDNA ancestor", Phil. Trans. R. Soc. Lond. B 337:167-175. Later Bradley et al. (1996) and Bailey et al. (1996) found the transversion/transition approach (presented above) to give mutation rates that are more in line with the general picture. However, as they do not explain the staggering lack of transitions between *Bison* and *Bos*, here the more straightforward approach is taken together with the majority of other researchers that estimate phylogenetic rates: 12% divergence between *Bos* and *Bison* in 1 Myr (palaeontology) when considering Troy's 240bp hypervariable region.

u. If it is assumed that the 6 polymorphic positions in *Bison* (all transitions) originally held non-*Bos* base pairs, then maximal divergence is 15% (Troys 240bp).

v. According to Loftus et al. (1994) average pairwise distance between Afro-European and Indian cattle is 7.86% in 375bp of the more variable region of the D-loop. Applying the phylogenetic rates estimated above yields a mid value 655 Kyr (=7.86% / 2 / 6%/Myr, min and max correspondingly).

w. See Loftus et al. (1994) for a short review of the traditional picture. Other references: Epstein (1971) "The origin of the domestic animals of africa", New York, Africana Publishing Corporation. - Clutton-Brock (1989) "[Cattle domestication]", pp. 200-206 in: Clutton-Brock (ed) The walking Larder: Patterns of domestication, pastoralism and predation, London, Unwin Hyman Ltd. - Epstein & Mason (1984) "Cattle", pp. 6-27 in: Mason (ed) Evolution of domesticated animals, London, Longman Group Limited. - Hesse (1997) "Cattle and oxen", pp. 442-443 in: Meyers (ed) The Oxford Encyclopedia of Archaeology in the Near East, New York, Oxford University Press. - Hesse (1997) "Animal husbandry", pp. 140-143 in: Meyers (ed) The Oxford Encyclopedia of Archaeology in the Near East, New York, Oxford University Press.

x. A former calculation (Loewe & Scherer (1997) "Mitochondrial Eve: the plot thickens", Trends Ecol. Evol. 12:422-423) subscribed to Loftus' et al. (1994) extrapolation of a human 12% phylogenetic substitution rate to their 915bp of cattle D-loop sequences. This led to an estimate of 23 substitutions in the last 210 Kyr (23 subst = 210 Kyr * 12%/Myr * 915bp). If these 23 substitutions occurred in 10Kyrs, a rate of 250%/Myr would result (23 bp / 915bp / 10 Kyr).
If mutation rates are directly inferred from phylogenetic rate and age estimates above, the following values result: Min: 3.8%*524Kyr/10Kyr = 199% - Mid: 6%*655Kyr/10Kyr = 393% - Max: 7.5%*1034Kyr/10Kyr = 776%. These appear to be too high. Rescaling expansion times for cattle computed by Troy et al. (2001) with a substitution rate of 38%/Myr with such a high inferred value (393%) yields not a particular convincing picture of cattle expansions:
Expansion of European cattle 986 years ago (416-1560) from 10200 years (T3, CI=4300-16100 years)
Expansion of African cattle 1010 years ago (425-1600) from 10450 years (T1, CI=4400-16500 years)
Expansion of Near-Eastern cattle 1270 years ago (532-2000) from 13100 years (T2, CI=5500-20700 years)
Expansion of all taurine cattle 1560 years ago (541-2040) from 16100 years (T, 95%CI=5600-21100 years). If these were the only alternatives, Troy et al. (2001) appear to propose the better scenario. However, a deeper look into archaeology proposes the following comprehensive picture:
The neolithic age, when cattle were domesticated about 10000 years ago, is known to have been quite wet. Archaeological remains show that people were feeding very much on pigs (Hesse (1997) "Pigs", pp. 347-348 in: Meyers (ed) The Oxford Encyclopedia of Archaeology in the Near East, New York, Oxford University Press. - Marcus (2000) "The view from Nebo: How archaeology is rewriting the bible and reshaping the Middle East", Boston, Black Bay Books., pp. 22-26). Pigs are known to require enough water and are difficult to herd. Thus, they need sedentary lifestyles and a rather wet climate (see Marcus (2000), ibid.). All this started to change about 5000 years ago, when climate became dryer. As Marcus (2000), ibid. reports, archaeological remains of that time show that people changed their food preferences away from pigs towards sheep, goat and cattle, leaving pigs only as cheap food for the poorer working class (see comparisons of bones from rich and poor people's garbage in urban sites of Mesopotamia). Thus with the advent of the rather dry middle bronze age (2000-1500 BCE.) pigs were declining and sheep, goats and cattle were increasing, not only as a source of food, but also as a source of labour (field ploughing agriculture is reported first in the 4th millennium BCE., Hesse, 1997, "Cattle and oxen", ibid.). The emergence of complex societies led to powerful nomadic kingdoms that controlled food resources for cities (Hesse, 1997 "Animal husbandry", ibid.). Thus people in the expanding cities had others herd livestock for them. As a consequence, growth of cities enlarged nomadic herds, and nomads of that time could be very influential depending on the size of their herds. The peak of this development was reached about 3000 years ago, where hardly any pork was eaten in the Near East (Marcus (2000), ibid.). This appears to be an archaeologically feasible date for a cattle expansion. If 4000 years b.p. is the early limit for this cattle expansion, 3000 is its most likely peak and 2400 years b.p. is its probable end, then expansion times estimated by Troy et al. (2001) can be used to infer substitution rates in Near-Eastern cattle:
Min: 38% * 5500 / 2400 = 87%, Mid: 38% * 13100 / 3000 = 166%, Max: 38% * 20700 / 4000 = 197%.
These values fit the range of observed values in humans. However, they would lead to an estimate of taurine-Indian cattle coalescence time of nearly 24 Kyr (=7.86% / 2 / 166%/Myr). A single domestication of cattle is nevertheless conceivable, if the history of Indian cattle is considered (for refs. see[w]):
After domestication (from *Bos primigenius namadicus*) in the Near East cattle were exported. One group was brought to the eastern fringes of the great salt desert of Iran prior to further eastward migration. There the arid-adapted physiology and hump developed that now characterise *Bos indicus* (see Epstein & Mason 1984, ibid.). As adaptation under such a high selective pressure depends presumably on relatively rare adaptive mutations, it is likely that all non-adapted individuals were lost and that individuals with a higher mutation rate would adapt first. This would (i) lead to a bottleneck after domestication and (ii) suggests a historically higher mutation rate in Indian cattle either due to a mutator phenotype or due to a reaction to environmental stress. Adaptive mutations are known to cause hitch-hiking events that reduce variability in populations. Thus, the Indian branch would be longer than other cattle lines, as only those that were in that small, fast evolving group would pass on the characteristic genetic signature known to belong to *Bos indicus* today. Combining both branches would lead to an older coalescence date. Assuming 10 Kyrs as the true domestication date, an only 4 fold higher mutation rate in Indian cattle (=(24Kyr+(24Kyr-10Kyr))/10Kyr) would lead to the pattern observed today. If this was not only a historic incident, the higher mutation rate should still be observable in Indian cattle today.

y. Luikart et al. (2001) "Multiple maternal origins and weak phylogeographic structure in domestic goats", Proc Natl Acad Sci U S A 98:5927-5932.

z. This was estimated from sequences of the entire mtDNA cytochrom b gene and palaeontological evidence that suggests 5-7 Myrs as divergence time between sheep and goats. Divergence between sheep and goats at third codon positions was 53.6% giving max 5.4% (53.6%/2/5Myr) and min3.8%/Myr.

aa. Using the 380 third codon positions divergence in mtDNA cytochrom b of 6 goats with 2 sheep as an outgroup. See Luikart et al. (2001).

ab. Luikart et al. (2001) - Zeder & Hesse (2000) "The initial domestication of goats (Capra hircus) in the Zagros mountains 10,000 years ago", Science 287:2254-2257. - Marean (2000) "Anthropology. Age, sex, and old goats", Science 287:2174-2175. - Pringle (1998) "Archaeology: The slow birth of agriculture", Science 282:1446-1450. - Mason (1984) "Goat", pp. 85-99 in: Mason (ed) Evolution of domesticated animals, London, Longman Group. - Zeder (1997) "Sheep and goats", pp. 23-25 in: Meyers (ed) The Oxford Encyclopedia of Archaeology in the Near East, Oxford University Press.

ac. Rates inferred for cytochrome B third codon positions. Min: 76.4% (= 3.8%*201Kyr/10Kyr) - Mid: 111.3% (=4.6%*242Kyr/10Kyr) - Max: 152.3% (=5.4%*282Kyr/10Kyr)

ad. Luikart et al. (2001) do not employ any phylogenetic rate here. This would lead to very bizarre conclusions, as the major primary expansion of domestic goats is 10.8 units mutational time into the past (for the 481 bp of the hypervariable D-loop analysed). Assuming 6% substitution rate as in cattle would date the major goat expansion to 187 Kyrs (10.8bp/2/481bp/6%/Myr). The secondary expansions at 6.6 and 2.3 units mutational time (114 Kyrs and 40 Kyrs, respectively) would still be well before initial domestication. If on the contrary the primary expansion is used to compute a substitution rate, then 10.8 mutations / 2 / 481bp must have occurred in 10 Kyrs leading to 112%/Myr, a number that fits nicely with the pedigree rates.
This, however, would not fit well with the extrapolation from the cytochrom b gene, as the control region usually has an elevated substitution rate. Further more, the date of the most recent common ancestor cannot be the date of the peak of the expansion at the same time. Expansion dates should always be younger. But when could the goat expansion have happened? If a similar reasoning as that for cattle is applied to goats, then expansion dates that overlap emergence of nomadic "kingdoms" in the Bronze Age can be envisioned (Hesse (1997) "Animal husbandry", ibid.). This seems feasible, as goats (after sheep) became the dominant livestock in the Near East (Hesse , 1997, "Cattle and oxen", ibid.) after the decline of pigs due to climatic and other reasons (Marcus ,2000, ibid.). Goats also played a major role in ancient Egypt and from 5500 b.p. up to the start of the Hyskos period at 3730 b.p. one type of goat was substituted by another (See Mason 1984, ibid.) If then 5000, 4200 and 3700 years b.p. are taken as the timeframe of goat expansion, the following mutation rates can be inferred:
Min: 225%/Myr (10.8bp/2/481bp/5.0Kyrs) - Mid: 267%/Myr (10.8bp/2/481bp/4.2Kyrs) - Max: 303%/Myr (10.8bp/2/481bp/3.7Kyrs)

ae. Vilà et al. (2001) "Widespread origins of domestic horse lineages", Science 291:474-477.

af. Minimal divergence (horse-donkey) = 14.3% / 3.9 Myrs divergence time according to molecular data = 3.7%/Myr / 2 = 1.8%/Myr.

ag. Vilà et al. (2001) use only mean divergence (horse-donkey = 16.1%) against 2 Myr or 3.9 Myrs (4.1% - 8.1% *divergence*);16.1/2.95Myr/2 = 2.7%.

ah. Maximal divergence (horse-donkey) = 19.1% / 2 Myrs divergence time of horses and stenoid equids from the fossil record = 9.55% /2 = 4.8%

ai. The minimum given by Vilà et al. (2001) is used (320 Kyrs). The most closely related sequences among the horses studied are separated for at least 0.2%/ 2/4.8%/Myr = 20Kyrs. However, this is no coalescence of all horses.

aj. Mean divergence between horse sequences was 2.6% (range 0.2 - 5%) after correcting for multiple hits and ignoring indels. 616 bp of the control region were analysed. 2.6%/2/2.7%/Myr=481Kyrs; Max given by Vilà et al. (2001 is used (other possiblitiy: 5%/2/1.8%/Myr=1388Kyrs).

ak. and incorporation of numerous wild matrilines according to Vilà et al. (2001) ibid.

al. Wapnish & Hesse (1997) "Equids", pp. 255-256 in: Meyers (ed) The Oxford Encyclopedia of Archaeology in the Near East, New York, Oxford University Press. - Mason, (ed, 1984) "Evolution of domesticated animals", London, Longman Group Limited.

am. Suggested by archaeological record, Clutten-Brock (1999) "A natural history of domesticated mammals". 2nd, Cambridge University Press.

an. 2.7% * 481 Kyrs / 6 Kyrs = 216%, applies to D-loop only; Min: 1.8% * 320 Kyrs / 6 Kyrs = 96%;

ao. Authors value used: 4.8% * 630 Kyr/6Kyr = 500%/Myr(Other max: 4.8% * 1388 Kyrs / 6 Kyrs=1110%, probably an outlyer).

ap. Giuffra et al. (2000) "The origin of the domestic pig: independent domestication and subsequent introgression", Genetics 154:1785-1791.

aq. Giuffra et al. (2000) use general rate of Brown et al. (1979) "Rapid evolution of animal mitochondrial DNA", Proc. Natl. Acad. Sci. U.S.A. 76:1967-1971.

ar. Sequences of cytochrom b from the most diverged clades of European and Asian pig have 1.45 ±0.12% differences. Giuffra et al. (2000) computed a lower estimate of 1% that attempts to take into account genetic diversity in the ancestral population. This yields a lower limit of 500Kyrs.
Mid : 1.33%/2/1%/Myr = 665 Kyrs; Max: 1.57%/2/1%/Myr = 785 Kyrs.

as. Pringle (1998) "Archaeology: The slow birth of agriculture", Science 282:1446-1450. However, archaeology has not yet decided, whether pigs were domesticated once or twice. Here a single domestication is assumed to compare the data. For further references see Hesse (1997) "Pigs", pp. 347-348 in: Meyers (ed) The Oxford Encyclopedia of Archaeology in the Near East, New York, Oxford University Press. - Mason, (ed, 1984) "Evolution of domesticated animals", London, Longman Group Limited.

at. Inferring substitution rates for the cytochrome c gene: Min: 1%*500Kyr / 11.5Kyr = 43%/Myr - Mid: 58%/Myr - Max: 68%/Myr.

au. Jaarola & Tegelstrom (1996) "Mitochondrial DNA variation in the field vole (Microtus agrestis): Regional population structure and colonization history", Evolution 50:2073-2085.

av. As no clock exists, the substitution rate of 2% (=4%/Myr divergence) in total mtDNA employed by Jaarola & Tegelstrom (1996) represents the "upper limit of the 'conventional' mammalian mtDNA clock and a lower limit for the recent rate estimates in *Mus (refs)*". Regarding the latter also see Martin & Palumbi (1993) "Body size, metabolic rate, generation time, and the molecular clock", Proc Natl Acad Sci U S A 90:4087-4091.

aw. Sequence divergence in total mtDNA RFLPs was 0.6% to 1%. Mid: 0.8%/2%/Myr/2 = 200Kyr

ax. At this time a land bridge was established between southern Sweden and Denmark. This is thought to have started large-scale colonisation by land mammals after glaciation had driven them from that area. See Jaarola & Tegelstrom (1996).

ay. 2% * 200 Kyrs / 11.2 Kyrs = 36%, applies to total mtDNA, primary data were RFLPs.

az. Petit et al. (1999) "No evidence of bottleneck in the postglacial recolonisation of Europe by the noctule bat (*Nyctalus noctula*)", Evolution 53:1247-1258.

ba. MMD = MisMatch Distribution. Indicates that the dates computed are demographic expansion dates estimated by the general approach of Harpending et al. (1993) "The genetic structure of ancient human populations", Current Antrophology 34:483-496. - Harpending et al. (1998) "Genetic traces of ancient demography", Proc. Natl. Acad. Sci. U.S.A. 95:1961-1967. - Schneider & Excoffier (1999) "Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA", Genetics 152:1079-1089.

bb. Petit et al. (1999) estimate, that the HVII region evolves 12.6 times faster than the ND1 gene for which they use the divergence rate of 0.5 - 2 % / Myr estimated by Lopez et al. (1997) "Rates of nuclear and cytoplasmic mitochondrial DNA sequence divergence in mammals", Mol. Biol. Evol. 14:277-286. Thus, they conclude that the divergence rate in the HVII region of the D-loop in noctule bats is 6.3-25.5 %/Myr. However, they continue to use only 20% (=10%/Myr subst. rate).

bc. Petit et al. (1999) give expansion dates for 3 groups. From top down "Ggh", "Grw", "Ros". Min+Max are 95% CI. Results from other groups were less clear.

bd. 10 000 C^{14} years ago forests of Europe recovered their previous range. Before that *Nyctalus noctula* was probably confined to forest refugia, as fossil remains are usually not found in Pleistocene cave deposits and all species of the genus *Nyctalus* are associated with forests. See refs in Petit et al. (1999).

be. All inferences as this: 10%/Myr phylogenetic rate * 39 Kyrs phylogenetic age estimate / 10 Kyrs inferred age estimate = 39% inferred rate/Myr.

bf. Ryder & Chemnick (1993) "Chromosomal and mitochondrial DNA variation in Orang Utans", J. Hered. 84:405-409. They used RFLPs.

bg. Bonatto & Salzano (1997) "A single and early migration for the peopling of the Americas supported by mitochondrial DNA sequence data", Proc. Natl. Acad. Sci. U.S.A. 94:1866-1871.

bh. Cavalli Sforza et al. (1994) "The history and geography of human genes", Princeton, Princeton University Press.

bi. Kolman et al. (1995) "Reduced mtDNA diversity in the Ngobe Amerinds of Panama", Genetics 140:275-283. The inferences are made from mismatch distributions of sequences from the control region of mtDNA.

bj. Excoffier & Schneider (1999) "Why hunter-gatherer populations do not show signs of pleistocene demographic expansions", Proc Natl Acad Sci U S A 96:10597-10602. They used published HV1 sequences from the control region of mtDNA.

bk. These averages have been estimated for Turkana (East Africa), Africa with Asia, America and Europe with Middle East (from top down).

Posters from International Conferences

Page A-39

Loewe & Scherer (1997)
"On the speed of genomic decay".
The Fifth Annual Meeting of the
Society for Molecular Biology and Evolution,
1-4 June 1997, Garmisch-Partenkirchen, Germany.

Page A-40

Loewe & Scherer (1998)
"Muller's ratchet in human mitochondrial DNA".
The Sixth Annual International Meeting of the
Society for Molecular Biology and Evolution,
17-20 June 1998, University of British Columbia, Vancouver, Canada.

Page A-41

Loewe & Scherer (1999)
"How many beneficial mutations stop Muller's ratchet?"
Evolution '99,
22-26 June 1999, University of Wisconsin, Madison, Wisconsin.

Page A-42

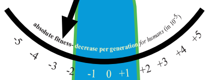
Loewe & Scherer (2001)
"Predicting extinctions due to Muller's ratchet in humans and bacteria",
8th Congress of the
European Society for Evolutionary Biology,
20-25 August 2001, Aarhus, Denmark.

Presented at The Fifth Annual Meeting of the Society for Molecular Biology and Evolution, 1-4 June 1997, Garmisch-Partenkirchen, Germany.

Speed of Genomic Decay

Laurence Loewe & Siegfried Scherer
email:
Institute for Microbiology FML
Technical University Munich
Vöttingerstr. 45
85354 Freising
Germany

Over time absolute fitness will



Why?

Mutation rate measurements (23) in man show:

$5*10^{-5}$ substitutions /site /generation in mtDNA

If nuclear DNA is 50x more stable: ≈ 3000 new mutations /newborn homo

With 95% neutral 3% SDM 2% advan.

30 SDM "netto" get fixed each generation (see 3,27)

With s ≈ 10^{-6} average absolute fitness decrease per SDM (13), in each generation

≈ $3*10^{-5}$ descendents less can be born in a given environment, assuming multiplicative fitness.

If this would continue, the existence of *homo* would be limited to some hundred thousands of generations.

Increase by[28,4]

Selection

of mutations advantageous in the actual environment will increase fitness with time (15). But the speed of fixation of new features has to stay below a certain speed-limit else the population will go extinct (29). Mutation rates put another constraint to the speed of fitness increase by selection, especially for the evolution of complex structures: their spontaneous emergence is just very improbable (26).

Genetic Drift

can increase fitness with time by fixing mutations that preadapt the evolutionary line for future environments (30). This contribution is limited by the frequency of these mutations. Because most mutations are disadvantageous, however, this is likely to apply for nearly neutral mutations as well (22).

Recombination

as such will not increase fitness, but plays a pivotal role in spreading new mutations and slowing down Mullers ratchet (20). Recombination can not make progeny fitter than the genes of their parents allow - except by rare combinations. It is constrained by the thightness of linkages.

Decrease by[13,19,21]

Mullers Ratchet

describes the contamination of any best class of organisms available by slightly deleterious mutations (SDM). Recombination limits the ratchets operations by producing progeny which is better adapted than it´s parents (20). Truncating selection may stop the decay at some equilibrium value (12).

Genetic Drift

may fix mutations with a selectioncoefficient s<1/2Ne (Ne=effective populationsize), because then random drift becomes stronger than selection (13). Slightly deleterious mutations with s<1/2Ne will accumulate almost like neutral mutations. Their destructive power is seen many generations later, as many of them combine their effect. If Ne exceeds the number of potentially deleterious loci, their destructive potential is to low to cause extinction (13).

Hitch-hiking

is a genetic drift process, which can fix slightly deleterious (or positive!) mutations occuring in an individual (chromosome) carrying a selected new feature.

Can the problem be solved?

- **Truncating selection**
  Epistatic interactions could make combined "slightly" deleterious mutations much worse than the sum of their independent effects, eliminating them by selection (3,12). In this case, however, at some stage of genome decay (nearly) *no* more slightly deleterious mutations can be tolerated. This makes (almost) *every* slightly deleterious Mutation a heavily disadvantageous one (19). This truncating selection seems to be unlikely.
- **Compensatory mutations**
  may not be too seldom to slow down decay or even revert the general decay trend if they are advantageous and frequently enough (1,7,30). However, most mutations *seem* to be deleterious - in any environment.
- **Lower mutation rates**
  in relevant regions of the genome (8-11,24,2). However, this would imply that (i) cells can distinguish relevant DNA from irrelevant (how?) or (ii) there are not many relevant sites or (iii) some mutation rates measured recently have to be interpreted differently.
- **Recombination rates**
  High rates of gene-transfer in bacteria (6,14) may slow down or stop the ratchets operation. However, regular recombination in higher metazoa mtDNA (19,17), in Y-chromosomes (25) and in linked genes are yet unknown. Mutational melt-down is possible even in some sexual populations (18).
- **Refuges**
  Spores resting for thousands of years or other refuges like deep rock habitats with extremly slow metabolism (5) may keep less contaminated procaryots and may allow recolonialization after extinction. But this very process is likely to slow down evolution of new adaptations (16) and would not save most eukaryots. Furthermore, it is unclear how such recolonialization processes necessary should happen on a regular basis.
- Any other suggestions?

Literatur: **(1)** Benson S (1997) Adaptive mutation: a general phenomenon or special case? BioEssays *19*(1), 9-11. **(2)** Charlesworth B (1996) Open questions: The good fairy godmother of evolutionary genetics. Current Biology *6*(3), 220. **(3)** Crow JF (1993) Mutation, mean fitness, and genetic load. In: Futuyma D & Antonovics J (ed.) Oxford Surveys in Evolutionary Biology *9*: 3-42. Oxford University Press. **(4)** Eigen M (1993) The origin of genetic information: viruses as models. Gene *135*,37-47. **(5)** Fredrickson JK & Onstott TC (1996) Microbes deep inside the earth. Sci. Am.(8), 42-47. **(6)** Guttman DS (1997) Recombination and colonality in natural populations of Escherichia coli. Trend. Ecol. Evol. *12*(1), 16-22. **(7)** Hartl DL & Taubes CH (1996) Compensatory nearly neutral mutations - selection without adaptation. J. theor. Biol. *182*(3), 303-309. **(8)** Holmquist GP & Filipski J (1994) Organization of mutations along the genome: a prime determinant of genome evolution. Trend. Ecol. Evol. *9*(2), 65-69. **(9)** Houle D, Hoffmaster DK, Assimacopoulos S & Charlesworth B (1992) The genomic mutation rate for fitness in *Drosophila*. Nature *359*, 58-60. **(10)** Keightley PD & Caballero A (1997) Genomic mutation rates for lifetime reproductive output and lifespan in Caenorhabditis elegans. Proc. Natl. Acad. Sci. USA *94*, 3823-3827. **(11)** Kibota TT & Lynch M (1996) Estimate of the genomic mutation rate deleterious to overall fitness in E. coli. Nature *381*, 694-696. **(12)** Kondrashov AS (1994) Muller´s ratchet under epistatic selection. Genetics *136*, 1469-1473. **(13)** Kondrashov AS (1995) Contamination of the genome by very slightly deleterious mutations: Why have we not died 100 times over? J. theor. Biol. *175*, 583-594. **(14)** Lan R & Reeves PR (1996) Gene transfer is a major factor in bacterial evolution. Mol. Biol. Evol. *13*(1), 47-55. **(15)** Lenski RE & Travisano M (1994) Dynamics of adaptation and diversification: A 10,000-generation experiment with bacterial populations. Proc. Natl. Acad. Sci. USA *91*, 6808-6814. **(16)** Liu Y-B & Tabashnik BE (1997) Experimental evidence that refuges delay insect adaptation to *Bacillus thuringiensis*. Proc. R. Soc. Lond. B *264*, 605-610. **(17)** Lunt DH & Hyman BC (1997) Animal mitochondrial DNA recombination. Nature *387*, 247.**(18)** Lynch M, Conery J & Bürger R (1995) Mutational meltdowns in sexual populations. Evolution *49*, 1067-1080. **(19)** Lynch M (1996) Mutation accumulation in transfer RNAs: Molecular evidence for mullers ratchet in mitochondrial genomes. Mol. Biol. Evol. *13*(1), 209-220. **(20)** Maynard Smith J (1978) The evolution of sex. Cambridge University Press. **(21)** Newman D & Pilson D (1997) Increased probability of extinction due to decreased genetic effective population size: experimental populations of Clarkia pulchella. Evolution *51*(2), 354-362. **(22)** Ohta T (1997) Role of random genetic drift in the evolution of interactive systems. J. Mol. Evol. *44*(Suppl. 1), S9-S14. **(23)** Parsons TJ, Muniec DS, Sullivan K, et al. (1997) A high observed substitution rate in the human mitochondrial DNA control region. Nature Genetics *15*, 363-368. **(24)** Peck JR & Eyre-Walter A (1997) Evolutionary genetics: The muddle about mutations. Nature *387*, 135-136. **(25)** Rice WR (1994) Degeneration of a nonrecombining Chromosome. Science *263*, 230-232. **(26)** Scherer S (1983) Basic functional states in the evolution of light-driven cyclic electron transport. J. theor. Biol. *104*, 289-299. **(27)** Sommer SS (1995) Recent human germ-line mutation: inferences from patients with hemophilia B. Trends in Genet. *11*, 141-147. **(28)** Szathmáry E & Maynard Smith J (1995) The major evolutionary transitions. Nature *374*, 227-232. **(29)** Worden RP (1995) A speed limit for evolution. J. theor. Biol. *176*, 137-152. **(30)** Zuckerkandl E (1997) Neutral and nonneutral mutations: the creative mix-evolution of complexity in gene interaction systems. J. Mol. Evol. *44*(Suppl. 1), S2-S8.

Mullers Ratchet in human mitochondrial DNA

Laurence Loewe & Siegfried Scherer, email:ll.ttt@lrz.tu-muenchen.de, Institute of Animal Sciences, Microbial Ecology Group, FML, Technische Universit t M nchen, Weihenstephaner Berg 3, D-85354 Freising, German
Presented at The Sixth Annual International Meeting of the Society for Molecular Biology and Evolution, 17-20 June 1998, University of British Columbia, Vancouver, Canada.

Fig. 1: Decay of absolute fitness by subsequent fixations of SDMs by Mullers Ratchet.

What is Mullers Ratchet?

If the slightly deleterious mutation (SDM) rate is high enough in finite populations, then these populations will increasingly suffer from a decrease of absolute fitness, since more and more SDMs will become fixed (Fig.1). The theory of Mullers Ratchet (Muller 1964; Felsenstein 1974) describes the conditions under which SDMs are accumulated (Fig. 2), i.e. when the Ratchet clicks (Maynard Smith 1978; Charlesworth et al. 1997). The most critical parameter is N_e, the size of the best class available in mutation-selection equilibrium. This can be calculated analytically: $N_0 = N_e * \exp(-U/s)$, where N_0 is the effective population size, **U** is the genomic mutation rate, and **s** is the selection coefficient describing the effect of each mutation (Haigh 1978). The Ratchet clicks if this best class becomes extinct by drift. This will not occur if there are more than 1000 individuals in this class. However, if there are fewer than 100 individuals the Ratchet will click slowly. By the time N_0 reaches 10, this rate is fast and still increases with decreasing N_0. Once absolute fitness decreases beyond the point where a population can reproduce itself, a mutational meltdown further increases the speed of the Ratchet (Gabriel 1993).

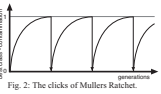


Fig. 2: The clicks of Mullers Ratchet.

Open Questions

Many questions arise, if one wishes to apply these results to the human population:

- Is the mutation rate measured in the Control Region of mtDNA really the same in the remainder of mtDNA as some data indicate (Howell et al. 1996)?
- Which fitness-function is the best description of structure-function relationships of mtDNA and does this change in decay?
- How do the complex multi-level population genetics influence clicks of the Ratchet, and thus extinction times?
- How will the picture change if overlapping generations are considered?
- What is the combined extinction time if mutations with many different effects are allowed, instead of only one type (Butcher 1995)?
- How far are mitochondria in the decay process (if at all)? Are there additional indicators besides tRNA-stability (Lynch 1997)?
- What is the effect distribution of mutation rates? Does it favour extinction or a solution for the apparent discrepancy of phylogenetic versus pedigree rates?
- .....

Resulting Extinction Times

Extinction times may either be measured in simulations, or estimated by measuring the time between two clicks and determining the maximal number of clicks a population can tolerate from the fitness-function:

s	U	N_0	T_{click}	$Clicks_{max}$	$T_{extinct}$
0.01	0.02	17000	never	142	never
0.01	0.4	$1e^{+13}$	7.6	142	1016
0.001	0.02	$5e^{-5}$	171	1450	250 000*
0.001	0.4	$5e^{-170}$	4.36	1450	5.142
0.0001	0.02	$4e^{-83}$	56	15000*	841.000*
0.0001	0.4	$<1e^{-300}$	4.6	15000*	69.000*

T_{click} (generations) has been measured in situations where the current population size (normally ca. 30.000) was larger than 70% of the habitats capacity (10.000). This was to avoid a bias towards smaller values that occurred during mutational meltdown (Gabriel et al. 1993). $Click_{max}$ is the total number of clicks observed until extinction. If extinctions could not be observed due to limitations in computational time, $Clicks_{max}$ (*) was estimated. $T_{extinct}$ is the mean of several runs except for values extrapolated from T_{click} and $Clicks_{max}$ (*).

Why are we still alive?

This question arises when some recent high mutation rate measurements in mtDNA are considered. The rates by themselves do not prove that we should have been extinct long ago. However, if the mutations appearing at these rates have effects between $0.01 > s > 10^{-6}$, then human long-term survival may be endangered. The trouble is that such mutation effects are hard to detect and cannot be excluded at present. Moreover, they seem rather likely, since - as it is often said - more mutations are deleterious than advantageous and more mutations have small effects than large ones.
But is this true?

Ratchet Effects on Fitness

Any SDMs fixed by the Ratchet can only be felt in combination with many other SDMs. The fitness-function, yet unknown, predicts how single SDMs combine their effects. If all SDM-effects are independent from each other, then multiplicative fitness can be assumed. However, epistatic interactions, decreasing or increasing the mutual effect of SDMs, thoroughly distort this simple picture (Fig. 1) and make the details of this function an open question.
Since the Ratchet can affect fitness only when it clicks, the conditions necessary for its operation are paramount for the evolution of fitness:

- Lack of recombination - no progeny better than its parents
- High mutation rate - more new mutations are introduced than selection can possibly remove ($U/s < 1$ for asexuals, $U < 1$ for sexuals)
- Small population size - best class readily dies out.
- Lack of truncating selection - SDMs appear at a constant rate.

Simulation Methods

The following simulation model was used to estimate extinction times:

- Generations were discrete. Probability of death after a simulation step was 1
- Prior to death, individuals reproduced identical copies of themselfes, except that a number of mutations were added (random-drawn, Poisson-distributed, expectation U)
- The Poisson-expectation of clones produced by each individual was the product of
  (a) the number of mutations the parent carried (using multiplicative fitness)
  (b) the maximal expectation of descendents (1 female => 4 females)
  (c) the quotient $\frac{CarryingCapacity}{PopulationSize}$, if the habitat was overcrowded. density dependence allowed up to 3 times more individuals than a habitats "normal capacity" under the conditions used. After each click, the distribution of mutations in the population was calculated. Simulations were run with parameter combinations that had been simulated earlier by

Man Ratchet

The operation of the Ratchet is normally associated with bottlenecked populations of asexual individuals (Andersson et al. 1996; Moran 1996). However, survival of organisms that rely on asexual genetic systems which are bottlenecked during gamelogenesis may be endangered too. Among these is the DNA of mitochondria in humans. Phylogenetic studies suggested that mtDNA mutates at higher rates than most nuclear DNA. Recent measurements of the intergenerational mtDNA mutation rate, however, indicated even higher rates and thus created problems for mtDNA molecular clocks (Howell et al. 1996; Parsons et al. 1997; Gibbons 1998). If these mutation rates are analyzed with the Ratchet theory, they seem to indicate that the Ratchet operates too fast in mitochondria. Since the carriers of the mutations observed were "normal" people, heavily deleterious mutations may be excluded from further analysis. The range of interesting parameters is:

- Ne: 10 000 to 100 000?
- U: min: 0.02 ($2*10^{-5}$ * 1 000 bp = "linked loci")
- U: max: 0.4 ($4*10^{-5}$ * 10 000 bp = "linked loci")
- s: 0.1 to 0.0001 or even less?

If N_0 is checked, it becomes clear that only those populations in this range of parameters, where $s > 0.1$ or where $s > 0.01$ and $U <$ 0.02, can withstand the onslaught of the Ratchet. This may help recalibrate the mtDNA clock. All others die off.

Potential Solutions:

Hotspots

Some mutations were observed at positions that mutate with elevated frequency. It had been suggested, that such hotspots explain the theoretical problems associated with the observations of the high intergenerational mutation rates (Pääbo 1996). These mutations are seen first in pedigrees because of their frequency, but introduce only phylogenetic noise in longterm studies, since a long sequence of mutations at a certain position may result in a single base substitution only. This could create the apparent contradiction between phylogenetic and pedigree rates. However, not all the mutations observed seemed to be of this type (Parsons et al. 1997; Gibbons 1998). If more neutral sites are discovered as hot spots, then the apparent danger of extinction is reduced.

Inhomogeneous Distribution of Mutation Rates

At present, it cannot be excluded that mutations arise at different rates in different populations (Parsons et al. 1997; Soodyall et al. 1997; Gibbons 1998). If this is so, then long-term survival of humans may not be endangered because populations with low mutation rates persist. From time-to-time they may give rise to populations with higher rates. Alternatively, mutation rates may be high in reality, but the affected positions in the sequence may be selectively *really* neutral. Hence, human mtDNA sequences would change faster than previously thought, but survival would not be endangered.

Epistatic Interactions leading to Truncating Selection and Compensatory Mutations

Finally, complicated interactions in the mitochondrial microcosm may lead to changes in the state of the mitochondrion, and thus make new mutations either very disadvantageous or positive. Both mutation effects will not contribute to an extinction risk. If the deleterious effects of new mutations increase with the number of mutations that happened before, then truncating selection (Crow 1997) may stop the Ratchet under very special conditions (Kondrashov 1994; Butcher 1995).
If new mutations are likely to compensate defects that have arisen because of earlier mutations (Stephan 1996; Schultz et al. 1997), then the Ratchet can be stopped or even reversed for the rare case that there are more advantageous mutations than deleterious ones.

References

Andersson DI & Hughes D (1996) Muller's ratchet decreases fitness of a DNA-based microbe. Proc. Natl. Acad. Sci.U.S.A. 93(2), 906-907.

Butcher D (1995) Muller's ratchet, epistasis and mutation effects. Genetics, 141(1), 431-437.

Charlesworth B & Charlesworth D (1997) Rapid fixation of deleterious alleles can be caused by Muller's Ratchet. Genet. Res. 70, 63-73.

Crow JF (1997) The high spontaneous mutation rate: Is it a health risk? Proc. Natl. Acad. Sci. U.S.A. 94, 8380-8386.

Felsenstein J (1974) The evolutionary advantage of recombination. Genetics 78, 737-756.

Gabriel W, Lynch M & Bürger R (1993) Muller's Ratchet and mutational meltdowns. Evolution 47(6), 1744-1757.

Gibbons A (1998) Calibrating the mitochondrial clock. Science 279, 28-29.

Haigh J (1978) The accumulation of deleterious genes in a population - Muller's Ratchet. Theor. Pop. Biol. 14, 251-267.

Howell N, Kubacka I & Mackey DA (1996) How rapidly does the human mitochondrial genome evolve? Am. J. Hum. Genet. 59, 501-509.

Kondrashov AS (1994) Muller's Ratchet under epistatic selection. Genetics 136, 1469-1473.

Loewe L & Scherer S (1997) Mitochondrial evo: the plot thickens. Trends Ecol. Evol. 12(11), 422-423.

Lunt DH & Hyman BC (1997) Animal mitochondrial DNA recombination. Nature 387, 247.

Lynch M (1997) Mutation accumulation in nuclear, organelle, and prokaryotic transfer RNA Genes. Mol. Biol. Evol. 14(9), 914-925.

Maynard Smith J (1978) The Evolution of Sex. New York, Cambridge University Press.

Moran NA (1996) Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. Proceedings of the National Academy of Sciences of the United States of America 93(7), 2873-2878.

Muller HJ (1964) The relation of recombination to mutation. Mutation Research 1, 2-9.

Pääbo S (1996) Mutational hot spots in the mitochondrial microcosm. Am. J. Hum. Genet. 59, 493-496.

Parsons TJ, Muniec DS, Sullivan K, Woodyatt N, Alliston-Greiner R, Wilson MR, Berry DL, Holland KA, Weedn VW, Gill P & Holland MM (1997) A high observed substitution rate in the human mitochondrial DNA control region. Nature Genet. 15(4), 363-368.

Schultz ST & Lynch M (1997) Mutation and extinction: The role of variable mutational effects, synergistic epistasis, beneficial mutations, and the degree of outcrossing. Evolution 51(5), 1363-1371.

Soodyall H, Jenkins T, Mukherjee A, Du Toit E & Stoneking M (1997) The Founding Mitochondrial DNA Lineages of Tristan da Cunha Islanders. Ann. J. Phys. Anthropol. 104(2), 157-166.

Stephan W (1996) The rate of compensatory evolution. Genetics 144(1), 419-426.

Stephan W, Chao L & Smale JG (1993) The advance of Muller's ratchet in a haploid asexual population: Approximate solutions based on diffusion theory. Genet. Res. 61(3), 225-231.

How many advantageous mutations stop Mullers Ratchet?

Laurence Loewe & Siegfried Scherer, email:ll.ttt@lrz.tu-muenchen.de, Institute of Animal Sciences, Microbial Ecology Group, FML, Technische Universität München, Weihenstephaner Berg 3, D-85354 Freising, Germany
Presented as Poster 839 at "Evolution '99", 22-26 June 1999, University of Wisconsin, Madison, Wisconsin.

Mullers Ratchet ...

... describes the accumulation of slightly deleterious mutations (**SDM**s) in finite populations (1). While advantageous mutations provide selectable raw material for evolution, deleterious ones are easily removed and thus do not threaten survival of populations with a large size **N**. However, if the selection coefficient **S** of a mutation is on the order of 1/N, selection looses its control over evolution and genetic drift gradually takes over. Such mutations may limit long-term survival of a population because they can be fixed by drift ***and*** have negative effects on fitness. Thus, to prevent decay of absolute fitness, these SDMs must never occur or always be removed or always be compensated.

is normally associated with small or asexual populations. However, deleterious mutation rates and effects are the most important parameters determining the Ratchets operation, since these can destroy any population, if they are high enough. Large, sexual populations just tend to tolerate higher rates.

... may threaten humans!

Nonrecombining genetic systems are important, even for sexual organisms (imagine humans without mitochondrial DNA or Y-Chromosomes). In the case of humans, this may doom to extinction on the long run. Combine ...

... suprisingly high deleterious (nuclear) genomic mutation rates (2),

... suprisingly high mutation rates observed in human mtDNA (3),

... the number of genetical diseases caused by deleterious mtDNA (4), i.e. the fact that mutations in mtDNA can hurt much, although not all do.

Now consider the fundamental biological experience that most mutations are rather harmful than hopeful and extrapolate this to those mutations whose effect can not be observed because it is too small. This leads to conditions that are ideal for Mullers Ratchet to operate. We measured potentially resulting extinction times for humans in computersimulations (5) and saw that ...

... the Ratchet is running for a large range of biologically reasonable parameter combinations (e.g.: mtDNA mutation rate U: $0.02<U<0.2$; S: $0.0001<S<0.01$; maximum expectation of number of offspring (=Fitness) $F_{max}=4$; N≈40 000). Extinction times observed under these conditions were between thousands and a million generations. (Only U = 0.02 with S=0.01 had no Ratchet running.)

... Purifiying selection alone is usually not enough to stop the Ratchet.

Critical parameters needed to stop Mullers Ratchet in this simulation study are marked with arrow.

Fig 1: Change of fitness (N = 10)

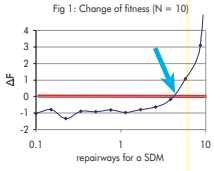


Fig 2: Change of fitness (N = 100)

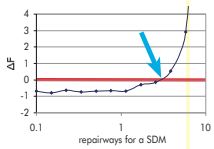
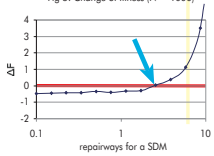


Fig 3: Change of fitness (N = 1000)



The Simulation Model

Stochastic individual based computer simulations were used modeling the following 'history' of a single population:

* create a population with carrying capacity **K** containing **N**=K individuals
* simulate 500 generations then measure fitnessdistribution $\mathbf{F_{start}}$
* simulate another 500 generations then measure $\mathbf{F_{end}}$, the final distribution of fitness and calculate the difference **ΔF** from the mean of F_{end} - F_{start}. ΔF was used to measure genomic decay implying that **F**, the maximum reproductive capacity of an individual (under no density constraints) is a reasonable indication of its success in evolution.

Each individual was characterized by the numbers of mutations it accumulated and by its overall genetic fitness **F** (defined as the expectation of offspring in a population without density constraints). Mutation counters were present for repairable and unrepairable SDMs, for compensatory mutations and for slightly advantageous mutations (**SAM**s) in order to allow observation of their relative contribution to genetic fitness. Individuals lifehistory was:

* Birth at generation g and immediate
* Mutation: Poisson derivates of the expected number of new mutations were added to the inherited mutations on the indivduals mutation counters. The new fitness F was calculated multiplicatively immediately after the generation of a mutation.
* Reproduction at g+1: generated Poisson(F) indivduals and made them accumulate mutations.
* Death at g+1: everyone dies (discrete generation).
* No recombination (as probably in mtDNA).

The simulation engine led to results comparable to those in literature and thus seems to be trustworthy.

Potential Solutions?

Several potential solutions may resolve the extinction paradox. For example:

Hot spots: Since the distribution of mutation rates along a genome and in a population is largely unknown, frequent changes at unimportant sites could occur.

Recombination: Combining the best and the worst, this allows selection to evaluate mutations in bundles; Selection would not see them, if they were alone.

Epistatic interactions leading to quasi-truncating selection: If new mutations have more pronounced effects than those accumulated before them, then this may stop genomic decay (6).

Different disributions of mutation effects: If mutations were either clearly deleterious (->removable) or really neutral (-> no harm) or advantageous often enough, this would stop the Ratchet.

This study was undertaken to investigate the level of positive mutations necessary to compensate for the fitness loss due to the Ratchet.

To focus on changes by compensatory mutations on a molecular level and on slightly advantageous mutations (**SAM**s), some

Results

parameters were kept constant: U_{sdm} = 0.1; S_{sdm}= -0.01; F_{max} = 4; initial population (N=K=free of mutations). In Figures, each dot represents the mean of 5 runs. For N≤10, populations often went extinct and thus were excluded from calculation of ΔF.

Fig 1-3: Compensatory repairways needed to stop the ratchet were calculated assuming a '**warranty**' of 0.2 (i.e.20% of the 10kb in mtDNA could change before potential repairways would be lost). Here ≈ 2-5 repairways were needed depending on N.

Fig 4-7: U_{sam} needed to stop the Ratchet depends on N. S_{sam} was -S_{sdm}. Larger populations could use positive mutations better.Here critical levels of U_{sam} were ≈ 2%-100% of U_{sdm}.

Advantageous Mutations ...

... can be classified into 3 categories (see description of the model):

(a) Compensatory mutations on a molecular level are defined here as the backmutation rate times the number of other potential molecular ways to repair the specific dammage done by a certain SDM. The number of repairable SDMs is limited, as repairways themselves are being destroyed even by netural or advantageous mutations.

(b) Compensatory QTL mutations could substitute fitness lost due to completely unrelated mutations until maximum potential of relevant traits is fully exploited.

(c) other advantageous mutations could lead to new capabilities.

See results for an example of (a) and (c) and ref (7) for some related studies.

... should be more frequent!

than most biologists would suspect from their experience with mutations, if they are to stop Mullers Ratchet on their own. Thus, other processes are likely to be pivotal. All experimental information on the distribution of mutation effects will be needed to find an answer. Further simulations should allow to check evolutionary consequences of a given observation and integrate other potential factors into the picture.

Fig 4: Change of fitness (N = 10)

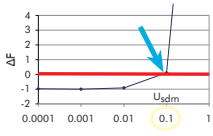


Fig 5: Change of fitness (N = 100)

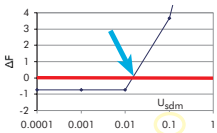


Fig 6: Change of fitness (N = 1000)

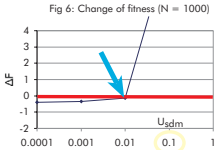
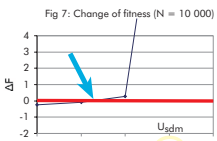


Fig 7: Change of fitness (N = 10 000)



References:

(1) **Maynard Smith J** (1978) The Evolution of Sex. New York, Cambridge University Press. **Muller** HJ (1964) The relation of recombination to mutational advance. Mutation Research 106:2-9. **Felsenstein J** (1974) The evolutionary advantage of recombination. Genetics 78, 737-756. **Haigh J** (1978) The accumulation of deleterious genes in a population - Muller's Ratchet. Theor. Pop. Biol. 14, 251-267. **Stephan W**, Chao L & Smale JG (1993) The advance of Muller's ratchet in a haploid asexual population: Approximate solutions based on diffusion theory. Genet. Res. 61(3), 225-231. **Gabriel W**, Lynch M & Bürger R (1993) Muller's Ratchet and mutational meltdowns. Evolution 47(6), 1744-1757. **Charlesworth B** & Charlesworth D (1997) Rapid fixation of deleterious alleles can be caused by Muller's ratchet. Genet. Res. 69, 213-218.

(2) **Crow JF** (1997) The high spontaneous mutation rate: Is it a health risk? Proc. Natl. Acad. Sci. U.S.A. 94, 8380-8386 **Eyre-Walker, A** & Keightley, PD (1999) High genomic deleterious mutation rates in hominids Nature 397, 344-347.

(3) **Parsons** TJ et al. (1997) A high observed substitution rate in the human mitochondrial DNA control region. Nature Genet. 15(4), 363-368. **Gibbons A.** (1998) Calibrating the mitochondrial clock. Science 279, 28-29. **Howell** N et al. (1996) Are new mutations in the human mitochondrial genome evolved? Am. J. Hum. Genet. 59, 501-509.

(4) see Science 268(1995):1402f.

(5) **Loewe L** & Scherer S (1998) Mullers Ratchet in human mitochondrial DNA. Poster on the VI SMBE meeting in Vancouver, June 1998.

(6) **Crow, J. F.** & Kimura, M. (1979) Efficiency of truncation selection. Proc. Natl. Acad. Sci. U.S.A. 76, 396-399

(7) **Poon A** & Otto SP (2000) Compensating for our load of mutations. Genetics 154(3), 411-417. **Schultz ST** & Lynch M (1997) Mutation and extinction: The role of variable mutational effects, synergistic epistasis, beneficial mutations, and the degree of outcrossing. Evolution 51(5), 1363-1371. **Stephan W** (1996) The rate of compensatory evolution. Genetics 144(1), 419-426.

Predicting Extinctions due to Mullers Ratchet in Humans and Bacteria

Laurence Loewe & Siegfried Scherer
email: Laurence.Loewe@evolutionary-research.net
Microbial Ecology Group
Department of Biosciences
Technische Universität München
Weihenstephaner Berg 3
85354 Freising
Germany

Presented at 8th Congress of the European Society for Evolutionary Biology, 20-25 August 2001, Aarhus, Denmark.

Introduction

With the advent of high throughput DNA sequencing it has become possible to observe mutation rates directly by comparing sequences from individuals with a known pedigree (Parsons et al. 1997; Sigurdardottir et al. 2000). When this was first done for human mitochondrial DNA, the rates observed were significantly higher than expected from phylogenetic studies and the question arose, how humans could escape Mullers Ratchet in their mitochondrial genomes (Howell et al. 1996). discrepancy between phylogenetic and pedigree mutation rates, mutational hot spots (Meyer et al. 1999) were invoked as well as long term selective removal of mutations (Hasegawa et al. 1998). The latter would require most mutations to be slightly deleterious which would either lead to their removal or drive Mullers Ratchet effectively. Only a detailed analysis can specify the mutational effects facilitating removal and those driving Mullers Ratchet.

The rate of Mullers Ratchet is notoriously difficult to predict. Therefore known analytic approaches (Haigh 1978, Gabriel et al. 1993, Gessler 1995, Gordo & Charlesworth 2000 a+b) are combined with simulations of an individual based model of Mullers Ratchet. This model was implemented by chosing a different population size regulation mechanism than previous models to assess sensitivity of results to such underlying mechanisms. Another goal was to observe various new details about Mullers Ratchet to gain new insight into its behaviour and the mechanisms behind it.

As Mullers Ratchet theory was mainly developed to understand the evolution of sex, a look at the situation of asexual haploid prokaryotes is possible with the same set of methods.

Mullers Ratchet is ...

a process that leads to repeated loss of the fittest class of individuals in a finite population. This is driven by stochastic accumulation of slightly deleterious muta-tions (SDM) in individuals with the least mutations: As soon as each of these best individuals has been contaminated with at least one new SDM, the Ratchet clicks round a notch and once again the good has become the best.

Global Computing is ...

a special way of massively distributed computing via the Internet. Global computing projects (e.g. like SETI@home) publicly distribute a background com-puting client that is fed with tasks from a central server. Results are collected and analyzed centrally, too. Participants usually are acknowledged in web statistics, if they want. Only problems that can be completely parallelized are suitable for this.

Mullers Ratchet in mitochondrial DNA may cause Extinctions in the Human Line

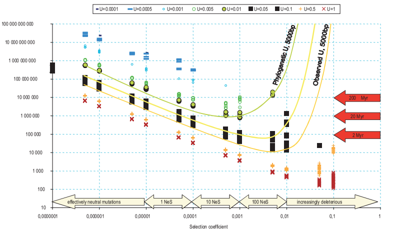


Figure 1: Mullers Ratchet in Humans. Each symbol stands for a single simulation with an effective population size of either 10 000 or 50 000 or 100 000.

Methods

Individual Based Model of Mullers Ratchet

While most other simulations of Mullers Ratchet have effective population sizes Ne fixed at their habitats carrying capacity K, this model allows for fluctuations. Ne is regulated by a density depen-dent culling factor (DDC) that equals K divided by the sum of the expected number of offspring of all individuals in the last gene-ration, but never exceeds 1. Iterating over *all* individuals in a gene-ration produces the next. ***Individual life history***:

➔ Birth. Here Poisson(U) new mutations are added to the child. Each mutation has a constant effect s on multiplicative fitness (0..1)
➔ Survival until consumption of resources of an adult is 100 %.
➔ Date transition (i.e. all new offspring produced before now)
➔ Survival until reproductive maturity is 100 %
➔ Produce Poisson(Vmax*W*Rmax*DDC) offspring, where ***Vmax*** is survival from fertilization (or birth, depends on Rmax) until adult resource consumption, ***Rmax*** is the maximal reproductive capacity (10 females for a 20 year generation in humans and 10^7 for 1 day of bacterial growth with ca. 1 generation each hour) and ***W*** is the parents current fitness
➔ Death (can happen only here, if no viability selection).

During a simulation K individuals are generated after start. Then all parameters listed are observed. Clicktimes of the Ratchet are measured only after the first click to reach equilibrium. A simulation is stopped, when either 500 clicks have been observed or Tend Generations have been reached ($= 10^5$-10^8) or when the population went extinct.

evolution@home containes these features and has been prepared to address advantageous mutations, two different Ratchettypes operating in the same population, and variable mutation effects in the near future. During computation it observes a considerable number of parameters, including length of various phases of a Ratchet's click, multiple clicks, actual sizes of the best classes and indicators for reaching equilibrium.

Why *evolution@home* global computing?

To facilitate a large scale search of parameter space via computation of many different parameter combinations, considerable computing power is needed. While some of the results on this poster could have been archived by analytic theory, the simulator is prepared to include several complications that can not be easily predicted. If one is not willing to stick with a very restricted set of parameter combinations, global computing is the only possibility, besides buying a super-computer. The infrastructure now under development (e.g. storing results in XML, setting priorities for runs, ...) is expected be helpful in future models too.

Mitochondrial Mutation rates

in humans have been estimated from phylogenies to range from 0.025 to 0.26 subst/bp/Myr and from pedigrees to range from 0.32 to 2.5 subst/bp/Myr. (Sigurdardottir et al. 2000; Parsons et al. 1997) Here we use 0.1 (Stoneking et al. 1992) and 1 (Parsons et al. 1998) as representative values. Assuming 20 years per generation these translate into U = 0.01 and U=0.1 for 5000 functional sites across mtDNA.

Bacterial Mutation Rates

are usually 1/300 per genome per generation (Drake et al. 1998). Of these, 6 % confer deleterious effects too large to drive the Ratchet (1.2 %, Kibota & Lynch 1996). The other 94% are probably not completely neutral. Most populations have a few percent mutators with mutation rates increased up to 1000 fold (LeClerc et al. 1996; Matic et al. 1997). This can lead to U=1 as found in many RNA-Viruses (Drake et al. 1998).

Mullers Ratchet may cause Extinction of Bacteria if they lack Recombination

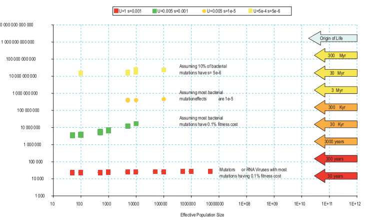


Figure 2: Mullers Ratchet in Bacteria. Four examples of realistic combinations of genomic mutation rate U per generation and selection coefficient s are given. Generations are assumed to be 1 day long under normal conditions, although doubling time under optimal conditions would be ca. 1 hour. Note that mutations with ca. 1% fitness cost as those found in mutation accumulation experiments are removed by selection and do not drive Mullers Ratchet.

Results

Extinctions Observed

Basically, both asexual genetic systems show the same picture: A certain range of biologically plausible combinations of mutation rate (U) and selection coefficient (s) leads to extinction in a time shorter than their assumed existence. Population size does not seem to stop that. Other parameter combinations, however, make Mullers Ratchet so slow, that it would not play a significant role for 5 billion years: Just make U and s low enough.

Combining the little information available on the distribution of mutational effects (dme) makes it hard to see, why certain deleterious mutations should occur, while all other mutations should be absolutely neutral or beneficial. If this were so, the nearly neutral theory (Ohta 1992) would be less successful as it deals mainly with very slightly deleterious mutations. Thus all selection coefficients can be expected in natural systems as you can always compromise a good thing in tenthousand little ways.

Human mitochondrial DNA

Most probable estimates of U are indicated by lines in Fig. 1. Measurements of U are mainly from the hypervariable control region. However, some observations indicate similar mutation rates for the rest of the genome (Denver et al. 2000; Howell et al. 1996). While this does not translate into phylogenetic substitution rates for functional regions, it is the relevant deleterious mutations rate for Mullers Ratchet.

Assuming that 5000 sites mutate at the conventional phylo-genetic rate is equivalent to assuming 500 sites with the average of all pedigree rates and leads to U= 0.01. Here the deadly range of s is quite narrow (ca. 0.0005-0.001) regarding extinction in 20 Myr. However, as the observed mutation rates probably apply to the whole genome, U=0.1 seems more likely. Here the deadly range is quite wide (s from 0.01 to $5x10^{-5}$). Only if most mutations have more than 1% effect, Mullers Ratchet stops and the selective removal hypothesis would solve the mutation rate conflict. If s is smaller than 10^{-3} to 10^{-5}, depending on U, the Ratchet is too slow to drive the human line to extinction in 20 Myr.

There is no reason to assume that mutations never have effects in the corresponding deadly ranges, thus the question arises: ***How could the human line survive 20 million years without going extinct due to the Ratchet in Mitochondria?***

Endosymbiontic Bacteria

As free living bacteria often exchange genes horizontally (Lan etal.1996), they probably escape the Ratchet by recombination. Endosymbionts like Buchnera, however, appear to have restricted opportunities to exchange genes, since they were sequestered in aphid hosts more than 100 Myr ago (Moran 1994). As the operation of Mullers Ratchet can be seen in their decreased ratio of synonymous to nonsynonymous substitutions, it appears unlikely that most mutation effects are deleterious enough to be removed quickly (s>1%). Thus endosymbionts appear to be another example for the genomic decay paradox, as can be seen in Figure 2.

The genomic decay paradox says that ...
some evolutionary lines appear to exist longer than their most probable extinction times predicted from current population genetics theory using best parameter estimates.

Potential Solutions

Recombination and horizontal gene transfer appear to solve the paradox for free living bacteria, but probably not for endosymbionts and human mitochondria as recombination is the big exception in these systems - if it occurs at all.

Recolonization of extinct mutator population niches can be expected to happen regularly. This is impossible for mitochondria.

Compensatory and advantageous mutations stop the decay, if frequent enough. If not, they slow it down in both systems.

Highly deleterious mutations prevent the Ratchet from operating, if common. Harsh selection for good mitochondria during early embryo stages or similar processes may be the reason.

Very small mutation effects could not stop the Ratchet, but make it irrelevent, if common and their effect is $<10^{-6}$ in mitochondria.

Synergistic epistasis could increase deleterious effects of mutations and thus stop the decay, if most mutations have equal effects.

Be part of evolution@home and solve the puzzle.

Simulator005 Input-parameters
(Abbreviation followed by ParameterID and explanation)
K 1 Carrying capacity of the population
Ri 2 Initial reproductive capacity
Ub 3 Genomic mutation rate for background mutations
Sb 4 Selection coefficient for background mutations
dmeb 5 Distribution of mutational effects (background)
Ur 6 Genomic mutation rate for the Ratchet observed
observed
dmer 8 Distribution of mutational effects (Ratchet)
Cend 9 Clicks to observe before worldhistory can stop
Tend 10 Last generation in standard worldhistory
rfcID 11 ID of RunFileCollection this simulation belongs to
rndseed 12 First seed used to initialize global RND generator

Simulator005 Output parameters
err 14 Errors since start of this worldhistory
intm 15 Result is intermediate (=1) or complete (=0)
timev 16 Total seconds needed for last evolve&observe-loop
timpr 17 Total amount of work (GigaIndividuals) predicted
timwh 18 Total seconds computing time for this single-run
Glo 19 Workunits in this run (in gigaindividuals)
minds 20 Performance during current single run in MINDS
Teq 21 Time from perfect start to exp(-U) fitness
Tclfd1 22 Time from perfect start to 1.FitnessDownClick
Tclfi1 23 Time from perfect start to 1.FitnessIncrClick
Tclb1 24 Time from perfect start to 1.background click
Tclr1 25 Time from perfect start to 1. Ratchet click
mmpr 26 Date where meltdown is predicted to start
Tmm 27 Date where meltdown started
Tex 28 Extinction time (single run)
wtim 29 Current world time in current simulation
aclfd 30 Absolute click time for fitness decreases
eclfd 31 Effective click time for fitness decreases
clsfd 32 Click size of fitness decreases
aclfi 33 Absolute click time for fitness increases
eclfi 34 Effective click time for fitness increases
clsfi 35 Click size of fitness increases
atcb 36 Click time (absolute) of background mutations
etcb 37 Click time (effective) of background mutations
tcbpr 38 Predicted click time of background mutations
clsb 39 Click size of background mutations
atcr 40 Click time (absolute) of Ratchet mutations
etcr 41 Click time (effective) of Ratchet mutations
tcrpr 42 Predicted click time of Ratchet mutations
clsr 43 Click size of Ratchet mutations
cl1 44 Absolute time for clicks that fix 1 mutation
cl2 45 Absolute time for clicks that fix 2 mutations
cl3 46 Absolute time for clicks that fix 3 mutations
cl4 47 Absolute time for clicks that fix 4 mutations
cl5 48 Absolute time for clicks that fix 5 mutations
T1a 49 Duration of phase 1a for Ratchet (n1 to 1.6*no)
T1b 50 Duration of phase 1b for Ratchet (1.6*no to no)
T2 51 Duration of phase 2 for Ratchet (no to 0)
Nt 52 Average total population size over worldhistory
Nod 53 Individuals with highest fitness (best class)
Noi 54 Individuals with lowest fitness (worst class)
irNo 55 Analytic best class size for Ratchet
Nor 56 Individuals with least Ratchet mutations
ibNo 57 Analytic best class size, background mutations
Nob 58 Individuals with least background mutations
Df 59 Distribution of overall genetic fitness in population
Dr 60 Number of Ratchet mutations per genome
Db 61 Number of background mutations per individual

References

Denver, D. R., K. Morris, et al. (2000). "High direct estimate of the mutation rate in the mitochondrial genome of Caenorhabditis elegans." Science 289(5488): 2342-4.

Drake, J. W., B. Charlesworth, et al. (1998). "Rates of spontaneous mutation." Genetics 148(4): 1667-1686.

Gabriel, W., M. Lynch, et al. (1993). "Mullers Ratchet and mutational meltdowns." Evolution 47(6): 1744-1757.

Gessler, D. D. G. (1995). "The constraints of finite size in asexual populations and the rate of the ratchet." Genetical Research 66(3): 241-253.

Gordo, I. and B. Charlesworth (2000). "The degeneration of asexual haploid populations and the speed of Muller's ratchet." Genetics 154(3): 1379-87.

Gordo, I. and B. Charlesworth (2000). "On the speed of Muller's ratchet." Genetics 156(4): 2137-40.

Haigh, J. (1978). "The accumulation of deleterious genes in a population - Muller's Ratchet." Theoretical Population Biology 14: 251-267.

Hasegawa, M., Y. Cao, et al. (1998). "Preponderance of slightly deleterious polymorphism in mitochondrial DNA: Nonsynonymous/synonymous rate ratio is much higher within species than between species." Molecular Biology and Evolution 15(11): 1499-1505.

Howell, N., I. Kubacka, et al. (1996). "How rapidly does the human mitochondrial genome evolve?" American Journal of Human Genetics 59: 501-509.

Kibota, T. T. and M. Lynch (1996). "Estimate of the genomic mutation rate deleterious to overall fitness in E. coli." Nature 381(6584): 694-696.

Lan, R. and P. R. Reeves (1996). "Gene transfer is a major factor in bacterial evolution." Molecular Biology and Evolution 13(1): 47-55.

LeClerc, J. E., B. Li, et al. (1996). "High mutation frequencies among Escherichia coli and Salmonella pathogens." Science 274(5290): 1208-1211.

Lynch, M., J. Blanchard, et al. (1999). "Perspective: Spontaneous deleterious mutation." Evolution . June, 1999; 53(3): 645-663.

Meyer, S., G. Weiss, et al. (1999). "Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA." Genetics 152(3): 1103-10.

Moran, N. A. (1996). "Accelerated evolution and Muller's rachet in endosymbiotic bacteria." Proceedings of the National Academy of Sciences of the United States of America 93(7): 2873-2878.

Ohta, T. (1992). "The nearly neutral theory of molecular evolution." Annual Review of Ecology and Systematics 23: 263-286.

Parsons, T. J. and M. M. Holland (1998). "Mitochondrial mutation rate revisited: hot spots and polymorphism - Response." Nature Genetics 18(2): 110-110.

Parsons, T. J., D. S. Muniec, et al. (1997). "A high observed substitution rate in the human mitochondrial DNA control region." Nature Genetics 15(4): 363-368.

Siguroardottir, S., A. Helgason, et al. (2000). "The mutation rate in the human mtDNA control region." American Journal of Human Genetics 66(5): 1599-1609.

Stoneking, M., S. T. Sherry, et al. (1992). "New approaches to dating suggest a recent age for human mtDNA ancestor." Philosophical Transactions of the Royal Society of London Series B Biological Sciences 337: 167-175.

Global Computing References:

Pearson, K. (2001). Internet-based Distributed Computing Projects. http://www.aspenleaf.com/distributed/.

Loewe, L.(2001). evolution@home. http://www.evolutionary-research.net

Glossary

Background selection occurs when many deleterious mutations are selectively removed from a chromosome. The important effect is that linked neutral mutations are removed with them.

Eco-ratchet[121]. This concept describes the repeated loss of temporarily unused but otherwise functional genes in slowly changing environments. If deleterious mutations knock out a gene that is not under selection in the current environment, but confers a benefit in an upcoming environment, then the eco-ratchet has clicked round a notch.

Effective neutrality describes mutations that do have an advantageous or deleterious effect, but the effect is too small for selection at the current population size. These mutations do accumulate like really neutral mutations. The border with effective neutrality is given by 1/effective population size.

Effective population size is a concept used to scale a large variety of natural settings to the Wright-Fisher model (random mating, diploid population of finite size without further structure) for better comparison. To make a complex theory short, it comprises all those individuals that actually contribute genetic material to the next generation.

Error of magnitude[121] is defined as the decadic logarithm of the fraction of an observed over an inferred value. Contrary to the relative error, it has the nice property of being symmetric for upward and downward deviations over many orders of magnitude.

Fitness balance[121] is a way of comparing all processes that might contribute to genomic decay: everything that increases long-term fitness is added on the left, everything that decreases it, is added on the right.

Fitness barrel[121] is a picture of an organism's genome that might eventually lead to a quantitative model of the eco-ratchet (see "A possible projection of the fitness barrel for a hypothetical organism." on page 326).

Divergence is the fraction of different sequence positions found when comparing the alignment of sequences of two evolutionary lines (eg. species). Assuming a date for the last common ancestor gives an approximate divergence rate (since the last common sequence). This is approximately twice the substitution rate in the corresponding lines.

121.These concepts have been newly defined in this work.

Drift. Genetic drift removes effectively neutral mutations from the population, because not all individuals of a generation produce offspring. Thus, the population loses all unique mutations with their death. The smaller the population, the stronger drift, the weaker selection and the larger the mutational effects that are effectively neutral. If mutation rates are not zero, drift leads to the fixation of neutral mutations, because it eventually removes all unmutated alleles.

Epistasis describes interactions between different mutations, where synergistic epistasis increases a mutational effect and antagonistic epistasis decreases it.

Genomic Decay Paradox[121] describes the situation where theoretical considerations suggest that a given evolutionary line should have become extinct already due to accumulation of slightly deleterious mutations, but obviously isn't. See Chapter 21 for definition and Chapter 26 for potential solutions.

Hitch-hiking often occurs, when an advantageous mutation is fixed: Linked mutations (neutral or deleterious) are fixed with it.

Neutral mutations are mutations that have absolutely no effect on any fitness-related trait of an individual, not even if many are combined. The important part of this definition is that the whole genome could be flooded with such mutations without fitness being affected. See also effective neutrality.

Muller's ratchet leads to the accumulation of slightly deleterious mutations, if purifying selection is weak enough. See this work for a review.

Mutation rate paradox[121] describes the observation that long-term phylogenetic substitution rates are often lower than short-term rates of various kinds. See Chapter 3 and 26 for definition and potential solutions.

Slightly deleterious mutations (SDMs) have no obvious effect, but still compromise the fitness of an individual. Can become deadly, if accumulated over the long term.

Substitution rate is the rate of fixed substitutions in a sequence over a given time period for a given evolutionary line (eg. a species). Usually assumed to be half of divergence rate.

Very slightly deleterious mutations (VSDMs) are effectively neutral mutations that can do long-term harm.

Abbreviations

2D 2-dimensional
3D 3-dimensional
API . . . . . Application Programming Interface
b.p. . . . . before present
BCE. . . . . Before Christian Era
bit . . . . . Smallest unit of information: 1x Yes/No
bp . . . . . base pairs
By = Byr . . . . . Billion years (10^9)
Byte . . . . . measure of memory; contains 8 bits;
ca. . . . . = approximately
C++ . . . . . An object-oriented programming language, successor of C
CI . . . . . Confidence interval
CPU . . . . . Central Processing Unit of a computer
CV . . . . . Coefficient of variation (= sd/mean)
Δ . . . . . a difference between two values
d . . . . . days
DB. . . . . DataBase
Diss . . . . . Dissertation
D-loop . . . . . control region of mtDNA
DSR . . . . . Dangerous Range of Selection coefficients (Muller's ratchet threat)
ε . . . . . = EEPSLION, software framework behind evolution@home
EE . . . . . Prefix of all classes in the framework described in this work
eg. . . . . = for example
EoM . . . . . error of magnitude
et al. . . . . and other authors
etc.. . . . . = and so forth with similar examples
FLOP . . . . . FLoating point OPeration = 1 calculation with 2 float numbers
FLOPS . . . . . FLoating point OPerations per Second, a CPU speed measure
GB. . . . . GigaBytes = 1024x1024x1024 Bytes, usually memory on PC
Gbit. . . . . Giga bit = 1024x1024x1024 bit, usually per second network speed
GC . . . . . global computing (=publicly distributed computing over Internet)
gen . . . . . generation
GFLOPS . . . . . Thousand MFLOPS CPU speed
GUI. . . . . Graphical User Interface
h . . . . . hours
HFFDB . . . . . Hierarchical Folder File DataBase
html . . . . . Hyper Text Mark-up Language = format for webpages
http://www... . . . . An Internet address on the world wide web

HV . . . . . Hypervariable region of the D-loop in mtDNA
ibid. . . . . = rest of this reference has been cited above
IBM. . . . . individual-based model
ie. . . . . = that is
InDel. . . . . Insertion or deletion in a sequence (often hard to distinguish)
INDS. . . . . INDividuals per Second, measures performance of ε simulations
KB. . . . . KiloBytes = 1024 Bytes, usually memory on PC
Kbit. . . . . Kilo bit = 1024 bit, usually per second network speed
Kbp . . . . . Kilobasepairs, 1000 bp sequence in genome
Ky = Kyr . . . . . Thousand years
lab . . . . . laboratory
μ . . . . . mean of a normal distribution
μl. . . . . microlitres
ml . . . . . millilitres
max . . . . . maximum of some value
MB . . . . . MegaBytes = 1024x1024 Bytes, usually memory on PC
Mbit . . . . . Mega bit = 1024x1024 bit, usually per second network speed
Mbp . . . . . Megabasepairs, 1000 000 bp sequence in genome
MDSR. . . . . Most Dangerous Range of Selection coefficients (Muller's ratchet threat)
MFLOPS. . . . . Million FLOPS CPU speed
mid . . . . . middle of some value, ie. mean or best guess / estimate
min . . . . . minimum of some value
min . . . . . minutes
MMD . . . . . MisMatch Distribution
MRCA . . . . . Most Recent Common Ancestor
mrr . . . . . multi-run-result (= mean, sd, etc. of many simulation results)
mtDNA. . . . . mitochondrial DNA
My = Myr . . . . . Million years
N. . . . . number of observations
$N_{e,crit}$ . . . . . critical smallest effective population size N_e that does not drive a population to extinction due to Muller's ratchet for a given combination of U and s in the assumed age of the line
netCDF. . . . . A self-descriptive file format for multi-dimensional array data
OS. . . . . operating system
PB . . . . . PetaBytes = 1024x1024x1024x1024x1024 Bytes
PC. . . . . personal computer
PDF. . . . . Portable Document Format of Adobe
Q_{10} . . . . . Factor that describes increase of reaction speed for 10°C increase in temperature
RAM . . . . . Random Access Memory of a computer
RFLP. . . . . Restriction Fragment Length Polymorphism

σ . . . . . standard deviation of a normal distribution
s = sec . . . . . seconds
SAM . . . . . slightly advantageous mutation
sd. . . . . standard deviation of a normal distribution
SDM . . . . . slightly deleterious mutation
SE . . . . . standard error of mean = $(sd)/(\sqrt{observations})$
sec . . . . . seconds
sex . . . . . synonym for regular recombination of genes
SMP . . . . . Symmetric MultiProcessing
srr . . . . . single-run-result (= one simulation result)
StDev . . . . . standard deviation
STL . . . . . Standard Template Library, helps with C++ programming
subst . . . . . substitutions
TB . . . . . TeraBytes = 1024x1024x1024x1024 Bytes
Tcl/Tk. . . . . A script programming language
TS . . . . . TimeSeries
ts . . . . . transition mutation
tv . . . . . transversion mutation
VSAM . . . . . very slightly advantageous mutation
VSDM. . . . . very slightly deleterious mutation
XML . . . . . eXtended Markup Language, great format for data exchange
See also list of symbols in Muller's ratchet theory (Table 6 on page 33).

References

Anonymous (1999) "Complex systems", Science 284:79-109.
Anonymous (1999) "Mitochondria", Science 283:1435-1438+1475-1498.
Anonymous (2001) "Nature insight: Complex systems", Nature 410:240-284.
Accelrys formerly GCG (2002) "Accelrys - Software for pharmaceutical, chemical and materials research" http://www.accelrys.com/
Adam G, Läuger P & Stark G (1988) "Physikalische Chemie und Biophysik". Zweite, völlig neu bearbeitete und erweiterte Auflage, Berlin, Springer-Verlag.
Adamatzky AI (1994) "Identification of cellular automata", Taylor&Francis.
Adami C (1998) "Introduction to Artificial Life", New York, Springer.
Adami C, Brown CT & Ofria C (2002) "Digital Life Laboratory: Software Avida" http://dllab.caltech.edu/avida
Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD et al. (2000) "The genome sequence of Drosophila melanogaster", Science 287:2185-2195.
Akashi H (1995) "Inferring weak selection from patterns of polymorphism and divergence at "Silent" sites in Drosophila DNA", Genetics 139:1067-1076.
Akashi H (1997) "Distinguishing the effects of mutational biases and natural selection on DNA sequence variation", Genetics 147:1989-1991.
Akashi H & Schaeffer SW (1997) "Natural selection and the frequency distributions of "silent" DNA polymorphism in Drosophila", Genetics 146:295-307.
Akashi H, Kliman RM & EyreWalker A (1998) "Mutation pressure, natural selection, and the evolution of base composition in Drosophila", Genetica 103:49-60.
Akashi H (1999) "Within- and between-species DNA sequence variation and the 'footprint' of natural selection", Gene 238:39-51.
Akashi H (1999) "Inferring the fitness effects of DNA mutations from polymorphism and divergence data: Statistical power to detect directional selection under stationarity and free recombination", Genetics 151:221-238.
Alberts B (1998) "The cell as a collection of protein machines: preparing the next generation of molecular biologists", Cell 92:291-294.
Allman J, Rosin A, Kumar R & Hasenstaub A (1998) "Parenting and survival in anthropoid primates: caretakers live longer", Proc. Natl. Acad. Sci. USA 95:6866-6869.
Alon U, Surette MG, Barkai N & Leibler S (1999) "Robustness in bacterial chemotaxis", Nature 397:168-171.
Alstad D (2002) "POPULUS" http://www.cbs.umn.edu/populus/
Amann RI, Ludwig W & Schleifer KH (1995) "Phylogenetic Identification and In Situ Detection of Individual Microbial Cells without Cultivation", Microbiol. Rev. 59:143-169.
Andersson DI & Hughes D (1996) "Muller's ratchet decreases fitness of a DNA-based microbe", Proc. Natl. Acad. Sci. USA 93:906-907.
Arctander P, Kat PW, Aman RA & Siegismund HR (1996) "Extreme genetic differences among populations of Gazella granti, Grant's gazelle, in Kenya", Heredity 76:465-475.
Aris-Brosou S & Excoffier L (1996) "The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism." Mol. Biol. Evol. 13:494-504.
Ashwood-Smith MJ (1965) "On the genetic stability of bacteria to freezing and thawing", Cryobiology 2:39-43.
Ashwood-Smith MJ & Grant E (1976) "Mutation induction in bacteria by freeze-drying", Cryobiology 13:206-213.
Ashwood-Smith MJ (1985) "Genetic damage is not produced by normal cryopreservation procedures involving either glycerol or dimethyl sulfoxide: a cautionary note, however, on possible effects of dimethyl sulfoxide", Cryobiology 22:427-433.
Awadalla P, Eyre-Walker A & Smith JM (1999) "Linkage disequilibrium and recombination in hominid mitochondrial DNA", Science 286:2524-2525.
Baake E (1995) "Diploid models on sequence space", Journal of Biological Systems 3:343-349.
Baake E & Wiehe T (1997) "Bifurcations in haploid and diploid sequence space models", J. math. Biol. 35:321-343.
Baake E & Gabriel W (1999) "Biological evolution through mutation, selection and drift: An introductory review", Annual Reviews of Computational Physics 7:203-264.
Baati L, Fabre-Gea C, Auriol D & Blanc PJ (2000) "Study of the cryotolerance of Lactobacillus acidophilus: effect of culture and freezing conditions on the viability and cellular protein levels", Int. J. Food. Microbiol. 59:241-247.
Bäck T (1996) "Evolutionary algorithms in theory and practice", New York, Oxford University Press.
Baigent S (2001) "Special issue: Modelling cell systems", Brief. Bioinformat. 2:221-288.
Bailey JF, Richards MB, Macaulay VA, Colson IB, James IT, Bradley DG, Hedges RE & Sykes BC (1996) "Ancient DNA suggests a recent expansion of European cattle from a diverse wild progenitor species", Proc. R. Soc. Lond. B Biol. Sci.

263:1467-1473.
Ball GL (1994) "Ecosystem modeling with GIS", Environmental Management 18:345-349.
Banke K (2002) "Open-Source-RDBMS von SAP: Erfaphrungswerte", iX:78-80.
Baquero F & Blazquez J (1997) "Evolution of antibiotic resistance", Trends Ecol. Evol. 12:482-487.
Barclay T, Eberl R, Gray J, Nordlinger J, Raghavendran G, Slutz D, Smith G, Smoot P, Hoffman J, Robb III N, Rossmeissl H, Duff B, Lee G, Mathesmier T, Sunne R, Stivers LA & Goodman K (1998) "The Microsoft TerraServer™" http://research.microsoft.com/~gray/Papers/MSR_TR_98_17_TerraServer.pdf
Barclay T, Gray J & Slutz D (2000) "Microsoft TerraServer: A Spatial Data Warehouse". Proceedings of the ACM SIGMOD, Austin, TX, ACM.
Barkai N & Leibler S (1997) "Robustness in simple biochemical networks", Nature 387:913-917.
Barlund I & Tattari S (2001) "Ranking of parameters on the basis of their contribution to model uncertainty", Ecol. Model. 142:11-23.
Bart J (1995) "Acceptance criteria for using individual-based models to make management decisions", Ecol. Applic. 5:411-420.
Barton JJ & Nackman LR (1994) "Scientific and Engineering C++: An Introduction with Advanced Techniques and Examples", Reading, Massachusetts, Addison-Wesley.
Barton NH & Charlesworth B (1998) "Why sex and recombination?" Science 281:1986-1990.
Baumann P, Baumann L, Lai CY, Rouhbakhsh D, Moran NA & Clark MA (1995) "Genetics, physiology, and evolutionary relationships of the genus Buchnera: intracellular symbionts of aphids", Annu. Rev. Microbiol. 49:55-94.
Baumgartner JW & Hazelbauer GL (1996) "Mutational analysis of a transmembrane segment in a bacterial chemoreceptor", J. Bacteriol. 178:4651-4660.
Baxevanis A & Quelette BFF, (eds, 2001) "Bioinformatics: A practical guide to the analysis of genes and proteins". 2nd, New York, John Wiley & Sons.
Begon M, Harper JL & Townsend CR (1996) "Ecology: Individuals, populations and communities". 3rd, Oxford, Blackwell Science.
Bell G (1982) "The Masterpiece of Nature: The Evolution and Genetics of Sexuality", Berkeley, University of California Press.
Bell G (1988) "Recombination and the immortality of the germ line", J. evol. Biol. 1:67-82.
Bennett AF, Dao KM & Lenski RE (1990) "Rapid evolution in response to high-temperature selection", Nature 346:79-81.
Bennett AF, Lenski RE & Mittler JE (1992) "Evolutionary adaptation to temperature: I. Fitness responses of Escherichia coli to changes in its thermal environment", Evolution 46:16-30.
Bennett AF & Lenski RE (1993) "Evolutionary adaptation to temperature: II. Thermal niches of experimental lines of Escherichia coli", Evolution 47:1-12.
Bennett AF & Lenski RE (1997) "Evolutionary adaptation to temperature: VI. Phenotypic acclimation and its evolution in Escherichia coli", Evolution 51:36-44.
Benton D (1996) "Bioinformatics. Principles and potential of a new multidisciplinary tool", Trends Biotechnol. 14:261-272.
Benton TG & Grant A (2000) "Evolutionary fitness in ecology: Comparing measures of fitness in stochastic, density-dependent environments", Evol. Ecol. Res. 2:769-789.
Benz J & Knorrenschild M (1997) "Call for a common model documentation etiquette", Ecol. Model. 97:141-143.
Benz J, Hoch R & Legovic T (2001) "ECOBAS - modelling and documentation", Ecol. Model. 138:3-15.
Berchtold S, Böhm C & Kriegel H-P (1998) "The Pyramid-Technique: Towards breaking the curse of dimensionality". Proc. Int. Conf. on Management of Data, ACM SIGMOD, Seattle, Washington.
Berchtold S, Böhm C, Jagadish HV, Kriegel H-P & Sander J (2000) "Independent Quantization: An index compression technique for high-dimensional data spaces". 16th International Conference on Data Engineering (ICDE), San Diego, CA.
Berendsen HJC (1998) "Protein folding - A glimpse of the holy grail?" Science 282:642-643.
Bergstrom CT & Pritchard J (1998) "Germline bottlenecks and the evolutionary maintenance of mitochondrial genomes", Genetics 149:2135-2146.
Berkeley Madonna (2002) "Berkeley Madonna Homepage" http://www.berkeleymadonna.com
Berman F, Wolski R, Figueira S, Schopf J & Shao G (1996) "Application-Level Scheduling on distributed heterogeneous networks (Technical Paper)". Proceedings of Supercomputing '96.
Berman F (2002) "Application Level Scheduling on the Computational Grid" http://apples.ucsd.edu
Bernardes AT (1996) "Mutation load and the extinction of large populations", Physica A 230:156-173.
Bernstein C & Bernstein H (1991) "Aging, Sex, and DNA Repair", San Diego, Academic Press.
Beukeboom LW, Weinzierl RP & Michiels NK (1995) "Amazon molly and Muller's Ratchet", Nature 375:111-112.
Beukeboom LW & Vrijenhoek RC (1998) "Evolutionary genetics and ecology of sperm-dependent parthenogenesis", J. evol. Biol. 11:755-782.
Beukeboom LW & van Batenburg FHD (1999) "The effect of paternal leakage on the rate of Muller's Ratchet". 7th Conference of the European Society for Evolutionary Biology, Barcelona, Spain.
Beveridge J (2000) "Transporting Data with XML: The C++ problem" http://www.xmlmag.com/upload/free/features/xml/2000/03sum00/jb0300/jb0300.asp
Birky CW, Jr. (1991) "Evolution and population genetics of organelle genes: Mechanisms and models", pp. 112-134 in:

Selander RK, Clark AG & Whittam TS (eds) Evolution at the Molecular Level, Sunderland, MA, Sindauer Associates, Inc.
Bisby FA (2000) "The quiet revolution: biodiversity informatics and the internet", Science 289:2309-2312.
Bjornstad ON & Grenfell BT (2001) "Noisy clockwork: time series analysis of population fluctuations in animals", Science 293:638-643.
Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, ColladoVides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B & Shao Y (1997) "The complete genome sequence of Escherichia coli K-12", Science 277:1453-1462.
Bleul A (2001) "Programmier-Extremisten: Softwareprojekte auf den Kern reduziert", c't:182-185.
Boerlijst MC, Bonhoeffer S & Nowak MA (1996) "Viral quasi-species and recombination", Proc. R. Soc. Lond. B Biol. Sci. 263:1577-1584.
Boerner GV, Yokobori SI, Moerl M, Doerner M & Paabo S (1997) "RNA editing in metazoan mitochondria: Staying fit without sex", FEBS Lett. 409:320-324.
Boguski MS (1998) "Bioinformatics - a new era", Trends Guide to Bioinformatics Supplement 1998:1-3.
Bonatto SL & Salzano FM (1997) "A single and early migration for the peopling of the Americas supported by mitochondrial DNA sequence data", Proc. Natl. Acad. Sci. USA 94:1866-1871.
Bonhoeffer S & Stadler PF (1993) "Error thresholds on correlated fitness landscapes", J. theor. Biol. 164:359-372.
Booch G (1994) "Object-oriented analysis and design. With applications". 2nd, Benjamin / Cummings Publishing Company.
Bossel H (1994) "Modeling and Simulation", Wellesley, MA, A. K. Peters.
Bosworth A (2002) "XML Magazine - END TAG XML Repositories: Do we need them?" XML Magazine.
Bowie JU, Reidhaar-Olson JF, Lim WA & Sauer RT (1990) "Deciphring the message in protein sequences: tolerance to amino acid substitutions", Science 247:1306-1310.
Box GEP & Cox DR (1964) "An analysis of transformations", J. Roy. Stat. Soc. Ser. B 26:211-243.
Boyce MS (1992) "Population viability analysis", Ann. Rev. Ecol. Syst. 23:481-506.
Bradley DG, MacHugh DE, Cunningham P & Loftus RT (1996) "Mitochondrial diversity and the origins of African and European cattle", Proc. Natl. Acad. Sci. U.S.A. 93:5131-5135.
Brenner S & Lewitter F, (eds, 1998) "Trends Guide to Bioinformatics". Trends Supplement 1998, Cambridge, Elsevier Trends Journals.
Brenner S (2000) "Genomics. The end of the beginning", Science 287:2173-2174.
Broadbent JR & Lin C (1999) "Effect of heat shock or cold shock treatment on the resistance of lactococcus lactis to freezing and lyophilization", Cryobiology 39:88-102.
Brommer JE (2000) "The evolution of fitness in life-history theory", Biol. Rev. Camb. Philos. Soc. 75:377-404.
Bronikowski AM, Bennett AF & Lenski RE (2001) "Evolutionary adaptation to temperature. VII. Effects of temperature on growth rate in natural isolates of Escherichia coli and Salmonella enterica from different thermal environments", Evolution 55:33-40.
Brook BW, O'Grady JJ, Chapman AP, Burgman MA, Akcakaya HR & Frankham R (2000) "Predictive accuracy of population viability analysis in conservation biology", Nature 404:385-387.
Brown WM, Prager EM, Wang A & Wilson AC (1979) "Rapid evolution of animal mitochondrial DNA", Proc. Natl. Acad. Sci. USA 76:1967-1971.
Bruegge B & Dutoit AH (2000) "Object-oriented software engeneering: Conquering complex and changing systems", Upper Saddle River, NJ, Prentice Hall.
Bull HJ, McKenzie GJ, Hastings PJ & Rosenberg SM (2000) "Evidence that stationary-phase hypermutation in the Escherichia coli chromosome is promoted by recombination", Genetics 154:1427-1437.
Bull HJ, McKenzie GJ & Rosenberg SM (2000) "Response to John Cairns: The contribution of transiently hypermutable cells to mutation in stationary phase", Genetics 156:925-926.
Burch CL & Chao L (1999) "Evolution by small steps and rugged landscapes in the RNA virus phi 6", Genetics 151:921-927.
Burch CL & Chao L (2000) "Evolvability of an RNA virus is determined by its mutational neighbourhood", Nature 406:625-628.
Bürger R & Hofbauer J (1994) "Mutation load and mutation-selection-balance in quantitative genetic traits", J. math. Biol. 32:193-218.
Butcher D (1995) "Muller's ratchet, epistasis and mutation effects", Genetics 141:431-437.
Butlin R, Schön I & Martens K (1998) "Asexual reproduction in nonmarine ostracods", Heredity 81:473-480.
Butlin RK (2000) "Virgin rotifers", Trends Ecol. Evol. 15:389-390.
Butlin R (2002) "Opinion - evolution of sex: The costs and benefits of sex: new insights from old asexual lineages", Nat Rev Genet 3:311-317.
Byers DL & Waller DM (1999) "Do plant populations purge their genetic load? Effects of population size and mating history on inbreeding depression", Ann. Rev. Ecol. Syst. 30:479-513.
Caballero A (1994) "Developments in the prediction of effective population size", Heredity 73 (Pt 6):657-679.
Caballero A (1995) "On the effective size of populations with separate sexes, with particular reference to sex-linked genes", Genetics 139:1007-1011.
Cairns J & Foster PL (1991) "Adaptive reversion of a frameshift mutation in Escherichia coli", Genetics 128:695-702.

Cairns J (2000) "The contribution of bacterial hypermutators to mutation in stationary phase", Genetics 156:923-923.
Calcott PH & Gargett AM (1981) "Mutagenicity of Freezing and Thawing", FEMS Microbiology Letters 10:151-155.
Calcott PH & Thomas M (1981) "Sensitivity of DNA-Repair Deficient Mutants of Escherichia-Coli to Freezing and Thawing", FEMS Microbiology Letters 12:117-120.
Cano RJ (1994) "Bacillus DNA in amber: A window to ancient symbiotic relationships?" ASM News 60:129-134.
Cano RJ & Borucki MK (1995) "Revival and identification of bacterial spores in 25- to 40-million-year-old Dominican amber", Science 268:1060-1064.
Carpenter SR (2002) "Ecological futures: Building an ecology of the long now", Ecology 83:2069–2083.
Casanova H, Legrand A, Zagorodnov D & Berman F (2000) "Heuristics for scheduling parameter sweep applications in Grid environments". Proceedings of the 9th Heterogeneous Computing Workshop (HCW'2000).
Casti JL (1996) "Confronting science's logical limits", Sci. Am.:78-81.
Caswell H & John AM (1992) "From the individual to the population in demographic models", pp. 36-61 in: DeAngelis DL & Gross LJ (eds) Individual-based models and approaches in ecology, New York, Chapman & Hall.
Cavalli Sforza LL, Menozzi P & Piazza A (1994) "The history and geography of human genes", Princeton, Princeton University Press.
Chao L, Levin BR & Stewart FM (1977) "A complex community in a simple habitat: an experimental study with bacteria and phage." Ecology 58:369-379.
Chao L (1988) "Evolution of sex in RNA viruses", J. theor. Biol. 133:99-112.
Chao L (1990) "Fitness of RNA virus decreased by Muller´s ratchet", Nature 348:454-455.
Chao L (1994) "Evolution of genetic exchange in RNA viruses", pp. 233-250 in: Morse SS (ed) The evolutionary biology of viruses, New York, Raven Press.
Chao L, Tran TT & Tran TT (1997) "The advantage of sex in the RNA Virus f6", Genetics 147:953-959.
Charlesworth B (1990) "Mutation-selection balance and the evolutionary advantage of sex and recombination", Genet. Res. 55:199-222.
Charlesworth B, Charlesworth D & Morgan MT (1990) "Genetic loads and estimates of mutation rates in highly inbred plant populations", Nature 347:380-382.
Charlesworth B (1991) "The evolution of sex chromosomes", Science 251:1030-1033.
Charlesworth D, Morgan MT & Charlesworth B (1992) "The effect of linkage and population size on inbreeding depression due to mutational load", Genet. Res. 59:49-61.
Charlesworth B (1993) "Directional selection and the evolution of sex and recombination", Genet. Res. 61:205-224.
Charlesworth B, Morgan MT & Charlesworth D (1993) "The effect of deleterious mutations on neutral molecular variation", Genetics 134:1289-1303.
Charlesworth D, Morgan MT & Chalesworth B (1993) "Mutation accumulation in finite populations", J. Hered. 84:321-325.
Charlesworth D, Morgan MT & Charlesworth B (1993) "Mutation accumulation in finite outbreeding and inbreeding populations", Genet. Res. 61:39-56.
Charlesworth B (1996) "Open questions: The good fairy godmother of evolutionary genetics", Curr. Biol. 6:220.
Charlesworth B & Charlesworth D (1997) "Rapid fixation of deleterious alleles can be caused by Muller's Ratchet", Genet. Res. 70:63-73.
Charlesworth B & Charlesworth D (2000) "The degeneration of Y chromosomes", Philos. Trans. R. Soc. Lond. B Biol. Sci. 355:1563-1572.
Chesser RK, Rhodes OE, Jr., Sugg DW & Schnabel A (1993) "Effective sizes for subdivided populations", Genetics 135:1221-1232.
Chicurel M (2000) "Mathematical biology. Life is a game of numbers", Nature 408:900-901.
Chow KC & Tung WL (1998) "Overexpression of dnaK/dnaJ and groEL confers freeze tolerance to Escherichia coli", Biochem Biophys Res Commun 253:502-505.
Clark MA, Moran NA & Baumann P (1999) "Sequence evolution in bacterial endosymbionts having extreme base compositions", Mol. Biol. Evol. 16:1586-1598.
Clark JS, Carpenter SR, Barber M, Collins S, Dobson A, Foley JA, Lodge DM, Pascual M, Jr RP, Pizer W, Pringle C, Reid WV, Rose KA, Sala O, Schlesinger WH, Wall DH & Wear D (2001) "Ecological forecasts: an emerging imperative", Science 293:657-660.
Clarke DK, Duarte EA, Elena SF, Moya A, Domingo E & Holland J (1994) "The red queen reigns in the kingdom of RNA viruses", Proc. Natl. Acad. Sci. USA 91:4821-4824.
Clegg JS, Willsie JK & Jackson SA (1999) "Adaptive significance of a small heat shock/a-crystallin protein (p26) in encysted embryos of the brine shrimp, Artemia franciscana", Am. Zoologist 39:836-847.
Clegg JS, Jackson SA & Popov VI (2000) "Long-term anoxia in encysted embryos of the crustacean, Artemia franciscana: viability, ultrastructure and stress proteins", Cell Tissue Res 301:422-446.
Clegg JS (2001) "Cryptobiosis--a peculiar state of biological organization", Comp Biochem Physiol B Biochem Mol Biol 128:613-624.
Clutten-Brock J (1999) "A natural history of domesticated mammals". 2nd, Cambridge, Cambridge University Press.
Clutton-Brock J (1989) "[Cattle domestication]", pp. 200-206 in: Clutton-Brock J (ed) The walking Larder: Patterns of domestication, pastoralism and predation, London, Unwin Hyman Ltd.

Cognitus (2002) "STELLA / ithink Homepage" http://www.cognitus.co.uk/bus-solution.html
Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N, Garnier T, Churcher C, Harris D, Mungall K, Basham D, Brown D, Chillingworth T, Connor R, Davies RM, Devlin K, Duthoy S, Feltwell T, Fraser A, Hamlin N, Holroyd S, Hornsby T, Jagels K, Lacroix C et al. (2001) "Massive gene decay in the leprosy bacillus", Nature 409:1007-1011.
Colston MJ & Hilson GR (1979) "The effect of freezing and storage in liquid nitrogen on the viability and growth of Mycobacterium leprae", J Med Microbiol 12:137-142.
Conery J & Lynch M (1999) "Genetic Simulation Library", Bioinformatics 15:85-86.
Conery J & Lynch M (2002) "Genetic Simulation Library" http://www.csi.uoregon.edu/projects/genetics/GSL/
Contamine V & Picard M (2000) "Maintenance and integrity of the mitochondrial genome: a plethora of nuclear genes in the budding yeast", Microbiol. Mol. Biol. Rev. 64:281-315.
Cooper VS & Lenski RE (2000) "The population genetics of ecological specialization in evolving Escherichia coli populations", Nature 407:736-739.
Cooper VS, Bennett AF & Lenski RE (2001) "Evolution of thermal dependence of growth rate of Escherichia coli populations during 20,000 generations in a constant environment", Evolution 55:889-896.
Coulson T, Mace GM, Hudson E & Possingham H (2001) "The use and abuse of population viability analysis", Trends Ecol. Evol. 16:219-221.
Crow JF & Kimura M (1970) "An Introduction to Population Genetics Theory", Edina, Burgess International Group Incorporated.
Crow JF & Kimura M (1979) "Efficiency of truncation selection", Proc. Natl. Acad. Sci. USA 76:396-399.
Crow JF (1993) "Mutation, mean fitness, and genetic load", pp. 3-42 in: Futuyma D & Antonovics J (eds) Oxford Surveys in Evolutionary Biology. 9, Oxford, Oxford University Press.
Crow JF (1994) "Advantages of sexual reproduction", Developmental Genetics 15:205-213.
Crow JF (1997) "The high spontaneous mutation rate: Is it a health risk?" Proc. Natl. Acad. Sci. USA 94:8380-8386.
Crow JF (1999) "You can know more than you can prove., personal communication.
Crow JF (1999) "The odds of losing at genetic roulette", Nature 397:293-294.
Crow JF (2000) "The origins, patterns and implications of human spontaneous mutation", Nat Rev Genet 1:40-47.
Current Protocols, (ed, 1996) "Current Protocols in Molceular Biology / Current Protocols in Immunology". updated all the time, Green Publishing Associates
John Wiley & Sons.
Cutler DJ (2000) "Understanding the overdispersed molecular clock", Genetics 154:1403-1417.
Cyranoski D (2002) "US and Vietnam join forces to count cost of Agent Orange", Nature 416:252-252.
Czaran T & Bartha S (1992) "Spatiotemporal dynamic models of plant populations and communities", Trends Ecol. Evol. 7:38-42.
Dalton R (2001) "Bilateral Vietnam study plans to assess war fallout of dioxin", Nature 413:442-442.
Dandekar T & Argos P (1996) "Identifying the tertiary fold of small proteins with different topologies from sequence and secondary structure using the genetic algorithm and extended criteria specific for strand regions", J. Mol. Biol. 256:645-660.
Davies EK, Peters AD & Keightley PD (1999) "High frequency of cryptic deleterious mutations in Caenorhabditis elegans", Science 285:1748-1751.
Davies JE & Roberts MC (2002) "Antimicrobial resistance: An ecological perspective", (http://www.asmusa.org/acasrc/pdfs/Colloquia/Antimicrobialrpt.pdf) Washington, D.C., American Academy of Microbiology.
Davis L, (ed, 1991) "Handbook of genetic algorithms", New York, Van Nostrand Reinhold.
de Ferro MIG, de Valladares RER & de Cardenas ILB (1999) "Physiological aspects and conservation of a Veillonella strain isolated from the oral cavity. Interaction with streptococci", Anaerobe 5:255-259.
de Visser JAGM, Hoekstra RF & Van Den Ende H (1997) "An experimental test for synergistic epistasis and its application in Chlamydomonas", Genetics 145:815-819.
de Visser J, Zeyl CW, Gerrish PJ, Blanchard JL & Lenski RE (1999) "Diminishing returns from mutation supply rate in asexual populations", Science 283:404-406.
de Visser JAGM (2002) "The fate of microbial mutators", Microbiology 148:1247-1252.
Dean AM, Dykhuizen DE & Hartl DL (1988) "Fitness effects of amino acid replacements in the beta- galactosidase of Escherichia coli", Mol. Biol. Evol. 5:469-485.
Dean AM (1989) "Selection and neutrality in lactose operons of Escherichia coli", Genetics 123:441-454.
Dean AM & Golding GB (1997) "Protein engineering reveals ancient adaptive replacements in isocitrate dehydrogenase", Proc. Natl. Acad. Sci. USA 94:3104-3109.
Dean AM (1998) "The molecular anatomy of an ancient adaptive event", Am. Sci. 86:26-37.
DeAngelis DL & Gross LJ, (eds, 1992) "Individual-based models and approaches in ecology", New York, Chapman & Hall.
deBoer JG & Glickman BW (1998) "The lacI gene as a target for mutation in transgenic rodents and Escherichia coli", Genetics 148:1441-1451.
Dedeine F, Vavre F, Fleury F, Loppin B, Hochberg ME & Bouletreau M (2001) "Removing symbiotic Wolbachia bacteria specifically inhibits oogenesis in a parasitic wasp", Proc. Natl. Acad. Sci. USA 98:6247-6252.

Denamur E, Bonacorsi S, Giraud A, Duriez P, Hilali F, Amorin C, Bingen E, Andremont A, Picard B, Taddei F & Matic I (2002) "High frequency of mutator strains among human uropathogenic Escherichia coli isolates", J. Bacteriol. 184:605-609.
Denver DR, Morris K, Lynch M, Vassilieva LL & Thomas WK (2000) "High direct estimate of the mutation rate in the mitochondrial genome of Caenorhabditis elegans", Science 289:2342-2344.
Diamond JM (1981) "Flightlessness and fear of flying in island species", Nature 293:507-508.
DKFZ Heidelberg (2002) "HUSAR" http://genome.dkfz-heidelberg.de
DMOZ (2001) "The Open Directory Project" http://dmoz.org
Domingo E & Holland JJ (1997) "RNA virus mutations and fitness for survival", Annu. Rev. Microbiol. 51:151-178.
Domingo E, Webster R & Holland J, (eds, 1999) "Origin and evolution of viruses", San Diego, Academic Press.
Domingo E (2000) "Viruses at the edge of adaptation", Virology 270:251-253.
Donnelly P (1991) "Comment on the growth and stabilization of populations", Stat. Sci. 6:277-279.
Donnelly P & Tavare S (1995) "Coalescents and genealogical structure under neutrality", Annu. Rev. Genet. 29:401-421.
Doolittle RF, Feng DF, Johnson MS & McClure MA (1989) "Origins and evolutionary relationships of retroviruses", Quart. Rev. Biol. 64:1-30.
Doyle B (2000) "Balance Software: Expat XML parser port to Macintosh" http://www.balancesoftware.com/people/brian/code/Expat_and_ParseXMLFile.sit.hqx
Drake JW (1991) "A constant rate of spontaneous mutation in DNA-based microbes", Proc. Natl. Acad. Sci. USA 88:7160-7164.
Drake JW (1993) "Rates of spontaneous mutation among RNA viruses", Proc. Natl. Acad. Sci. USA 90:4171-4175.
Drake JW (1993) "General antimutators are improbable", J. Mol. Biol. 229:8-13.
Drake JW, Charlesworth B, Charlesworth D & Crow JF (1998) "Rates of spontaneous mutation", Genetics 148:1667-1686.
Duarte E, Clarke D, Moya A, Domingo E & Holland J (1992) "Rapid fitness losses in mammalian RNA virus clones due to Muller's ratchet", Proc. Natl. Acad. Sci. USA 89:6015-6019.
Dunning JB, Stewart DJ, Danielson BJ, Noon BR, Root TL, Lamberson RH & Stevens EE (1995) "Spatially explicit population models: current forms and future uses", Ecol. Applic. 5:3-11.
Durrett R & Levin SA (1994) "Stochastic spatial models: a user's guide to ecological applications", Philos. Trans. R. Soc. Lond. B Biol. Sci. 343:329-350.
Dykhuizen D & Hartl DL (1980) "Selective neutrality of 6PGD allozymes in E. coli and the effects of genetic background", Genetics 96:801-817.
Dykhuizen DE & Hartl DL (1983) "Selection in chemostats", Microbiol. Rev. 47:150-168.
Dykhuizen DE, de Framond J & Hartl DL (1984) "Selective neutrality of glucose-6-phosphate dehydrogenase allozymes in Escherichia coli", Mol. Biol. Evol. 1:162-170.
Dykhuizen DE, Dean AM & Hartl DL (1987) "Metabolic flux and fitness", Genetics 115:25-32.
Dykhuizen DE (1990) "Experimental studies of natural selection in bacteria", Ann. Rev. Ecol. Syst. 21:373-398.
Dytham C (1999) "Choosing and using statistics: A biologist's guide", Oxford, Blackwell Science.
Ebert TA (1999) "Plant and animal populations - Methods in demography", San Diego, Academic Press.
Echtle K (1988) "Ansätze zur Fehlertoleranz von Rechensystemen", HMD Handbuch der modernen Datenverarbeitung 25:3-13.
Eddy S (2002) "Sean Eddy Lab Homepage + HMMER" http://HMMER.wustl.edu/
Edwards JL, Lane MA & Nielsen ES (2000) "Interoperability of biodiversity databases: biodiversity information on every desktop", Science 289:2312-2314.
Edwards JS, Ibarra RU & Palsson BO (2001) "In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data", Nat. Biotechnol. 19:125-130.
Eigen M (1971) "Selforganization of matter and the evolution of biological macromolecules", Naturwissenschaften 58:465-523.
Eigen M & Schuster P (1977) "The Hypercycle - A principle of natural self-organization. Part A: Emergence of the Hypercycle", Naturwissenschaften 64:541-565.
Eigen M & Schuster P (1978) "The Hypercycle - A principle of natural self-organization. Part C: The realistic Hypercycle", Naturwissenschaften 65:341-369.
Eigen M & Schuster P (1978) "The Hypercycle - A principle of natural self-organization. Part B: The abstract Hypercycle", Naturwissenschaften 65:7-41.
Eigen M (1993) "The origin of genetic information: viruses as models", Gene 135:37-47.
Elena SF, Cooper VS & Lenski RE (1996) "Punctuated evolution caused by selection of rare beneficial mutations", Science 272:1802-1804.
Elena SF & Lenski RE (1997) "Long-term experimental evolution in Escherichia coli .7. Mechanisms maintaining genetic variability within populations", Evolution 51:1058-1067.
Elena SF & Lenski RE (1997) "Test of synergistic interactions among deleterious mutations in bacteria", Nature 390:395-398.
Elena SF, Davila M, Novella IS, Holland JJ, Domingo E & Moya A (1998) "Evolutionary dynamics of fitness recovery from the debilitating effects of Muller's ratchet", Evolution 52:309-314.
Elena SF, Ekunwe L, Hajela N, Oden SA & Lenski RE (1998) "Distribution of fitness effects caused by random insertion

mutations in Escherichia coli", Genetica 103:349-358.
Ellison AM & Bedford BL (1995) "Response of a wetland vascular plant community to disturbance: A simulation study", Ecol. Applic. 5:109-123.
Elting A & Huber W (2001) "Schnellverfahren: Mit Extreme Programming immer im Plan?" c't:186-191.
EMBnet (2002) "EMBOSS Homepage" http://www.uk.embnet.org/software/EMBOSS
Endy D, You L, Yin J & Molineux IJ (2000) "Computation, prediction, and experimental tests of fitness for bacteriophage T7 mutants with permuted genomes", Proc. Natl. Acad. Sci. USA 97:5375-5380.
Endy D & Brent R (2001) "Modelling cellular behaviour", Nature 409:391-395.
Enea M & Salemi G (2001) "Fuzzy approach to the environmental impact evaluation", Ecol. Model. 136:131-147.
Epstein H (1971) "The origin of the domestic animals of africa", New York, Africana Publishing Corporation.
Epstein H & Mason IL (1984) "Cattle", pp. 6-27 in: Mason IL (ed) Evolution of domesticated animals, London, Longman Group Limited.
Escarmis C, Davila M, Charpentier N, Bracho A, Moya A & Domingo E (1996) "Genetic lesions associated with Muller's ratchet in an RNA virus", J. Mol. Biol. 264:255-267.
ESRI ESRI (1998) "ESRI Shapefile Technical Description. An ESRI White Paper" http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf
European Bioinformatics Institute (2002) "ClustalW" http://www.ebi.ac.uk/clustalw
Ewald PW (1994) "Evolution of mutation rate and virulence among human retroviruses", Philos. Trans. R. Soc. Lond. B Biol. Sci 346:333-343.
Ewens WJ (1979) "Mathematical population genetics", Berlin, Springer Verlag.
Ewens WJ (1989) "The Effective Population Sizes in the Presence of Catastrophes", pp. 9-25 in: Feldman MW (ed) Mathematical evolutionary theory, Princeton, New Jersey, Princeton University Press.
Excoffier L & Schneider S (1999) "Why hunter-gatherer populations do not show signs of pleistocene demographic expansions", Proc. Natl. Acad. Sci. USA 96:10597-10602.
Excoffier L (2002) "Arlequin's home on the web" http://lgb.unige.ch/arlequin
Eyre-Walker A, Gaut RL, Hilton H, Feldman DL & Gaut BS (1998) "Investigation of the bottleneck leading to the domestication of maize", Proc. Natl. Acad. Sci. USA 95:4441-4446.
Eyre-Walker A & Keightley PD (1999) "High genomic deleterious mutation rates in hominids", Nature 397:344-347.
Eyre-Walker A (2000) "Do mitochondria recombine in humans?" Philos. Trans. R. Soc. Lond. B Biol. Sci. 355:1573-1580.
Fedak G, Germain C, Néri V & Cappello F (2001) "XtremWeb: A generic global computing system", pp. 582-588 (http://www.lri.fr/~fedak/XtremWeb/Gcpd.ps) in: Buyya R & Mohay G (eds) 1rst International Workshop on Global and Peer-to-Peer Computing on Large Scale Distributed Systems at the 1rst IEEE/ACM International Symposium on Cluster Computing and the Grid - CCGrid2001, Brisbane, Australia, IEEE Press.
Feil EJ & Spratt BG (2001) "Recombination and the population structures of bacterial pathogens", Annu. Rev. Microbiol. 55:561–590.
Felsenstein J (1974) "The evolutionary advantage of recombination", Genetics 78:737-756.
Felsenstein J (2002) "PHYLIP Homepage" http://evolution.genetics.washington.edu/phylip.html
Fenster CB, Galloway LF & Chao L (1997) "Epistasis and its consequences for the evolution of natural populations", Trends Ecol. Evol. 12:282-286.
Fernández Murga ML, Font de Valdez G & Disalvo AE (2000) "Changes in the surface potential of Lactobacillus acidophilus under freeze-thawing stress", Cryobiology 41:10-16.
Fijalkowska IJ, Dunn RL & Schaaper RM (1993) "Mutants of Escherichia coli with increased fidelity of DNA replication", Genetics 134:1023-1030.
Fijalkowska IJ & Schaaper RM (1993) "Antimutator mutations in the alpha subunit of Escherichia coli DNA polymerase III: Identification of the responsible mutations and alignment with other DNA polymerases", Genetics 134:1039-1044.
Finkel SE & Kolter R (1999) "Evolution of microbial diversity during prolonged starvation", Proc. Natl. Acad. Sci. USA 96:4023-4027.
Finkel SE, Zinser ER & Kolter R (2000) "Long-term survival and evolution in the stationary phase", pp. 231-238 in: Storz G & Hengge-Aronis R (eds) Bacterial stress responses, Washington, D.C., ASM Press.
Fischman J (1995) "Have 25-Million-year-old bacteria returned to life?" Science 268:977-977.
Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb J-F, Dougherty BA, Merrick JM, McKenney K, Sutton GG, FitzHugh W, Fields C, Gocayne JD, Scott J, Shirley R, Liu L-I, Glodeck A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom E, Cotton MD et al. (1995) "Whole-Genome Random Sequencing and Assembly of Haemophilus influenzae Rd", Science 269:496-512.
Fonseca F, Beal C & Corrieu G (2001) "Operating conditions that affect the resistance of lactic acid bacteria to freezing and frozen storage", Cryobiology 43:189-198.
Foster I & Kesselmann C (1998) "The Grid: Blueprint for a new computing infrastructure", Morgan Kaufmann Publishers.
Foster PL (1998) "Adaptive mutation: Has the unicorn landed?" Genetics 148:1453-1459.
Foster PL (1999) "Mechanisms of stationary phase mutation: a decade of adaptive mutation", Annu Rev Genet 33:57-88.
Foster PL & Rosche WA (1999) "Mechanisms of mutation in nondividing cells. Insights from the study of adaptive mutation in Escherichia coli", Ann N Y Acad Sci 870:133-145.

Foster JA (2001) "Evolutionary computation", Nat. Rev. Genet. 2:428-436.
Fraile A, Escriu F, Aranda MA, Malpica JM, Gibbs AJ & Garciaarenal F (1997) "A century of tobamovirus evolution in an australian population of Nicotiana glauca", J. Virol. 71:8316-8320.
Frankham R & Ralls K (1998) "Conservation biology - Inbreeding leads to extinction", Nature 392:441-442.
Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelly JM, Fritchman JL, Weidman JF, Small KV, Sandusky M, Fuhrmann J, Nguyen D, Saudeck TRUDM, Phillips CA, Merrick HM, Tomp J-F, Bott BADDF, Hu P-C, Lucier TS, Peterson SN, Smith HO et al. (1995) "The Minimal Gene Complement of Mycoplasma genitalium", Science 270:397-403.
Fredrickson JK & Onstott TC (1996) "Microbes deep inside the earth", Sci. Am. August:42-47.
Fredrickson JK & Onstott TC (1996) "Leben im Tiefengestein", Spektrum der Wissenschaft Dezember:66-71.
Fry JD, Keightley PD, Heinsohn SL & Nuzhdin SV (1999) "New estimates of the rates and effects of mildly deleterious mutation in Drosophila melanogaster", Proc. Natl. Acad. Sci. USA 96:574-579.
Fuhrman JA (1999) "Marine viruses and their biogeochemical and ecological effects", Nature 399:541-548.
Fukatsu T (1994) "Endosymbiosis of aphids with microorganisms: A model case of dynamic endosymbiotic evolution", Plant Species Biology 9:145-154.
Funchain P, Yeung A, Stewart JL, Lin R, Slupska MM & Miller JH (2000) "The consequences of growth of a mutator strain of Escherichia coli as measured by loss of function among multiple gene targets and loss of fitness", Genetics 154:959-970.
Funk DJ, Wernegreen JJ & Moran NA (2001) "Intraspecific variation in symbiont genomes: Bottlenecks and the aphid-buchnera association", Genetics 157:477-489.
Gabriel W & Burger R (1992) "Survival of small populations under demographic stochasticity", Theor. Popul. Biol. 41:44-71.
Gabriel W, Lynch M & Bürger R (1993) "Mullers Ratchet and mutational meltdowns", Evolution 47:1744-1757.
Gabriel W & Burger R (2000) "Fixation of clonal lineages under Muller's ratchet", Evolution 54:1116-1125.
Gamma E, Helm R, Johnson R & Vlissides J (1995) "Design Patterns", Reading, Massachusetts, Addison-Wesley.
Gamma E, Helm R, Johnson R & Vlissides J (1996) "Entwurfsmuster. Elemente wiederverwendbarer objectorientierter Software". 1, Bonn, Addison-Wesley.
Garcia-Dorado A (1997) "The rate and effects distribution of viable mutation in Drosophila: Minimum distance estimation", Evolution 51:1130-1139.
Garwin L (1995) "In praise of interdisciplinarity", Nature 376:547.
Gathmann FO & Williams DD (1998) "Inter-site: a new tool for the simulation of spatially realistic population dynamics", Ecol. Model. 113:125-139.
Gavrilets S (1997) "Evolution and speciation on holey adaptive landscapes", Trends Ecol. Evol. 12:307-312.
Gavrilets S, Acton R & Gravner J (2000) "Dynamics of speciation and diversification in a metapopulation", Evolution 54:1493-1501.
Gehring WJ & Ikeo K (1999) "Pax 6: mastering eye morphogenesis and eye evolution", Trends Genet. 15:371-377.
Gerrish PJ & Lenski RE (1998) "The fate of competing beneficial mutations in an asexual population", Genetica 103:127-144.
Gershenfeld N (1999) "The nature of mathematical modeling", Cambridge, Cambridge University Press.
Gessler DDG (1995) "The constraints of finite size in asexual populations and the rate of the ratchet", Genet. Res. 66:241-253.
Gessler DDG & Xu SZ (1998) "An embarrassment of riches: the stochastic generation of beneficial mutations", Genetica 103:145-155.
Gibbons A (1995) "Human evolution: The mystery of humanity's missing mutations", Science 267:35-36.
Gibbons A (1998) "Calibrating the mitochondrial clock", Science 279:28-29.
Gibson S (2001) "The strange tale of the Denial of Service attacks against grc.com" http://media.grc.com:8080/files/grcdos.pdf - http://grc.com/dos/grcdos.htm
Gill P, Ivanov PL, Kimpton C, Piercy R, Benson N, Tully G, Evett I, Hagelberg E & Sullivan K (1994) "Identification of the remains of the Romanov family by DNA analysis", Nat. Genet. 6:130-135.
Gillespie JH (1991) "The Causes of Molecular Evolution", New York, Oxford University Press.
Gillman M & Hails R (1997) "An introduction to ecological modelling - putting practice into theory", Oxford, Blackwell Science.
Giraud A, Matic I, Tenaillon O, Clara A, Radman M, Fons M & Taddei F (2001) "Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut", Science 291:2606-2608.
Giraud A, Radman M, Matic I & Taddei F (2001) "The rise and fall of mutator bacteria", Curr. Opin. Microbiol. 4:582-585.
Giuffra E, Kijas JM, Amarger V, Carlborg O, Jeon JT & Andersson L (2000) "The origin of the domestic pig: independent domestication and subsequent introgression", Genetics 154:1785-1791.
Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H & Oliver SG (1996) "Life with 6000 genes", Science 274:546-567.
Gold T (1992) "The deep, hot biosphere", Proc. Natl. Acad. Sci. USA 89:6045-6049.
Golding B, (ed, 1994) "Non-neutral evolution: Theories and molecular data", New York, Chapman & Hall.
Golding GB & Dean AM (1998) "The structural basis of molecular adaptation", Mol. Biol. Evol. 15:355-369.

Golle P & Mironov I (2001) "Uncheatable distributed computations" http://crypto.Stanford.EDU/~pgolle/papers/distr.ps
Goodfellow JM, (ed, 1995) "Computer modelling in molecular biology", Weinheim, VCH Verlagsgesellschaft.
Goodman MF & Fygenson DK (1998) "DNA polymerase fidelity: From genetics toward a biochemical understanding", Genetics 148:1475-1482.
Goodman CS & Coughlin BC (2000) "Introduction. The evolution of evo-devo biology", Proc. Natl. Acad. Sci. USA 97:4424-4425.
Gordo I & Charlesworth B (2000) "On the speed of Muller's ratchet", Genetics 156:2137-2140.
Gordo I & Charlesworth B (2000) "The degeneration of asexual haploid populations and the speed of Muller's ratchet", Genetics 154:1379-1387.
Gordo I & Charlesworth B (2001) "The speed of Muller's ratchet with background selection, and the degeneration of Y chromosomes", Genet. Res. 78:149-161.
Grant F (2001) "Statistica goes from strength to strength", Scientific Computing World:32-34.
Graur D & Pupko T (2001) "The Permian bacterium that isn't", Mol. Biol. Evol. 18:1143-1146.
Green RF & Noakes DLG (1995) "Is a Little Bit of Sex as Good as a Lot?" J. theor. Biol. 174:87-96.
Greenblatt CL, Davis A, Clement BG, Kitts CL, Cox T & Cano RJ (1999) "Diversity of microorganisms isolated from amber", Microbial Ecology 38:58-68.
Grout B, Morris J & McLellan M (1990) "Cryopreservation and the Maintenance of Cell-Lines", Trends Biotechnol. 8:293-297.
Gustafson JL & Snell QO (1995) "HINT: A new way to measure computer performance". Proceedings of the 28th Hawaii International Conference on System Sciences, Wailea, Maui, Hawaii.
Gustafson JL & Todi R (1998) "Conventional benchmarks as a sample of the performance spectrum". Hawaii International Conference on System Sciences, http://www.scl.ameslab.gov/Publications/HICSS98/HICSS98.pdf.
Gustafson JL, Snell Q & Todi R (2001) "HINT Web site" http://www.scl.ameslab.gov/HINT
Hadjimarcou MI, Kokoska RJ, Petes TD & Reha-Krantz LJ (2001) "Identification of a mutant DNA polymerase delta in Saccharomyces cerevisiae with an antimutator phenotype for frameshift mutations", Genetics 158:177-186.
Haefner JW (1992) "Parallel computers and individual-based models: An overview", pp. 126-164 in: DeAngelis DL & Gross LJ (eds) Individual-based models and approaches in ecology, New York, Chapman & Hall.
Hagen JB (2000) "The origins of bioinformatics", Nat. Rev. Genet. 1:231-236.
Haigh J (1978) "The accumulation of deleterious genes in a population - Muller's Ratchet", Theor. Popul. Biol. 14:251-267.
Haldane JBS (1937) "The effect of variation on fitness", Am. Nat. 71:337-349.
Hall BG (1994) "On alternatives to selection-induced mutation in the Bgl operon of Escherichia coli", Mol. Biol. Evol. 11:159-168.
Hall BG (2001) "Predicting Evolutionary Potential. I. Predicting the Evolution of a Lactose-PTS System in Escherichia coli", Mol. Biol. Evol. 18:1389-1400.
Hall BG (2001) "Phylogenetic trees made easy: A how-to manual for molecular biologists", Sinauer Associates.
Hallam TG, Trawick TL & Wolff WF (1996) "Modeling effects of chemicals on a population: Application to a wading bird nesting colony", Ecol. Model. 92:155-178.
Halvorson HO & Monroy A, (eds, 1985) "The Origin and Evolution of Sex", New York, Alan R. Liss Inc.
Hanski I & Gilpin ME, (eds, 1997) "Metapopulation biology: Ecology, genetics and evolution", San Diego, Academic Press.
Hanski I (1998) "Metapopulation dynamics", Nature 396:41-49.
Harms V (1998) "Biomathematik Statistik und Dokumentation. Eine leichtverständliche Einführung". 7, Kiel, Germany, Harms Verlag.
Harms U (2001) "Cluster Computing. The Grid: Verteiltes Rechnen im Internet", iX:160-163.
Harpending HC, Sherry ST, Rogers AR & Stoneking M (1993) "The genetic structure of ancient human populations", Curr. Antrophol. 34:483-496.
Harpending HC, Batzer MA, Gurven M, Jorde LB, Rogers AR & Sherry ST (1998) "Genetic traces of ancient demography", Proc. Natl. Acad. Sci. USA 95:1961-1967.
Harris RS, Longerich S & Rosenberg SM (1994) "Recombination in adaptive mutation", Science 264:258-260.
Harrison M & with other contributors (1997) "Tcl/Tk Tools". 1, Cambridge, O'Reilly & Associates.
Harrison M & McLennan M (1998) "Effektiv Tcl/Tk programmieren". 1, Bonn, Addison-Wesley.
Hartl DL & Dykhuizen DE (1981) "Potential For Selection Among Nearly Neutral Allozymes Of 6- Phosphogluconate Dehydrogenase In Escherichia-Coli", Proc. Natl. Acad. Sci. USA 78:6344-6348.
Hartl DL, Dykhuizen DE & Dean AM (1985) "Limits of adaptation: The evolution of selective neutrality", Genetics 111:655-674.
Hartl DL (1989) "The physiology of weak selection", Genome 31:183-189.
Hartl DL, Moriyama EN & Sawyer SA (1994) "Selection intensity for codon bias", Genetics 138:227-234.
Hartl DL & Taubes CH (1996) "Compensatory nearly neutral mutations: Selection without adaptation", J. theor. Biol. 182:303-309.
Hasegawa M, Cao Y & Yang ZH (1998) "Preponderance of slightly deleterious polymorphism in mitochondrial DNA: Nonsynonymous/synonymous rate ratio is much higher within species than between species", Mol. Biol. Evol. 15:1499-1505.

Hastings IM (1989) "Potential germline competition in animals and its evolutionary implications", Genetics 123:191-198.
Hastings IM (1991) "Germline-selection: Population genetic aspects of the sexual/asexual life cycle", Genetics 129:1167-1176.
Hastings IM (1992) "Population genetic aspects of deleterious cytoplasmic genomes and their effect on the evolution of sexual reproduction", Genet. Res. 59:215-225.
Haubold B & Wiehe T (2001) "Statistics of divergence times", Mol. Biol. Evol. 18:1157-1160.
Hazen RM & Roedder E (2001) "Biogeology. How old are bacteria from the Permian age?" Nature 411:155-156.
Hedges SB, Bogart JP & Maxson LR (1992) "Ancestry of unisexual salamanders", Nature 356:708-710.
Heino M & Hanski I (2001) "Evolution of migration rate in a spatially realistic metapopulation model", Am. Nat. 157:495-511.
Heistermann J (1994) "Genetische Algorithmen. Theorie und Praxis evolutionärer Optimierung", Stuttgart, B.G.Teubner Verlagsgesellschaft.
Herrmann B & Hummel S, (eds, 1994) "Ancient DNA: Recovery and analysis of genetic material from paleontological, archaeological, museum, medical, and forensic specimens", New York, Springer.
Hess K & Philipp W (2001) "Bell's theorem and the problem of decidability between the views of Einstein and Bohr", Proc. Natl. Acad. Sci. USA 98:14228–14233.
Hesse B (1997) "Cattle and oxen", pp. 442-443 in: Meyers EM (ed) The Oxford Encyclopedia of Archaeology in the Near East, New York, Oxford University Press.
Hesse B (1997) "Animal husbandry", pp. 140-143 in: Meyers EM (ed) The Oxford Encyclopedia of Archaeology in the Near East, New York, Oxford University Press.
Hesse B (1997) "Pigs", pp. 347-348 in: Meyers EM (ed) The Oxford Encyclopedia of Archaeology in the Near East, New York, Oxford University Press.
Hiendleder S, Mainz K, Plante Y & Lewalski H (1998) "Analysis of mitochondrial DNA indicates that domestic sheep are derived from two different ancestral maternal sources: no evidence for contributions from urial and argali sheep", J. Hered. 89:113-120.
Higgins K & Lynch M (2001) "Metapopulation extinction caused by mutation accumulation", Proc. Natl. Acad. Sci. USA 98:2928-2933.
Higgs PG (1994) "Error thresholds and stationary mutant distributions in multi-locus diploid genetics models", Genet. Res. 63:63-78.
Higgs PG & Woodcock G (1995) "The accumulation of mutations in asexual populations and the structure of genealogical trees in the presence of selection", J. math. Biol. 33:677-702.
Higgs PG (1998) "Compensatory neutral mutations and the evolution of RNA", Genetica 103:91-101.
Hillis WD (1993) "Why physicists like models and why biologists should", Curr. Biol. 3:79-81.
Holland JH (1975) "Adaptation in Natural and Artificial Systems. An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence", Ann Arbor, The University of Michagan Press.
Holland JH (1992) "Genetic Algorithms", Sci. Am. July:44-50.
Hopkins JC & Leipold RJ (1996) "On the dangers of adjusting the parameters values of mechanism-based mathematical models", J. theor. Biol. 183:417-427.
Horai S, Hayasaka K, Kondo R, Tsugane K & Takahata N (1995) "Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs", Proc. Natl. Acad. Sci. USA 92:532-536.
Horn M, Wagner M, Möller K-D, Schmid EN, Fritsche TR, Schleifer K-H & Michel R (2000) "Neochlamydia hartmannellae gen. nov., sp. nov. (Parachlamydiaceae), an endoparasite of the amoeba Hartmannella vermiformis", Microbiology 146:1231-1239.
Horn M & Wagner M (2001) "Dasein im Verborgenen: Bacterien, die in Acanthamöben leben", Biologie in unserer Zeit 31:160-168.
Houle D, Kondrashov AS, Yampolsky LY, Caldwell S & Steponkus PL (1997) "The effect of cryopreservation on the lethal mutation rate in Drosophila melanogaster", Genet. Res. 69:209-213.
Howell N, Kubacka I & Mackey DA (1996) "How rapidly does the human mitochondrial genome evolve?" Am. J. Hum. Genet. 59:501-509.
Howell N & Smejkal CB (2000) "Persitent heteroplasmy of a mutation in the human mtDNA control region: Hypermutation as an apparent consequence of simple-repeate expansion/contraction", Am. J. Hum. Genet. 66:1589-1598.
Huai HY & Woodruff RC (1997) "Clusters of identical new mutations can account for the "overdispersed" molecular clock", Genetics 147:339-348.
Huai HY & Woodruff RC (1998) "With the correct concept of mutation rate, cluster mutations can explain the overdispersed molecular clock", Genetics 149:467-469.
Huang W, Petrosino J, Hirsch M, Shenkin PS & Palzkill T (1996) "Amino acid sequence determinants of beta-lactamase structure and activity", J. Mol. Biol. 258:688-703.
Huelsenbeck JP, Bull JJ & Cunningham CW (1996) "Combining data in phylogenetic analysis", Trends Ecol. Evol. 11:152-158.
Huelsenbeck JP & Rannala B (1997) "Phylogenetic methods come of age: testing hypotheses in an evolutionary context", Science 276:227-232.

Hugenholtz P, Goebel BM & Pace NR (1998) "Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity", J. Bacteriol. 180:4765-4774.
Huisman J & Weissing FJ (2001) "Fundamental unpredictability in multispecies competition", Am. Nat. 157:488-494.
humanIT (1998) "InfoZoom: Der Knowledge Browser" http://www.humanIT.de/
Humphries HC, Coffin DP & Lauenroth WK (1996) "An individual-based model of alpine plant distributions", Ecol. Model. 84:99-126.
Hurst LD & McVean GT (1996) "Evolutionary Genetics: ... and scandalous symbionts", Nature 381:650-651.
Hurst LD & McVean GT (1998) "Do we understand the evolution of genomic imprinting?" Curr. Opin. Genet. Dev. 8:701-708.
Huston M, Deangelis D & Post W (1988) "New computer models unify ecological theory", BioScience 38:682-691.
Hutchison CA, Peterson SN, Gill SR, Cline RT, White O, Fraser CM, Smith HO & Venter JC (1999) "Global transposon mutagenesis and a minimal Mycoplasma genome", Science 286:2165-2169.
Hütt M-T (2001) "Datenanalyse in der Biologie - Eine Einführung in Methoden der nichtlinearen Dynamik, fraktalen Geometrie und Informationstheorie", Berlin, Springer.
Hyman M & Vaddadi P (2000) "Effektive C++-Techniken", Bonn, Galileo Press.
Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R & Hood L (2001) "Integrated genomic and proteomic analyses of a systematically perturbed metabolic network", Science 292:929-934.
Ihaka R & Gentleman R (1996) "R: A Language for Data Analysis and Graphics", Journal of Computational and Graphical Statistics 5:299-314.
Imhof M & Schlotterer C (2001) "Fitness effects of advantageous mutations in evolving Escherichia coli populations", Proc. Natl. Acad. Sci. USA 98:1113-1117.
Inchausti P & Halley J (2001) "Investigating long-term ecological variability using the Global Population Dynamics Database", Science 293:655-657.
Ingman M, Kaessmann H, Paabo S & Gyllensten U (2000) "Mitochondrial genome variation and the origin of modern humans", Nature 408:708-713.
Innan H & Stephan W (2001) "Selection Intensity Against Deleterious Mutations in RNA Secondary Structures and Rate of Compensatory Nucleotide Substitutions", Genetics 159:389-399.
Inouye T & IBM TPF Development (1999) "Introducing Folders and Pockets - Outcomes Research (FAP-ORES)" http://www-4.ibm.com/software/ts/tpf/news/nv5n4/v5n4a13.htm
Isernhagen R (2000) "Softwaretechnik in C und C++: Modulare, objektorientierte und generische Programmierung". 2. korrigierte, München, Carl Hanser Verlag.
Ivanov PL, Wadhams MJ, Roby RK, Holland MM, Weedn VW & Parsons TJ (1996) "Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij Romanov establishes the authenticity of the remains of Tsar Nicholas II", Nat. Genet. 12:417-420.
Jaarola M & Tegelstrom H (1996) "Mitochondrial DNA variation in the field vole (Microtus agrestis): Regional population structure and colonization history", Evolution 50:2073-2085.
Jacobson I, Christerson M, Jonsson P & Övergaard G (1998) "Object-oriented software engineering". 11th, Harlow, England, ACM press.
Jager HI, Cardwell HE, Sale MJ, Bevelhimer MS, Coutant CC & VanWinkle W (1997) "Modelling the linkages between flow management and salmon recruitment in rivers", Ecol. Model. 103:171-191.
Jazin EE, Cavelier L, Eriksson I, Oreland L & Gyllensten U (1996) "Human brain contains high levels of heteroplasmy in the noncoding regions of mitochondrial DNA", Proc. Natl. Acad. Sci. USA 93:12382-12387.
Jazin E, Soodyall H, Jalonen P, Lindholm E, Stoneking M & Gyllensten U (1998) "Mitochondrial mutation rate revisited: hot spots and polymorphism", Nat. Genet. 18:109-110.
Jenuth JP, Peterson AC, Fu K & Shoubridge EA (1996) "Random genetic drift in the female germline explains the rapid segregation of mammalian mitochondrial DNA", Nat. Genet. 14:146-151.
Jenuth JP, Peterson AC & Shoubridge EA (1997) "Tissue-specific selection for different mtDNA genotypes in heteroplasmic mice", Nat. Genet. 16:93-95.
Josuttis N (1996) "Die C++ Standardbibliothek", Bonn, Addison Wesley.
Judson OP (1994) "The rise of the individual-based model in ecology", Trends Ecol. Evol. 9:9-14.
Judson OP & Normark BB (1996) "Ancient asexual scandals", Trends Ecol. Evol. 11:41-46.
Judson OP & Normark BB (2000) "Evolutionary genetics. Sinless originals", Science 288:1185-1186.
Kai Software (2001) "Kai C++ is the best multi-platform ISO C++ compiler." http://www.kai.com
Kapp OH, Moens L, Vanfleteren J, Trotman CNA, Suzuki T & Vinogradov SN (1995) "Alignment of 700 globin sequences: Extent of amino acid substitution rate and its correlation with variation in volume", Protein Science 4:2179-2190.
Karp PD (1996) "Database links are a foundation for interoperability", Trends Biotechnol. 14:273-279.
Kauffman S (1987) "Towards a general theory of adaptive walks on rugged landscapes", J. theor. Biol. 128:11-46.
Kawata M (1995) "Effective population size in continuously distributed population", Evolution 49:1046-1054.
Kawecki TJ (1998) "Red queen meets Santa Rosalia: Arms races and the evolution of host specialization in organisms with parasitic lifestyles", Am. Nat. 152:635-651.
Keefe AD & Szostak JW (2001) "Functional proteins from a random-sequence library", Nature 410:715-718.

Keightley PD (1994) "The distribution of mutation effects on viability in Drosophila melanogaster", Genetics 138:1315-1322.
Keightley PD (1996) "Nature of deleterious mutation load in Drosophila", Genetics 144:1993-1999.
Keightley PD & Caballero A (1997) "Genomic mutation rates for lifetime reproductive output with lifespan in Caenorhabditis elegans", Proc. Natl. Acad. Sci. USA 94:3823-3827.
Keightley PD (1998) "Inference of genome-wide mutation rates and distributions of mutation effects for fitness traits: A simulation study", Genetics 150:1283-1293.
Keightley PD & Ohnishi O (1998) "EMS-induced polygenic mutation rates for nine quantitative characters in Drosophila melanogaster", Genetics 148:753-766.
Keightley PD & Eyre-Walker A (1999) "Terumi Mukai and the riddle of deleterious mutation rates", Genetics 153:515-523.
Keightley PD & Bataillon TM (2000) "Multigeneration maximum-likelihood analysis applied to mutation-accumulation experiments in Caenorhabditis elegans", Genetics 154:1193-1201.
Keightley PD, Davies EK, Peters AD & Shaw RG (2000) "Properties of ethylmethane sulfonate-induced mutations affecting life-history traits in Caenorhabditis elegans and inferences about bivariate distributions of mutation effects", Genetics 156:143-154.
Keilin D (1959) "The Leeuwenhoek Lecture: The problem of anabiosis or latent life: history and current concept", Proc. R. Soc. Lond. B Biol. Sci. 150:149-191.
Kemp B, Porter SJ & Dawson JF (1998) "Population partitioning in genetic algorithms", Electronics Letters 34:1928-1929.
Kennedy MJ, Reader SL & Swierczynski LM (1994) "Preservation records of micro-organisms: evidence of the tenacity of life", Microbiology 140:2513-2529.
Kibota TT & Lynch M (1996) "Estimate of the genomic mutation rate deleterious to overall fitness in E. coli", Nature 381:694-696.
Kim J, Alizadeh P, Harding T, Hefner-Gravink A & Klionsky DJ (1996) "Disruption of the yeast ATH1 gene confers better survival after dehydration, freezing, and ethanol shock: potential commercial applications", Appl. Environ. Microbiol. 62:1563-1569.
Kim WS, Khunajakr N & Dunn NW (1998) "Effect of cold shock on protein synthesis and on cryotolerance of cells frozen for long periods in Lactococcus lactis", Cryobiology 37:86-91.
Kimura M, Maruyama T & Crow JF (1963) "The mutation load in small populations", Genetics 48:1303-1312.
Kimura M & Maruyama T (1966) "The mutational load with epistatic gene interactions in fitness", Genetics 54:1337-1351.
Kimura M (1983) "The neutral theory of molecular evolution", Cambridge, Cambridge University Press.
Kimura M (1995) "Limitations of Darwinian selection in a finite population", Proc. Natl. Acad. Sci. USA 92:2343-2344.
Kingsolver JG, Hoekstra HE, Hoekstra JM, Berrigan D, Vignieri SN, Hill CE, Hoang A, Gibert P & Beerli P (2001) "The strength of phenotypic selection in natural populations", Am. Nat. 157:245-261.
Kirschner AKT, Eiler A, Zechmeister TC, Velimirov B, Herzig A, Mach R & Farnleitner AH (2002) "Extreme productive microbial communities in shallow saline pools respond immediately to changing meteorological conditions", Environ. Microbiol. 4:546-555.
Klösgen W & Zytkow J, (eds, 2001) "Handbook of data mining and knowledge discovery", Oxford, Oxford University Press.
Koch AL (1997) "Microbial physiology and ecology of slow growth", Microbiol. Mol. Biol. Rev. 61:305-318.
Koehler CM, Lindberg GL, Brown DR, Beitz DC, Freeman AE, Mayfield JE & Myers AM (1991) "Replacement of bovine mitochondrial DNA by a sequence variant within one generation", Genetics 129:247-256.
Kolman CJ, Bermingham E, Cooke R, Ward RH, Arias TD & Guionneau Sinclair F (1995) "Reduced mtDNA diversity in the Ngobe Amerinds of Panama", Genetics 140:275-283.
Komaki K & Ishikawa H (1999) "Intracellular Bacterial Symbionts of Aphids Possess Many Genomic Copies per Bacterium", J. Mol. Evol. 48:717–722.
Kondrashov AS (1988) "Deleterious mutations and the evolution of sexual reproduction", Nature 336:435-440.
Kondrashov AS & Crow JF (1988) "King's formula for the mutation load with epistasis", Genetics 120:853-856.
Kondrashov AS (1993) "Classification of hypotheses on the advantage of amphimixis", J. Hered. 84:372-387.
Kondrashov AS (1994) "Muller's Ratchet under epistatic selection", Genetics 136:1469-1473.
Kondrashov AS (1994) "Mutation load under vegetative reproduction and cytoplasmic inheritance", Genetics 137:311-318.
Kondrashov AS & Houle D (1994) "Genotype-environment interactions and the estimation of the genomic mutation rate in Drosophila melanogaster", Proc. R. Soc. Lond. B Biol. Sci. 258:221-227.
Kondrashov AS (1995) "Contamination of the Genome by Very Slightly Deleterious Mutations: Why Have We Not Died 100 Times Over?" J. theor. Biol. 175:583-594.
Kondrashov AS (1997) "Evolutionary genetics of life cycles", Ann. Rev. Ecol. Syst. 28:391-435.
Koonin EV (2001) "Primer: Computational genomics", Curr. Biol. 11:R155-R158.
Koonin EV, Makarova KS & Aravind L (2001) "Horizontal gene transfer in Procaryotes: Quantification and classification", Annu. Rev. Microbiol. 55:709–742.
Krause RM (1992) "The Origin of Pagues: Old and New", Science 257:1073-1078.
Krebs H (1979) "On asking the right kind of question in biological research". Molecular Mechanisms of Biological Recognition: Proceedings of the Sixth Ahron Katzir-Katchalsky Conference in conjunction with the Minerva Symposia in Biology, Göttingen and Braunlage/Harz, September 24-30, 1978, Elsevier/North-Holland Biomedical Press.
Kreft J-U, Booth G & Wimpenny JWT (1998) "BacSim, a simulator for individual-based modelling of bacterial colony

growth", Microbiology 144:3275-3287.
Kriegel H-P (2001) "Access Methods for High-Dimensional Data Spaces" http://www.dbs.informatik.uni-muenchen.de/Forschung/Index
Kriegel H-P (2001) "Knowledge Discovery in Databases" http://www.dbs.informatik.uni-muenchen.de/Forschung/KKD
Krumholz LR (2000) "Microbial communities in the deep subsurface", Hydrogeology Journal 8:4-10.
Kunz BA, Ramachandran K & Vonarx EJ (1998) "DNA sequence analysis of spontaneous mutagenesis in Saccharomyces cerevisiae", Genetics 148:1491-1505.
Kyrpides N (2002) "GOLD (TM): Genomes OnLine Database Homepage" http://wit.integratedgenomics.com/GOLD/
L'Ecuyer P (1988) "[Long period random number generator]", Communications of the ACM 31:742-774.
Lacy RC & Ballou JD (1998) "Effectiveness of selection in reducing the genetic load in populations of Peromyscus polionotus during generations of inbreeding", Evolution 52:900-909.
Lambert JD & Moran NA (1998) "Deleterious mutations destabilize ribosomal RNA in endosymbiotic bacteria", Proc. Natl. Acad. Sci. USA 95:4458-4462.
Lambert LH, Cox T, Mitchell K, Rosselo-Mora RA, Del Cueto C, Dodge DE, Orkand P & Cano RJ (1998) "Staphylococcus succinus sp. nov., isolated from Dominican amber", International Journal of Systematic Bacteriology 48:511-518.
Lan R & Reeves PR (1996) "Gene transfer is a major factor in bacterial evolution", Mol. Biol. Evol. 13:47-55.
Lande R (1998) "Risk of population extinction from fixation of deleterious and reverse mutations", Genetica 103:21-27.
Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J et al. (2001) "Initial sequencing and analysis of the human genome", Nature 409:860-921.
Langton CG, (ed, 1989) "Artificial Life: The proceedings of an interdisciplinary workshop on the synthesis and simulation of living systems held September 1987 in Los Alamos, New Mexico", Redwood City, California, Addison-Wesley.
Law A & Kelton D (1991) "Simulation modelling and analysis". 2nd., New York, McGraw-Hill.
Lawson J (1999) "Operational code authentication" http://www.distributed.net/source/specs/opcodeauth.html
LeClerc JE, Li B, Payne WL & Cebula TA (1996) "High mutation frequencies among Escherichia coli and Salmonella pathogens", Science 274:1208-1211.
LeClerc JE & Cebula TA (1997) "Highly variable mutation rates in commensal and pathogenic Escherichia coli - Reply", Science 277:1834-1834.
Lee JK, Lascoux M & Nordheim EV (1996) "Number of lethal equivalents in human populations: how good are the previous estimates?" Heredity 77 (Pt 2):209-216.
Lee CH, Gilbertson DL, Novella IS, Huerta R, Domingo E & Holland JJ (1997) "Negative effects of chemical mutagenesis on the adaptive behavior of vesicular stomatitis virus", J. Virol. 71:3636-3640.
Lemmon AR (2002) "EvoTutor - Learning through interactive simulation" http://www.evotutor.org/
Lenski RE & Levin BR (1985) "Constraints on the Coevolution of Bacteria and Virulent Phage - a Model, Some Experiments, and Predictions for Natural Communities", Am. Nat. 125:585-602.
Lenski R (1988) "Dynamics of interactions between bacteria and virulent bacteriophage", Adv. Microb. Ecol. 10:1-44.
Lenski RE (1988) "Experimental studies of pleiotropy and epistasis in Escherichia coli: I. Variation in competitive fitness among mutants resistant to virus T4", Evolution 42:425-432.
Lenski RE (1988) "Experimental studies of pleiotropy and epistasis in Escherichia coli: II. compensation for maladaptive effects associated with resistance to virus T4", Evolution 42:433-440.
Lenski RE, Rose MR, Simpson SC & Tadler SC (1991) "Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2000 generations", Am. Nat. 138:1315-1341.
Lenski RE & Travisano M (1994) "Dynamics of adaptation and diversification: A 10,000-generation experiment with bacterial populations", Proc. Natl. Acad. Sci. USA 91:6808-6814.
Lenski RE, Ofria C, Collier TC & Adami C (1999) "Genome complexity, robustness and genetic interactions in digital organisms", Nature 400:661-664.
Leroi AM, Bennett AF & Lenski RE (1994) "Temperature acclimation and competitive fitness: An experimental test of the beneficial acclimation assumption", Proc. Natl. Acad. Sci. USA 91:1917-1921.
Leroi AM, Lenski RE & Bennett AF (1994) "Evolutionary adaptation to temperature. III. Adaptation of Escherichia coli to a temporally varying environment", Evolution 48:1222-1229.
Levin BR, Stewart FM & Chao L (1977) "Resource limited growth, competition and predation: a model and some experimental studies with bacteria and bacteriophage", Am. Nat. 111:3-24.
Levin SA, Grenfell B, Hastings A & Perlson AS (1997) "Mathematical and computational challenges in population biology and ecosystems science", Science 275:334-343.
Levin BR, Perrot V & Walker N (2000) "Compensatory mutations, antibiotic resistance and the population genetics of adaptive evolution in bacteria", Genetics 154:985-997.
Levins R (1966) "The strategy of model building in population biology", Am. Sci. 54:421-431.
Li WH & Nei M (1977) "Persistence of common alleles in two related populations or species", Genetics 86:901-914.
Li WH (1997) "Molecular evolution", Sunderland, MA, Sinauer Associates Incorporated.
Liberles DA, Schreiber DR, Govindarajan S, Chamberlin SG & Benner SA (2001) "The Adaptive Evolution Database

(TAED)", Genome Biol. 2:RESEARCH0028.
Liberles D (2002) "TAED: The Adaptive Evolution Database Homepage" http://www.sbc.su.se/~liberles/TAED.html
Lim DV (1998) "Microbiology". 2nd ed., Boston, McGraw-Hill.
Lindahl T (1993) "Instability and decay of the primary structure of DNA", Nature 362:709-715.
LION Bioscience (2002) "SRS, Bioscout" http://www.lion-bioscience.com/
Lippman SB (1993) "C++ Primer". 2, Reading, Massachusetts, Addison-Wesley.
Lipson H & Pollack JB (2000) "Automatic design and manufacture of robotic lifeforms", Nature 406:974-978.
Liu JG (2001) "Integrating ecology with human demography, behavior, and socioeconomics: Needs and approaches", Ecol. Model. 140:1-8.
Loeb DD, Swanstrom RS, Everitt L, Manchester M, Stamper SE & Hutchison CA, III (1989) "Complete mutagenesis of the HIV-1 protease", Nature 340:397-400.
Loewe L (1997) "Mutation Rates, Muller's Ratchet, Genetic Load and Eve: Young or Dead?" First international workshop on human mitochondrial DNA, Washington, D.C.
Loewe L & Scherer S (1997) "Mitochondrial Eve: the plot thickens", Trends Ecol. Evol. 12:422-423.
Loewe L & Scherer S (1997) "On the Speed of Genomic Decay". The Fifth Annual Meeting of the Society for Molecular Biology and Evolution, Garmisch-Partenkirchen, Germany.
Loewe L & Scherer S (1998) "Mullers Ratchet in human mitochondrial DNA". The Sixth Annual International Meeting of the Society for Molecular Biology and Evolution, University of British Columbia, Vancouver, Canada.
Loewe L & Scherer S (1998) "On the Speed of Genomic Decay". Annual Meeting of the Society for the Study of Evolution, University of British Columbia, Vancouver, Canada.
Loewe L & Scherer S (1999) "How many beneficial mutations stop Mullers Ratchet?" Evolution '99, University of Wisconsin, Madison, Wisconsin.
Loewe L & Scherer S (1999) "How many beneficial mutations are needed to stop Mullers Ratchet?" Seventh Congress of the European Society for Evolutionary Biology, Barcelona, Spain.
Loewe L (2000) "How many beneficial mutations are needed to stop Mullers Ratchet in mtDNA?" Spacial Ecology Programme: Workshop on Population Extinction, Tvärminne Zoological Station, Finnland, Division of Population Biology, University of Helsinki.
Loewe L & Scherer S (2001) "Predicting extinctions due to Mullers Ratchet in humans and bacteria". 8th Congress of the European Society for Evolutionary Biology, Aarhus, Denmark.
Loewe L (2002) "evolution@home: Experiences with work units that span more than 7 orders of magnitude in computational complexity", pp. 425-431 (http://www.evolutionary-research.net/Science/Papers/2002/Loewe2002-EaHworkunits.pdf) in: Bal HE, Löhr K-P & Reinefeld A (eds) 2nd International Workshop on Global and Peer-to-Peer Computing on Large Scale Distributed Systems at the 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid2002), Berlin, Germany, IEEE Computer Society.
Loewe L (2002) "The mutation rate paradox: Pedigree versus 'archaeological' versus phylogenetic mutation rates in mitochondrial DNA", Under revision.
Loewe L (2002) "The evolution@home website" http://www.evolutionary-research.net
Loewe L (2002) "Global computing for bioinformatics", Brief. Bioinformat. 3:in press.
Loftus RT, MacHugh DE, Bradley DG, Sharp PM & Cunningham P (1994) "Evidence for two independent domestications of cattle", Proc. Natl. Acad. Sci. USA 91:2757-2761.
Loi P, Ptak G, Barboni B, Fulka Jr. J, Cappai P & Clinton M (2001) "Genetic rescue of an endangered mammal by cross-species nuclear transfer using post-mortem somatic cells", Nat. Biotechnol. 19:962 - 964.
Lopez JV, Culver M, Stephens JC, Johnson WE & O'Brien SJ (1997) "Rates of nuclear and cytoplasmic mitochondrial DNA sequence divergence in mammals", Mol. Biol. Evol. 14:277-286.
Luikart G, Gielly L, Excoffier L, Vigne JD, Bouvet J & Taberlet P (2001) "Multiple maternal origins and weak phylogeographic structure in domestic goats", Proc. Natl. Acad. Sci. USA 98:5927-5932.
Luz H & Vingron M (2001) "Online Lectures on Bioinformatics" http://www.dkfz-heidelberg.de/tpi/bioinfo/index.html
Lynch M & Gabriel W (1990) "Mutation load and the survival of small populations", Evolution 44:1725-1737.
Lynch M, Butcher RBD & Gabriel W (1993) "The mutational meltdown in asexual populations", J. Hered. 84:339-344.
Lynch M & Jarrell PE (1993) "A method for calibrating molecular clocks and its application to animal mitochondrial DNA", Genetics 135:1197-1208.
Lynch M, Conery J & Burger R (1995) "Mutational meltdowns in sexual populations", Evolution 49:1067-1080.
Lynch M, Conery J & Bürger R (1995) "Mutation accumulation and the extinction of small populations", Am. Nat. 146:489-518.
Lynch M (1996) "Mutation accumulation in transfer RNAs: Molecular evidence for Muller's ratchet in mitochondrial genomes", Mol. Biol. Evol. 13:209-220.
Lynch M (1997) "Mutation accumulation in nuclear, organelle, and prokaryotic transfer RNA Genes", Mol. Biol. Evol. 14:914-925.
Lynch M & Walsh B (1998) "Genetics and analysis of quantitative traits", Sunderland, Massachusetts, Sinauer Associates.
Lynch M, Blanchard J, Houle D, Kibota T, Schultz S, Vassilieva L & Willis J (1999) "Perspective: Spontaneous deleterious mutation", Evolution 53:645-663.

Lynch M, Pfrender M, Spitze K, Lehman N, Hicks J, Allen D, Latta L, Ottene M, Bogue F & Colbourne J (1999) "The quantitative and molecular genetic architecture of a subdivided species", Evolution 53:100-110.
Lynch M & Conery JS (2000) "The evolutionary fate and consequences of duplicate genes", Science 290:1151-1155.
Lynch M & Force A (2000) "The probability of duplicate gene preservation by subfunctionalization", Genetics 154:459-473.
Maddison DR (2002) "The Tree of Life Homepage" http://phylogeny.arizona.edu/
Maddison W & Maddison D (2002) "Mesquite - A modular system for evolutionary analysis" http://mesquiteproject.org/
Maindonald J & Braun J (2002) "Data analysis and graphics using R: An example-based approach", Cambridge, Cambridge University Press.
Maley CC & Caswell H (1993) "Implementing I-State Configuration Models for Population- Dynamics - an Object-Oriented Programming Approach", Ecol. Model. 68:75-89.
Mallet J (2001) "Holy Landscapes", Science 291:602-602.
Malyshev S (2000) "NetCDF, udunits and CPP for Macintosh" http://crga.atmos.uiuc.edu/~sergey/soft/index.html
Mao EF, Lane L, Lee J & Miller JH (1997) "Proliferation of mutators in a cell population", J. Bacteriol. 179:417-422.
Maple (2002) "Maple Homepage" http://www.maplesoft.com
Marcus AD (2000) "The view from Nebo: How archaeology is rewriting the bible and reshaping the Middle East", Boston, Black Bay Books.
Marean CW (2000) "Anthropology. Age, sex, and old goats", Science 287:2174-2175.
Markiewicz P, Kleina LG, Cruz C, Ehret S & Miller JH (1994) "Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence", J. Mol. Biol. 240:421-433.
Marshall E (1996) "Hot property: Biologists who compute", Science 272:1730-1732.
Martens K, (ed, 1998) "Sex and parthenogenesis: Evolutionary ecology of reproductive modes in non-marine ostracods." Leiden, Netherlands, Backhuys Publishers.
Martens K & Schon I (2000) "Parasites, predators and the Red Queen", Trends Ecol. Evol. 15:392-393.
Martin AP & Palumbi SR (1993) "Body size, metabolic rate, generation time, and the molecular clock", Proc. Natl. Acad. Sci. USA 90:4087-4091.
Martinez MA, Pezo V, Marlière P & Wain-Hobson S (1996) "Exploring the functional robustness of an enzyme by in vitro evolution", EMBO J. 15:1203-1210.
Maruyama T & Kimura M (1980) "Genetic variability and effective population size when local extinction and recolonization of subpopulations are frequent", Proc. Natl. Acad. Sci. USA 77:6710-6714.
Mason IL, (ed, 1984) "Evolution of domesticated animals", London, Longman Group Limited.
Mason IL (1984) "Goat", pp. 85-99 in: Mason IL (ed) Evolution of domesticated animals, London, Longman Group Limited.
Mathworks (2002) "MathLab Homepage" http://www.mathworks.com
Matic I, Radman M, Taddei F, Picard B, Doit C, Bingen E, Denamur E & Elion J (1997) "Highly variable mutation rates in commensal and pathogenic Escherichia coli", Science 277:1833-1834.
Matin A, Auger EA, Blum PH & Schultz JE (1989) "Genetic basis of starvation survival in nondifferentiating bacteria", Annu. Rev. Microbiol. 43:293-316.
Matsumoto M & Nishimura T (1998) "Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator", ACM Transactions on Modeling and Computer Simulation 8:3-30.
Matzke MA & Matzke AJ (2000) "Cloning problems don't surprise plant biologists", Science 288:2318.
Maughan H, Birky Jr CW, Jr., Nicholson WL, Rosenzweig WD & Vreeland RH (2002) "The paradox of the "Ancient" bacterium which contains "Modern" protein- coding genes", Mol. Biol. Evol. 19:1637-1639.
May RM (1976) "Simple mathematical models with very complicated dynamics", Nature 261:459-467.
May RM (1976) "Models for single populations", pp. 4-25 in: May RM (ed) Theoretical ecology: principles and applications, Oxford, Blackwell Scientific Publications.
Maynard Smith J & Haigh J (1974) "The hitchhiking effect of a favorable gene", Genet. Res. 23:23-35.
Maynard Smith J (1978) "Some consequences of sex and recombination - II. Muller's ratchet", pp. 33-36 in: Maynard Smith J (ed) The Evolution of Sex, New York, Cambridge University Press.
Maynard Smith J (1978) "The Evolution of Sex", New York, Cambridge University Press.
Mayr E (1997) "The objects of selection", Proc. Natl. Acad. Sci. USA 94:2091-2094.
McClelland M, Florea L, Sanderson K, Clifton SW, Parkhill J, Churcher C, Dougan G, Wilson RK & Miller W (2000) "Comparison of the Escherichia coli K-12 genome with sampled genomes of a Klebsiella pneumoniae and three salmonella enterica serovars, Typhimurium, Typhi and Paratyphi", Nucleic Acids Res. 28:4974-4986.
McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P, Courtney L, Porwollik S, Ali J, Dante M, Du F, Hou S, Layman D, Leonard S, Nguyen C, Scott K, Holmes A, Grewal N, Mulvaney E, Ryan E, Sun H, Florea L, Miller W, Stoneking T, Nhan M, Waterston R et al. (2001) "Complete genome sequence of Salmonella enterica serovar Typhimurium LT2", Nature 413:852-856.
McDonald K & Sun DW (1999) "Predictive food microbiology for the meat industry: a review", Int. J. Food. Microbiol. 52:1-27.
McGlade J, (ed, 1999) "Advanced Ecological Theory", Oxford, Blackwell Science.
McKenzie GJ & Rosenberg SM (2001) "Adaptive mutations, mutator DNA polymerases and genetic change strategies of

pathogens", Curr. Opin. Microbiol. 4:586-594.
Meffe GK & Carroll CR, (eds, 1997) "Principles of conservation biology". 2nd, Sunderland, MA, Sinauer Associates.
Mehl H (1994) "Methoden verteilter Simulation", Braunschweig, Germany, Vieweg.
Meinl P (1988) "Fehlertolerante Anwendungssysteme - Möglichkeiten und Probleme", HMD Handbuch der modernen Datenverarbeitung 25:14-23.
Meir E (1996) "EcoBeaker 1.0 - An ecological simulation program. EcoBeaker Laboratory Guide and the EcoBeaker Program Manual", Sunderland, MA, Sinauer Associates, Inc.
Meir E (2002) "Welcome to Ecobeaker 2.0 - Ecology teaching software" http://www.ecobeaker.com/
Melzer AL & Koeslag JH (1991) "Mutations do not accumulate in asexual isolates capable of growth and extinction: Muller's Ratchet re-examined", Evolution 45:649-655.
Mesquite Software Inc. (2001) "CSIM18: The simulation engine preferred by systems designers, programmers and educators" http://www.mesquite.com
Meyer S, Weiss G & von Haeseler A (1999) "Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA", Genetics 152:1103-1110.
Meyers S (1995) "Effektiv C++ programmieren". 2. korrigierte, Bonn, Addison-Wesley.
Michalewicz Z (1992) "Genetic Algorithms + Data Structures = Evolution Programs", Berlin, Springer-Verlag.
Michod RE & Levin Bruce R, (eds, 1988) "The Evolution of Sex: An Examination of Current Ideas", Sunderland, Sinauer Associates Incorporated.
Miller JH (1996) "Spontaneous mutators in bacteria: Insights into pathways of mutagenesis and repair", Annu. Rev. Microbiol. 50:625-643.
Miller JH, Suthar A, Tai J, Yeung A, Truong C & Stewart JL (1999) "Direct selection for mutators in Escherichia coli", J. Bacteriol. 181:1576-1584.
Miller JH, Funchain P, Clendenin W, Huang T, Nguyen A, Wolff E, Yeung A, Chiang JH, Garibyan L, Slupska MM & Yang H (2002) "Escherichia coli Strains (ndk) Lacking Nucleoside Diphosphate Kinase Are Powerful Mutators for Base Substitutions and Frameshifts in Mismatch-Repair-Deficient Strains", Genetics 162:5-13.
Mindell DP (1991) "Aligning DNA sequences: Homology and phylogenetic weighting", pp. 73-89 in: Miyamoto MM & Cacraft J (eds) Phylogenetic analysis of DNA sequences, New York, Oxford University Press.
Misawa K & Tajima F (1997) "Estimation of the amount of DNA polymorphism when the neutral mutation rate varies among sites", Genetics 147:1959-1964.
Mitchell M & Taylor CE (1999) "Evolutionary computation: An overview", Ann. Rev. Ecol. Syst. 30:593-616.
Mitchison A (1993) "Will We Survive? As host and pathogen evolve together, will the immune system retain the upper hand?" Sci. Am. September:102-108.
MITOMAP (1999) "MITOMAP Mitochondrial Genomics Database" http://www.gen.emory.edu/mitomap.html
Mitrani I (1982) "Simulation techniques for discrete event systems", Cambridge, Cambridge University Press.
Miyamoto MM & Cacraft J (1991) "Phylogenetic analysis of DNA sequences", New York, Oxford University Press.
Mode CJ (1985) "Stochastic processes in demography and their computer implementation", Berlin, Springer-Verlag.
Modelkinetix (2002) "ModelMaker Homepage" http://www.modelkinetix.com/modelmaker
Molofsky J (1994) "Population dynamics and pattern formation in theoretical populations", Ecology 75:30-39.
Mongold JA, Bennett AF & Lenski RE (1996) "Evolutionary adaptation to temperature: IV. Adaptation of Escherichia coli at a niche boundary", Evolution 50:35-43.
Mongold JA, Bennett AF & Lenski RE (1999) "Evolutionary adaptation to temperature. VII. Extension of the upper thermal limit of Escherichia coli", Evolution 53:386-394.
Mooij WM & Boersma M (1996) "An object-oriented simulation framework for individual-based simulations (OSIRIS): Daphnia population dynamics as an example", Ecol. Model. 93:139-153.
Moran NA, Munson MA, Baumann P & Ishikawa H (1993) "A molecular click in endosymbiotic bacteria is calibrated using the insect hosts", Proc. R. Soc. Lond. B Biol. Sci. 253:167-171.
Moran NA (1996) "Accelerated evolution and Muller's rachet in endosymbiotic bacteria", Proc. Natl. Acad. Sci. USA 93:2873-2878.
Morey DF (1994) "The early evolution of the domestic dog", Am. Sci. 82:336-347.
Morita RY (2000) "Is H2 the universal engery source for long-term survival?" Microbial Ecology 38:307-320.
Morse SS, (ed, 1994) "The evolutionary biology of viruses", New York, Raven Press.
Mourier T, Hansen AJ, Willerslev E & Arctander P (2001) "The human genome project reveals a continuous transfer of large mitochondrial fragments to the nucleus", Mol. Biol. Evol. 18:1833-1837.
Moxon ER, Rainey PB, Nowak MA & Lenski RE (1994) "Adaptive evolution of highly mutable loci in pathogenic bacteria", Curr. Biol. 4:24-33.
Moya A, Rodriguez Cerezo E & Garcia Arenal F (1993) "Genetic structure of natural populations of the plant RNA virus tobacco mild green mosaic virus", Mol. Biol. Evol. 10:449-456.
Mukai T (1964) "The genetic structure of natural populations of Drosophila melanogaster. I. Spontaneous mutation rate of polygenes controlling viability", Genetics 50:1-19.
Muller HJ (1950) "Our load of mutations", Am. J. Hum. Genet. 2:111-176.
Muller HJ (1964) "The relation of recombination to mutational advance", Mut. Res. 1:2-9.

Murray BG (1997) "Population Dynamics of Evolutionary Change: Demographic Parameters as Indicators of Fitness", Theor. Popul. Biol. 51:180-184.
Nachman MW & Crowell SL (2000) "Estimate of the mutation rate per nucleotide in humans", Genetics 156:297-304.
Najjar YM, Basheer IA & Hajmeer MN (1997) "Computational neural networks for predictive microbiology: I. Methodology", Int. J. Food. Microbiol. 34:27-49.
Nakatsu CH, Korona R, Lenski RE, DeBruijn FJ, Marsh TL & Forney LJ (1998) "Parallel and divergent genotypic evolution in experimental populations of Ralstonia sp." J. Bacteriol. 180:4325-4331.
National Center for Biotechnology Information (2002) "Entrez - PubMed" http://www.ncbi.nlm.nih.gov/entrez
Nealson K & Ghiorse WC (2001) "Geobiology: Exploring the interface between the biosphere and the geosphere", (http://www.asmusa.org/acasrc/pdfs/Colloquia/geobiology.pdf) Washington, D.C., American Academy of Microbiology.
Nei M (1992) "Age of the Common Ancestor of Human Mitochondrial DNA", Mol. Biol. Evol. 9:1176-1178.
Nei M & Takahata N (1993) "Effective population size, genetic diversity, and coalescence time in subdivided populations", J. Mol. Evol. 37:240-244.
Nelson C (2000) "Tcl/Tk Programmer's Reference". 1, Berkeley, CA, Osbourne/McGraw-Hill.
Nichol S (1996) "RNA Viruses: Life on the edge of catastrophe", Nature 384:218-219.
Nicholson WL, Munakata N, Horneck G, Melosh HJ & Setlow P (2000) "Resistance of Bacillus endospores to extreme terrestrial and extraterrestrial environments", Microbiol. Mol. Biol. Rev. 64:548-572.
NMSR (2002) "About SIMEX" http://www.nmsr.labmed.umn.edu/nmsr/simex/
Nobis G (1979) "Der älteste Haushund lebte vor 14 000 Jahren", Umschau 79:610.
Normark BB (1999) "Evolution in a putatively ancient asexual aphid lineage: Recombination and rapid karyotype change", Evolution 53:1458-1469.
Novella IS, Duarte EA, Elena SF, Moya A, Domingo E & Holland JJ (1995) "Exponential increases of RNA virus fitness during large population transmissions", Proc. Natl. Acad. Sci. USA 92:5841-5844.
Novella IS, Elena SF, Moya A, Domingo E & Holland JJ (1995) "Size of genetic bottlenecks leading to virus fitness loss is determined by mean initial population fitness", J. Virol. 69:2869-2872.
Novella IS, Elena SF, Moya A, Domingo E & Holland JJ (1996) "Repeated transfer of small RNA virus populations leading to balanced fitness with infrequent stochastic drift", Mol. Gen. Genet. 252:733-738.
Novella IS, Quer J, Domingo E & Holland JJ (1999) "Exponential fitness gains of RNA virus populations are limited by bottleneck effects", Virology 73:1668-1671.
Nowak M & Schuster P (1989) "Error thresholds of replication in finite populations mutation frequencies and the onset of Muller's ratchet", J. theor. Biol. 137:375-396.
Nunney L (1999) "The effective size of a hierarchically structured population", Evolution 53:1-10.
Nurse P (1997) "Reductionism: The ends of understanding", Nature 387:657-657.
Nusse HE & Yorke JA (1996) "Basins of attraction", Science 271:1376-1380.
O'Neill RV (1973) "Error analysis of ecological models", Deciduous Forest Biome. Memo Report 71-15.
Ochman H (1999) "Bacterial evolution: Jittery genomes", Curr. Biol. 9:R485-486.
Ochman H, Elwyn S & Moran NA (1999) "Calibrating bacterial evolution", Proc. Natl. Acad. Sci. USA 96:12638-12643.
Ochman H, Lawrence JG & Groisman EA (2000) "Lateral gene transfer and the nature of bacterial innovation", Nature 405:299-304.
Ohashi J & Tokunaga K (2000) "Sojourn times and substitution rate at overdominant and linked neutral loci", Genetics 155:921-927.
Ohta T (1987) "Very slightly deleterious mutations and the molecular clock", J. Mol. Evol. 26:1-6.
Ohta T (1992) "The nearly neutral theory of molecular evolution", Ann. Rev. Ecol. Syst. 23:263-286.
Ohta T (1993) "An examination of the generation-time effect on molecular evolution", Proc. Natl. Acad. Sci. USA 90:10676-10680.
Ohta T & Gillespie JH (1996) "Development of neutral and nearly neutral theories", Theor. Popul. Biol. 49:128-142.
Ohta T (1998) "Evolution by nearly-neutral mutations", Genetica 103:83-90.
Oliver A, Canton R, Campo P, Baquero F & Blazquez J (2000) "High frequency of hypermutable Pseudomonas aeruginosa in cystic fibrosis lung infection", Science 288:1251-1254.
Omlin M & Reichert P (1999) "A comparison of techniques for the estimation of model prediction uncertainty", Ecol. Model. 115:45-59.
Omlin M, Brun R & Reichert P (2001) "Biogeochemical model of Lake Zurich: sensitivity, identifiability and uncertainty analysis", Ecol. Model. 141:105-123.
Orr HA & Irving S (2001) "Complex epistasis and the genetic basis of hybrid sterility in the Drosophila pseudoobscura Bogota-USA hybridization", Genetics 158:1089-1100.
Osawa S (1995) "Evolution of the genetic code", Oxford, Oxford University Press.
Otto SP & Orive ME (1995) "Evolutionary consequences of mutation and selection within an individual", Genetics 141:1173-1187.
Otto SP & Hastings IM (1998) "Mutation and selection within the individual", Genetica 103:507-524.
Otto SP & Lenormand T (2002) "Resolving the paradox of sex and recombination", Nat Rev Genet 3:252-261.
Ousterhout JK (1994) "Tcl and the Tk Toolkit", Reading, MA, Addison-Wesley.

Ousterhout JK (1995) "Tcl und Tk: Entwicklung grafischer Benutzerschnittstellen für das X Window System", Bonn, Addison-Wesley.
Ozawa A & Freter R (1964) "Ecological mechanism controlling growth of Escherichia coli in continuous flow cultures and in the mouse intestine", J. Infect. Dis. 114:235-242.
Pääbo S (1996) "Mutational hot spots in the mitochondrial microcosm", Am. J. Hum. Genet. 59:493-496.
Page B (1991) "Diskrete Simulationen -Eine Einführung mit Modula 2", Berlin, Springer-Verlag.
Pagel M (2000) "Phylogenetic-evolutionary approaches to bioinformatics", Brief. Bioinformat. 1:117-130.
Painter PR & Marr AG (1968) "Mathematics of microbial populations", Annu. Rev. Microbiol. 22:519-548.
Palumbi SR (2001) "Humans as the world's greatest evolutionary force", Science 293:1786-1790.
Pamilo P, Nei M & Li WH (1987) "Accumulation of mutations in sexual and asexual populations", Genet. Res. 49:135-146.
Pandey A & Mann M (2000) "Proteomics to study genes and genomes", Nature 405:837-846.
Panoff JM, Thammavongs B & Gueguen M (2000) "Cryoprotectants lead to phenotypic adaptation to freeze-thaw stress in Lactobacillus delbrueckii ssp. bulgaricus CIP 101027T", Cryobiology 40:264-269.
Papadopoulos D, Schneider D, Meier Eiss J, Arber W, Lenski RE & Blot M (1999) "Genomic evolution during a 10,000-generation experiment with bacteria", Proc. Natl. Acad. Sci. USA 96:3807-3812.
ParaSoft (2001) "CodeWizard" http://www.parasoft.com/cplus.htm
Parcy F, Nilsson O, Busch MA, Lee I & Weigel D (1998) "A genetic framework for floral patterning", Nature 395:561-566.
Parkes RJ, Cragg BA & Wellsbury P (2000) "Recent studies on bacterial populations and processes in subseafloor sediments: A review", Hydrogeology Journal 8:11-28.
Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, Churcher C, Mungall KL, Bentley SD, Holden MT, Sebaihia M, Baker S, Basham D, Brooks K, Chillingworth T, Connerton P, Cronin A, Davis P, Davies RM, Dowd L, White N, Farrar J, Feltwell T, Hamlin N, Haque A et al. (2001) "Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18", Nature 413:848-852.
Parson W, Parsons TJ, Scheithauer R & Holland MM (1998) "Population data for 101 Austrian Caucasian mitochondrial DNA d-loop sequences: Application of mtDNA sequence analysis to a forensic case", International Journal of Legal Medicine 111:124-132.
Parsons TJ, Muniec DS, Sullivan K, Woodyatt N, Alliston-Greiner R, Wilson MR, Berry DL, Holland KA, Weedn VW, Gill P & Holland MM (1997) "A high observed substitution rate in the human mitochondrial DNA control region", Nat. Genet. 15:363-368.
Parsons TJ & Holland MM (1998) "Mitochondrial mutation rate revisited: hot spots and polymorphism - Response", Nat. Genet. 18:110-110.
Paulsson J (2002) "Multileveled selection on plasmid replication", Genetics 161:1373-1384.
Pearson K (2002) "Internet-based Distributed Computing Projects, http://www.aspenleaf.com/distributed/.
Peccoud J (1995) "Automating molecular biology: A question of communication", Bio/Technology 13:741-745.
Peck JR & Waxman D (2000) "What's wrong with a little sex?" J. evol. Biol. 13:63-69.
Perna NT, Plunkett G, 3rd, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Posfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G et al. (2001) "Genome sequence of enterohaemorrhagic Escherichia coli O157:H7", Nature 409:529-533.
Peruski Jr. LF & Peruski AH (1997) "The Internet and the new biology: Tools for genomic and molecular research", Washington, D.C., American Society for Microbiology.
Pesole G, Sbisa E, Preparata G & Saccone C (1992) "The evolution of the mitochondrial D-loop region and the origin of modern man", Mol. Biol. Evol. 9:587-598.
Peter S (1997) "Prototypischer Entwurf und Implementierung einer Simulationssprache für evolutive Vorgänge in der Biologie", Lehrstuhl für Systemanalyse, Fachbereich Informatik, Universität Dortmund.
Peters N (1990) "Evolution without selection: Quantitative aspects of the eye rudimentation in cave fishes", Mémoires de Biospéologie Tome XVII:43-48.
Peters N (1992) "Constraint rather than chance: Regressive and reconstructive evolutionary process in cave fishes progress the same rate", Mitteilungen aus dem Hamburgischen Zoologischen Museum und Institut 89:97-113.
Peters N, Schacht V, SChmidt W & Wilkens H (1993) "Gehirnproportionen und Ausprägungsgrad der Sinnesorgane von Astyanax mexicanus (Pisces, Characinidae) - Ein Vergleich zwischen dem Flußfisch und seinen Höhlenderivaten "Anoptichtys"", Z. zool. Syst. Evolut.-forsch 31:144-159.
Peters N (1996) "Über die Ursachen der Gehirnreduktion bei den Haustieren", Verh. naturwiss. Ver. Hamburg NF 35:237-257.
Peters AD & Keightley PD (2000) "A test for epistasis among induced mutations in Caenorhabditis elegans", Genetics 156:1635-1647.
Petit E, Excoffier L & Mayer F (1999) "No evidence of bottleneck in the postglacial recolonization of Europe by the noctule bat (Nyctalus noctula)", Evolution 53:1247-1258.
Phelps TJ, Murphy E, Pfiffner M & White D (1994) "Comparison between geochemical and biological estimates of subsurface microbial activity", Microbial Ecology 28:335.
Popper K (1963) "Conjectures and Refutations".

Popper KP (1989) "Logik der Forschung". 9th, Tübingen.
Postgate J & Priest FG (1995) "Putative oligocene spores", Microbiology 141:2763-2764.
Potts M (1994) "Desiccation tolerance of prokaryotes", Microbiol. Rev. 58:755-805.
Potts M (1999) "Mechanisms of desiccation tolerance in cyanobacteria", Eur J Phycol 34:319-328.
PowerSim (2002) "PowerSim - Custom Simulator Solutions" http://powersim.com/
Prata S (1993) "Künstliches Leben: Evolution auf dem PC erleben", München, te-wi Verlag.
Press WH, Teukolsky SA, Vetterling WT & Flannery BP (1992) "Numerical recipies in C". 2nd, Cambridge, Cambridge University Press.
Primack RB (1998) "Essentials of Conservation Biology". 2nd, Sunderland, MA, Sinauer Associates.
Pringle H (1998) "Archaeology: The slow birth of agriculture", Science 282:1446-1450.
Prügel-Bennett A (1997) "Modelling evolving populations", J. theor. Biol. 185:81-95.
Prusinkiewicz P & Lindenmayer A (1990) "The Algorithmic Beauty of Plants", New York, Springer-Verlag.
Queitsch C, Sangster TA & Lindquist S (2002) "Hsp90 as a capacitor of phenotypic variation", Nature 417:618-624.
R-system (2002) "R: A Language for Data Analysis and Graphics" http://www.r-project.org/
Raff RA (1996) "The shape of life: genes, development, and the evolution of animal form", Chicago, University of Chicago Press.
Railsback SF (2001) "Concepts from complex adaptive systems as a framework for individual-based modelling", Ecol. Model. 139:47-62.
Raines P (1998) "Tcl/Tk kurz und gut". 1, Cambridge, O'Reilly Verlag.
Raines P & Tranter J (1999) "Tcl/Tk in a nutshell. A desktop quick reference". 1, Sebastopol, CA, O'Reilly & Associates.
Rainey PB & Travisano M (1998) "Adaptive radiation in a heteregeneous environment", Nature 394:69-72.
Rainey PB (1999) "Evolutionary genetics: The economics of mutation", Curr. Biol. 9:R371-R373.
Rapaport DC (1995) "The art of molecular dynamics simulations", Cambridge, Cambridge University Press.
Ray T (2002) "Tierra Homepage" http://www.isd.atr.co.jp/~ray/tierra
Raymond E (2001) "The cathedral and the bazar" http://www.tuxedo.org/~esr/writings/cathedral-bazaar/
Rechenberg I (1994) "Evolutionsstrategie ´94", Stuttgart, Fridrich Frommann Verlag.
Rechenberg I (1996) "Vorwort des Herausgebers", pp. 5-7 in: Koch-Schwessinger G (ed) Wasserstoffproduktion durch Purpurbakterien, Stuttgart, frommann-holzboog.
Regan HM, Lupia R, Drinnan AN & Burgman MA (2001) "The currency and tempo of extinction", Am. Nat. 157:1-10.
Reha-Krantz LJ (1998) "Regulation of DNA polymerase exonucleolytic proofreading activity: Studies of bacteriophage T4 "antimutator" DNA polymerases", Genetics 148:1551-1557.
Rennell D, Bouvier SE, Hardy LW & Poteete AR (1991) "Systematic mutation of bacteriophage T4 lysozyme", J. Mol. Biol. 222:67-88.
Rice WR (1994) "Degeneration of a nonrecombining chromosome", Science 263:230-232.
Rice WR (2002) "Experimental tests of the adaptive significance of sexual recombination", Nat Rev Genet 3:241-251.
Rigler FH & Peters RH (1995) "Reductionism versus holism: An old problem rejuvenated by the computer", pp. 95-115 in: Rigler FH & Peters RH (eds) Science and limnology, Oldendorf/Luhe, Germany, Ecology Institute.
Riley MS, Cooper VS, Lenski RE, Forney LJ & Marsh TL (2001) "Rapid phenotypic change and diversification of a soil bacterium during 1000 generations of experimental evolution", Microbiology 147:995-1006.
Rispe C & Moran NA (2000) "Accumulation of deleterious mutations in endosymbionts: Muller's ratchet with two levels of selection", Am. Nat. 156:425-441.
Rogers AR (1995) "Genetic evidence for a pleistocene population explosion", Evolution 49:608-615.
Rogers AR (2001) "Order emerging from chaos in human evolutionary genetics", Proc. Natl. Acad. Sci. USA 98:779-780.
Rosche WA & Foster PL (1999) "The role of transient hypermutators in adaptive mutation in Escherichia coli", Proc. Natl. Acad. Sci. USA 96:6862-6867.
Rose MR & Lauder GV, (eds, 1996) "Adaptation", San Diego, Academic Press.
Rosenberg SM, Longerich S, Gee P & Harris RS (1994) "Adaptive mutation by deletions in small mononucleotide repeats", Science 265:405-407.
Rosenberg SM, Harris RS & Tokelson J (1995) "Molecular handles on adaptive mutation", Mol. Microbiol. 18:185-189.
Rosenberg SM (1997) "Mutation for survival", Curr. Opin. Genet. Dev. 7:829-834.
Rosenberg SM, Thulin C & Harris RS (1998) "Transient and heritable mutators in adaptive evolution in the lab and in nature", Genetics 148:1559-1566.
Rosin CD, Belew RK, Morris GM, Olson AJ & Goodsell DS (1999) "Coevolutionary analysis of resistance-evading peptidomimetic inhibitors of HIV-1 protease", Proc. Natl. Acad. Sci. USA 96:1369-1374.
Ross T & McMeekin TA (1994) "Predictive microbiology", Int. J. Food. Microbiol. 23:241-264.
Rossello-Mora R & Amann R (2001) "The species concept for prokaryotes", FEMS Microbiol. Rev. 25:39-67.
Rowe G (1994) "Theoretical Models in Biology. The Origin of Life, the Immune System, and the Brain", Oxford, Clarendon Press.
Rowe L & Houle D (1996) "The lek paradox and the capture of genetic variance by condition dependent traits", Proc. R. Soc. Lond. B Biol. Sci. 263:1415-1421.
Rozen DE & Lenski RE (2000) "Long-term experimental evolution in Escherichia coli. VIII. Dynamics of a balanced

polymorphism", Am. Nat. 155:24-35.
Ryan FJ, Nakada D & Schneider MJ (1961) "Is DNA replication a necessary condition for spontaneous mutation?" Z. Vererbungsl. 92:38-41.
Ryan FJ, Okada T & Nagata T (1963) "Spontaneous mutation in spheroplasts of Escherichia coli", J Gen Microbiol 30:193-199.
Ryder ML (1984) "Sheep", pp. 63-85 in: Mason IL (ed) Evolution of domesticated animals, London, Longman Group Limited.
Ryder OA & Chemnick LG (1993) "Chromosomal and mitochondrial DNA variation in Orang Utans", J. Hered. 84:405-409.
SAAM (2002) "SAAM II" http://www.saam.com/
Sadoglu P (1967) "The selective value of eye and pigment loss in Mexican cave fish", Evolution 21:541-549.
Saether BE (1997) "Environmental stochasticity and population dynamics of large herbivores: A search for mechanisms", Trends Ecol. Evol. 12:143-149.
Sala M & Wain-Hobson S (1999) "Drift and conservatism in RNA virus evolution: Are they adapting or merely changing?" pp. 115-140 in: Domingo E, Webster R & Holland J (eds) Origin and evolution of viruses, San Diego, Academic Press.
Sambrook J, Fritsch EF & Maniatis T (1989) "Molecular Cloning: A Laboratory Manual. Second Edition". 2, Cold Spring Harbour, Cold Spring Harbor Laboratory Press.
Santiago E & Caballero A (1995) "Effective size of populations under selection", Genetics 139:1013-1030.
SAP (2002) "SAP-DB Homepage" http://www.sapdb.org/
Schaaper RM (1993) "The mutational specificity of two Escherichia coli dnaE antimutator alleles as determined from lacI mutation spectra", Genetics 134:1031-1038.
Schaaper RM (1998) "Antimutator mutants in bacteriophage T4 and Escherichia coli", Genetics 148:1579-1585.
Schader M & Kuhlins S (1996) "Programmieren in C++: Einführung in den Sprachstandard". 4. neubearbeitete und erweiterte, Berlin, Springer Verlag.
Schartl M, Nanda I, Schlupp I, Wilde B, Epplen JT, Schmid M & Parzefall J (1995) "Incorporation of subgenomic amounts of DNA as compensation for mutational load in a gynogenetic fish", Nature 373:68-71.
Schartl M, Wilde B, Schlupp I & Parzefall J (1995) "Evolutionary origin of a parthenoform, the Amazon molly Poecilia formosa, on the basis of a molecular genealogy", Evolution 49:827-835.
Scheffer M, Baveco JM, Deangelis DL, Rose KA & Vannes EH (1995) "Super-Individuals a Simple Solution for Modeling Large Populations on an Individual Basis", Ecol. Model. 80:161-170.
Scheppach J (2001) "Jetzt wird aus jedem Computer ein globales Supergehirn!" P.M.:80-86.
Scherer S (1990) "The protein molecular clock: time for a reevaluation", Evol. Biol. 24:83-106.
Scherer S & Neuhaus K, (eds, 2002) "Life at low temperatures". The Prokaryotes: An evolving electronic resource for the microbiological community, 3rd Edition. latest update release 3.9 (March 2002), New York.
Schiewe MC (1991) "The science and significance of embryo cryopreservation", Journal of Zoo and Wildlife Midicine 22:6-22.
Schild H & Rammensee HG (2000) "Perfect use of imperfection", Nature 404:709-710.
Schneider G, Schrodl W, Wallukat G, Muller J, Nissen E, Ronspeck W, Wrede P & Kunze R (1998) "Peptide design by artificial neural networks and computer-based evolutionary search", Proc. Natl. Acad. Sci. USA 95:12179-12184.
Schneider S & Excoffier L (1999) "Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA", Genetics 152:1079-1089.
Schneider D, Duperchy E, Coursange E, Lenski RE & Blot M (2000) "Long-term experimental evolution in Escherichia coli. IX. Characterization of insertion sequence-mediated mutations and rearrangements", Genetics 156:477-488.
Schu P & Reith M (1995) "Evaluation of Different Preparation Parameters for the Production and Cryopreservation of Seed Cultures with Recombinant Saccharomyces-Cerevisiae", Cryobiology 32:379-388.
Schultz ST & Lynch M (1997) "Mutation and extinction: The role of variable mutational effects, synergistic epistasis, beneficial mutations, and the degree of outcrossing", Evolution 51:1363-1371.
Schulze-Kremer S (1996) "Molecular bioinformatics: Algorithms and applications", Berlin, Walter de Gruyter.
Schulze-Kremer S (1996) "Evolutionary computation", pp. 211-271 in: Schulze-Kremer S (ed) Molecular bioinformatics: Algorithms and applications, Berlin, Walter de Gruyter.
Schuster P (1996) "How does complexity arise in evolution", Complexity:22-30.
Sedgewick R (1992) "Algorithmen in C++", Bonn, Addison-Wesley.
Sequeira RA, Olson RL & McKinion JM (1997) "Implementing generic, object-oriented models in biology", Ecol. Model. 94:17-31.
Setlow P (1992) "I will survive: Protecting and repairing spore DNA", J. Bacteriol. 174:2737-2741.
Setlow B & Setlow P (1995) "Small, acid-soluble proteins bound to DNA protect Bacillus subtilis spores from killing by dry heat", Appl. Environ. Microbiol. 61:2787-2790.
Setlow P (1995) "Mechanisms for the prevention of damage to DNA in spores of Bacillus species", Annu. Rev. Microbiol. 49:29-54.
Setubal J & Meidanis J (1997) "Introduction to computational molecular biology", Boston, PWS Publishing Company.
Shabalina SA, Yampolsky LY & Kondrashov AS (1997) "Rapid decline of fitness in panmictic populations of Drosophila

melanogaster maintained under relaxed natural selection", Proc. Natl. Acad. Sci. USA 94:13034-13039.
Shaffer ML (1997) "Population Viability Analysis: Determining natures's share", pp. 215-217 in: Meffe GK & Carroll CR (eds) Principles of conservation biology, 2nd, Sunderland, MA, Sinauer Associates.
Sherry ST, Harpending HC, Batzer MA & Stoneking M (1997) "Alu evolution in human populations: Using the coalescent to estimate effective population size", Genetics 147:1977-1982.
Shigenobu S, Watanabe H, Hattori M, Sakaki Y & Ishikawa H (2000) "Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS", Nature 407:81-86.
Shin T, Kraemer D, Pryor J, Liu L, Rugila J, Howe L, Buck S, Murphy K, Lyons L & Westhusin M (2002) "A cat cloned by nuclear transplantation", Nature 415:859-859.
Shugart HH, Smith TM & Post WM (1992) "The potential for application of individual-based simulation models for assessing the effects of global change", Ann. Rev. Ecol. Syst. 23:15-38.
Shukla J (1998) "Predictability in the midst of chaos: A scientific basis for climate forecasting", Science 282:728-731.
Siegert H-J (1991) "Simulation zeitdiskreter Systeme", München, Oldenbourg Verlag.
Sigurdardottir S, Helgason A, Gulcher JR, Stefansson K & Donnelly P (2000) "The mutation rate in the human mtDNA control region", Am. J. Hum. Genet. 66:1599-1609.
Silvertown J, Holtier S, Johnson J & Dale P (1992) "Cellular automaton models of interspecific competition for space - the effect of pattern on process", J. Ecol. 80:527-534.
Slade PF (2001) "Simulation of 'hitch-hiking' genealogies", J. math. Biol. 42:41-70.
Slothower RL, Schwarz PA & Johnston KM (1996) "Some guidelines for implementing spatially explicit, individual based ecological models within location-based raster GIS" http://www.sbg.ac.at/geo/idrisi/gis_environmental_modeling/sf_papers/slothower_roger/sf23.html
Small C (1997) "A Tool for Constructing Safe Extensible C++ Systems". Third USENIX Conference on Object-Oriented Technologies (COOTS), Portland, Oregon.
Smallen S, Cirne W, Frey J, Berman F, Wolski R, Su M-H, Kesselman C, Young S & Ellisman M (2000) "Combining Workstations and Supercomputers to Support Grid Applications: The Parallel Tomography Experience". Proceedings of the 9th Heterogenous Computing Workshop.
Smart J (2002) "wxWindows Home: Cross-Platform GUI Library" http://www.wxWindows.org
Smith JT (1999) "C++ Toolkit for Engineers and Scientists". 2nd, Heidelberg, Springer.
Smith J (2000) "Nice work - but is it science? Untestable ecological theory won't help solve environmental problems." Nature 408:293-293.
Sniegowski PD, Gerrish PJ & Lenksi RE (1997) "Evolution of high mutation rates in experimental populations of E. coli", Nature 387:703-705.
Snowling SD & Kramer JR (2001) "Evaluating modelling uncertainty for model selection", Ecol. Model. 138:17-30.
Sokal RR & Rohlf FJ (1981) "Biometry: The principles and practice of statistics in biological research. Second Edition", San Francisco, California, W.H. Freeman & Co.
Solter D (2000) "Mammalian cloning: advances and limitations", Nat Rev Genet 1:199-207.
Soodyall H, Jenkins T, Mukherjee A, du Toit E, Roberts DF & Stoneking M (1997) "The founding mitochondrial DNA lineages of Tristan da Cunha Islanders", Am. J. Phys. Anthropol. 104:157-166.
Souza V, Turner PE & Lenski RE (1997) "Long term experimental evolution in Escherichia coli. 5. Effects of recombination with immigrant genotypes on the rate of bacterial evolution", J. evol. Biol. 10:743-769.
Spedding V (2001) "XML to take science by storm", Scientific Computing World:15-22.
Spengler SJ (2000) "Bioinformatics in the information age", Science 287:1221-1222.
Spolsky CM, Phillips CA & Uzzell T (1992) "Antiquity of clonal salamander lineages revealed by mitochondrial DNA", Nature 356:706-708.
Stahl DA & Tiedje JM (2002) "Microbial ecology and genomics: a crossroads of opportunity", (http://www.asmusa.org/acasrc/pdfs/MicroEcoReport.pdf) Washington, D.C., American Academy of Microbiology.
Staley JT, Castenholz RW, Colwell RR, Holt JG, Kane MD, Pace NR, Salyers AA & Tiedje JM (1997) "The Microbial World: Foundation of the biosphere", (http://www.asmusa.org/acasrc/pdfs/Colloquia/microbialworld.pdf) Washington, D.C., American Academy of Microbiology.
Staley JT & Reysenbach A-L (2001) "Biodiversity of Microbial Life: Foundation of Earth's Biosphere", New York, Wiley.
Starfield AM & Bleloch AL (1986) "Building models for conservation and wildlife management", London, Collier Macmillan.
Stearns SC (1992) "The evolution of life histories", Oxford, Oxford University Press.
Stenseth NC & Maynard Smith J (1984) "Coevolution in ecosystems: Red Queen evolution or stasis?" Evolution 38:870-880.
Stephan W, Chao L & Smale JG (1993) "The advance of Muller's ratchet in a haploid asexual population: Approximate solutions based on diffusion theory", Genet. Res. 61:225-231.
Stephan W (1996) "The rate of compensatory evolution", Genetics 144:419-426.
Stephan W & Kim Y (2002) "Recent applications of diffusion theory to population genetics", pp. 72-93 in: Slatkin M & Veuille M (eds) Modern developments in theoretical population genetics, Oxford, Oxford University Press.
Stevens A (2001) "C Programming: It's good work when you can find it", Dr. Dobb's J.:121-124.
Stöck M & Lamatsch DK (2002) "Triploide Wirbeltiere: Wege aus der Unfruchtbarkeit oder Eingeschlechtigkeit", Naturw.

Rdsch. 55:349-358.
Stoneking M, Sherry ST, Redd AJ & Vigilant L (1992) "New approaches to dating suggest a recent age for human mtDNA ancestor", Philos. Trans. R. Soc. Lond. B Biol. Sci 337:167-175.
Stothard JR (1997) "Phylogenetic inference with RAPDs: Some observations involving computer simulation with viral genomes", J. Hered. 88:222-228.
Strauss E (1999) "Can mitochondrial clocks keep time?" Science 283:1435, 1437-1438.
Striegnitz J & Veldhuizen T (2001) "LRZ Tutorial: Writing efficient programs with C++" http://www.lrz-muenchen.de/services/compute/hlrb/manuals/lecture_notes/HPC++1.ps.gz
Stroustrup B (1994) "Design and Evolution of C++", AT&T Bell Labs.
Stroustrup B (1998) "Die C++ Programmiersprache". 3., aktualisierte und erweiterte, Bonn, Addison-Wesley.
Sugg DW & Chesser RK (1994) "Effective population sizes with multiple paternity", Genetics 137:1147-1155.
Sugihara G & May RM (1990) "Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series", Nature 344:734-741.
Suzuki M, Baskin D, Hood L & Loeb LA (1996) "Random mutagenesis of Thermus aquaticus DNA polymerase I: Concordance of immutable sites in vivo with the crystal structure", Proc. Natl. Acad. Sci. USA 93:9670-9675.
Swetina J & Schuster P (1982) "Self-replication with errors. A model for polynucleotide replication", Biophysical Chemistry 16:329-345.
Swofford D (2002) "PAUP 4.0" http://paup.csit.fsu.edu/
Szafraniec K, Borts RH & Korona R (2001) "Environmental stress and mutational load in diploid strains of the yeast Saccharomyces cerevisiae", Proc. Natl. Acad. Sci. USA 98:1107-1112.
Szathmáry E & Maynard Smith J (1995) "The major evolutionary transitions", Nature 374:227-232.
Taddei F, Matic I & Radman M (1995) "cAMP-dependent SOS induction and mutagenesis in resting bacterial populations", Proc. Natl. Acad. Sci. USA 92:11736-11740.
Taddei F, Radman M, Maynard-Smith J, Toupance B, Gouyon PH & Godelle B (1997) "Role of mutator alleles in adaptive evolution", Nature 387:700-702.
Takahata N & Slatkin M (1983) "Evolutionary dynamics of extranuclear genes", Genet. Res. 42:257-265.
Takahata N (1993) "Allelic genealogy and human evolution", Mol. Biol. Evol. 10:2-22.
Talia D (1998) "Cellular automata thrive on parallel systems", Scientific Computing World:21-22.
Tamura K & Nei M (1993) "Estimation of the Number of Nucleotide Substitutions in the Control Region of Mitochondrial DNA in Humand and Chimpanzees", Mol. Biol. Evol. 10:512-526.
Tanaka Y, Yoh M, Takeda Y & Miwatani T (1979) "Induction of Mutation in Escherichia-Coli by Freeze-Drying", Appl. Environ. Microbiol. 37:369-372.
Tang SZ, Li Y & Wang YH (2001) "Simultaneous equations, error-in-variable models, and model integration in systems ecology", Ecol. Model. 142:285-294.
Tannenbaum AS (1995) "Distributed operating systems", Englewood Cliffs, New Jersy, Prentice Hall.
Taylor HM & Karlin S (1998) "An introduction to stochastic modeling". 3rd., San Diego, Academic Press.
Templeton A (2002) "Out of Africa again and again", Nature 416:45-51.
Textor V (2001) "Programmierung eines VBA-Systems zur Auswertung von Mutationsakkumulationsexperimenten mit Mikroorganismen in Flüssigkultur und Bestimmung der Nachweisgrenzen für Fitnessveränderungen", Institut für Mikrobiologie, Forschungszentrum für Milch und Lebensmittel, Technische Universität München.
The Arabidopsis Genome Initiative (2000) "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana", Nature 408:796-815.
The C. elegans Sequencing Consortium (1998) "Genome sequence of the nematode Caenorhabditis elegans: A platform for investigating biology", Science 282:2012-2018.
Thompson JN (1998) "Rapid evolution as an ecological process", Trends Ecol. Evol. 13:329-332.
Ting CT, Tsaur SC, Wu ML & Wu CI (1998) "A rapidly evolving homeobox at the site of a hybrid sterility gene", Science 282:1501-1504.
Toquenaga Y & Wade MJ (1996) "Sewall Wright meets Artificial Life: the origin and maintenance of evolutionar novelty", Trends Ecol. Evol. 11:478-482.
Torkelson J, Harris RS, Lombardo MJ, Nagendran J, Thulin C & Rosenberg SM (1997) "Genome-wide hypermutation in a subpopulation of stationary-phase cells underlies recombination-dependent adaptive mutation", EMBO J. 16:3303-3311.
Torsvik V, Goksoyr J & Daae FL (1990) "High diversity of DNA of soil bacteria", Appl. Environ. Microbiol. 56:782-787.
Trammell K (1996) "Work flow without fear", Byte:55-60.
Travisano M, Mongold JA, Bennett AF & Lenski RE (1995) "Experimental tests of the roles of adaptation, chance, and history in evolution", Science 267:87-90.
Travisano M, Vasi F & Lenski RE (1995) "Long-term experimental evolution in Escherichia coli. III. Variation among replicate populations in correlated responses to novel environments", Evolution 49:189-200.
Travisano M (1997) "Long-term experimental evolution in Escherichia coli. VI. Environmental constraints on adaptation and divergence", Genetics 146:471-479.
Travisano M & Rainey PB (2000) "Studies of adaptive radiation using model microbial systems", Am. Nat. 156:S35-S44.
Travisano M (2001) "Evolution: Towards a genetical theory of adaptation", Curr. Biol. 11:R440-442.

Troy CS, MacHugh DE, Bailey JF, Magee DA, Loftus RT, Cunningham P, Chamberlain AT, Sykes BC & Bradley DG (2001) "Genetic evidence for Near-Eastern origins of European cattle", Nature 410:1088-1091.
Turner PE, Cooper VS & Lenski RE (1998) "Tradeoff between horizontal and vertical modes of transmission in bacterial plasmids", Evolution 52:315-329.
Unidata Program Center, Russ Rew, Glenn Davis, Steve Emmerson & Davies H (1997) "NetCDF User's Guide for C: An access interface for self-describing, portable data". Version 3, http://www.unidata.ucar.edu/packages/netcdf/, Unidata Program Center.
United Nations Organization (1992) "Rio Declaration: Agenda 21", New York, UNO, Department of Public Information, Project Manager for Sustainable Development.
van Tongeren OFR (1995) "Data-Analysis or Simulation-Model - a Critical-Evaluation of Some Methods", Ecol. Model. 78:51-60.
Van Valen L (1973) "A new evolutionary law", Evol. Theory 1:1-30.
Vassilieva LL, Hook AM & Lynch M (2000) "The fitness effects of spontaneous mutations in Caenorhabditis elegans", Evolution 54:1234-1246.
Velicer GJ, Kroos L & Lenski RE (1998) "Loss of social behaviors by Myxococcus xanthus during evolution in an unstructured habitat", Proc. Natl. Acad. Sci. USA 95:12376-12380.
Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C et al. (2001) "The sequence of the human genome", Science 291:1304-1351.
Vigilant L, Stoneking M, Harpending H, Hawkes K & Wilson AC (1991) "African populations and the evolution of human mitochondrial DNA", Science 253:1503-1507.
Vilà C, Savolainen P, Maldonado JE, Amorim IR, Rice JE, Honeycutt RL, Crandall KA, Lundeberg J & Wayne RK (1997) "Multiple and ancient origins of the domestic dog", Science 276:1687-1689.
Vilà C, Leonard JA, Gotherstrom A, Marklund S, Sandberg K, Liden K, Wayne RK & Ellegren H (2001) "Widespread origins of domestic horse lineages", Science 291:474-477.
Visscher PM, Smith D, Hall SJ & Williams JL (2001) "A viable herd of genetically uniform cattle", Nature 409:303.
von Dassow G, Meir E, Munro EM & Odell GM (2000) "The segment polarity network is a robust developmental module", Nature 406:188-192.
von Neumann J & Burks AW (1966) "Theory of self-reproducing automata", Urbana, University of Illinois Press.
Vorobyova E, Soina V, Gorlenko M, Minkovskaya N, Zalinova N, Mamukelashvili A, Gilichinsky D, Rivkina E & Vishnivetskaya T (1997) "The deep cold biosphere: facts and hypothesis", FEMS Microbiology Reviews 20:277-290.
Vreeland RH, Rosenzweig WD & Powers DW (2000) "Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal", Nature 407:897-900.
Vucetich JA, Waite TA & Nunney L (1997) "Fluctuating population size and the ratio of effective to census population size", Evolution 51:2017-2021.
W3C (2000) "Extensible Markup Language (XML) 1.0" http://www.w3.org/TR/REC-xml
Wagner GP & Gabriel W (1990) "Quantitative variation in finite parthenogenetic populations: What stops Muller's ratchet in the absence of recombination?" Evolution 44:715-731.
Wagner GP & Krall P (1993) "What is the difference between models of error thresholds and Muller's ratchet?" J. math. Biol. 32:33-44.
Wagner GP & Altenberg L (1996) "Complex adaptations and the evolution of evolvability", Evolution 50:967-976.
Wakeley J (1993) "Substitution Rate Variation Among Sites in Hypervariable Region 1 of Human Mitochondrial DNA", J. Mol. Evol. 37:613-623.
Wakeley J (1996) "The excess of transitions among nucleotide substitutions: New methods of estimating transition bias underscore its significance", Trends Ecol. Evol. 11:158-163.
Wallace DC (1999) "Mitochondrial diseases in man and mouse", Science 283:1482-1488.
Wallace DC, Brown MD & Lott MT (1999) "Mitochondrial DNA variation in human evolution and disease", Gene 238:211-230.
Walters C, Roos EE, Touchell DP, Stanwood PC, Towill L, Wiesner L & Eberhart SA (1998) "Refrigeration can save seeds economically", Nature. vol.395 pp.758-758.
Wapnish P & Hesse B (1997) "Equids", pp. 255-256 in: Meyers EM (ed) The Oxford Encyclopedia of Archaeology in the Near East, New York, Oxford University Press.
Waterman MS, (ed, 1995) "Introduction to computational biology: Maps, sequences and genomes", London, Chapman & Hall.
Webster T & Francis A (2000) "Tcl/Tk für Dummies. Gegen den täglichen Frust mit Tcl/Tk". 1, Bonn, MITP-Verlag.
Weismann A (1893) "Die Allmacht der Naturzüchtung. Eine Erwiderung an Herbert Spencer", Jena, Verlag von Gustav Fischer.
Welch BB (2000) "Practical programming in Tcl and Tk". 3rd, Upper Saddle River, New Jersey, Prentice Hall.
Welch DM & Meselson M (2000) "Evidence for the evolution of Bdelloid Rotifers without sexual reproduction or genetic exchange", Science 288:1211-1215.

Wen J, Chen X & Bowie JU (1996) "Exploring the allowed sequence space of a membrane protein", Nat. Struct. Biol. 3:141-148.
Whiting RC & Buchanan RL (1997) "Predictive Modeling", pp. 728-739 in: Doyle MP, Beuchat LR & Montville TJ (eds) Food microbiology: Fundamentals and frontiers, Washington, D.C., ASM Press.
Whitlock MC, Phillips PC, Moore FBG & Tonsor SJ (1995) "Multiple fitness peaks and epistasis", Ann. Rev. Ecol. Syst. 26:601-629.
Whitlock MC & Barton NH (1997) "The effective size of a subdivided population", Genetics 146:427-441.
Whitlock MC (2002) "Selection, load and inbreeding depression in a large metapopulation", Genetics 160:1191-1202.
Whitman WB, Coleman DC & Wiebe WJ (1998) "Prokaryotes: The unseen majority", Proc. Natl. Acad. Sci. USA 95:6578-6583.
Whittam TS (1992) "Population Biology: Sex in the soil", Curr. Biol. 2:676-678.
Whittam TS (1996) "Genetic variation and evolutionary processes in natural populations of Escherichia coli", pp. 2708-2720 in: Neidhardt FC, Curtiss III R, Ingraham JL, Lin ECC, Low KB, Magasanik B, Reznikoff WS, Riley M, Schaechter M & Umbarger HE (eds) Escherichia coli and Salmonella: cellular and molecular biology. 2, 2, Washington D.C., ASM Press.
Wilke CO, Wang JL, Ofria C, Lenski RE & Adami C (2001) "Evolution of digital organisms at high mutation rates leads to survival of the flattest", Nature 412:331-333.
Wilson WG (1998) "Resolving discrepancies between deterministic population models and individual-based simulations", Am. Nat. 151:116-134.
Wilson WG (2000) "Simulating ecological and evolutionary systems in C", Cambridge, Cambridge University Press.
Wissel C (1989) "Theoretische Ökologie: Eine Einführung", Berlin, Springer-Verlag.
Wolf JB, Brodie III ED & Wade MJ, (eds, 2000) "Epistasis and the evolutionary process", New York, Oxford University Press.
Wolfram Research (2002) "Mathematica Homepage" http://www.wolfram.com
Wright S (1931) "Evolution in mendelian populations", Genetics 16:97-159.
Wuethrich B (1998) "The asexual life", Science 281:1981-1981.
wxWindows community, Smart J & al. e (2002) "wxWindows Home: Cross-Platform GUI Library" http://www.wxWindows.org/
Xu WZ, Fukuhara J, Yamamoto K, Yomo T & Urabe I (1994) "Random mutagenesis of glutamine synthetase from Escherichia coli: Correlation between structure, activity, and fitness", Journal of Fermentation and Bioengineering 77:252-258.
Yamamoto Y & Jeffery WR (2000) "Central role for the lens in cave fish eye degeneration", Science 289:631-633.
Yedid G & Bell G (2001) "Microevolution in an electronic microcosm", Am. Nat. 157:465-487.
Yeiser B, Pepper ED, Goodman MF & Finkel SE (2002) "SOS-induced DNA polymerases enhance long-term survival and evolutionary fitness", Proc. Natl. Acad. Sci. USA 99:8737-8741.
Yi TM, Huang Y, Simon MI & Doyle J (2000) "Robust perfect adaptation in bacterial chemotaxis through integral feedback control", Proc. Natl. Acad. Sci. USA 97:4649-4653.
Yousten AA & Rippere KE (1997) "DNA similarity analysis of a putative ancient bacterial isolate obtained from amber", FEMS Microbiol. Lett. 152:345-347.
Zeder MA (1997) "Sheep and goats", pp. 23-25 in: Meyers EM (ed) The Oxford Encyclopedia of Archaeology in the Near East, New York, Oxford University Press.
Zeder MA & Hesse B (2000) "The initial domestication of goats (Capra hircus) in the Zagros mountains 10,000 years ago", Science 287:2254-2257.
Zeyl C & Bell G (1997) "The advantage of sex in evolving yeast populations", Nature 388:465-468.
Zeyl C & DeVisser J (2001) "Estimates of the rate and distribution of fitness effects of spontaneous mutation in Saccharomyces cerevisiae", Genetics 157:53-61.
Zeyl C, Mizesko M & de Visser J (2001) "Mutational meltdown in laboratory yeast populations", Evolution 55:909-917.
Zhang Y & Ryder OA (1995) "Different Rates of Mitochondrial DNA Sequence Evolution in Kirk's Dik-dik (Madoqua kirkii) Populations", Molecular Phylogenetics and Evolution 4:291-297.

Curriculum vitae

Personal Details:

Name:	Laurence Loewe
Born:	6th May 1969, Cape Town
Nationality:	German
Marital status:	married

Education and Qualifications:

1975 - 1988	School. Final exam (Abitur) in biology and physics at high school "Gymnasium Fridericianum Erlangen", Germany
1987 - 1991	"Jugend forscht" competition, 5 years in a row. Research project investigating the effects of physical discharge parameters on productivity in Miller-experiments on chemical evolution.
1990 - 1995	Diploma in Biology (Dipl.-Biol.), Universität Konstanz, Germany. Thesis on specific mutagenesis in the active center of human medium chain Acyl-CoA Dehydrogenase and characterisation of its effect on catalytic activity.
1998 - 2002	Doctoral Thesis in the Microbial Ecology Group, Department of Biosciences, Technische Universität München, Germany: "Evolutionary Bioinformatics: Predicting genetic stability of asexual genomes by global computing"

Working Experience:

1988 - 1990	National service (Zivildienst) at a social therapeutic welfare institution
1989	Practical course in bioinformatics and electronmicroscopy at the German Cancer Research Center (DKFZ) Heidelberg, Germany
1992	Working student (HiWi), Universität Hohenheim, Germany
1992 + 1994	Working student (HiWi), Universität Konstanz, Germany
1996 - 1997	Scholarship for research on the project “Evolution of biological information” by Wort und Wissen e.V., Konstanz, Germany
1998 - 2002	Researcher at the Microbial Ecology Group, Department of Biosciences, Technische Universität München, Germany
2003	Visiting scientist at the IWR Technical Simulation Group, Interdisciplinary Centre for Scientific Computing, University of Heidelberg, Germany
2003 -	Post-doctoral research fellow at the Institute of Cell, Animal and Population Biology, University of Edinburgh, Scotland, UK

Prizes:

1987 "Jugend Forscht" Bayern, Germany, 3rd place (Landeswettbewerb München). Paper on the significance of primeval earth simulations for the theory of chemical evolution.

1991 "Jugend Forscht" Baden-Würtemberg, Germany, 2nd place (Landeswettbewerb Stuttgart). Paper on the productivity of high voltage discharges in Miller experiments.

Original publications

Loewe L (2002) "evolution@home: Experiences with work units that span more than 7 orders of magnitude in computational complexity", 425-431. 2nd International Workshop on Global and Peer-to-Peer Computing on Large Scale Distributed Systems at the 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid2002), 21-24 May, Berlin, Germany, IEEE Computer Society (see http://www.evolutionary-research.net/ for download).

Loewe L, Textor V & Scherer S (2003) "High deleterious genomic mutation rate in stationary phase of *Escherichia coli*", Science 302:1558-1560.

Loewe L (2004) "The mutation rate paradox: Pedigree versus 'archaeological' versus phylogenetic mutation rates in mitochondrial DNA", under revision.

Loewe L (2004) "Muller's ratchet in mtDNA may cause extinctions in mammals", in prep.

Loewe L (2004) "Muller's ratchet may contribute to the uncultivable majority of bacteria", in prep.

Further publications based on ideas from this dissertation will be listed on the corresponding webpage of this dissertation (see http://www.evolutionary-research.net/ for download).

Reviews, peer reviewed

Loewe L & Scherer S (1997) "Mitochondrial Eve: The plot thickens", Trends Ecol. Evol. 12:422-423.

Loewe (2002) "Global computing for bioinformatics", Brief. Bioinformat. 3:377-388.

Loewe L (2004) "Why do we need evolutionary bioinformatics?", under revision.

Loewe L (2004) "Muller's ratchet reviewed", in prep.

Invited Talks

Loewe L (1997) "Mutation rates, Muller's Ratchet, genetic load and mitochondrial Eve". First international workshop on human mitochondrial DNA, Washington, D.C.

Loewe L (2001) "evolution@home: New frontiers in global computing and evolutionary bioinformatics" EMBnet Annual General Meeting, Vienna, Austria.

Talks at international conferences

Loewe L & Scherer S (1999) "How many beneficial mutations are needed to stop Muller's ratchet?" Seventh Congress of the European Society for Evolutionary Biology, Barcelona, Spain.

Loewe L (2000) "How many beneficial mutations are needed to stop Muller's ratchet in mtDNA?" Spacial ecology workshop on extinction, Tvärminne Zoological Station, Finland.

Loewe L (2002) "evolution@home: Experiences with work units that span more than 7 orders of magnitude in computational complexity", 2nd International Workshop on Global and Peer-to-Peer Computing on Large Scale Distributed Systems at the 2nd IEEE/ACM Internatl. Symp. on Cluster Computing and the Grid, Berlin, Germany.

Loewe L (2002) "Quantifying the threat from Muller's ratchet in mtDNA?" Extinction Thresholds Conference, Helsinki, Finland.

Loewe L, Textor V & Scherer S (2003) "High deleterious genomic mutation rate in stationary phase of *Escherichia coli*" Population Genetics Group Meeting, University of Sussex, UK.

Posters at international conferences

Loewe L & Scherer S (1997) "On the speed of genomic decay". The 5th Annual International Meeting of the Society for Molecular Biology and Evolution, Garmisch-Partenkirchen, Germany.

Loewe L & Scherer S (1998) "Muller's ratchet in human mitochondrial DNA". The 6th Annual International Meeting of the Society for Molecular Biology and Evolution, University of British Columbia, Vancouver, Canada.

Loewe L & Scherer S (1998) "On the speed of genomic decay". Annual Meeting of the Society for the Study of Evolution, Vancouver, British-Columbia, Canada.

Loewe L & Scherer S (1999) "How many advantageous mutations stop Muller's ratchet?" Evolution '99, University of Wisconsin, Madison, Wisconsin.

Loewe L & Scherer S (2001) "Predicting extinctions due to Muller's ratchet in humans and bacteria". Eighth Congress of the European Society for Evolutionary Biology, Aarhus, Denmark.

Diploma theses supervised:

Peter S (1997) "Prototypischer Entwurf und Implementierung einer Simulationssprache für evolutive Vorgänge in der Biologie", Lehrstuhl für Systemanalyse, Fachbereich Informatik, Universität Dortmund.

Textor V(2001) " Programmierung eines VBA Systems zur Auswertung von Mutations-Akkumulations Experimenten mit Mikroorganismen in Flüssigkultur und Bestimmung der Nachweisgrenzen für Fitnessveränderungen.", Microbial Ecology Group, Department of Biosciences, Technische Universität München.

Book contributions

Loewe L (1991) "Chemische Evolutie", Amersfoort, Netherlands, Amersfoortse Studies no. 8, ISBN: 90-70145-20-0.

Loewe L & Scherer S (1996) "Molekulare Evolution", pp. 75-84 in: Scherer S (ed) Entstehung der Photosynthese - Grenzen molekularer Evolution bei Bakterien?, Neuhausen-Stuttgart, Hänssler.

Loewe L (1998) "Grundbegriffe der molekularen Evolution", pp. 96-107 in: Junker R & Scherer S (eds) Evolution: Ein kritisches Lehrbuch, 4th ed., Gießen, Weyel Lehrmittelverlag.

Scherer S & Loewe L (2001) "Probleme bei der Erklärung molekularer Maschinen durch Evolution", pp. 161-186 in: Weingartner P (ed) Evolution als Schöpfung? Ein Streitgespräch zwischen Philosophen, Theologen und Naturwissenschaftlern, Stuttgart, Kohlhammer Verlag.

Popular science contributions:

Loewe L (1996) ""Megaevolution" - Makroevolution - Mikroevolution", Stud. Int. J. 3:71-75.

Loewe L (1997) "Genome im Überblick", Stud. Int. J. 4:3-13.

Loewe L & Scherer S (1997) "Mutter Eva in Bewegung", Stud. Int. J. 4:58-65.

Loewe L (1998) "Skandalöse Symbionten", Stud. Int. J. 5:36-37.

Loewe L (1998) "Neue Ähnlichkeiten zwischen Nilpferden, Kühen und Walen", Stud. Int. J. 5:86-89.

Fehrer J & Loewe L (1999) "Evolution in ökologischen Zeitskalen", Stud. Int. J. 6:41-42.

Loewe L (2000) "Geschwindigkeitsbegrenzungen für adaptive Evolution", Stud. Int. J. 7:31-33.

Website maintained

http://www.evolutionary-research.net/

Contact Information

Laurence Loewe

email: Laurence.Loewe@evolutionary-research.net (will remain)

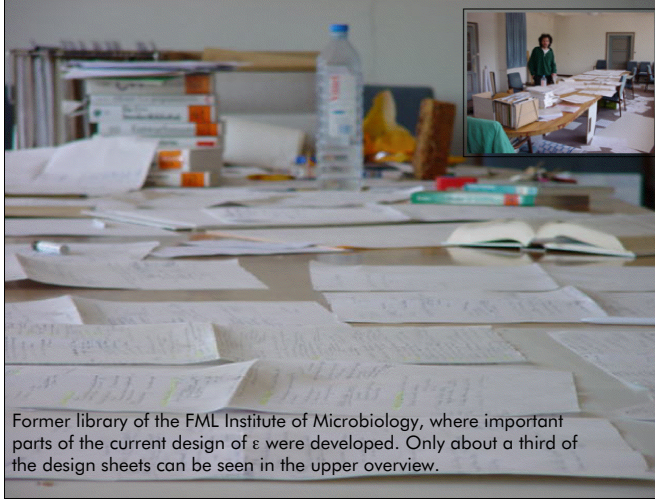
My new work address:

Institute of Cell, Animal and Population Biology, University of Edinburgh
Ashworth Laboratories, Kings Buildings, West Mains Road, Edinburgh EH9 3JT, UK
Tel.: +44 (131) 6 50 7330

Where most of this work was conducted:

Microbial Ecology Group, Department of Biosciences, Technische Universität München
Weihenstephaner Berg 3, 85354 Freising, Germany
Tel.: +49 (8161) 71 3851, Fax.: +49 (8161) 71 4492
Private: Prechtlstr. 8a, 85354 Freising, Germany
Tel.: +49 (8161) 230 116 - email: Laurence.Loewe@web.de

Acknowledgements



Former library of the FML Institute of Microbiology, where important parts of the current design of ε were developed. Only about a third of the design sheets can be seen in the upper overview.

I will probably never be able to complete this section of my thesis. Too many have helped me to finish this work by their encouragement, the insight they shared in discussions and by their practical help.

First of all I thank Siegfried Scherer for having the courage to allow me to conduct this work: This turned out to be hard for you and it was hard for me, but I would have never known this exciting field, if you had not allowed me to choose this risky path. Thank you for the freedom to develop my own ideas about evolution and for organising the funding I needed to pursue them. Ausdrücklichen herzlichen Dank an die Mitarbeiter der mikrobiologischen Routine: Sie haben durch ihre tägliche Arbeit dazu beigetragen, daß unser Institut über genügend finanzielle Mittel verfügt hat, um mir diese Arbeit zu ermöglichen. Then a big thank you to Stephan Peter and Volker Textor. Stephan implemented the first design of ε at a time where I still had to learn C++: I learned a lot from your work. Volker implemented the fitness-measurement-analysis software in VBA and did the largest foundational part of the experiments reported in here: A very important part of this work would be missing without your efforts. Thank you to Jeff Blanchard and Mike Lynch, who supplied me with the bacterial strains used in this study. The development of the lines in this study was supported by a National Institutes of Health grant to Michael Lynch.

Thank you to Eberhard Bertsch for being second examiner of this thesis and for accompanying this work over the years from a distance: Your advice was pivotal in setting me on the IT-path I have chosen. Thank you to my parents for supplying the PowerBook I used over the initial years for my daily work and for supplying the Windows 95 PC I needed so badly for making my first steps in C++. Thank you to my aunt Fiat Loewe and to my wife Martha Loewe for the iBook with Dreamwaver, R and enough RAM: I *could not* have completed writing without it! Thank you also to Mr. Alkofer, Mr. Melzer, Mr. Schöffmann and the others, working on the demolition of the old FML facilities: You allowed me to use the old 72m^2 FML library for a key part of the design of ε. I would not have been able to write Chapter 10 or start evolution@home without the months of intense interconnection of diverse simulation-related topics in that room. Nachdem in meinem neuen FML Büro nicht genug Platz war, um ein solch kompliziertes Design zusammenzuschreiben, bin ich Wilhelm und Inge Schmid ganz besonders dankbar: Ohne die 6 Monate in dem geräumigen Dachboden Eures Gästehauses wäre diese Arbeit wahrscheinlich nie fertig geworden. Es war einfach so gut, den ersten Teil der Diss bei Euch zu schreiben und Ihr habt mich über die Maßen ermutigt. Wer immer mal etwas Ruhe oder einfach nur eine Übernachtung in Memmingen braucht, soll ruhig mal bei Euch vorbeischauen (http://www.ghs-mm.de)!

More people than I can list have helped the development of the professional side of this work by discussing its various biological and information technological aspects with me and sometimes only by small signs that pointed me in the right direction. From the FML Institute of Microbiology: Siegfried Scherer and his wife Sigrid Hartwig-Scherer, Klaus Neuhaus, Clara Kaplan, Ralf Mayr, Helene Oberreuter and her husband Joachim Charzinski, Monika Ehling-Schulz and her husband Stephan Schulz, Felix von Stetten, Sonja von Stetten, Martin Loessner, Markus Zimmer, Mareike Wenning, Herbert Seiler, Natasa Vukov, Patrick Schiwek and others. From the LMU Zoological Institute groups whose journal clubs I visited: Ecology: Wilfried Gabriel, Beate Nürnberger, Jonathan Jeschke and others. Evolutionary Biology: Wolfgang Stephan, Ellen Baake, Yuseob Kim, David de Lorenzo, Thomas Städler and others. From various conferences and other occasions: Michael Heisig, Gabriel Wittum, Markus Kirkilionis, Franck Cappello, Cécile Germain and others from their group, Hans-Peter Kriegel, Silvio Macedo, Walter Weber, Hans-Jürgen Apell, Matthias Brehm, Ingo Krause, Martin Bachmaier, Michael Weber, Jessica Meyer-Eiss, Nigel Crompton, Peter van der Veen, Reinhard Junker, Zoltan Tacacz, Hartmut & Claudia Gabriel, Michael Lynch, David Houle, Laurent Excoffier, Tom Parsons, Svante Pääbo, Adam Eyre-Walker, Günther Wagner, Jeff Blanchard, Damian Gessler, Carl Bergstrom, Nancy Moran, Isa Schön, Dunja Lamatsch, David Haigh, Laurence Hurst, Girogio Bernardi, Brian Charlesworth, Mike Whitlock, James Crow, Alex Kondrashov and others... (unintentionally incomplete list, not in order of importance of contributions). Thank you to David Askew for proofreading: You did a great job.

I want to thank those ca. 200 non-anonymous and all anonymous participants of evolution@home: Your contributions of more than 16 CPU years computing time have been of great use for the biological part of this work and your devotion to the task have been very encouraging. Special thanks go to the top individual contributors as of Dec 2002: RabeRudi, Seth A. Keel, Paranoia Retnek, Lord Spagthorpe, arswitchman switchman, John_F_Kennedy and to the top contributing groups Rechenkraft.de, arswitchman, Paranoia, delightfulstrawberry, Victor Ferreira, Thomas Nokleby, and all others.

I want to further thank my family and friends for the ongoing personal support in an adventure that seemed to never end. Thank you Vati, Mutti, Ingrid, Rainer, Fiat, Elisabeth, Ralf, Ursi, Wilhelm, Inge, Andy, Reinhard, Margit, Kristina, Silke, Leo, Doris, Stephanie, Christian, Marion, Matthias, Robert, Sonja, Siegfried, Jessica, Matthias, Robert, Anja, and others.

Extraordinarily great thanks go to Martha: You showed the strength to hold on under very frustrating circumstances and encouraged me to find my way, although it cost you very much. Your rewards will not disapoint you. Thank you, my best friend, for always being there, for going with me to the cinema, supporting and encouraging me, even in my darkest moments. Thank you for your inspiration, for setting me on this track and for helping me to reach the goal.