DEPARTMENT OF PHYSICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Dissertation

Prebiotic self-assembly of polynucleotides via templated ligation: sequence selection, replication and emergent phenomena

Tobias Johannes Göppel

Fakultät für Physik
Technische Universität München

Prebiotic self-assembly of polynucleotides via templated ligation: sequence selection, replication and emergent phenomena

Tobias Johannes Göppel

Vollständiger Abdruck der von der Fakultät für Physik der Technischen Universität München zur Erlangung eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:
Prof. Dr. Friedrich C. Simmel
Prüfende der Dissertation:
1. Prof. Dr. Ulrich Gerland
2. Prof. Dr. Dieter Braun

Die Dissertation wurde am 03.02.2022 bei der Technischen Universität München eingereicht und durch die Fakultät für Physik am 04.03.2022 angenommen.

Abstract

The *RNA world* is the principal hypothesis to explain the emergence of living systems on the early Earth. It states that the separation between the two roles in information transfer – DNA and RNA storing the genetic information and proteins carrying out the encoded functions – observed in all forms of life only occurred at a later stage of evolution. Before that, according to the RNA world, RNA oligonucleotides combined these two roles and acted as both, the carriers of genetic information as well as catalytic entities, promoting their own replication and providing functionality. Hence, the RNA world offers an elegant solution to the dilemma which came first, the information storing polymer or the functional protein. Yet, the RNA world does not explain the origin of the first functional RNA molecules called ribozymes. While recent experiments revealed potential prebiotic pathways to synthesize nucleotides, the mechanisms assembling these building blocks into functional RNA entities remain still unclear.

What is the origin of the RNA world? A starting point to answer this question is to unveil the growth dynamics on the smallest scale, i.e., in systems where mononucleotides and short oligomers collectively evolved into longer strands.

In this collective growth, *templated polymerization*, and *templated ligation* play central roles. In the first process, a primer bound to a template strand becomes extended by nucleotides in a step-wise fashion. In the second process, two strands of any length adjacently hybridized on a third strand get joined covalently. Both processes are intrinsically sequence-selective in a two-fold way: First, the (de)hybridization dynamics on the template strands favor binding of complementary (oligo)nucleotides. Secondly, the ligation step is discriminative since non-complementary nucleotides in the vicinity of the ligation site stall the formation of a new covalent bond. The present thesis aims to shed light on the growth and sequence dynamics resulting from templated polymerization and ligation. To this end, we study three distinct self-assembly scenarios.

In the first scenario, we focus on the growth via templated polymerization, also called primary extension. Given that a weakly catalytically active oligonucleotide emerges by chance from a random polymerization process in the background, the maintenance of its functionality crucially depends on the accurate copying of its sequence before it degrades. Modern cells achieve error rates as low as 10^{-9} with sophisticated enzymes capable of *kinetic proofreading*. In contrast, experiments probing enzyme-free copying of RNA and DNA find error ratios on the order of 10%. With this low intrinsic copying fidelity, plausible replication scenarios require an accuracy-enhancing mechanism. We propose that *kinetic error filtering* could drastically increase the likelihood of producing exact copies of nucleic acid sequences. The mechanism exploits the observation that initial errors in templated polymerization of both DNA and RNA are likely to trigger a cascade of consecutive errors and significantly stall the downstream extension. We incorporate these characteristics into a mathematical model with experimentally estimated parameters and leverage this model to probe to what extent accurate and faulty polymerization products can be kinetically discriminated. Limiting the time window for polymerization prevents the completion of erroneous strands resulting in a pool in which full-length products show enhanced accuracy. This comes at the price of a concomitant reduction in yield. However, the yield rate and the fidelity can be increased simultaneously by kinetic error filtering if the templates are not too long. With cyclically varying environments, e.g., temperature cycles in hydrothermal systems, repeated copying attempts could produce exact copies of sequences as long as 50mers within their lifetime, facilitating the maintenance of (weakly) catalytically active oligonucleotides.

In the second scenario, we drop the idealized assumption of having one distinct template strand with one specific primer statically bound at one specific end surrounded by mononucleotides only. Instead, we assume a pool composed of strands of various lengths without distinct roles. Now, all strands serve as a template, primer, and substrate for the extension at the same time. Within that pool, arbitrary complexes containing multiple strands assemble and disassemble continuously by hybridization and dehybridization. Some of these complexes may give rise to the ligation of two bound strands. In the second scenario, we also make an essential simplifying assumption: The sequence dependence of the (de)hybridization and ligation dynamics is treated in a mean-field picture. In this picture, the stability of the binding between two strands increases exponentially with the length of their overlap while the ligation rate becomes constant. We show that a competition of time scales in the self-assembly process generically leads to non-monotonic strength distributions with two distinct length scales. The first length scale characterizes the onset of a strong non-equilibrium regime and is visible as a local minimum. The dynamics in this regime is governed by extension cascades, where the elongation of a bound strand with a short building block is more likely than its dehybridization. The second length scale appears as a local concentration maximum and reflects a balance between degradation and dehybridization of completely hybridized double strands in a hetero catalytic extension-reassembly process. Analytical arguments and extensive numerical simulations allowed us to predict and control these emergent length scales. Experimental non-monotonic strand-length distributions confirming our theory are obtained in a setup with thermal cycling using random DNA sequences. The second scenario emphasizes the role of structure-forming processes for the earliest stage of prebiotic evolution. The accumulation of strands with a typical length reveals a possible starting point for higher-order self-organization that ultimately leads to a self-replicating evolving system.

In the third scenario, we go one step further and discard the mean-field approach employed in the second scenario and treat sequences explicitly. For simplicity, we use a binary system composed of two complementary nucleotides only. In this explicit treatment of sequences, non-complementary nucleotide pairs are discriminated energetically and decrease the stability of complexes formed of multiple strands. In particular, this also affects those complexes in which ligations can occur. To compute the stability of multi-strand complexes, we employ a nearest-neighbor energy model building on so-called stacking interactions between neighboring nucleotide pairs. Moreover, the rate for ligation is not uniform anymore. Non-complementary nucleotide pairs close to the ligation site slow down the concatenation of two adjacent strands, i.e., lead to kinetic discrimination as in the first scenario. How do sequence-dependent energetic and kinetic properties affect the structure formation in sequence space? Starting from a symmetric pool of mononucleotides and a few short oligomers, can fluctuations at early stages or small sequence-dependent energetic biases break the symmetry and trigger the selection of certain sequence motives? Our main finding in the third scenario is that energetic discrimination between correctly and incorrectly paired nucleotides is not sufficient by itself to promote substantial self-enhanced sequence selection. However, if non-complementary strand termini at the ligation site slow down the ligation step strong enough, distinct patterns in sequence space arise.

Zusammenfassung

Die *RNA Welt* ist die populärste Hypothese zur Erklärung des Ursprungs lebender Systeme auf der frühen Erde. Sie besagt, dass die Rollenteilung im Hinblick auf den Transfer von Information – DNA und RNA speichern die genetische Information und Proteine führen die codierten Funktionen aus – welche in allen Lebensformen zu beobachten ist, erst zu einem späteren Zeitpunkt der Evolution enstanden ist. Zuvor haben nach der RNA Welt RNA Oligonukleotide diese zwei Rollen in sich vereint und sowohl als Träger der genetischen Information als auch als katalytische Einheiten agiert, welche ihre eigene Replikation bewerkstelligen und Funktionalität bereitstellen. Die RNA Welt bietet folglich einen eleganten Ausweg aus dem Dilemma, was zuerst entstand: das informationstragende Polymer oder das funktionale Protein. Die RNA Welt Hypthese gibt jedoch keinen Aufschluss über den Ursprung der ersten funktionalen RNA Moleküle, auch Ribozyme genannt. Neuere Experimente haben zwar präbiotisch plausible Reaktionspfade hin zur Synthese von Nukleotiden aufgezeigt, jedoch liegen die Mechanismen, welche diese Bausteine in funktionale RNA-Einheiten assemblieren, weiterhin im Dunkeln. Worin liegt also der Ursprung der RNA Welt?

Ein möglicher Ansatz, diese Frage zu beantworten, besteht darin, die Wachstumsdynamik auf der untersten Ebene, d. h. in Systemen, in welchen Mononukleotide und kurze Oligonukleotide kollektiv hin zu längeren Strängen evolvieren, genauer zu untersuchen. In diesem kollektiven Wachstum spielen *templierte Polymerisation* und *templierte Ligation* eine zentrale Rolle. In dem ersten Prozess wird ein Primer, welcher an einen Templatstrang gebunden ist, schrittweise durch Mononukleotide verlängert. In dem zweiten Prozess werden zwei Stränge beliebiger Länge, welche direkt nebeneinander auf einem dritten Strang hybridisiert sind, kovalent verbunden. Beide Prozesse sind intrinsisch sequenzselektiv auf zweifache Weise: Erstens bevorzugt die (De)Hybridisierungsdynamik auf dem Templatstrang das Binden komplementärer (Oligo)Nukleotide. Zweitens ist der Ligationsschritt diskriminierend, da nicht komplementäre Nukleotide in der Umgebung der Ligationsstelle die Bildung neuer kovalenter Bindungen verlangsamen. Die hier vorliegende Arbeit möchte die Wachstums- und Sequenzdynamik genauer untersuchen, welche aus templierter Polymerisation und Ligation resultiert. Dieses Ziel vor Augen studieren wir drei unterschiedliche Szenarien zur Selbstassemblierung.

Im ersten Szenario fokussieren wir uns auf das Wachstum durch templierte Polymerisation, auch Primerverlängerung genannt. Nehmen wir an, dass ein schwach katalytisches Oligonukleotid aus einer im Hintergrund ablaufenden zufälligen Ligation ohne Templat entsteht, so hängt der Erhalt von dessen Funktionalität von der akkuraten Replikation der Sequenz, bevor diese degradiert, ab. Moderne Zellen erreichen durch hoch entwickelte Enzyme, fähig zum *kinetischen Korrekturlesen*, extrem niedrige Fehlerraten von ca. 10^{-9}. Im Gegensatz dazu werden in Experimenten mit RNA und DNA zur nicht-enzymatischen Vervielfältigung Fehlerraten der Größenordnung 10% gemessen. Ausgehend von dieser niedrigen Güte bei der Replikation verlangen plausible Vervielfältigungsszenarien nach einem Mechanismus, welcher die Güte erhöht. Wir schlagen vor, dass das *kinetische Ausfiltern von Fehlern* die Wahrscheinlichkeit, eine exakte Kopie zu produzieren, drastisch erhöht. Der Mechanismus nutzt die Beobachtung aus, dass initiale Fehler bei der templierten Ligation von DNA und RNA Kaskaden von nachfolgenden Fehlern auslösen und die weitere Verlängerung signifikant verlangsamen. Wir beziehen diese Charakteristika in unser mathematisches Modell mit ein und untersuchen, bis zu welchem Grad akkurate und fehlerhafte Polymerisierungsprodukte kinetisch unterschieden werden können. Eine Begrenzung des Zeitfensters für die Polymerisierung verhindert die Vervollständigung fehlerhafter Stränge. Dies führt dazu, dass Produkte der vollen Länge eine erhöhte Güte aufweisen. Diese hat jedoch eine deutlichen Reduktion des Ertrags zur Folge. Allerdings können Ertrag pro Zeit und Güte simultan durch das kinetische Ausfiltern von Fehlern gesteigert werden, sofern die Templatstränge nicht zu lang sind. Sich zyklisch verändernde Umgebungsbedingungen, z.B. induziert durch Temperaturzyklen in hydrothermalen Systemen, resultieren in wiederholten Replikationsversuchen. Dadurch können exakte Kopien von Strängen, bestehend aus bis zu 50 Nukleotiden, innerhalb von deren Lebenszeit generiert werden, was den Erhalt von (schwach) katalytisch aktiven Oligonukleotiden erleichtert.

In dem zweiten Szenario lassen wir die idealisierende Annahme eines Systems bestehend aus einem bestimmten Templatstrang an welchen an einem spezifischen Ende ein bestimmter Primer gebunden ist, umgeben von Mononukleotiden, fallen. Stattdessen nehmen wir einen Pool, welcher sich aus Strängen beliebiger Längen ohne bestimmte Rolle zusammensetzt, an. Nun dienen alle Stränge gleichermaßen als Templat, Primer und Substrat für die Verlängerung. Innerhalb dieses Pools bilden sich kontinuierlich Komplexe aus mehreren Strängen und zerfallen wieder. Manche dieser Komplexe ermöglichen die Ligation zweier gebundener Stränge. Auch in diesem zweiten Szenario nehmen wir eine zentrale Vereinfachung vor. Die Sequenzabhängigkeit der (De)Hybridisierungsdynamik wird in einem Molekularfeld Bild behandelt. In diesem Bild wächst die Bindestabilität zwischen zwei Strängen exponentiell mit deren Überlapp an, wohingegen die Ligationsrate einen konstanten Wert annimmt. Wir zeigen, dass eine Kompetition zwischen den Zeitskalen des Assemblierungsprozesses generisch zu einer nicht monotonen Längenverteilung der Stränge führt. Die erste Längenskala charakterisiert den Beginn eines Regimes fern des Gleichgewichts und manifestiert sich als lokales Minimum. Die Dynamik innerhalb dieses Regimes wird durch Extensionkaskaden bestimmt, in welchen die Verlängerung eines Stranges mit kurzen Bausteinen wahrscheinlicher wird als dessen Dehybridisierung. Die zweite Längenskala erscheint als lokales Maximum und spiegelt die Balance zwischen Dehybridisierung und Ausfluss vollständig hybridisierter Doppelstränge in einem heterokatalytischen Verlängerungs- und Reassmeblierungsprozess wider. Analytische Argumente und extensive numerische Simulationen ermöglichen es uns, die emergenten Längenskalen vorherzusagen und zu kontrollieren. Experimentelle, nicht monotone Stranglängenverteilungen, welche unsere Theorie bestätigen, konnten wir in einem Aufbau mit Temperaturzyklen unter der Verwendung von zufälligen DNA Sequenzen erzielen. Das zweite Szenario unterstreicht die Rolle von Strukturbildung bereits auf der untersten Ebene. Die Akkumulierung von Strängen einer typischen Länge könnte einen Ausgangspunkt darstellen für eine nächsthöhere Form der Selbstorganisation, welche schließlich in selbstreplizierenden und evolvierenden Systemen aufgeht.

In dem dritten Szenario gehen wir wieder einen Schritt weiter und verwerfen den Ansatz des Molekularfelds aus dem zweiten Szenario und behandeln Sequenzen nun explizit. Einfachheitshalber nehmen wir hier ein binäres System, bestehend aus zwei komplementären Nukleotiden an. In der expliziten Behandlung der Sequenzen werden nicht komplementäre Nukleotide energetisch diskriminiert, da diese die Stabilität von Komplexen, bestehend aus mehreren Strängen, herabsetzten. Um die Stabilität von Komplexen zu bestimmen, verwenden wir ein Energiemodell basierend auf der Stapelwechselwirkung benachbarter Nukleotidpaare. Des Weiteren ist die Ligationsrate nicht mehr uniform. Nicht komplementäre Nukleotide in der Nähe der Ligationstelle verlangsamen das Verbinden zweier direkt benachbarter Stränge und ziehen folglich eine kinetische Diskriminierung nach sich wie schon im ersten Szenario. Wie beeinflussen die sequenzabhängigen energetischen und kinetischen Eigenschaften die Bildung von Struktur im Sequenzraum? Ausgehend von einem symmetrischen Pool, zusammengesetzt aus Mononukleotiden und wenigen kurzen Oligomeren—können Fluktuationen zu frühen Zeitpunkten oder kleine, sequenzabhängige energetische Differenzen die Symmetrie brechen und die Selektion von bestimmten Sequenzmotiven auslösen? Unsere zentrales Ergebnis in diesem Szenario besteht darin, dass die energetische Diskriminierung zwischen komplementären und nicht komplementären Nukleotiden alleine nicht ausreicht, um eine substanzielle Sequenzselektion hervorzubringen. Wenn jedoch nicht komplementäre Strangenden an der Ligationsstelle den Ligationsschritt zusätzlich verlangsamen, manifestieren sich eindeutige Muster im Sequenzraum.

Contents

1. Introduction

1.1. What is life?

"*What is life?*" This seemingly simple question was asked by the Nobel Prize-winning physicist Erwin Schrödinger on the cover of his famous book in 1944 [1]. Since then, numerous attempts have been made to answer this question [2]. However, a unique definition of what *life* or *alive* means still does not exist [3]. Moreover, we do not even know whether universal laws that are inherent for all forms of biological life on Earth or elsewhere exist [2]. In his article "The Seven Pillars of Life" [4], the former editor of *Science*, Daniel Koshland, reports about a conference where the world's scientific elite gathers to work out a definition for life in a humorous way. After a long debate, the scientists seemed to converge on a common solution: "The ability to reproduce—that is the essential characteristic of life, said one statesman of science. Everyone nodded in agreement that the essential of a life was the ability to reproduce, until one small voice was heard. Then one rabbit is dead. Two rabbits—a male and female—are alive but either one alone is dead. At that point, we all became convinced that although everyone knows what life is there is no simple definition of life." According to astrophysicist Christopher McKay, "it is not all that surprising that we do not have a fundamental understanding of what life is" since we have "only one example of life—life on Earth" [3].

Scientists from different communities might have different ideas about what life or being alive could mean. In 1986 the evolutionary theorist Richard Dawkins wrote that "It is information, words, instructions... If you want to understand life, don't think about vibrant, throbbing gels and oozes, think about information technology [5]". Most scientists agree that an essential aspect of life is its capability to evolve and adapt. In their review article "Life is Physics: Evolution as a Collective Phenomenon Far From Equilibrium" [6] from 2011, Nigel Goldenfeld and Carl Woese emphasize that "evolution is a fundamental physical process" that should be treated as "a problem in non-equilibrium statistical mechanics." This statement is in agreement with the perception of life from Munich-based physicists of the Mast- and Braun-lab, saying that "Living systems must be out of equilibrium in order to create structures, enable replication to persist against decay, and to give rise to Darwinian evolution [7]."

In 1994 the NASA Exobiology Discipline Working Group proposed a working definition of life as “a self-sustaining chemical system capable of Darwinian evolution [8].” NASA’s working definition is taken up either explicitly or implicitly in many biology-, chemistry- or physics-oriented articles on the *origins of life* to describe what life resembles [9, 10, 11, 12, 13]. There seems to be some consensus within the community that this working definition, while not perfect, is sufficient to capture the most important aspects of life. This working definition is also the basis for the thesis presented here.

1.2. How did life emerge?

Finding potential answers to the question of how life could have emerged on the early Earth is one of the greatest and most exciting scientific challenges. What are the chemical, geological, and physical boundary conditions that triggered the emergence of the first self-replicating and living systems? “Modern” life as we know it is the result of an ongoing mutation, selection, and replication process. What did the very first steps in the transition from dead to living matter look like? And where on our planet did they occur? Scientists from different disciplines such as biology, chemistry, geology, astronomy, and physics, as well as philosophy, are gathering behind these questions.

If we decide to search for answers to the question about the origins of life, we will have to accept that our answers can only be formulated as hypotheses. The hypotheses may seem more or less plausible. However, none of the hypotheses put forward will ever be conclusively proven [14, 15] because many of the earliest fossils of “ancient” life, which would be necessary to definitely verify or reject a given hypothesis, have been irretrievably lost in the course of time [16, 17]. This prospect may seem dreary at first glance. However, the idea that every plausible theory of the origin of life might have been (or might be) realized elsewhere in the universe, if not on Earth, may reconcile us with the search for potential answers. Organic molecules, which are essential for life on the planet Earth, have already been found in meteorites, cosmic clouds, and protoplanetary discs [18, 19, 20, 21]

A universal trait of “modern” life is the *central dogma of molecular biology* (see Fig.1.1). The central dogma of molecular biology states that genetic information of any living being is stored in the particular sequence of its nucleic acids. This information can be transferred to sequences of amino acids, i.e., translated into functional proteins. However, information, in the sense of a defined sequence of amino acids, cannot be transmitted from one protein to another, nor can it be translated back to a sequence of nucleic acids [23].

How long can we trace back the existence of “modern” life implementing the central dogma on Earth? The oldest reliable microfossils of such forms of life date back to about 3.3-3.5 billion years [7, 24]. This makes “modern life” almost as old as our plant,

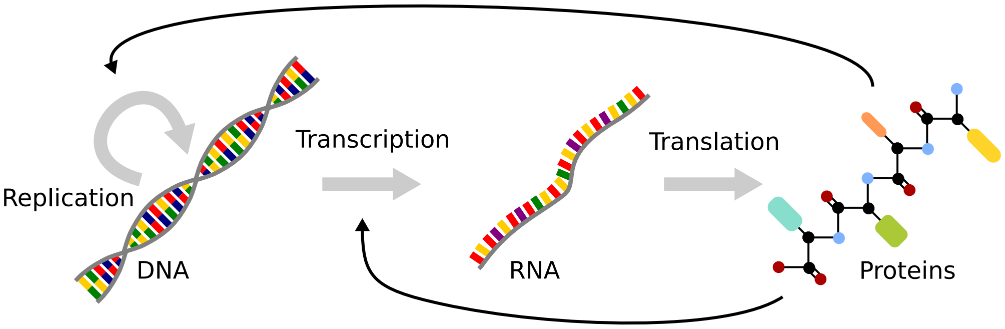


Figure 1.1.: Central dogma of molecular biology, protein biosynthesis, and DNA replication. Genetic information is first transcribed from DNA into RNA and then translated into sequences of amino acids, i.e., proteins. Information cannot be exchanged between proteins or transferred back into DNA. During protein biosynthesis, a polymerase enzyme transcribes sections, i.e., gens on the DNA into RNA. According to the RNA sequence, a new protein is assembled at the ribosomes. During DNA replication, polymerase enzymes assemble (complementary) copies of the DNA strands. The blueprints for the polymerase enzymes are contained in the DNA sequence. This figure has been newly created for this thesis. It is based on Fig. 1.2.1 in Ref. [22].

which was formed about 4.5 billion years ago. The universe is estimated to be 13.7 billion years old. Presumably, about 0.2 billion years after its formation, the planet Earth carried water which is considered essential for any form of life for the first time [7, 13]. After the first half billion years of its existence, the Earth became the target of a large number of heavy impacts by asteroids. The size of these asteroids was probably so large that their impacts triggered the evaporation of the oceans and sterilization of the Earth's surface [25]. Hence, one could conclude that life must have emerged (or reemerged after getting extinct by the heavy bombardment) between 4 and 3.5 billion years ago [13]. After that, the transition from the simplest unicellular "modern" forms of life towards complex multicellular organisms took roughly ten times longer than the emergence of the first unicellular organisms. All in all, the emergence of life, despite its enormous complexity, was an astonishingly fast and effective process. The transition to "complex modern" life is the result of Darwinian evolution, i.e., the continuous interplay of mutation, adaptation, and selection pressure implemented, e.g., by the competition for limited resources. Our knowledge about the evolutionary processes, that set in after the first single-cell organisms were formed, comes from the comparison of DNA sequences of various species and classes of proteins [7].

What did "ancient" life on the early Earth look like before the first simple yet "modern" single-celled organisms based on the central dogma of molecular biology started to populate the Earth? How did "ancient" life evolve to "modern" life? And how did this "ancient" life emerge from dead matter in the first place? These questions cannot be answered by means of a simple biological back-extrapolation [7], such as phylogenetic attempts to reconstruct the tree of life [26]. However, we can put forward plausible scenarios based on chemical, physical, and geological arguments and speculate about boundary contentions that must be met for life to emerge [7]. Three such boundary conditions are formulated in Ref. [7]. They read:

1. Life can only exist far away from equilibrium. Due to the second law of thermodynamics, this is an indispensable condition. Only a system that is coupled to some active driving mechanism that keeps it away from the equilibrium state permanently allows for a low local entropy. In a state of high entropy, complex structures of life can neither be created nor maintained.

2. The basic building blocks for more complex biomolecules must have formed earlier. Then the basic units must have been made available in our "origin of life setup." It is highly likely that the concentrations of the basic building blocks were quite small initially. As we will see in the following, there is a lot of evidence for a RNA-based prebiotic chemistry.

3. The transition from prebiotic chemistry to self-replicating and evolving biomolecular structures most likely occurred in a closed system. Molecules need to react over a long time to form complex structures. Thus, there is a need for some trapping mechanism that prevents the molecules and the reaction products from diffusing away, as they would do in an open system. At the same time, there must be some feeding mechanism that brings fresh molecules into the system in order to prevent the system from relaxing to an equilibrium state. A porous volcanic rock structure could fulfill both requirements. It is sufficiently permeable to allow for a steady influx of basic units whilst efficiently hindering longer molecules from leaving the system, i.e., from getting lost. If the rock structure is exposed to a temperature gradient, mechanisms that lead to high local concentrations of molecules are available [27, 28, 29].

1.3. The RNA world

In "modern" cellular life implementing the central dogma of molecular biology, the storage of genetic information is largely separated from function. DNA double strands are the carriers of genetic information. During protein biosynthesis, DNA double strands are partially opened by proteins with helicase activity, and the genetic information is transcribed into RNA strands by a polymerase enzyme. These RNA sequences, in turn, are then translated into sequences of amino acids at the risbosome which then fold into proteins [30]. The proteins carry out different functions to maintain the metabolism, to protect the cell against environmental stress, and to replicate the genetic information during cell division with high accuracy (see Section 2.4 in Chapter 2). According to the central dogma, the blueprints for the proteins with helicase activity as well as polymerase enzymes transcribing sections from DNA sequence into RNA sequences, or copying a given DNA strand into a new one, are also contained in the DNA. Hence, enzymes cannot exist without chains of nucleic acids, and vice versa. Therefore, thinking about the emergence of life on the early Earth, a dilemma of the chicken-and-egg type arises. Which came first, the nucleic acids carrying the genetic instructions for the formation of enzymes or the enzymes that copy and assemble nucleic acids?

The so-called *RNA world hypothesis* formulated in the sixties provides a solution to the dilemma raised above [31, 32, 33, 34, 35, 36]. It states that DNA and proteins were absent in "ancient" forms of life and attributes a dual role to RNA molecules. The RNA world hypothesis proposes that RNA oligonucleotides acted as the (only) carriers of genetic information and as functional molecules able to promote their own replication at the same time. When it was formulated, the existence of catalytic RNA complexes, so-called ribozymes, was purely speculative [37]. It was 15 years later when the first ribozyme was actually discovered [38]. Since then, many RNA sequences, which, via Watson-Crick base-pairing, fold in secondary structures that show catalytic activity, have been found or engineered in the laboratory by in vitro evolution. Some of these RNA structures are able to assist the assembly of larger RNA oligomers from shorter stands and single ribonucleotides (see Refs. [39, 40, 41, 42, 43, 44] for examples for recent experiments investigating ribozyme assisted assembly). In Ref. [28], Salditt et al. present a high yielding prebiotically plausible realization of a PCR [45] set-up. Copying of RNA strands starting from RNA primers is driven by an RNA polymerase developed early in Ref. [46]. Separation of product and template strand is not achieved by periodic changes of the global temperature as in a "classical" PCR situation. Instead, strands are moved inside the reaction chamber by laminar convection induced by a temperature gradient. Replication and strand separation occurs at cold and hot regions of the chamber, respectively. Also, as a result of the temperature gradient, the RNA polymerases accumulate in a cold region of the chamber where they do not suffer from fast degradation due to hydrolysis. Natural realizations of reaction chambers subjected to steep temperature gradients are found in porous rocks in the vicinity of hydrothermal vents on the ground of the ocean. The results of Salditt et al. "demonstrate a size-selective pathway for autonomous RNA-based replication in natural nonequilibrium conditions." Despite the recent progress in engineering "performant" ribozymes by in vitro evolution, a functional RNA complex that could replicate itself autonomously within a pool of single ribonucleotides and short and unstructured RNA oligomers is still missing [11]. Identifying such an RNA sequence would be a major scientific advance and provide further evidence for the RNA world hypothesis.

The discovery that the ribosomes, the macromolecular complexes in the cells of all living systems where amino acids are assembled into proteins, are for large parts formed of RNA [47, 48, 49] strongly supports the RNA world hypothesis. Although the ribosome contains protein compounds as well, the catalytic core unit where the formation of peptide bonds between different amino acids occurs is entirely made out of RNA [50]. Moreover, a study where significant parts of the ribosomal proteins were removed by proteinases indicates that a ribosome can still keep a considerable level of its catalytic activity [51]. For that reason, one can say that "the ribosome is a ribozyme [52]." Therefore, "the smoking gun is seen in the structure of the contemporary ribosome" [37]. In essence, RNA can store, assemble, copy, modify (mutations during copying, catalyzed cleavage [53]), and translate genetic information into proteins. Thus, it seems not too far-fetched that the "modern" DNA-RNA-protein world emerged from an "ancient" RNA world. The ribosome is also interpreted as "a missing link in the evolution of life" [49] or as a "molecular fossil" [54] from the ancient RNA world.

1.4. Prior to the RNA world

Recent experimental work identified potential reaction pathways, allowing for the synthesis of ribonucletotides from prebiotically plausible precursors [55, 56, 57]. It is believed that these single ribonucleotides can polymerize into relatively short RNA oligmoers with random sequences [58, 59, 60]. However, it is still unclear how the first ribozymes could have occurred from a mixture of single ribonucleotides and relatively short and unstructured RNA oligomers [42, 61, 59, 28, 62, 63]. The smallest functional RNA complexes known today are roughly 30 to 100 nucleotides long [64, 53, 65]. However, ribozymes complex enough, such that they are capable of, e.g., assisting replication, are likely to be at least 150 nucleotides long [46, 66, 67]. For polymers of a length between 30 to 150 nucleotides, not less than 10^{18} to 10^{90} distinct sequences are possible. Yet, the subset of catalytically active sequences within that enormous pool is assumed to be marginal [58]. Hence, the spontaneous formation of a functional sequence from a random pool of single nucleotides and short oligmoers seems improbable [58, 68]. And even if a weakly functional RNA complex is created by chance, it would probably be lost quickly due to degradation and cannot evolve further if there are no mechanisms available to replicate its sequence accurately. This leads us to the question, how the RNA world emerged?

It is proposed that an unstructured pool of initial RNA oligomers with a high sequence entropy has to undergo some sort of pre-selection or filtering process that reduces the overall sequence entropy of the pool before a Darwinian-evolution-like search for ribozymes with increasing catalytic abilities could kick-start [58, 59].

Two purely chemical processes known as *enzyme-free templated polymerization* and *enzyme-free templated ligation* might have played a central role in such an initial phase of selection. In the first process, a so-called primer strand that is bound to a longer so-called template strand becomes extended by single nucleotides in a step-wise fashion. This process is the nonenzymatic analog to copying a DNA strand by a polymerase in modern cell replication. In the second process, two strands of any length adjacently hybridized on a third strand get joined covalently. Both processes have been realized experimentally [69, 70, 71, 72, 73, 74, 75, 68]. The two ligation processes are intrinsically sequence-selective in a two-fold way. First, the (de)hybridization dynamics on the template strands favor the binding of complementary (oligo)nucleotides. Primarily non-complementary overlaps lead to unstable configurations that dissociate quickly [76]. Later on, we will refer to this form of discrimination as *thermodynamic discrimination*. Second, the ligation step is selective since non-complementary nucleotides in the vicinity of the ligation site stall the formation of a new covalent bond [77, 78]. Hence, non-complementary strand termini are unlikely to be extended. We will refer to this phenomenon as *kinetic discrimination*. As a result, the products of the two ligation processes are likely to result in partial complementary copies of the templating strand. Repeating the process over and over again could lead to a self-assembly and copying dynamics that might increase the abundance of certain sequence motifs and lower the overall entropy of the sequence pool. As strands become more similar, i.e., complementary, the probability of forming stable complexes in which other ligation reactions could occur increases, and the growth dynamics will become even more enhanced. However, misincorporations frequently occur during non-enzymatic copying and self-assembly self-selection might subvert sequence selection process.

However, it is challenging to experimentally investigate the enzyme-free self-assembly processes from a random pool of mononucleotides and short oligmomers into longer and potentially more structured strands. Generally, the experiments require long times, while the reaction yields remain low and undesired side products obscure the results. Moreover, the question of how to track the evolution of the whole sequence pool resulting from templated polymerization simultaneously remains an unsolved technical problem [79, 80, 81, 62, 63].

Two recent proof-of-principle experiments demonstrated that self-assembly via templated ligation indeed leads to the formation of long strands with reduced sequence entropy via cooperative modes of growth. These experiments used DNA strands of length twelve or twenty as starting material and employed a ligase enzyme to speed up the assembly dynamics [58, 59].

1.5. Aims of this thesis

In this thesis, we approach replication and self-assembly via enzyme-free templated polymerization and ligation from a theoretical point of view through mathematical modeling and large-scale computer simulations. Our theoretical models allow us, under certain simplifying assumptions, to explore growth and replication regimes that are not (yet) accessible experimentally. We study three different scenarios. In the first scenario, we consider an idealized templated-polymerization process, where a template strand gets copied repeatedly by the extension of a primer strand in a step-wise fashion and probe under which conditions accurate replication is possible. In the second scenario, we assume a pool composed of strands of various lengths without distinct roles. All strands can serve as a template, primer, and substrate. To focus on emergent properties and different dynamical growth regimes in the self-assembly process, we treat the sequence dependence in a mean-field picture. The third scenario also assumes a mixed pool. This time, we make the sequence dependence of the different steps in templated polymerization and ligation explicit, study the sequence dynamics in detail, and investigate how thermodynamic and kinetic discrimination can break the symmetry in sequence space of unbiased initial pool. No simulation tools were available to study the complex self-assembly processes emerging from mixed pools in the second and the third scenario in detail. Therefore, we made a considerable effort to fill the gap and developed a suitable software framework in the first place. Currently, this software framework is employed in four follow-up research projects by master and Ph.D. students of the Gerland group.

This thesis is structured as follows: In Chapter 2 we review some key papers for enzyme-free copying and self-assembly and provide the background information that we will need to motivate our theoretical models later on. Chapter 3 is dedicated to the first replication scenario. Chapter 4 describes the simulation method that we developed to investigate the self-assembly dynamics from mixed pools in detail. Chapter 5 and Chapter 6 consider to the second and the third assembly scenario, respectively. In Chapter 7 we summarize our results and presents ideas for follow-up projects.

2. Background

2.1. Properties of ribonucleic acids

2.1.1. Basic building blocks of ribonucleic acids

In this section, we briefly review the structural properties of ribonucleic acids (RNAs). The section is based on the corresponding chapters in Refs. [30, 82].

In "modern" life, RNA appears as mRNA (messenger RNA), tRNA (transfer RNA), and rRNA (ribosomal RNA). The mRNA molecules transmit the genetic information that was first transcribed from the DNA in the nucleus to the ribosomes, where the protein biosynthesis is carried out. In contrast to mRNA, tRNA and rRNA do not encode for genetic information. The role of tRNA is to transport the correct amino acid to the corresponding codon on the mRNA during the translation process. The rRNA constitutes the active core of the ribozymes in the cytoplasm, assembling amino acids into proteins via the formation of peptide bonds.

RNA strands are composed of by ribonucleotides that are linked via phosphodiester bonds. Phosphodiester bonds belong to the class of covalent bonds [83]. The basic building blocks of each ribonucleotide are the nucleobase, the ribose (5-carbon sugar molecule), and the phosphate group. The nucleobase and the phosphate group of a ribonucleotide are attached to the carbon atom at 1′- and 3′-positions of the ribose molecule, respectively. Moreover, the phosphate group is also attached to the carbon atom at the 5′ position of the ribose molecule belonging the following nucleotide (see Fig. 2.1). In case of DNA the sugar molecule is a 2-deoxyribose. While other nucleobases exist, modern life solely uses the four nucleobases adenine (A), cytosine (C), guanine (G), and uracil (U) as basic RNA building blocks. The single characters are commonly used to refer to the different nucleobases. In DNA, thymine (T) is used instead of uracil. A, C, G, U, and T are called natural or canonical nucleobases.

Nucleobases that do not appear in living organisms are referred to as non-canonical or non-natural nucleobases [84, 85, 86, 87]. Why nature only employs four nucleobases, i.e., four different "characters" to encode genetic information, is an open question [88, 89]. In principle, using a larger number of "characters" would allow for more efficient storage of genetic information since the length of the total genome could be significantly shorter. Indeed, so-called Hachimoji DNA and RNA systems containing eight types of

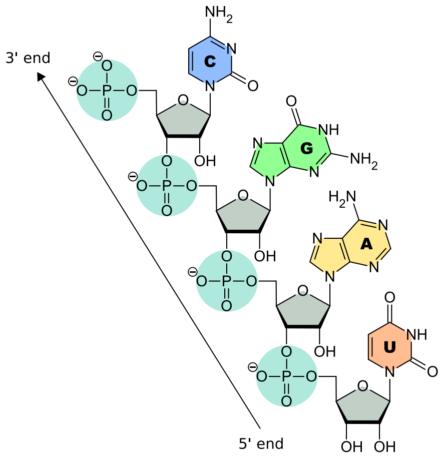


Figure 2.1.: An RNA oligomer composed of the four canonical nucleic bases C, G, A, and U linked via a sugar (gray) and phosphate backbone (turquoise). The 5′ and 3′ end determine the direction of the RNA strand. The figure is adapted from Wikimedia Commons: 'RNA-Nucleobases' published under public domain [91].

nucleobases have been proposed as a way to store digital data [90]. The fact that the modern genetic code uses only four types of nucleobases is also described as a "frozen accident [34]." With our current knowledge, we can not exclude that "ancient life" used other nucleobases than modern life (see Section 6.4.3).

The sugar backbone has a "loose" phosphate group at one end (see Fig. 2.1). This end is called the 3′ end, while the other end is called the 5′ end. These two distinct ends define the direction of an RNA strand. During translation, the mRNA strand is synthesized, starting with the 5′ end. For this reason, we say that an RNA strand points from the 5′ to the 3′ end and starts with the 5′ when we write down its sequence.

2.1.2. Secondary structures of ribonucleic acids

In the nuclei of living systems, DNA is generically found in a double-stranded helical conformation. In contrast, RNA typically appears in single-stranded form (mRNA, tRNA). Nonetheless, RNA strands can form double-stranded helical structures as well [92, 76]. The structures of typical DNA and RNA helices differ slightly. Typically, an RNA helix is more compact and also more stable than a DNA helix and counts approximately ten nucleotides per helical turn. In the generic case, an RNA helix is formed by two complementary strands pointing in opposite directions. U is the complementary nucleobase to A, and G is the complementary nucleobase to C. In a helical structure, complementary nucleobases are linked via hydrogen bonds, i.e., non-covalent bonds. Two complementary bases linked by one (A and C) or two (C and G)

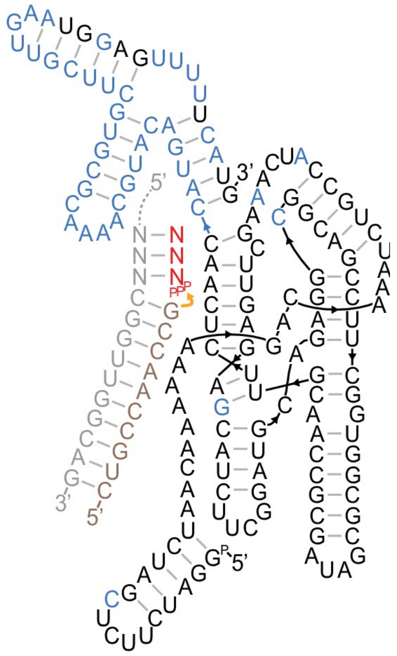


Figure 2.2.: Secondary structure containing loops and single-stranded segments of the 153 nucleotides long t5 ribozyme capable to extend a primer strand (light brown) bound to a template strand (light gray) with an oligmor of length three (red). Here, NNN stands for an arbitrary triplet sequence. This figure is adapted from Fig. 4 in Ref. [40] published under CC BY 4.0 License.

hydrogen bonds form so-called Watson–Crick base pairs. The two non-complementary nucleobases G and U can form so-called wobble pairs. These wobble pairs have thermodynamic properties similar to Watson–Crick base pairs. However, they perturb the regular helical structure when introduced in a double strand [93].

The bending stiffness of RNA strands is characterized in terms of the so-called persistence length. RNA-oligomers, which are significantly shorter than the persistence length, can be considered stiff and modeled as rigid rods. In contrast, RNA polymers that are much longer than the persistence length are considered flexible and are modeled as a worm-like chain. In between, the RNA molecule is called semi-flexible, and more complex models are needed to describe its behavior. The exact persistence length depends on the properties of the solvent and the temperature [94]. The persistence length of double-stranded RNA and DNA is of the order of 150 base pairs, with RNA being somewhat stiffer than DNA, while the persistence length of single-stranded RNA and DNA only corresponds to a few nucleotides [95, 96, 97, 98, 99, 100].

Double-stranded helical structures are not the only secondary structure that RNA strands can take. Far more complex structures, including single-stranded segments and "non-linear" elements such as loops and branches, are possible and even required for RNA strands to become catalytically active see Fig. 2.2. In Section 2.1.3 we will have a closer look at the energetic properties of RNA secondary structures and their thermal stability.

In our theoretical models presented in Chapters 3–6, we will treat nucleic-acids strands in a coarse-grained way. In particular, we will not explicitly treat the different components of the nucleotides. Instead, our basic units will be the "entire" nucleotides. In this picture, oligomers are chains of nucleotides that are linked covalently. Moreover, single nucleotides can be interpreted as strands with length one.

We will end this section with a short remark on chirality of nucleic acids. Since ribose is a chiral molecule, the entire nucleotide is chiral as well. In the laboratory, RNA nucleotides can be synthesized in the D- and L-form. However, living systems only employ nucleotides in the D-form, resulting in fully homochiral nucleic acid strands [101]. It is still a puzzle why biology "chose" to use the D-form for nucleotides during the course of evolution [102, 103, 104] (see also Section 7). Since the general laws of physics and chemistry as we know them do not depend on chirality, a mirror-symmetric biology could be possible in theory [105].

2.1.3. Energetic properties of secondary structures

Every RNA secondary structure is associated with a *Gibbs free energy*. The total Gibbs free energy $\Delta\mathcal{G}_{\mathrm{tot}}$ of a given secondary structure is the sum of the Gibbs free energies that are assigned to the different elements within the secondary structure [106] and an initiation term (apart from minor correction terms resulting from specific transitions between the individual elements [76]). If the complex is formed by only one strand, the initiation term is zero. The different elements of the secondary structure can be *double-stranded helical segments*, *dangling ends*, different kinds of *loops*, and *branching points* [76] (see Fig. 2.2 and Fig. 2.3). In the literature, these elements are also called *structural motifs*. With that, total Gibbs free energy $\Delta\mathcal{G}_{\mathrm{tot}}$ can be written as

$$\Delta\mathcal{G}_{\mathrm{tot}} \approx \sum_{m \,\in\, \substack{\text{structural} \\ \text{motifs}}} \Delta\mathcal{G}_m + \mathcal{G}_{\mathrm{init}}. \tag{2.1}$$

In general, the more negative the Gibbs free energy is, the more stable, i.e., favorable is the secondary structure or the considered element. Typically, the energetic contributions for double-stranded helical segments are negative, while "non-linear" structural motifs typically lead to a positive contribution, i.e., an energetic penalty [76]. The above statements also apply to DNA secondary structures [107]. The energy contributions

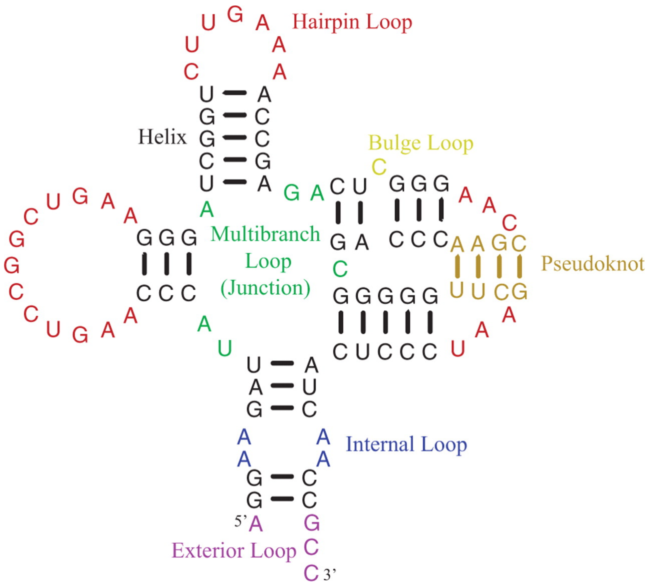


Figure 2.3.: An RNA secondary structure showing all "linear" and "non-linear" elements, i.e., structural motifs for which experimentally measured and extrapolated energy parameters exit is the *Turner* nearest neighbor parameter sets [76]. This figure is adapted from Fig. 2 in Ref. [76] published under CC-BY-NC-SA-2.0-UK License.

associated with many different structural motifs in various sequence contexts have been determined in extensive experiments for both RNA and DNA from melting curve experiments via van't Hoff analysis methods under standardized conditions [108]. Moreover, empirical rules have been found to extrapolate the existing experimental data to new configurations and other conditions (temperature, ionic concentration, etc.). The experimental data are tabulated and documented in the *Turner* and *SantaLucia* database and for RNA and DNA [76, 107]. The databases are the foundation of different software packages for the prediction of RNA and DNA secondary structures such as the *ViennaRNA Package* or *NUPACK* [109, 110].

The "linear" double-stranded helical segment is the simplest structural motif and is also called *stem* or *helix* in the literature [111, 76]. The energy of double-stranded helical segments is due to two contributions: The first contribution comes from the hydrogen bonds, which form between complementary nucleobases or nucleobases forming wobble pairs. The second contribution is the result of so-called stacking interactions between neighboring base pairs [112, 113]. This stacking interaction is a quantum chemical effect and can be explained by means of orbital theory. The contributions cannot be clearly separated from each other [114]. Therefore, they are summarized into so-called *block energies* that are assigned to blocks of neighboring base pairs. These block energies are motif-dependent, i.e., they depend on the type of the pairing bases and their arrangement. With that, the Gibbs free energy of a double-stranded segment is obtained as the sum of the block energies for all nearest-neighbor blocks within the segment. The overall Gibbs free energy of a double-stranded segment is obtained by adding correction terms accounting for special configurations at the ends or for symmetry properties.

In the following, we will present some explicit calculations of Gibbs free energy of RNA secondary structures. The examples are adapted from the tutorial available on the Turner database [76]. These examples will illustrate the general approach for the calculation. However, they will not cover all possible details. Thereby, we will use the parameters for standard conditions, i.e., for 37 °C degrees.

We will start with the simple example of a double-stranded helical segment with a terminal mismatch as shown below.

$$\begin{array}{cccc} 5' & A{-}G{-}C{-}C{-}G{-}U{-}\ \mathrm{G} & & 3' \\ & \cdot\quad\cdot\cdot\quad\cdot\cdot\quad\cdot\cdot\quad\cdot\cdot\quad\cdot & & \\ 3' & U{-}C{-}G{-}G{-}C{-}A{-}\ \mathrm{A} & & 5' \end{array}$$

In this representation a $\cdot$ symbolizes a hydrogen bond between two complementary nucleobases which are part of a helical structure, whereas a $-$ stands for a covalent bond. Our "notation" introduced here differs from the "notation" used in Fig. 2.2 and Fig. 2.3. Our notation will turn out useful in Chapter 6. The Gibbs free energy $\Delta\mathcal{G}^\circ_{\text{tot}}$ for this configuration is given by

$$\begin{aligned}\Delta\mathcal{G}^\circ_{\text{tot}} =& 2\times\Delta\mathcal{G}^\circ\begin{pmatrix}A{-}G\\ \cdot\quad\cdot\cdot\\ U{-}C\end{pmatrix} + 2\times\Delta\mathcal{G}^\circ\begin{pmatrix}G{-}C\\ \cdot\cdot\quad\cdot\cdot\\ C{-}G\end{pmatrix} + \Delta\mathcal{G}^\circ\begin{pmatrix}C{-}C\\ \cdot\cdot\quad\cdot\cdot\\ G{-}G\end{pmatrix}\\ &+\Delta\mathcal{G}^\circ_{\text{non}}\begin{pmatrix}U{-}\ \mathrm{G}\\ \cdot\qquad\\ A{-}\ \mathrm{A}\end{pmatrix} + \Delta\mathcal{G}^\circ_{\text{AU-penalty}} + \Delta\mathcal{G}^\circ_{\text{init}}.\end{aligned}\tag{2.2}$$

The first line of Eq. (2.1.3) is the sum of the block energies over all nearest neighbor blocks, whereas the second line contains three energetic correction terms. Note that "complete" blocks, that can be transformed into each other by rotation, are identical and thus have the same energy. The first term $\Delta\mathcal{G}^\circ_{\text{non}}$ in the second line is given to the last block ending with a non-complementary base pair. Surprisingly, non-complementary base pairs at the terminus weakly stabilize the complex. Hence, the correction term is negative. The second term is an energetic penalty $\Delta\mathcal{G}^\circ_{\text{AU-penalty}}$ taking into account the

AU pair at the other end, which has a destabilizing effect on the duplex. The last term $\Delta\mathcal{G}^\circ_{\text{init}}$ also is a penalty term and occurs since the complex contains two strands. In a thermodynamic picture, the initiation penalty accounts for the loss of system entropy due to the fusion of separate entities into one new complex. In a kinetic picture, the penalty can be explained by a probability smaller than one that two strands coming into contact actually form a new complex (see Chapter 4). The ° is a standard notation to emphasize that the energy values for standard conditions are used. Plugging in the numerical values from the Turner database into Eq. (2.2), we obtain

$$\begin{aligned}\Delta\mathcal{G}^\circ_{\text{tot}} =& -2.08\,\frac{\text{kcal}}{\text{mol}} \times 2 - 3.42\,\frac{\text{kcal}}{\text{mol}} \times 2 - 2.36\,\frac{\text{kcal}}{\text{mol}} \\ & -1.10\,\frac{\text{kcal}}{\text{mol}} + 0.45\,\frac{\text{kcal}}{\text{mol}} + 4.90\,\frac{\text{kcal}}{\text{mol}} \\ =& -9.11\,\frac{\text{kcal}}{\text{mol}}.\end{aligned} \tag{2.3}$$

For standard conditions the value of $-9.11\frac{\text{kcal}}{\text{mol}}$ for the Gibbs free energy $\Delta\mathcal{G}^\circ_{\text{tot}}$ corresponds to $-14.8\,k_{\text{B}}T$.

The next example for a secondary structure shown below is a double-stranded helical segment flanked by two so-called *dangling ends*. Dangling ends are single-stranded overhangs [76]. In contrast, endpoints of double-stranded helical segments without any overhangs are called a *blunt ends* [30].

$$\begin{array}{ccccccccccccccc} 5' & A & - & G & - & C & - & A & - & C & - & G & - & C & 3' \\ & & & \cdot\cdot & & \cdot\cdot & & \cdot & & \cdot\cdot & & \cdot\cdot & & & \\ 3' & & & C & - & G & - & U & - & G & - & C & & & 5' \end{array}$$

The Gibbs free energy $\Delta\mathcal{G}^\circ_{\text{tot}}$ for this configuration is given by

$$\begin{aligned}\Delta\mathcal{G}^\circ_{\text{tot}} =& \Delta\mathcal{G}^\circ \begin{pmatrix} G-C \\ \cdot\cdot \\ C-G \end{pmatrix} + \Delta\mathcal{G}^\circ \begin{pmatrix} C-A \\ \cdot\cdot\;\;\cdot\cdot \\ G-U \end{pmatrix} + \Delta\mathcal{G}^\circ \begin{pmatrix} A-C \\ \cdot\cdot\;\;\cdot\cdot \\ U-G \end{pmatrix} + \Delta\mathcal{G}^\circ \begin{pmatrix} C-G \\ \cdot\cdot\;\;\cdot\cdot \\ G-C \end{pmatrix} \\ & + \Delta\mathcal{G}^\circ_{5'} \begin{pmatrix} A-C \\ \cdot\cdot \\ G \end{pmatrix} + \Delta\mathcal{G}^\circ_{3'} \begin{pmatrix} G-C \\ \cdot\cdot \\ C \end{pmatrix} + \Delta\mathcal{G}^\circ_{\text{init}}.\end{aligned} \tag{2.4}$$

Again, the first line of Eq. (2.2) accounts for the nearest neighbor blocks, whereas the second line accounts for the energetic correction terms. The dangling ends at the 5′ and

$3'$ positions have stabilizing effects. Hence, $\mathcal{G}^\circ_{5'}$ and $\mathcal{G}^\circ_{5'}$ take negative values. With the energy values from the Turner database, Eq. (2.4) becomes

$$\begin{aligned}\Delta\mathcal{G}^\circ_{\text{tot}} =& -3.42\,\frac{\text{kcal}}{\text{mol}} - 2.11\,\frac{\text{kcal}}{\text{mol}} - 2.24\,\frac{\text{kcal}}{\text{mol}} - 2.36\,\frac{\text{kcal}}{\text{mol}}\\ & -0.4\,\frac{\text{kcal}}{\text{mol}} - 0.2\,\frac{\text{kcal}}{\text{mol}} + 4.90\,\frac{\text{kcal}}{\text{mol}}\\ =& -6.6\,\frac{\text{kcal}}{\text{mol}}.\end{aligned} \tag{2.5}$$

The value of $-6.6\frac{\text{kcal}}{\text{mol}}$ for the Gibbs free energy $\Delta\mathcal{G}^\circ_{\text{tot}}$ is equivalent to $-10.7\,k_{\text{B}}T$ under standard conditions.

Next, we a look at a secondary structure containing a symmetric internal loop as sketched below.

$$\begin{array}{lccc} 5' & C-U & \diagup^{\,C-C\,}\diagdown & G-G \\ & \cdot\cdot\ \ \cdot & & \cdot\cdot\ \ \cdot\cdot \\ 3' & G-A & \diagdown_{\,C-A\,}\diagup & C-C \end{array}$$

In this configuration, the loop is formed by non-complementary nucleotides. Complementary nucleotides can also form loops. However, this configuration is usually less favorable than a closed helical structure. For this secondary structure the Gibbs free energy $\Delta\mathcal{G}^\circ_{\text{tot}}$ reads

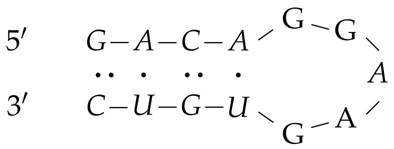
$$\Delta\mathcal{G}^\circ_{\text{tot}} = \Delta\mathcal{G}^\circ \begin{pmatrix} C-U \\ \cdot\cdot\quad\cdot\cdot \\ G-A \end{pmatrix} + \Delta\mathcal{G}^\circ \begin{pmatrix} G-C \\ \cdot\cdot\quad\cdot\cdot \\ C-G \end{pmatrix} + \Delta\mathcal{G}^\circ_{2\times 2\text{int.loop}} + \Delta\mathcal{G}^\circ_{\text{AU-closure}} + \Delta\mathcal{G}^\circ_{\text{init}}. \tag{2.6}$$

The first and the second term in Eq. (2.6) are the block energies for the "complete" nearest neighbor blocks. The third term, i.e., $\Delta\mathcal{G}^\circ_{2\times 2\text{int.loop}}$, penalizes the 2 × 2 energetically. Hence it is positive. Moreover, the fourth term, i.e., $\Delta\mathcal{G}^\circ_{\text{AU-closure}}$, is an energetic penalty associated with the closing base pair formed by an A and a U. Plugging in the numerical values from the Turner database into Eq. (2.6), we obtain

$$\begin{aligned}\Delta\mathcal{G}^\circ_{\text{tot}} =& -2.08\,\frac{\text{kcal}}{\text{mol}} - 3.42\,\frac{\text{kcal}}{\text{mol}} + 2.24\,\frac{\text{kcal}}{\text{mol}} + 0.7\,\frac{\text{kcal}}{\text{mol}} + 4.90\,\frac{\text{kcal}}{\text{mol}}\\ =& 1.44\,\frac{\text{kcal}}{\text{mol}}.\end{aligned} \tag{2.7}$$

For standard conditions the value of $1.44\frac{\text{kcal}}{\text{mol}}$ for the Gibbs free energy $\Delta\mathcal{G}^\circ_{\text{tot}}$ corresponds to 2.6 $k_\mathrm{B}T$. The Gibbs free energy $\Delta\mathcal{G}^\circ_\mathrm{tot}$ of the given secondary structure is positive and therefore energetically unfavorable. In this example, we consider a short symmetric internal loop. The Turner database provides equations that allow calculating the energy contributions for larger and asymmetric loops. Typically the energy penalty for a loop grows with its size and asymmetry.

In the last example, we consider an RNA strand that forms a so-called hairpin loop, as shown below.



The Gibbs free energy $\Delta\mathcal{G}^\circ_\mathrm{tot}$ for this configuration is given by

$$\begin{aligned}\Delta\mathcal{G}^\circ_\mathrm{tot} =& 2 \times \Delta\mathcal{G}^\circ \begin{pmatrix} G-A \\ \cdot\cdot \quad \cdot\cdot \\ C-U \end{pmatrix} + \Delta\mathcal{G}^\circ \begin{pmatrix} A-C \\ \cdot\cdot \quad \cdot\cdot \\ U-G \end{pmatrix} \\ &+ \Delta\mathcal{G}^\circ_\text{AU-penalty} + \Delta\mathcal{G}^\circ_{\mathrm{AU}\to\mathrm{G\,G}} + \Delta\mathcal{G}^\circ_\text{G G first} + \Delta\mathcal{G}^\circ_\text{hairpin 5}.\end{aligned} \tag{2.8}$$

Again, the first line of Eq. (2.8) accounts for the nearest neighbor blocks. The second line accounts for the hairpin loop and for the other energetic correction terms. The first correction term $\Delta\mathcal{G}^\circ_\text{AU-penalty}$ is an energetic penalty occurring because the double-stranded helical structure ends with a base pair formed by an A and a U. The first correction term is an energetic penalty occurring because the double-stranded helical structure ends with a base pair formed by an A and a U. Empirically, it has been found out that specific transition into the loop shown here has a stabilizing effect. The second term and third term, i.e., $\Delta\mathcal{G}^\circ_{\mathrm{AU}\to\mathrm{G\,G}}$ and $\Delta\mathcal{G}^\circ_\text{G G first}$ account for this stabilizing effect. The last term in Eq. (2.8) is an energetic penalty associated with the formation of a hairpin loop of length five. With the energies values from the Turner database, Eq. (2.8) becomes

$$\begin{aligned}\Delta\mathcal{G}^\circ_\mathrm{tot} =& -2.11\,\frac{\mathrm{kcal}}{\mathrm{mol}} \times 2 - 2.24\,\frac{\mathrm{kcal}}{\mathrm{mol}} \\ &+ 0.45\,\frac{\mathrm{kcal}}{\mathrm{mol}} - 0.8\,\frac{\mathrm{kcal}}{\mathrm{mol}} - 0.8\,\frac{\mathrm{kcal}}{\mathrm{mol}} + 5.7\,\frac{\mathrm{kcal}}{\mathrm{mol}} \\ =& -1.9\,\frac{\mathrm{kcal}}{\mathrm{mol}}.\end{aligned} \tag{2.9}$$

The value of $-1.9\frac{\mathrm{kcal}}{\mathrm{mol}}$ for the Gibbs free energy $\Delta\mathcal{G}^\circ_\mathrm{tot}$ is equivalent to $-3.5\,k_\mathrm{B}T$ under standard conditions. According to the turner database, the energetic penalty for the formation of a hairpin loop increases monotonously with its size if it contains more than ten nucleotides. For hairpin loops containing less than ten nucleotides, the energetic penalty is a nonmonotonic function of the loop size.

In the above examples, we assumed a temperature of 37 °C since this is the temperature for which the energy parameters are determined experimentally. For temperatures other the 37 °C the energy values can be extrapolated as shown in the following. The Gibbs free energy $\Delta\mathcal{G}^\circ$ can be split into an enthalpic and an entropic contribution, i.e.,

$$\Delta\mathcal{G}^\circ = \Delta\mathcal{H}^\circ + T\Delta\mathcal{S}^\circ. \tag{2.10}$$

Here, the temperature T has the unit K. The Tuner and SantaLucia databases do not only provide measurements for $\Delta\mathcal{G}^\circ$ but also for $\Delta\mathcal{H}^\circ$. Assuming that $\Delta\mathcal{H}^\circ$ and $\Delta\mathcal{S}^\circ$ are temperature independent, we can use Eq. (2.10) to derive an extrapolation formula for $\Delta\mathcal{G}^\circ$ at temperatures other then 37 °C. For the Gibbs free energy at 37 °C we now explicitly write $\Delta\mathcal{G}^\circ_{37}$. Otherwise we write $\Delta\mathcal{G}^\circ_T$. From Eq. (2.10) we obtain

$$\Delta\mathcal{S}^\circ = \frac{\Delta\mathcal{H}^\circ - \Delta\mathcal{G}^\circ_{37}}{310.15\,\mathrm{K}}. \tag{2.11}$$

Plugging Eq. (2.11) into Eq. (2.10), we get

$$\Delta\mathcal{G}^\circ_T = \Delta\mathcal{H}^\circ + T\frac{\Delta\mathcal{H}^\circ - \Delta\mathcal{G}^\circ_{37}}{310.15\,\mathrm{K}}. \tag{2.12}$$

Generally, the assumption of $\Delta\mathcal{H}^\circ$ and $\Delta\mathcal{S}^\circ$ being independent of temperature is not exactly true. The extrapolation formula Eq. (2.12) is therefore expected to give reasonable results only in a certain interval around 37 °C. According to the Turner database, the interval ranges from 10 °C to 60 °C.

2.1.4. Base pairing probability and minimum free energy conformation

In the examples in Section 2.1.3, we have derived the total Gibbs free energies for several secondary structures. However, a single strand or set of strands can fold into multiple different secondary structures. The number of possible configurations increases exponentially with the number of nucleotides N.

To determine the most likely secondary structure, i.e., the secondary structure with the lowest Gibbs free energy, one has to consider all possible conformations. Computationally, this task is challenging. In 1981 Zuker et al. presented an efficient algorithm based on dynamic programming techniques making it possible to find the optimal configuration in $\mathcal{O}(N^3)$ steps. The algorithm is commonly referred to as the *Zuker algorithm*.

If one is interested in the probability that two given nucleotides form a pair, one has to compute the full equilibrium partition function for the ensemble of all possible configurations. McCaskill et al. developed a computational method to derive the partition function that also scales with $\mathcal{O}(N^3)$ [106]. As well as the Zuker algorithm, the *McCaskill algorithm* is based on dynamic programming, avoiding costly recursions.

2.2. Experimental realizations of enzyme-free templated polymerization and ligation

2.2.1. Enzyme-free template-directed polymerization

From a conceptional point of view, enzyme-free template directed polymerization is the simplest form of purely chemical assembly and replication. In this scenario, template strands are "provided" with a short primer strands. The primer strands are stably bound at the 5′ end of the template and become extended in a step-wise fashion by mononucleotides from the surrounding solution. The extensions to the primer strands correspond to a complementary copies of the template sequence. In practice, performing and analyzing enzyme-free template directed polymerization experiments is challenging [68, 80]. Up to date, no enzyme-free RNA system capable of undergoing a full replication cycle, i.e., producing a complementary copy of the complementary copy of the initial template sequence successfully, has been found.[1]

In "modern" cells, template polymerization is assisted by highly sophisticated enzymes capable of kinetic proofreading (see Section 2.4 of this chapter, and Chapter 3). These enzymes effectively increase the affinity for the incorporation of correctly pairing nucleotide and thus facilitate the replication process.

In the absence of enzymes, base pairing between the incoming and the templating nucleotide is the only mechanism to "attract" monomers to the template strand [68]. The formation of a new phosphodiester bond, joining an "ordinary" incoming nucleotide (nucleoside monophosphates) to the primer strand, would be an endergonic reaction, and hence unlikely to occur. Therefore, some form of activation driving the bond-forming reaction is required. The activation comes in the form of a so-called *leaving group* attached to the phosphate group at the 5′ position. The removal of the leaving group via hydrolysis releases the energy that is necessary to form the new bond as depicted in Fig. 2.4 [68]. Leaving groups enabling ample reactivity that are commonly used are, e.g., imidazolide (Im), 2-methylimidazole (MeIm), and oxyazabenzotriazole

[1] This statement was made multiple times by Prof. Dr. Dr. Clemens Richert from the University of Stuttgart in the context of the evaluation of the *CRC 235 "Emergence of Life"* in January 2022. This statement is also found in the corresponding proposal for the next funding period.

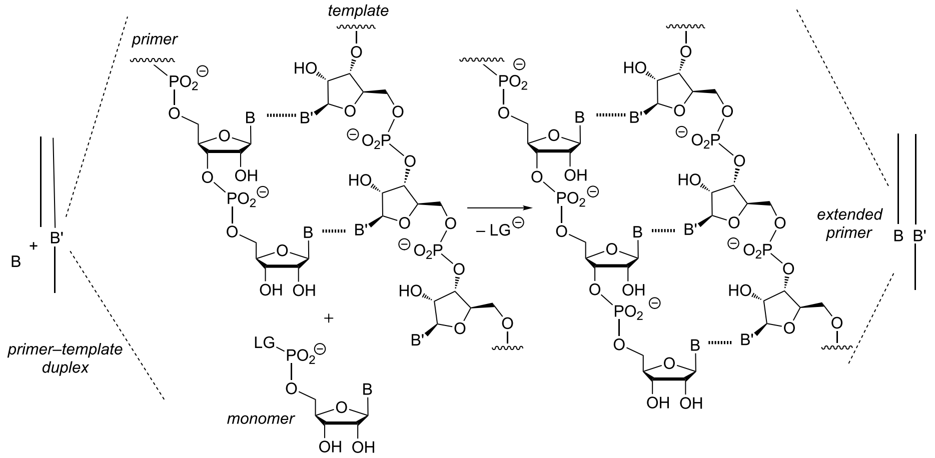


Figure 2.4.: One copying step in a purely chemical templated polymerization process. B is the nucleobase of the incoming activated nucleotide. B′ is the templating nucleotide at the ligation site. LG corresponds to the leaving group. The removal of the leaving group provides the energy required for the formation of a new phosphodiester bond. This figure is adapted from Fig. 1 in Ref. [68] published under CC BY 4.0 License.

(OAt) [115]. In contrast, "naturally" activated nucleotides, i.e., nucleoside triphosphates, can not polymerize spontaneously on template strands.

The extension process is a multiple step process. The dehybridization rate for a mononucleotide, hybridized adjacent to the primer terminus, is assumed to be much larger than the rate for the formation of the new phosphodiester bond [116]. Hence, every extension step is likely to be proceeded by multiple hybridizations and dehybridizations events. As a result, the "occupancy" of the position next to the primer terminus can be described by the ratio of the rates for dehybridization and hybridization, i.e., the corresponding dissociation constants [117, 115] (see Fig. 2.5). Knowing both, the dissociation constant and the overall extension rate, the rate constant for the bare ligation step can be estimated [116, 117, 115].

Early pioneering work on enzyme-free copying of RNA templated was carried out by Orgel and his group in the seventies and eighties. Orgel and his co-workers were able to demonstrate that it is possible to copy a 14-nucleotide template sequence consisting of G and C in principle. However, the overall yield of copying products was poor (less than 2%) [80]. Since then, the experiments and the method of analysis have steadily improved and provided many new insights. Nonetheless, accurately copying mixed sequences containing all four nucleotides with high yield is still a challenge.

Systematic studies probing a large number of different template sequences and primer termini have shown that the rate of extension depends strongly on the sequence

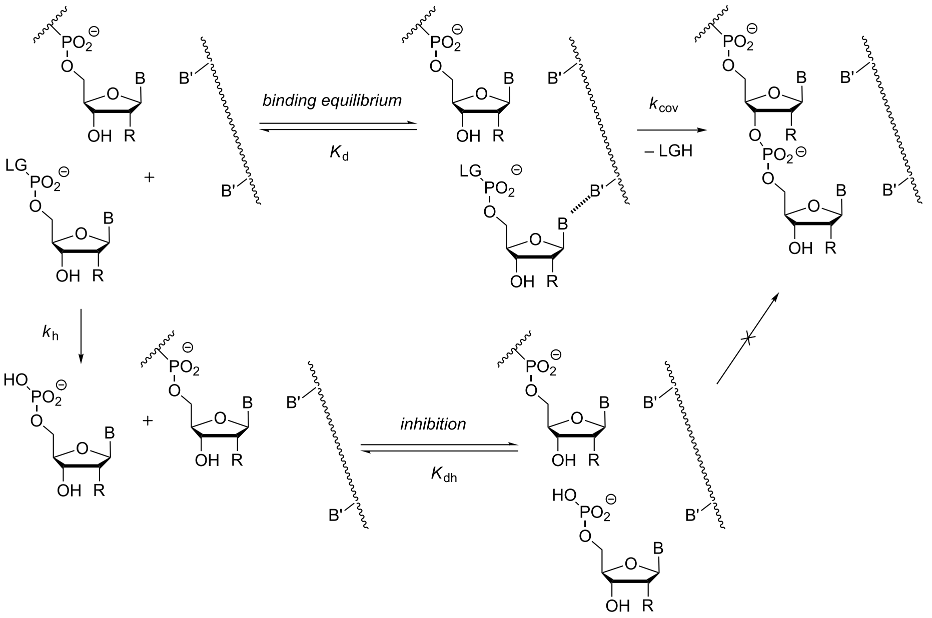


Figure 2.5.: The formation of a new covalent bond occurs at a slow rate k_{conv}. Therefore, every copying step is proceeded by multiple hybridizations and dehybridizations events characterized by a dissociation constant. Monomers in solution eventually lose their leaving group via hydrolysis and therefore can no longer be incorporated into the growing primer strand. However these "spent" nucleotides can still bind onto the template and thereby prevent "reactive" monomers from reaching the extension side. This figure adapted from Fig. 4 in Ref. [68] published under CC BY 4.0 License.

context. For typical conditions, dissociation constants measured via NMR spectroscopy range between $\sim$ 10 mM for C as a templating base and more than 500 mM for A as a templating base. Also, the activation chemistry being used has a significant impact and can lead to context depended on variations of the dissociation constant of a factor of 2-10. The rate constant for the bare ligation step also shows some variations depending on the sequence context and the activation chemistry. However, these variations are smaller than those of the dissociation constants. The time scale for the bare ligation step is typically of the order of hours [115]. Moreover, the Richert group could show that so-called micro helpers, i.e., short oligonucleotides binding one position downstream of the ligation site, could significantly enhance the overall incorporation rate. Micro helpers provide an additional stacking interaction, and thus increase its binding affinity for the incoming nucleotide [117].

Activated mononucleotides eventually lose their leaving group via hydrolysis when they are in solution. These mononucleotides can still bind to the template. However, they cannot be incorporated anymore into the growing primer strand. "Sitting" at the ligation site, these “spent” mononucleotides prevent “reactive” monomers from reaching the extension side. This blocking of the ligation site leads to a slowdown of the copying process over time [117, 115]. In Ref. [70] it is demonstrated that *in situ* activation can be used to circumvent the inhibition problem. As an alternative to *in situ* activation, immobilized primer-template complexes in an open system could be used. In such an open system, “spend” mononucleotides are washed out while “fresh” activated mononucleotides are provided via an influx [69].

Another important experimental observation is that the accuracy of the copying process is strongly sequence-dependent. Systemic assays probing different non-complementary nucleotide configurations at the 3′-terminus of the primer revealed that mismatches in the vicinity of the ligation site significantly slow down the extension reaction. This effect is called post-mismatch *stalling* and was originally discovered in enzymatic copying before it was observed in enzyme-free systems [118]. Typically, an initial mismatch stalls the copying speed by one to two orders of magnitude. A second mismatch then slows down the extension by another factor of six [77, 78]. In addition, mismatches at the primer terminus also strongly impact the subsequent copying fidelity. If no mismatch is present in proximity to the ligation site, the average probability for a misincorporation is about 0.17 [119] for solutions containing all four nucleotides. After a first mismatch, the error probability increases roughly threefold [78]. For DNA systems, the error probability is approximately 0.8 [119]. After a first mismatch, it increases more than sevenfold[78]. Systematic measurements of the copying accuracy following error clusters of size two are missing, but the existing data and estimates based on RNA on folding software suggest that the error probability increases with the size of the preceding error cluster [78].

The high intrinsic error probabilities are problematic when it comes to the enzyme-free copying of longer template sequences [80]. In comparison, “modern” enzyme-based cellular copying achieves error probabilities as low as 10^{-9} [30, 82, 120] (see also Section 2.4 in this chapter). The genetic information in a short “*gene*” would disappear into the noise during the multiple rounds of copying [121]. Given the high error probability, how was an accurate transmission of *“genetic”* information achieved on the Early Earth? As we will discuss in Chapter 3, stalling and error cascades in combination cause slow growing primers to have high error fractions and open the door to a prebiotic error reduction mechanism

2.2.2. Enzyme-free template-directed ligation

Recently the Richert group demonstrated that the extension of a primer strand is also possible with *in situ* activated dimers and trimers [68]. In the experiments, one extension step was performed under “clean” conditions. “Clean” conditions

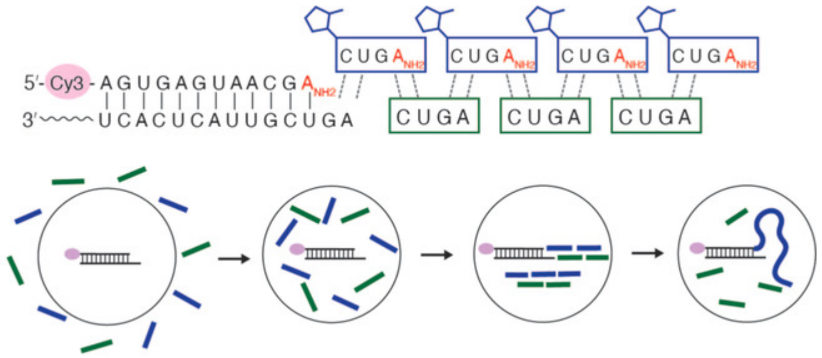

5'- Cy3 - A G U G A G U A A C G A
NH2
3' U C A C U C A U U G C U G A
C U G A
NH2
C U G A

Figure 2.6.: Experimental realization of self-assembly via template-directed ligation on short splint strands in solution (top) and in vesicles permeable to tetramers (bottom). To speed up the reactions, the nucleotides at the 3′ ends of the primer and the substrate strands were replaced by chemically modified nucleotides (3′-amino-2′,3′-dideoxyribo-nucleotide) which are highlighted in red. This figure is adapted from Fig. 5 in Ref. [73] published under CC BY-NC 4.0 License.

here mean that the solution surrounding the primer-template complex contained exactly one dimer motif or one trimer motif, namely, the one that is complementary to the templating section on the template strand. All sixteen dimer motifs and three exemplary trimer motifs have been probed. Additionally, an exploratory competition assay was conducted using solutions with one dimer motif (CG) and one monomer (C). Although the monomer concentration was tenfold larger than the dimer concentration in the exploratory competition assay, the majority of the product had incorporated the dimer. Their larger binding affinity explains the better "performance" of the dimers. Moreover, cyclization of dimers was identified as an important site reaction that effectively reduces the concentration of "reactive" dimers and trimers in the system. The cyclization reaction is sequence-dependent and affects some dimer motifs more than others. The strength of the cyclization effect has to be taken into account in future experiments seeking the template sequence which is most readily copied. So far, data for experiments employing mixtures of dimer motifs, from which error probabilities for the copying of dimers could be extracted, have not been published. Such data sets would be very useful for future modeling attempts.

Moreover, the Szostak group could show in Ref. [73] that primer extension also works with tetramers and decamers. The study contains two examples for one-step experiments where the primer strand gets extended by either a tetramer or a decamer with one specific sequence. Moreover, two examples for multiple-step extension processes are given. In the first multiple-step extension process, a templating sequence repeating the same tetramer motif eight times is copied by complementary tetramers from the solution. In the second process, a templating sequence, where four distinct

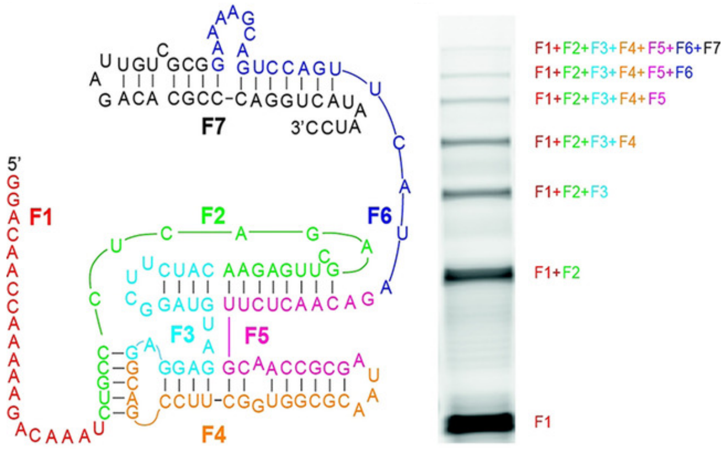


Figure 2.7.: Assembly of an RNA polymerase ribozyme, composed of 150 nucleotides from seven fragments. activated by prebiotically plausible imidazole-based chemistry on non-activated splint strands. This figure is adapted from Fig. 4 in Ref. [63] published under CC BY-NC 4.0 License.

tetramer motifs follow each other, is copied in a mixed solution containing the four complementary motifs. To enable high reaction rates, the nucleotides at the 3′ ends of the primer and substrate strands were replaced by chemically modified nucleotides (3′-amino-2′,3′-dideoxyribo-nucleotidez). These modifications are prebiotically not plausible. Therefore, the authors emphasize that their system has to be viewed as a model system for the replication dynamics of RNA strands on larger time scales under prebiotically more plausible conditions.

Ref. [73] also contains an interesting experimental realization of a multi-step assembly process. In this assembly process, the ligation step is carried out on a so-called *splint* strands. Splint strands are relatively short non-activated templating strands that only partially overlap with the substrate strands that are to be ligated. Due to the short overlap, dissociation occurs quickly. The assembly process was realized in solution and in vesicles which are permeable to tetramers as sketched in Fig. 2.6. Assuming that the assembled sequence corresponds to a ribozyme sequence, it could not fold into its catalytic secondary structure if the ligation steps were carried out on a long reverse template sequence. The assembled sequence would remain bound to the template sequence if its dissociation were not induced by a strong increase in temperature or other significant changes of the environmental conditions [122, 123, 124, 125, 126, 28]. Hence, ligation on splints could be one way to overcome the inherent inhibition problem. For chemically modified nucleotides at the primer and substrate 3′ termini (see above), ligation products containing up to 200 nucleotides were found both in

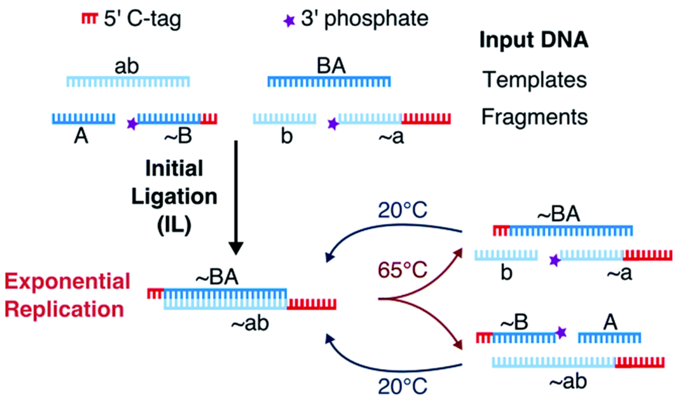


Figure 2.8.: Sketch of the purely chemical cross-replication scheme for 12mer DNA fragments activated with EDC on 24mer templating strands designed by Edeleva et al. Temperature oscillations drive the separation of the product and templating strands. For analysis purposes the fragments *a* and *B* contained polycytidine tags at their 5′-ends (shown in red in the figure and symbolized by a ∼ symbol in the text of the figure). This figure is adapted from Fig. 4 in Ref. [62] published under CC BY-NC 3.0 License.

solution and in vesicles after one day under optimized conditions. Again, the authors stress that the assays are to be understood as a proof-of-principle experiments showing that it is possible to assemble long oligomers by means of enzyme-free templated ligation in the absence of long template strands.

Similar approaches have been used by the Holliger group and Szostak group to demonstrate the non-enzymatic assembly of functional ribozymes, from shorter RNA fragments in so-called *one-pot reactions*, in two recent independent publications [63, 72].

The Holliger group studied the assembly of an RNA polymerase ribozyme, composed of 150 nucleotides from seven fragments with lengths ranging between 20 and 30 nucleotides on split strands of the same lengths, employing a prebiotically plausible imidazole-based activation chemistry [127]. However, the yield of fully assembled ribozymes was rather poor (∼ 0.5%) and can be explained by hydrolysis of the activation group in an aqueous solution and inhibition due to folding of the involved RNA strands.

In contrast, the Szostak group investigated the assembly of a 52 nucleotides long RNA ligase ribozyme from five different oligomers, containing ten to twelve nucleotides on slightly shorter splint strands. As in Ref. [73] the 3′ ends of the initial strands were replaced by chemically modified nucleotides (3′-amino-2′,3′-dideoxyribo-nucleotide). With that, the overall assembly process showed a high efficiency under optimized conditions. Moreover, in order to facilitate the dissociation of the splint strands after a successful ligation such that the ribozyme can fold into its secondary structure easily, three different strategies were explored. These strategies consist in introducing G-U wobble pairs, replacing G with the non-canonical nucleobase I (inosine), or using DNA-splints. All three strategies have in common that they weaken the binding between substrate and splint strands.

We close this section by briefly describing a model system for purely chemical continuous cross-replication with activated DNA oligonucleotides set up by Edeleva et al. In this model system, DNA strands are used. Moreover, the chemical activation of oligonucleotides is carried out in situ by means of EDC. The model is sketched in Fig. 2.8. The initial solution contains template strands with two distinct sequences denoted by ab and BA, as well as shorter DNA strands containing sequences of length 12 that are complementary to either the $3'$ or $5'$ subsection of one of the templating sequences denoted by A, B, a and b. In the reaction scheme, sequence ab serves as a template for the ligation of the subsequences A and B to form sequence BA. In turn sequence BA serves as a template for the ligation of sequence ab from the subsequences a and b. Temperature cycles are employed to drive the separation of template and product strands. The thermally driven cross-replication dynamics gave rise to significant amplification of the templating sequences. Moreover, the cross-replication dynamics could sustain repeated dilution of template strands, implemented via replacing a certain part of the solution with a mixture containing only short ''feeding" fragments A, B, a, and b. Edeleva et al. also carefully analyzed side reactions that led to inhibition of the cross-replication dynamics. The insights gained in the analysis of the inhibiting side reactions were used to set up a mathematical model for the replication dynamics. This mathematical model showed that the replication dynamics are highly sensitive to variations of the rate constants for the chemical reactions as well as the the strength of feeding and dilution. The cross-replication system proposed by Edeleva et al. is a purely chemical one-pot realization of an autocatalytic set, where DNA oligonucleotides mutually catalyze their formations [128]. The idea of the autocatalytic set as a chemical intermediate on the way to Darwinian evolution and "modern" life was popularized by Kauffmann life in the eighties [129].

The examples of enzyme-free template-directed assembly via ligation, briefly described in this section, are all relatively recent. They are based on a large body of previous work going back until the nineteen eighties (see Ref. [130, 74, 75, 131] for more "historical" examples). To the best knowledge of the author of this thesis, all existing experimental studies have in common that they employ initial pools, containing one or only a few different oligonucleotides, with precisely designed sequences to limit the size of the product space. Experimentally probing the enzyme-free self-assembly in a random pool composed of mononucleotides and short fragments remains a challenge since the assays require long time, while the reaction yields are low and undesired side products easily obscuring the results. Moreover, it is an unsolved problem how to track the evolution of the whole sequence pool simultaneously [79, 80, 81, 62, 63].

2.2.3. Ribozyme assisted polymerization and ligation

In order to broaden the overview of enzyme-free assembly of RNA strands, we will give some examples of ribozyme-assisted templated polymerization and ligation in this section. These examples do not cover the whole body of the existing world. They are only meant to give an idea of what has already been achieved in the field. A more extended overview can be found in Ref.[9].

The first example, i.e., the study by Salditt et al. [28], was already briefly mentioned in the introduction of this thesis. In this study, a realization of a PCR [45] set-up, which is prebiotically plausible, is analyzed. In short, modern PCR systems typically work as follows: A solution containing few templates strands, specific primer strands, and single nucleotides in high abundance as well as polymerize enzymes is given into a thermocycler. The thermocycler alters the temperature of the sample in a periodic fashion. During the cold phase, primer strands bind to the template strands and become extended by single nucleotides by the polymerizing enzyme. In the subsequent hot phase, the template and product strands dissociate. During the next cold phase, the template strand and the product strand can serve as template strands. The process is repeated multiple times, leading to exponential amplification of the template strand. In their prebiotic analog, Salditt et al. study the replication of an RNA template sequence, that is 35 nucleotides long, assisted by an RNA polymerase composed of 210 nucleotides. The RNA polymerase has already been used in earlier work by the Joyce group [46]. The RNA polymerase does not require "artificially" activated nucleotides as needed for purely chemical polymerization (see Section 2.2.1). Instead, it can work with "naturally" activated nucleotides, i.e., nucleoside triphosphates. In the prebiotic scenario considered by Salditt et al, dissociation of product and template strands is not achieved by periodic changes of the global temperature as in a "modern" PCR application. Instead, a localized heat source in reaction volume creates a temperature gradient which is constant in time. As a result, a laminar convection flow occurs that transports the RNA molecules from the hot to the cold regions in a cyclic fashion. The effective temperature profile experienced by the RNA strands is similar to the temperature profile created by a thermocycler. Moreover, the temperature gradient induces a thermophoretic movement of the RNA molecules away from the heat source along the temperature gradient. The thermophoretic drift is size-dependent and increases with length. Therefore, the RNA polymerases accumulate in the cold region of the reaction volume where they are protected from degradation due to hydrolysis occurring at high temperatures. In summary, primer strands bound to template strands are extended by the RNA polymerases in the cold region and then separated from the template strand in the hot region. The set-up is prebiotically plausible since natural realizations of reaction volumes subjected to steep temperature gradients are found in porous rocks in the vicinity of hydrothermal vents on the ground of the ocean. As far as we know, such hydrothermal systems were abundant on the early Earth [132]. The results of Salditt et al. "demonstrate a size-selective pathway for autonomous RNA-based replication in natural nonequilibrium conditions."

Another recent work demonstrating the capacity of ribozymes was carried out by Attwater et al. [40]. In this study, a 153 nucleotides long ribozyme called the *t5 ribozyme* (see Fig. 2.2) is demonstrated to be able to synthesize copies of various RNA strands using trinucleotide triphosphates called triplet building blocks. The t5 ribozyme is the result of several rounds of *in vitro* evolution. While the *t5 ribozyme* cannot create a copy of another t5 ribozyme or of other strands of a comparable length, it can replicate fragments of its own sequence. In the study, several important observations are made. First, a longer primer strand to initiate the replication process is not necessarily required by the t5 ribozyme. Instead, the t5 ribozyme can create its "own" primer at an arbitrary position on the template strand via the ligation of adjacent triplet building blocks. Extension of the newly formed primer composed of six nucleotides can then occur in both directions, i.e., in the canonical 5′–3′ direction as well as in reverse 3′–5′ direction. Hence, the triplet building block system could be a way to circumvent the so-called *primer problem*. Moreover, the fidelity of the ligation process, assisted by the t5 ribozyme using triplet building blocks, showed a higher fidelity than those employing the most accurate known mononucleotide RNA polymerase ribozymes. Fidelities larger than 95% could be achieved. This result is surprising since the correct triplet building block has to be chosen out of all 64 possible triplet building blocks. Interestingly, the fidelity depends not only on the templating triplet but also on the total concentration of triplet building blocks in the solution. This phenomenon is explained by the formation of duplexes of (partially) complementary triplet building blocks. Triplet building blocks, that are rich in C and G, are more likely to form duplexes in solution than triplet building blocks containing mostly A and U. Hence, the effective concentration of free C- and G-rich triplet building blocks decreases. As a result, the strongly binding C- and G- rich triplet building blocks are less likely to supplant A- and U-rich triplet building blocks in the competition for the binding site. Since duplex formation is a second-order reaction, it is strongly concentration dependent. The study also demonstrated the capacity of triplet building blocks to "invade" and unfold secondary structures. Secondary structures, such as hairpins, can prevent the sequence from being copied. Even highly evolved polymerase ribozymes fail at copying inhibited sequences [46]. However, Attwater et al. show that using triplet building blocks, t5 ribozyme can actually copy sequences that are predicted to form highly stable self-blocking secondary structures. This finding will be relevant in future studies to circumvent the *inhibition problem* [80].

Another recent key study was carried out by Mutschler et al. [42]. In this study, it is

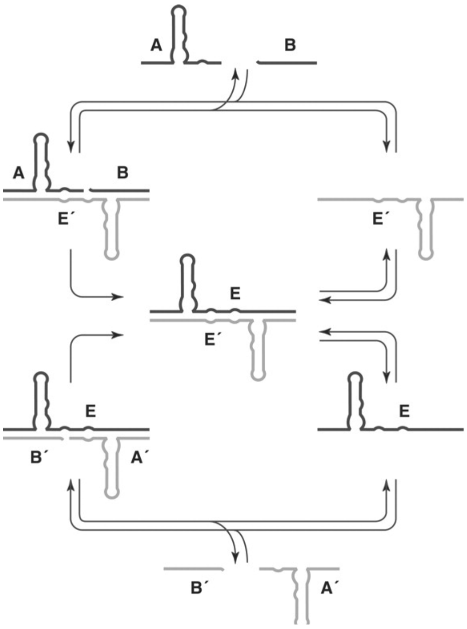


Figure 2.9.: Sketch of the cross-replicating dynamics for the RNA enzymes denoted by E and E′. The RNA ligase E′ (gray) carries out the ligation of the two fragments A and B (black), resulting in the RNA ligase E. The RNA ligase E, in turn, catalyzes the formation of the RNA ligase E′ via the ligation of the fragments A′ and B′. The dissociation of the product complex formed by E and E′ enables new rounds of replication. This figure is taken from T. A. Lincoln and G. F. Joyce. "Self-Sustained Replication of an RNA Enzyme." In: Science 323.5918 (Feb. 2009) [43]. Reprinted with permission from AAAS.

demonstrated that an RNA polymerase, that is 225 nucleotides long, can be assembled from fragments not longer than 30 nucleotides by a short hairpin ribozyme that is itself fragmented into three strands not longer than 21 nucleotides. The assembly process was driven by strong variations of the thermal and ionic conditions, and effective concentrations as result of so-called freeze-thaw cycles with temperatures ranging from −30 °C to 37 °C. The freezing process enables invasion of, and strand replaced in, secondary structures which would suffer from inhibition at constant temperatures. In particular, the site of the (fragmented) hairpin ribozyme that catalysis the ligation reaction of fragments forming the RNA polymerase needs to be "freed" to not stall the assembly process. The exact mechanistic basis of the driving via freeze-thaw cycles is not yet known. However, the study provides impressive evidence that it does work. By the use of iterative freeze-thaw cycles, the yield of full-length products could be increased almost ten-fold in comparison to a control experiment performed at a constant temperature. Moreover, it was observed that RNA RNA polymerase activity of the whole pool, comprising partially and fully assembled RNA polymerase molecules, significantly exceeded the RNA polymerase activity that one could expect by only taking into account complete ligation products containing all fragments. This observation suggests that partial ligation products can also contribute to the polymerase activity in a collaborative manner.

We close this section by briefly describing a self-sustained replication system of two RNA enzymes designed by Lincoln and Joyce [43]. In the study, the original sequences of the two RNA ligases, denoted by E and E' obtained by previous *in vitro* evolution experiments, are split into fragments A and B, and A' and B', respectively. In solution, the fragments A and B, and A' and B', respectively, can be assembled by complete RNA ligases E and E' initially present at small concentrations. Eventually, the product strand dissociates from the catalyzing RNA ligase. After their separation, both the product strand and the original RNA ligase, can catalyze further ligation reactions. The cross-replicating process is sketched in Fig. 2.9. Interestingly, the replication dynamics precedes well under isothermal conditions at 42 °C. Apparently, the binding between ligation products and catalyzing enzymes is weak enough that cyclic variations of the reaction conditions are not required to induce the separation of the product strand from the catalyzing RNA ligase. The concentration of the RNA ligases E and E' first grow exponentially, and saturate as the pool is depleted in the fragments A and B, and A' and B'. The authors show that the replication dynamics is well described by a logistic growth equation. Moreover, replication is sustained in serial dilution experiments, where 4% of the solution from one reaction mixture is transferred to the next. In this scenario, only the first reaction mixture contains RNA ligases E and E'. Under serial dilution, overall amplifications of larger than 10^8-fold could be achieved. The system designed by Lincoln and Joyce is an impressive experimental RNA-based realization of an autocatalytic set.

2.3. Proof of principle experiments for sequence selection in self-assembly via templated ligation

As stated in Section 2.2.2, all existing experimental studies on enzyme-free assembly via templated ligation have in common that they start with initial pools that contain one or only a few different oligonucleotides with precisely designed sequences, aiming to limit the number of possible product sequences. More realistic scenarios must not limit the sequence space of product strands *a priori* to allow for the spontaneous formation of cooperative network dynamics, that break the symmetry in sequence.

Two recent model experiments, carried out in the Braun group, demonstrate that self-assembly via templated ligation from non-constrained pools of building blocks,

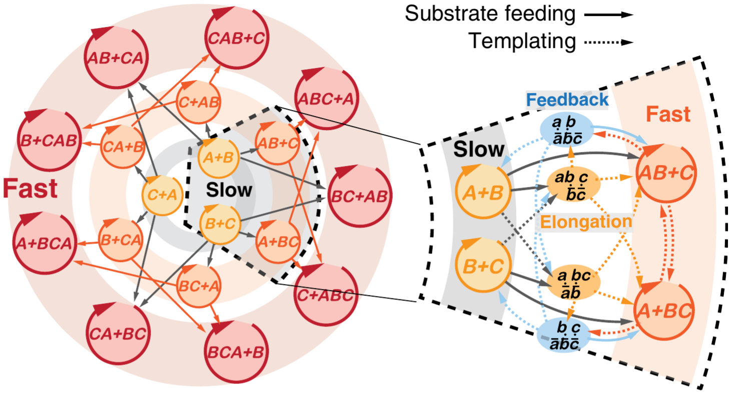


Figure 2.10.: Sketch of a cooperative ligation network emerging from three template strands AB, CA, and BC. A dashed arrow, pointing from a first sequence to a second sequence, signifies that the first sequence can serve as a template for the ligation of the second sequence. A solid arrow signifies that the first sequence can be used as a substrate in a ligation reaction that is templated by the second sequence. The highlighted section shows how the network extends towards longer sequences by elongation via templated ligation. More possibilities exist for longer strands to participate in a reaction, either as a template or substrate strand. Larger overlaps lead to higher binding stability. As a result, the effective elongation speed increases. This figure is adapted from Fig. 3 in Ref. [59] published under CC BY 4.0 License.

indeed, gives rise to cooperative dynamics favoring the formation of certain sequence motifs while suppressing the creation of others. Both experiments share that they use longer DNA strands as basic building blocks. Moreover, to speed up the assembly process, a ligase enzyme catalyzes the ligation reaction, and temperature cycles drive the separation of template and product strands. The two studies are very extensive and include detailed theoretical modeling. Here we can only briefly present some of the key insights. For more details, the reader is referred to the original papers [58, 59].

In the first study [59], Toyabe and Braun designed three pairs of complementary sequences of length 20 with equal binding characteristics. The sequence pairs are denoted by $A = \{a, \bar{a}\}$, $B = \{b, \bar{b}\}$, and $C = \{c, \bar{c}\}$. As in the original paper, we will only use the capital letters A, B, and C to illustrate the dynamics. The emerging cooperative dynamics is best explained for a simple scenario where we start with a solution containing all building blocks A, B, and C, as well as the template strands AB, CA, and BC in small concentrations. Initially, these template strands will experience exponential growth via templated ligation. We now introduce the notation of the *sequence motif*. A sequence motif simply corresponds to the succession of two letters from the "alphabet" $\{A, B, C\}$. Sequence motifs are written in angle brackets, e.g., $\langle AB \rangle$. With that, every initial template, composed of two building blocks carries exactly one sequence motif that is identical to its sequence. Since the template strands replicate exponentially in the beginning, the corresponding sequence motifs also replicate exponentially in the beginning. However, the two sequence motifs $\langle AB \rangle$ and $\langle BC \rangle$ transition to a cooperative mode of growth quickly (see highlighted region in Fig. 2.10). The strands AB and BC can form a staggered complex where only the B-parts overlap. The subsequent hybridization and ligation of either an A or a C building block lead to a new strand ABC carrying the two sequence motifs $\langle AB \rangle$ and $\langle BC \rangle$. The new strand ABC now acts as a template for the ligation of the basic building blocks, A, and C, with the larger building blocks AB and CB. Importantly, the dissociation constant for the larger building blocks is more than 40 times smaller than the dissociation constant for the basic building blocks. Hence, the effective processes $AB + C \rightarrow ABC$ and $A + BC \rightarrow ABC$ occur much faster then $C + B \rightarrow BC$. As a result, the cooperating sequence motifs $\langle AB \rangle$ and $\langle BC \rangle$ replicate quicker than the sequence motif $\langle BC \rangle$, which is excluded from the cooperative mode of growth. Toyabe and Braun show that the cooperating sequence motifs can sustain against serial dilution (mimicking molecular degradation), while the excluded motif approaches extinction. Note that the competition for basic building blocks A, B, and C is symmetric. Every sequence motif competes with a second motif for one specific building block. Since the binding properties of the motifs are identical, other cooperative pairs could be formed. For instance, in a system initialized with the template strands BC, CA, and BA, $\langle BC \rangle$ and $\langle CA \rangle$ would be the cooperative motifs. The simple scenarios considered so far, only contained three different template strands initially. The template strands were chosen such that no strands containing more than three basic building blocks could emerge. However, cooperative networks of sequence motifs also emerge from initial pools containing more than three template strands. Toyabe and Braun, for instance, analyzed a system where the motifs $\langle AB \rangle$, $\langle BC \rangle$, $\langle CA \rangle$, $\langle CB \rangle$, $\langle BA \rangle$, and $\langle AC \rangle$ are present. The motifs $\langle AB \rangle$, $\langle BC \rangle$, and $\langle CA \rangle$ cooperate via the formation of the periodic sequence ...$ABCABC$... while the motifs $\langle CB \rangle$, $\langle BA \rangle$, and $\langle AC \rangle$ can form the common periodic sequence ...$CBACAB$... (see Fig. 2.10). It is shown that the cooperative network, that has a higher initial concentration, can sustain against the serial dilution, while the network with a lower initial concentration becomes extinct. Hence, the formation of a cooperative network leads to frequency-dependent replication. Moreover, for the cooperating sequence motifs the replication becomes faster than exponential. Interestingly, the strand-length distributions, determined from the experiments starting with six different template strands, show heavy tails. In addition, Toyabe and Braun use computational modeling to extrapolate their experiments to scenarios where all nine possible template strands are present initially. The initial concentrations are almost similar and only subjected to small fluctuations of 5%. Moreover, Toyabe and Braun use computational modeling to extrapolate their experiments to scenarios where all nine possible template strands are present initially. The initial concentrations are almost similar and only subjected to small fluctuations of 5%. However, these small fluctuations are sufficient to trigger the emergence of one dominant cooperative network formed by three sequence motifs. Over the course of time all motifs, that are not part of the dominant cooperative network, die out.

While in the first study, Toyabe and Braun utilized six different basic building blocks, Kudella et al. extended the number of distinct basic building blocks to 1024 in the second study [58]. The 1024 distinct basic building blocks are given by all possible binary 12mer sequences that can be formed from the complementary deoxyribonucleotide A and T. Longer strands, emerging from the random 12mer pool after multiple temperature cycles, showed highly structured sequences featuring two apparent characteristics. First, longer sequences either showed a high content of A nucleotides or T nucleotides, i.e., they could be classified as either A-type or T-type sequences. For random binary sequences, one would expect that the distribution of the average fraction of A nucleotides in sequences of a given length has a binomial shape with a maximum at 0.5. However, Kudella et al. showed that the average fraction of A nucleotides in sequences composted of two or more basic building blocks follows a bimodal distribution with two distinct maxima at A contents of 0.3 and 0.7. The first feature is explained by a reduced tendency for A-type or T-type sequences to form secondary inhibitory structures such as hairpins. While sequences with a balanced A and T content are likely to form hairpins, A- and T-rich sequences are predominantly unfolded and serve as efficient templates with the opposite imbalance towards A or T. This explanation is supported by theoretical modeling. Second, the alternating motif ATAT is the dominant sequence motif occurring at the ligation junction. This feature probably stems from a small synthesis bias for the 12mer building blocks to end with an AT at the $3'$ terminus. A ligation junction, resulting from the ligation of an AT motif at the $3'$ terminus of the “left” substrate strand and an arbitrary motif at the $5'$ terminus of the “right” substrate strand, can serve as the templating sequence for a new ligation junction in another replication cycle. Since the AT motif is self-complementary, the new templating sequence will select for an AT motif at the $5'$ terminus of the “right” substrate strand. Hence, over the course of time, this bias becomes self-amplifying and manifests itself as a distinct pattern of the form ATAT at the ligation junction. This explanation is supported by a theoretical model as well. The two apparent characteristics, i.e., the elevated A- or T-content and the ATAT pattern at the ligation junction are not the only structural properties of the pool of longer sequences. An analysis of the basic building blocks, strung together in longer strands, revealed a strong sequence correlation between the building blocks in the centers of different strands, and between the building blocks at the $3'$ and $5'$ ends of different strands, respectively. Kudella et al. conclude that: “Despite its minimalism, the studied system contains all elements, necessary for Darwinian evolution: out of equilibrium conditions, transmission of sequence information from template to substrate strains, reliable reproduction of a subset of oligomer products and the possibility to select from the long fast-growing sequences in the process. At the dawn of life, such pre-Darwinian dynamics would have pushed prebiotic systems toward lower entropy states."

2.4. Principals of kinetic proofreading

Replication of DNA in "modern" organisms is accomplished by a highly elaborate enzymatic "replication machinery" that does not only allow for a high reaction speed of approximately 1000 nucleotides per second but also for a remarkably low error fraction of up to 10^{-9} [30, 82, 120]. The high accuracy is achieved by a process knwon as *kinetic proofreading*, which was first described conceptually by John Hopfield [133] and independently by Jacques Ninio [134]. The principles of kinetic proofreading do not only apply to the copying of genetic information but to various biochemical reactions where a high specificity is needed [133, 135, 136, 137, 138].

In the following, we will take a closer look at the process of kinetic proofreading to contrast it with *kinetic error filtering*, a prebiotically plausible precursor mechanism to reduce the fraction of misincorporations in replication products, later on in Chapter 3. We thereby closely follow the original work from Hopfield in Ref. [133] and the more recent presentation of Rao and Peliti in Ref. [139].

2.4.1. Michaelis-Menten scheme for templated polymerization

We start the discussion by looking at a simple Michaelis-Menten-like scheme for the enzymatic template-directed extension of a polymer given below.

$$\& \underset{k_r^-}{\overset{k_r^+}{\rightleftharpoons}} \&r \xrightarrow{F} \& \quad \text{and} \quad \& \underset{k_w^-}{\overset{k_w^+}{\rightleftharpoons}} \&w \xrightarrow{F} \& \tag{2.13}$$

Here, the & symbol stands for a growing polymer with a generic sequence that is attached to a processive copying enzyme moving along the template strand during the replication process. The complex formed by the template strand, the growing strand, and the enzyme can "catch" a *right* nucleotide matching the template strand, or a *wrong* nucleotide that does fit the template strand from the surrounding solution to form the intermediate states &r or &w. Being in the generic initial state &, the rate for the combined process of catching the right nucleotide r and forming the intermediate states &r is k_r^+. The rate for the equivalent process involving the wrong nucleotide is k_w^+. The intermediate states either transition back to the initial state with rates k_r^- or k_w^-, or go over to the final state with rate F. The rate for the finalizing incorporation step is assumed to be independent of whether the intermediate state involves a right or a wrong nucleotide. If the intermediate state reacts to the final state, the enzyme moves forward by one step. This final state then corresponds to the initial state for the next extension. For the next step, it does not matter if the last incorporation was a mismatch or not. Hence, in Scheme. (2.13), the last reaction step appears as a simple reset step. The probabilities $P_{\&}$, $P_{\&r}$ and $P_{\&w}$ of being in the generic initial, the right, or the wrong intermediate state can be described by the following set of master equations [140]

$$\begin{aligned}
\partial_t P_{\&} &= \left(F + k_r^-\right) P_{\&r} + \left(F + k_w^-\right) P_{\&w} - \left(k_r^+ + k_w^+\right) P_{\&}, && (2.14)\\
\partial_t P_{\&r} &= k_r^+ P_{\&} - \left(F + k_r^-\right) P_{\&r}, && (2.15)\\
\partial_t P_{\&w} &= k_w^+ P_{\&} - \left(F + k_w^-\right) P_{\&w}, && (2.16)\\
1 &= P_{\&} + P_{\&r} + P_{\&w}. && (2.17)
\end{aligned}$$

Here, the last equation corresponds to the constraint that all probabilities have to sum up to one. In the stationary state, where $\partial_t P_{\&} = \partial_t P_{\&r} = \partial_t P_{\&w} = 0$, the above set of equation can be solved easily. The resulting stationary-state solution is characterized by the *error fraction* f_e which is defined as the average fraction of misincorporations per extension step. From the stationary-state solutions of Eqs. (2.14)-(2.17) we obtain:

$$f_e = \frac{P_{\&w}}{P_{\&r} + P_{\&w}} = \frac{k_w^+ \left(F + k_r^-\right)}{k_w^+ \left(F + k_r^-\right) + k_r^+ \left(F + k_w^-\right)}. \tag{2.18}$$

To simplify Eq. (2.18) we now assume that the rates for the formation of the right and the wrong intermediate state are identical, i.e., $k_r^+ = k_w^+$. Hence, discrimination between right and wrong nucleotides solely proceeds via the rates for the back transition to the initial state. If the intermediate state contains a mismatch, the back transition to the initial state occurs more rapidly, such that

$$\frac{k_w^-}{k_r^-} = \frac{K_r}{K_w} = e^{\gamma} \text{ with } \gamma > 0. \tag{2.19}$$

Here, K_r and K_w are the dissociation constants of the intermediate state. Moreover, γ corresponds to the difference in the Gibbs free energies associated with the binding of the wrong and the right nucleotide next to the primer terminus expressed in units of $k_{\mathrm{B}}T$. Attributing the binding energy to the rates for unbinding k_r^- and k_w^- is a common kinetic assumption, and has been confirmed experimentally [141, 142]. With that,

Eq. (2.18) simplifies to

$$\Rightarrow f_e = \frac{F/k_r^- + 1}{2F/k_r^- + 1 + e^\gamma}. \tag{2.20}$$

From Eq. (2.20) we see that the lowest error fraction, i.e., the highest accuracy, is achieved in the limit of a vanishingly low rate for the final incorporation step, i.e., for $F \to 0$. In this limit, we obtain

$$f_e \to f_0 = \frac{1}{1 + e^\gamma} \sim e^{-\gamma}. \tag{2.21}$$

For a finite rate for the last reaction step, we have $f_e > f_0$. Moreover, the error fraction in Eq. (2.20) grows monotonically with an increasing rate F for the final incorporation step. Hence, there is a trade-off between polymerization speed and accuracy.

2.4.2. Hopfield scheme for templated polymerization

In the simple Michaelis-Menten-like scheme for the enzymatic template-directed extension discussed in Section 2.4.1, the error fraction is limited to $e^{-\gamma}$. However, typical binding energies for RNA and DNA nucleotides are of the order of a few $k_\mathrm{B}T$ [107, 76]. Hence, the resulting lower bound for the error fraction would be much larger than what is actually observed in nature. Hopfield's idea is now to introduce an intermediate state in the reaction scheme to iterate the same kind of discrimination state, such that the scheme for the right incorporation on the left-hand side of of Scheme (2.13) becomes

$$\begin{tikzcd}[column sep=large]
\& \arrow[r, harpoon, shift left, "k_r^+"] & \&r^* \arrow[l, harpoon, shift left, "k_r^-"] \arrow[r, harpoon, shift left, "h^+"] & \&r \arrow[l, harpoon, shift left, "h^-"] \arrow[r, "F", "\text{incorporation} \equiv \text{reset}"'] \arrow[d, harpoon, shift left, "l_r^-"] & \&. \\
 & & \& \arrow[u, harpoon, shift left, "l_r^+"] &
\end{tikzcd} \tag{2.22}$$

An equivalent reaction scheme for wrong incorporation, i.e., the right part of Scheme (2.13), exists. Scheme (2.22) assumes that the reactions from the first intermediate state $\&r^*$ or $\&w^*$ to the second intermediate states $\&r$ or $\&w$ are non-specific, i.e., do not distinguish between right and wrong nucleotides. Therefore, if we would not consider the "vertical" branch in the reaction scheme connecting the second intermediate state with the initial state, no gain in replication accuracy could be achieved. In the literature, the first and the second "horizontal" reaction pathways leading to intermediate states $\&r^*$ or $\&w^*$ and $\&r$ or $\&w$ are referred to as the *first* and the *second* *intermediate forming pathways*, respectively. In contrast, the “vertical” reaction pathway is called the *proofreading pathway*. To increase the copying fidelity, the vertical pathway and, in particular, the vertical exit channel from the second intermediate state is crucial, as we will see in the following.

The set of master equations describing the probabilities of being in either the initial state or in one of the intermediate states in the full Hopfield scheme are given by

$$\begin{aligned}
\partial_t P_{\&} &= k_r^- P_{\&r^*} + k_w^- P_{\&w^*} + \left(F + l_r^-\right) P_{\&r} + \left(F + l_w^-\right) P'_{\&w} && (2.23)\\
&\quad - \left(k_r^+ + k_w^+ + l_r^+ + l_w^+\right) P_{\&} && (2.24)\\
\partial_t P_{\&r^*} &= k_r^+ P_{\&} + k_r^- P_{\&r} - \left(k_r^- + h^+\right) P'_{\&r^*} && (2.25)\\
\partial_t P_{\&w^*} &= k_w^+ P_{\&} + k_w^- P_{\&w} - \left(k_w^- + h^+\right) P'_{\&w^*} && (2.26)\\
\partial_t P_{\&r} &= h^+ P_{\&r*} + l_r^+ P_{\&} - \left(h^- + l_r^- + F\right) P'_{\&r} && (2.27)\\
\partial_t P_{\&w} &= h^+ P_{\&w*} + l_w^+ P_{\&} - \left(h^- + l_w^- + F\right) P'_{\&w} && (2.28)\\
1 &= P_{\&} + P_{\&r^*} + P_{\&w^*} + P_{\&r} + P_{\&w}. && (2.29)
\end{aligned}$$

Solving the system in the stationary state, we obtain the error fraction

$$f_e = \left(1 + \frac{\left[k_r^- l_r^+ + (k_r^+ + l_r^+)\, h^+\right]\left[(F + l_w^-)\, h^+ + (F + l_w^- + h^-)\, k_w^-\right]}{\left[\left(F + l_r^-\right) h^+ + \left(F + l_r^- + h^-\right) k_r^-\right]\left[k_w^- l_w^+ + \left(k_w^+ + l_w^+\right) h^+\right]}\right)^{-1}. \qquad (2.30)$$

As written in Eq. (2.22), the reaction scheme is subject to the equilibrium constraints

$$l_r^- k_r^+ h^+ = l_r^+ k_r^- h^- \quad \text{and} \quad l_w^- k_w^+ h^+ = l_w^+ k_w^- h^-. \qquad (2.31)$$

Plugging in the constraints in Eq. (2.30), we find an expression for the error fraction, which is identical to the error fraction obtained in the Michaelis-Menten-like scheme in Eq. (2.18). In other words, we do not obtain any increase in the copying fidelity from the branched reaction scheme, including the additional intermediate state in its current form. In order to enhance the fidelity of the replication process, the constraints in Eq. (2.31) have to be eliminated by introducing some form of chemical driving of the forward reaction. In cells, the chemical driving would typically be implemented by coupling the forward reaction from the first to the second intermediate state with rate h^+ to a dephosphorylation reaction of the form $\mathrm{ATP} \rightarrow \mathrm{AMP} + \mathrm{PP_i}$. If the concentration of AMP and PP_i is kept at a low level, the back reaction becomes negligible, i.e., $h^- \approx 0$. If we further assume that the intermediate states $\&r$ and $\&w$ are states with high energy, we can neglect the “shortcut” reactions from the initial state to the second intermediate state and set $l_r^+, l_w^+ \approx 0$. Again, taking the limit of a slow final incorporation step $F \rightarrow 0$,

Eq. (2.30) becomes

$$f_e = \left(1 + \frac{P_{\&r}}{P_{\&w}}\right)^{-1} = \left(1 + \frac{k_w^+ l_r^- \left(k_r^- + h^+\right)}{k_r^+ l_w^- \left(k_w^- + h^+\right)}\right)^{-1}. \tag{2.32}$$

As for the Michaelis-Menten-like scheme, we assume identical rates for the formation of the first intermediate states $\&r^*$ and $\&w^*$, i.e., $k_w^+ = k_r^+$ (and also $l_r^+ = l_w^+$ while $l_r^+, l_w^+ \approx 0$). These assumption imply that the discrimination solely resides in the back reactions and that

$$\frac{k_w^-}{k_r^-} = \frac{l_w^-}{l_r^-} = e^\gamma. \tag{2.33}$$

If we now assume that $h^+ \ll k_r^-$ as well, such that the first reaction quickly reaches an equilibrium, we end up with an error fraction

$$f_e = \left(1 + e^{2\gamma}\right)^{-1} \approx e^{-2\gamma} = \left(f_0\right)^2, \tag{2.34}$$

that is twice as large as for the simple Michaelis-Menten-like scheme. If we relax the assumptions of a fast equilibrating first intermediate state, and also allow for larger values of the rate for the final incorporation step, we recover the trade-off between polymerization speed and copying accuracy that we have already seen in the simple Michaelis-Menten scheme and obtain an error fraction [133]

$$f_e \geq \left(f_0\right)^2. \tag{2.35}$$

One may interpret the Hopfield reaction scheme with $l_r^+, l_w^+ \approx 0$ and $l_r^-, l_w^- \gg F$ described above as a biochemical ratchet performing two directed steps, each verifying with a certain fidelity whether the *right* reaction pathway is followed. The overall enhancement of specificity is obtained by multiplying the specificity achieved in the two individual steps. In real enzymatic copying of genetic information, kinetic proofreading is more complex than in this simplified example. However three distinct stages of discrimination can be identified and interpreted in the light of the abstract scheme sketched above. First of all, a matching nucleotide sitting directly downstream to the primer's 3′-end is typically bound to the template strand more tightly than an incorrect one, and will therefore dissociate slower. This can be interpreted as the first intermediate forming step. The transition from the first to the second intermediate corresponds to the formation of a covalent bond attaching the nucleotide to the terminus of the primer. In this transition, the incoming chemically activated nucleotide itself is hydrolyzed to provide the energy that is necessary to drive the reaction. The second stage of error control is called the *exonucleolytic proofreading* step. In case of a

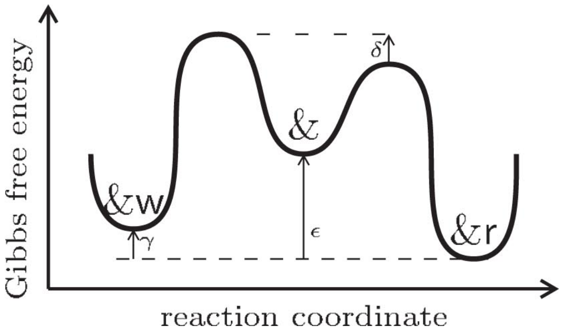


Figure 2.11.: Free energy landscape for the generalized Michaelis-Menten scheme for templated polymerization. ϵ corresponds to the energy, that represents the energetic difference between the initial state and the intermediate state involving the right nucleotide. γ denotes the difference in the energies of the *right* and *wrong* intermediate states, whereas δ denotes the difference in the energetic barriers to overcome in the reactions. If $\delta > \gamma$, the scheme operates in the *kinetic regime* otherwise it operates in the *energetic regime*. The figure is adapted with permission from [139].

mismatched residue at the terminus of the growing strand, the polymerase undergoes a conformational change instead of moving forward [82]. The nucleotide is now exposed to the polymerase's exonucleolytic site, which catalyzes the clipping of the wrongly incorporated nucleotide [30, 143]. Accidental cleavage of the correct nucleotide follows with a very low probability [133]. Essentially the exonucleolytic site can be seen as a "delete key" which removes the last "character" giving the system another try to add the matching one [82]. Once the correct nucleotide is incorporated, the polymerase moves forward enable the next extension. This step can be interpreted as the final step in the abstract scheme discussed above.

The original Hopfield reaction scheme discussed above can be extended to lower the error fraction even more. This extension consists of the introduction of another intermediate state with high energy such that the replication scheme takes the form

$$\begin{tikzcd}
\& \arrow[r, harpoon, shift left, "k_r^+"] & \&r^* \arrow[l, harpoon, shift left, "k_r^-"] \arrow[r, "h^+"] & \&r^{**} \arrow[r, "m^+"] \arrow[d, "l_r^-"] & \&r \arrow[r, "F", "\text{incorporation} \equiv \text{reset}"'] \arrow[d, "n_r^-"] & \& \\
 & & \& & \& &
\end{tikzcd} \tag{2.36}$$

The introduction of the additional intermediate state decreases the lower bound for the error fraction in the limit $F \to 0$ down to

$$f_e \approx (f_0)^3 . \tag{2.37}$$

Note that in Eq. (2.36) we did not draw the reactions with rates that are negligible. Moreover, generalizing Eq. (2.36) to a reaction scheme involving N high energy intermediates results in a lower bound for the error fraction that is given by

$$f_e \approx (f_0)^{N+1} . \tag{2.38}$$

In real biological systems, the additional intermediate steps can be thought of as conformational changes of the enzyme that alter the thermodynamic properties locally, such that the absolute value of the free energy associated with the hybridized nucleotide becomes higher. In these intermediate stages, the wrong nucleotide still has higher chance of detaching from the template strand. Hence, by going through the conformation change, the concatenating enzyme "double-checks" the incoming nucleotide before it actually catalyzes the covalent bond formation [30].

2.4.3. Energetic and kinetic error discrimination

In our derivation of the lower bound Eqs. (2.21) for the error fractions in the Michaelis-Menten copying scheme, we assumed that the discrimination between right and wrong nucleotides resides solely in the rates for the back reactions and that the rates for the forward reactions are identical, i.e., $k_w^-/k_r^- = K_r/K_w = e^\gamma$ and $k_w^+ = k_r^+$. This choice of the kinetic parameters corresponds to an energy landscape with equal barrier heights and an energy difference γ separating the valleys corresponding to the &r- and &w-state [139]. Since the lower bound for the error fraction is determined by the energies of the intermediate states, we say that the reaction scheme operates in the *energetic regime* [139].

In this section, we also allow for an energy difference δ between the heights of the barriers which have to be overcome in the forward and backward reactions (see Fig. 2.11), such that

$$\frac{k_w^+}{k_r^+} = e^{-\delta} \quad \text{and} \quad \frac{k_w^-}{k_r^-} = e^{\gamma-\delta}. \tag{2.39}$$

The explicit rates for formation and decomposition resulting from the energy landscape depicted in Fig. 2.11 then read

$$k_r^+ = k_0 e^{\delta+\epsilon}, \quad k_w^+ = k_0 e^{\delta}, \quad k_r^- = k_0 e^{\delta} \quad \text{and} \quad k_w^- = k_0 e^{\gamma}, \tag{2.40}$$

where k_0 corresponds to the inverse of overall time scale associated with first reaction

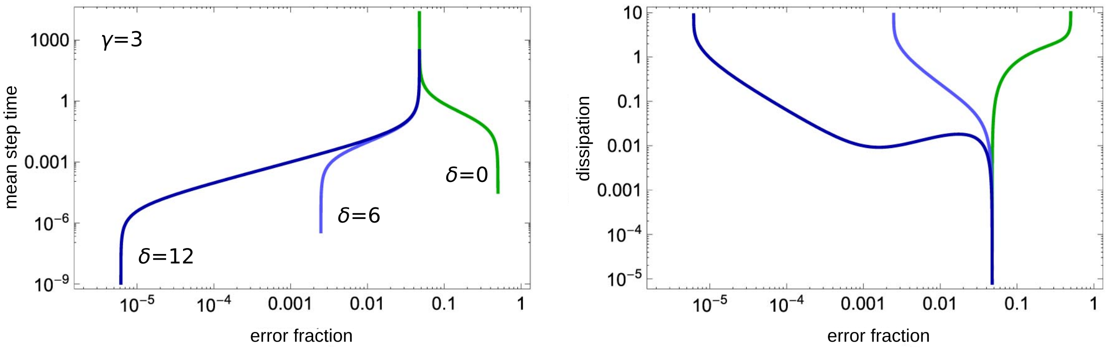


Figure 2.12.: Trade-offs between copying speed, accuracy, and dissipated energy for the generalized Michaelis-Menten scheme for templated polymerization for $\epsilon = 10$. The mean step time corresponds to the time needed to complete one extension step on average. The blue curves correspond to a reaction scheme operating in the kinetic regime with $\delta > \gamma$ while the green curve corresponds to a scheme working in the energetic regime with $\delta > \gamma$. In the kinetic regime, the lowest error rate is obtained in the $F \to \infty$ limit. The error fraction is a monotonically decreasing function of F. For low error fractions, the amount of dissipated energy diverges. In contrast, the smallest error fraction is achieved in the $F \to 0$ limit in the energetic regime. For these quasi-static conditions, dissipation is absent. The figure is adapted with permission from [139].

pathway. In this more general scenario, the error fraction is given by

$$f_e = \frac{e^{\delta} + \frac{F}{k_0}}{\left(e^{\gamma} + 1\right) e^{\delta} + \left(e^{\gamma} + 1\right) \frac{F}{k_0}}. \tag{2.41}$$

Now, the lower bound for the error fraction f'_0 takes the form

$$f'_0 = \frac{1}{1 + e^{\max\{\delta,\gamma\}}} \approx e^{-\max\{\delta,\gamma\}}. \tag{2.42}$$

Depending on whether $\delta > \gamma$ or $\delta < \gamma$, the replication scheme either operates in the *kinetic* or *energetic regime*. In the kinetic regime, the lowest error rate is achieved for a diverging rate for the final incorporation step, i.e., $F \to \infty$, which is necessary to drive the process forming the intermediate far out of equilibrium [139]. In this case, the error fraction becomes monotonically decreasing for increasing F. In other words, the faster the process gets, the more accurate it becomes. Driving the copying process far out of equilibrium is inevitably coupled to the dissipation of energy which has to be provided by the environment in some form. In contrast, in the energetic regime, the smallest error fraction is achieved for a vanishing final incorporation rate $F \to 0$ as discussed in Section 2.4.1. Under these quasi-static conditions, no dissipation occurs. The trade-offs

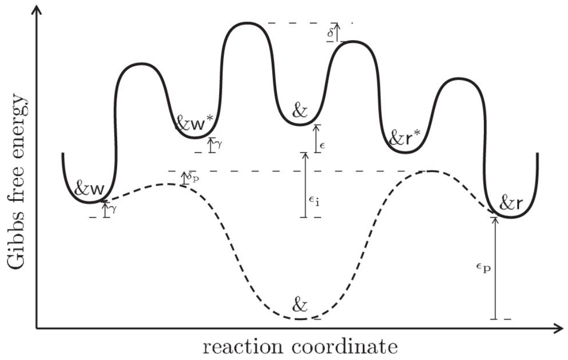


Figure 2.13.: Free energy landscape for the generalized Hopfield scheme for templated polymerization. The proofreading pathway, i.e., the reaction from the second high energy intermediate state to the initial state, is symbolized by the dashed curve. γ, δ, and δ_P are the energetic and kinetic discrimination constants allowing for discrimination of right and wrong nucleotides in the first pathway as well as in the proofreading pathway. The free energy that is required to complete a full cycle & $\rightarrow$ &r^* $\rightarrow$ &r $\rightarrow$ & or & $\rightarrow$ &w^* $\rightarrow$ &w $\rightarrow$ & sums up to $\epsilon + \epsilon_i + \epsilon_p$. The sum $\epsilon + \epsilon_i + \epsilon_p$ can be interpreted as the energy that the enzyme has to take up from the environment to move back to the initial state when a nucleotide gets rejected in the last intermediate state (see [139] for details). The figure is adapted with permission from [139].

between copying speed, accuracy, and dissipated energy arising in the energetic and the kinetic regime of generalized Michaelis-Menten scheme are sketched in Fig. 2.12 for different values of the *discrimination constants* γ and δ. The distinction between kinetic and energetic regimes is not only hypothetical. In Ref [144] experimentally measured kinetic rates appearing in two distinct DNA replicating systems were analyzed. The first system employs a *T*7 polymerase while the second one uses a *Pol*γ polymerase. The authors of that study argue that the copying process can be mapped onto the simple Michales-Menten scheme, and further show that the *T*7 polymerase works in the energetic and *Pol*γ in the kinetic replication regime.

In the Hopfield scheme, the energy landscape can be modified such that discrimination of wrong nucleotides appears also in the forward reaction from the initial state to the first intermediate state as well. For instance, for the energy landscape depicted in Fig. 2.13 we have

$$\frac{k_w^+}{k_r^+} = e^{-\delta},\ \frac{k_w^-}{k_r^-} = e^{\gamma-\delta},\ \frac{l_w^+}{l_r^+} = e^{\delta_P} \text{ and } \frac{l_w^-}{l_r^-} = e^{\delta_P-\gamma}. \tag{2.43}$$

The above relations only specify the ratios of corresponding rates involving a right or a wrong nucleotide. They do not specify the prefactors (equivalent to the k_0 or $k_0 e^{\epsilon}$ terms

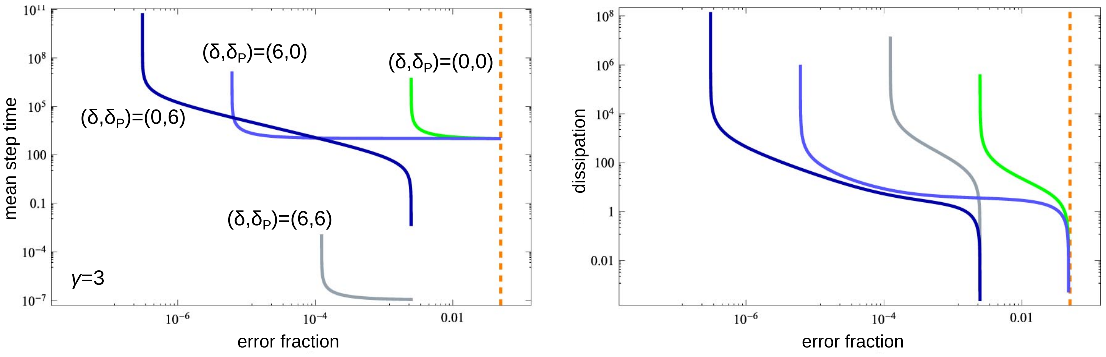


Figure 2.14.: Trade-offs between copying speed, accuracy, and dissipated energy for the generalized Hopfield scheme for templated polymerization for $\epsilon = 10$. The dark blue curve corresponds to a reaction scheme with kinetic discrimination in the first pathway and in the proofreading pathway. The light blue and the grey curve stand for reaction schemes where kinetic discrimination appears, either in the first or proofreading pathway. The green curve corresponds to a system operating in an entirely energetic discrimination regime. The dashed orange lines correspond to the equilibrium error fraction achievable in the simple Michalis-Menten scheme for $F \to 0$. The perfectors of the reaction rates, which are left unspecified by the relations in Eq. (2.43), are chosen to minimize the error fraction. In this way, the trade-offs recover the minimal error rate that can be obtained. The figure is adapted with permission from [139].

in Eqs. 2.40) of these rates. One can show, that the error fraction in this system obeys

$$f_e \geq \frac{1}{1 + e^{\max\{\gamma,\delta\}+\gamma+\delta_P}} \approx e^{-\max\{\gamma,\delta\}-\gamma-\delta_P}. \tag{2.44}$$

The equation for the error fraction shows that the discrimination constants, that describe the energetic differences in the heights of the barriers and the valleys associated with the first intermediate forming reaction pathway, never appear together. One can show that in this system, the smallest error fraction is always obtained in the limit of a vanishing rate for the final incorporation step $F \to \infty$. The free energy which is dissipated in the second reaction pathway of the copying process is entirely used to increase the accuracy of the final copy. Different trade-offs between copying speed, accuracy, and dissipated energy for different exemplary combinations of the discrimination constants γ, δ and δ_P are shown in Fig 2.14.

2.4.4. Further reactions schemes leading to enhanced accuracy and new trade-offs

Since the publication of John Hopfield and Jacques Nino's seminal work in the mid-1970s, numerous variants of the original reaction schemes, as well as new topologies for reaction networks, have been proposed. A large number of theory papers exist that investigate the thermodynamic properties of these networks and classify different working regimes [145, 146, 147, 148, 149, 150, 151, 136, 152]. The working regimes resulting from different topologies of the proofreading networks are partially mutually exclusive. Moreover, they are characterized by different trade-offs between the speed of product formation, accuracy, as well as the energy needed to reduce errors and to drive the reaction in the forward direction, i.e., to the product side. New mathematical formalisms were required to elucidate the coupling between the increase of accuracy and energy consumption and to find optimal conditions.

The paper by Galstyan et al. bringing up the idea of *proofreading through spatial gradients* [153] might be particularly interesting in the context of the origins of life and early replication mechanisms. Therefore, the general idea shall briefly be discussed in the following. In proofreading through spatial gradients, the enhancement of accuracy does not rely on sophisticated structural properties of the copying enzyme, such as an exonucleolytic side, that leads to the formation and "double-checking" of intermediate states as in "traditional" proofreading. Such highly evolved copying "machines" most likely did not exist on early Earth. Instead, the spatial proofreading scheme is based on the time delay between the binding of the substrate and the final reaction step in which the product is formed. This time delay appears naturally if the two reactions occur at different locations in space. The separation of the two reaction steps in space can be achieved by spatial concentration gradients. The concentration of substrate molecules should be peaked in one location, and the concentration of some chemical activator species necessary in the final reaction step should have a sharp maximum at another location. Hence, the intermediate "loaded" with substrate molecule in the first location has to travel through space to the second location by diffusion in order to transition to the final state. If we assume that the wrong substrate has a higher rate for dissociation, the number of incorrect intermediates arriving at the second location is reduced. If the distance between the two locations is long enough, such that the traveling of the intermediate takes longer than the average time scale for substrate dissociation, remarkably low error fractions can be achieved. As in other "traditional" proofreading schemes, proofreading through spatial gradients gives rise to a trade-off between accuracy and the speed at which products are formed. Generally speaking, the longer it takes the "loaded" intermediate to travel from the first location to the second location, the higher the accuracy but the lower the product yield. As discussed above, proofreading is a process that is out of equilibrium and therefore coupled to the consumption of energy. In typical enzymatic proofreading mechanisms, it is the enzyme that is consuming this energy. In contrast, in proofreading through spatial gradients, energy is “consumed” by the system to maintain the spatial concentrations profile.

3. A kinetic error filtering mechanism for enzyme-free copying of nucleic acid sequences*

Accurate copying of nucleic acid sequences is essential for self-replicating systems. As discussed in Section 2.4 of Chapter 1, modern cells achieve error ratios as low as 10^{-9} with sophisticated enzymes capable of kinetic proofreading. In contrast, experiments probing enzyme-free copying of RNA and DNA as potential prebiotic replication processes find error ratios on the order of 10%. Given this low intrinsic copying fidelity, plausible scenarios for the spontaneous emergence of molecular evolution require an accuracy-enhancing mechanism. Here, we study a 'kinetic error filtering' scenario that dramatically boosts the likelihood of producing exact copies of nucleic acid sequences. The mechanism exploits the observation that initial errors in template-directed polymerization of both DNA and RNA are likely to trigger a cascade of consecutive errors and significantly stall downstream extension. We incorporate these characteristics into a mathematical model with experimentally estimated parameters, and leverage this model to probe to what extent accurate and faulty polymerization products can be kinetically discriminated. Limiting the time window for polymerization prevents completion of erroneous strands, resulting in a pool in which full-length products show an enhanced accuracy. This comes at the price of a concomitant reduction in yield. However, the yield rate and the fidelity can be simultaneously increased by kinetic error filtering, if the templates are not too long. Within cyclically varying environments, e.g., temperature cycles in hydrothermal systems, repeated copying attempts could produce exact copies of sequences as long as 50mers within their lifetime, facilitating the emergence and maintenance of catalytically active oligonucleotides.

*The chapter is adapted from the manuscript: "A kinetic error filtering mechanism for enzyme-free copying of nucleic acid sequences", by Tobias Göppel, Benedikt Obermayer, Irene A. Chen and Ulrich Gerland, which has been submitted for publication to *eLife*. The author of this thesis is the only first author of the manuscript. The manuscript is currently in the second round of the review process. The manuscript and the reviewers' comments are available on the bioRχive (see Ref. [154]).

3.1. Introduction

Accurate copying of genetic information is essential for the emergence of living systems [155, 156, 157]. In extant cells, template-directed polymerization of polynucleotides is catalyzed by sophisticated enzymatic machineries, which mitigate and correct copying errors [158]. A key enzymatic mechanism is kinetic proofreading, an on-the-fly correction scheme that uses chemical energy to perform multiple discrimination steps between correct and incorrect nucleotides [133, 134, 135], enabling remarkably low error ratios, e.g., between 10^{-10} and 10^{-8} per base pair for DNA replication. The principals of kinetic proofreading are described in detail in Section 2.4 of the introduction. Life must have emerged without complex enzymes [34, 33, 31]. Enzyme-free copying of short information-carrying polymers such as RNA or DNA strands has been studied extensively [10, 159]. In particular, non-enzymatic template-directed polymerization has become an established experimental model system to investigate prebiotic modes of copying: A short strand bound to a longer 'template' strand is sequentially extended at its 3'-end with single nucleotides or short oligomers [160, 119, 71, 81, 161, 68], producing a (partial) complementary copy of the template. Lacking an inherent correction mechanism, errors during this copying process are frequent [78, 116, 77, 119]. Experiments in the presence of all four bases suggest that not even genetic information as short as ten nucleotides could be maintained by non-enzymatic template-directed polymerization [121].

How could accurate copying of genetic information be achieved without complex enzymes? A possible precursor to kinetic proofreading, which actively corrects errors right after they occur (see Fig. 3.1A), is a passive error filtering mechanism, in which erroneous copies are not corrected, but preferentially eliminated or separated based on their physicochemical properties. How could such error filtering arise in a prebiotically plausible scenario? A key experimental observation is that the speed of template-directed polymerization strongly depends on the sequence context [116, 71, 117]. Mismatches at the 3'-terminus of a partial copy slow down the extension reaction by one or two orders of magnitude [77], and facilitate the incorporation of further non-complementary nucleotides, leading to error clusters [78]. The first effect, called post-mismatch stalling, was originally discovered in enzymatic copying before it was observed in enzyme-free systems [118]. In combination, post-mismatch stalling and error clustering cause erroneous partial copies to grow slowly, opening the door to an error filtering mechanism based on kinetic discrimination (see Fig. 3.1B): With a limited time window for copying, only copies with no or few errors can reach full length, such that any physicochemical process that is length-selective can achieve error filtering (see Fig. 3.2A).

The two prerequisites for kinetic error filtering can both be provided by non-

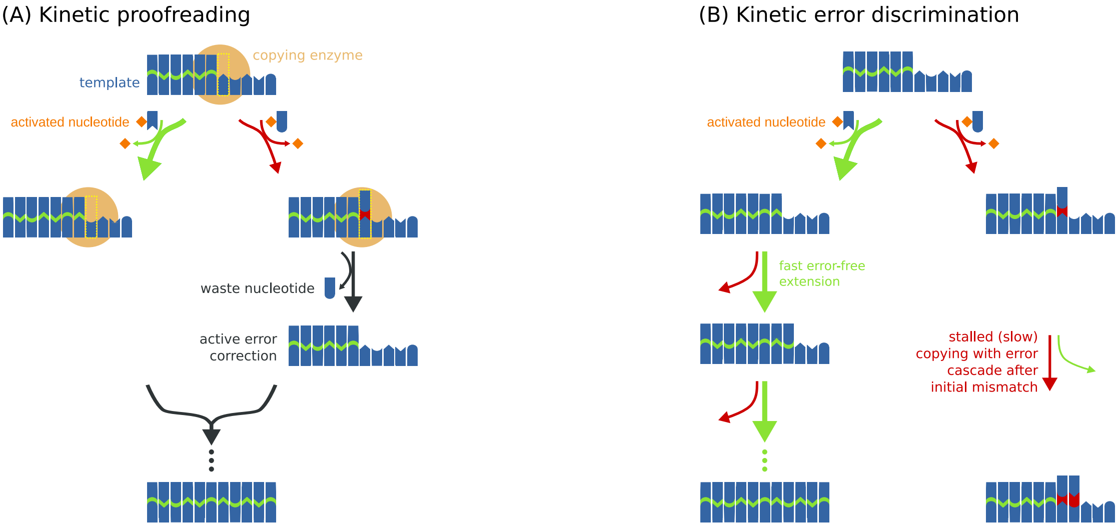


Figure 3.1.: Kinetic proofreading versus kinetic error discrimination. (A) Kinetic proofreading adds an error correction step on top of the thermodynamic discrimination between correct and incorrect nucleotides. A polymerase with proofreading ability can remove a covalenty attached mismatch, allowing for a second chance to incorporate the correct nucleotide. The error correction is coupled to the consumption of chemical energy. (B) In contrast, errors occurring during non-enzymatic copying remain. The accuracy is controlled by only one discrimination step, and cannot exceed the thermodynamic limit set by the intrinsic discrimination free energy. However, an initial error typically triggers a cascade of consecutive errors and kinetically stalls the speed of downstream extension, allowing for kinetic error discrimination: If the polymerization process is stopped after a limited time, accurate copies reach full length, whereas erroneous strands remain as short waste products.

equilibrium environments, e.g., on the early Earth: (i) A limited time window for the copying process emerges when the ambient temperature, pH, or molecular concentrations change periodically, from conditions that promote base-pairing to conditions that favor dissociation of hybridized strands [122, 123, 124, 125, 126, 28]. (ii) Length-selective physical properties, e.g. transport in thermal gradients [27, 162], accumulation on mineral surfaces [163], or retention within lipid vesicles [164], can cause a preferential loss of shorter strands. Both of these conditions can be simultaneously met in hydrothermal systems [162, 126, 28].

If kinetic error filtering is a plausible accuracy-enhancing mechanism, by how much could it boost the accuracy? Would it not reduce the yield of the copying process such as to annihilate its beneficial effects? And, most importantly, could kinetic error filtering be sufficiently effective to support the spontaneous emergence and maintenance of catalytically active oligonucleotides by template-directed polymerization? These questions intrinsically require a quantitative analysis, which we provide here.

Prior work [77] studied the beneficial effect of post-mismatch stalling on Eigen's error threshold [155], within a coarse-grained mutation-selection model of two replicators competing in an environment with constant carrying capacity. In contrast, we consider a primordial scenario, in which the accuracy and yield of a primitive copying process must be sufficient to form at least one accurate copy for a template, on average, before the template is destroyed, e.g., by hydrolysis [71]. If this condition is not met, then any accidental discovery of a weakly catalytic sequence by random assembly will be lost again before it can further evolve. We base our analysis on a quantitative model rooted in data from primer-extension experiments with DNA and RNA, including also the effect of error clustering [78]. Using this model, we explicitly study the stochastic kinetics of template-directed polymerization in cyclic environments that offer only limited time windows for polymerization. We first characterize the fidelity-yield trade-off that emerges within a single such time window. Our subsequent analysis then reveals that cyclic environments can effectively break this fidelity-yield trade-off. This permits kinetic error filtering to facilitate the emergence and maintenance of catalytically active oligonucleotides, by significantly increasing the sequence length for which correct copies can be obtained within the lifetime of a template.

3.2. Results

3.2.1. Kinetic error filtering versus kinetic proofreading

Correcting errors right after they occur is a natural solution to the problem of high intrinsic error rates. The evolutionary origin of kinetic proofreading is not clear, but extant cells use this principle not only in their processive copying enzymes mediating transcription [158], replication [165], and translation [137], but also in non-processive enzymes such as tRNA synthetase [135] or the T-cell receptor complex [138]. However, even a minimal, non-processive copying scheme with proofreading would require an enzyme that can perform a conformational transition coupled to energy release, in addition to catalyzing backbone bond formation (Fig. 3.1A). Therefore, it appears highly unlikely that a gratuitous proofreading mechanism would be available to primitive prebiotic replicating systems.

Without error correction, errors escaping the intrinsic thermodynamic discrimination will remain, unless erroneous copies are preferentially removed from the system. Kinetic error filtering is a two-step mechanism for such a preferential removal: First, it kinetically suppresses the formation of erroneous full length copies by limiting the copying process to a finite time window in a cyclic environment. Second, it preferentially leaks shorter strands out of the system, thereby removing the strands that contain most errors. From a thermodynamic perspective, kinetic error filtering is

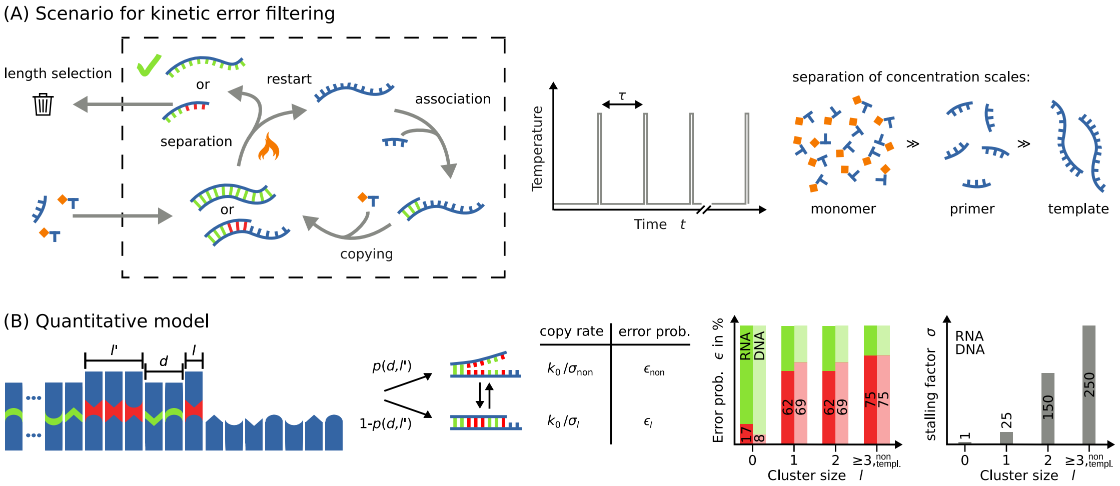


Figure 3.2.: Scenario and model for kinetic error filtering. (A) To support kinetic error filtering, a suitable environment must provide limited time windows for non-enzymatic copying and length-selective transport or adhesion of polynucleotides. We envisage a scenario, in which e.g. the ambient temperature changes periodically, leaving only time windows of typical duration τ for the association between complementary strands and template-directed polymerization. Furthermore, we posit that the system leaks shorter oligonucleotides, and thereby preferentially removes erroneous copies. Conversely, monomers and short 'primer' oligonucleotides can also readily enter the system from the external environment. (B) Our quantitative model for non-enzymatic copying distinguishes only between matching and mismatching base pairs. The length l of an error cluster at the extension site determines the stalling factor σ_l and the error fraction ϵ_l (numerical values shown in the bar plots). Additionally, the distance d to the preceding error cluster and its size l' affect the extension mode: The next extension occurs at the reduced speed and fidelity of a non-templated process with the probability $p(d, l')$ for an unbound terminus.

driven by (a part of) the free energy dissipated in the environment. This is in contrast to kinetic proofreading, where an enzyme couples the dissipation of chemical energy to error correction [133, 134].

3.2.2. Model

To function, kinetic error filtering requires a suitable non-equilibrium environment. We will consider a scenario of the type illustrated in Fig. 3.2A: A leaky compartment is embedded in an aqueous environment providing a mixture of chemically reactive nucleotides and their polymerization products. Longer sequences have an increased residence time within this compartment, due to their charge or physical size. We do not make any assumption about the specific mechanism mediating the retention of longer sequences; it could be based on surface interactions [166], size-dependent transport through the compartment boundary [164], or bulk transport effects such as thermophoresis and convection [29, 162]. While spontaneous polymerization produces oligonucleotides with a statistical distribution of chain lengths [70], with longer oligomers much less (in general, exponentially less [60]) likely than monomers, stochastic fluctuations may occasionally lead to a long 'template' sequence within such a compartment. We assume that the physico-chemical conditions (temperature, pH, or salt concentrations) in the vicinity of a template display a cyclic variation, such that the hybridization of short oligomers acting as 'primer' sequences only occurs within time windows of typical duration τ. The cyclic variation may arise from internal convection cycles [28] or from external periodic variations. To be more concrete, we will consider the case of temperature cycles for our model. The key assumption is that (partial) copies separate from their template at the end of a cycle, and that the probability for rebinding in the next cycle is low, since short primer molecules are much more abundant.

A submerged porous rock exposed to a temperature gradient could provide one natural realization of a suitable non-equilibrium environment. Within a pore, the interplay of convection and thermophoresis leads to a flow field, in which molecules move and experience periodic temperature changes, as has been demonstrated experimentally with a controlled lab setup [28]. In this case, polynucleotides of length 35 experienced temperature cycles featuring a short peak, during which double strands dehybridize. The copying time windows τ within such thermal flow chambers are controlled by the chamber geometry [27].

In the quantitative analysis presented next, we will see that efficient kinetic error filtering imposes conditions on the timescale τ depending on the template length L. One might then object that this scenario for kinetic error filtering requires "fine-tuning" of conditions. However, natural compartments and pores come in a broad range of sizes and geometries, and this natural variation can produce a correspondingly broad range of timescales τ. As a consequence, the massively parallel nature of natural experiments eliminates the potential fine-tuning issue.

3.2.3. Template-directed polymerization in a limited time

To analyze kinetic error discrimination within a finite time τ, we use a mathematical model based on experimental characterizations of non-enzymatic template-directed polymerization [78, 77, 119]. Within this model, template-directed integration of monomers proceeds at a basal extension rate k_0 in the absence of any mismatches. This rate defines the basal extension timescale $t_0 = 1/k_0$, which serves as the elementary time unit for this study, since the actual experimental timescale depends on the precise chemical conditions, including the type of leaving group used for the chemical activation of nucleotides [115]. In typical experiments, t_0 is on the order of one hour.

We parameterize the probability ϵ for a copying error and the stalling factor σ as a function of the local structure of the template-copy complex at the extension site Fig. 3.2B). This structure is described by (i) the number $l \geq 0$ of successive mismatches directly at the extension site, (ii) the size $l' \geq 0$ of the next error cluster further upstream of the extension site, and (iii) the distance $d > 0$ to this next error cluster. Based on the values d and l', we estimate the probability $p(d, l')$ that a terminus following a series of mismatches is in an unbound dangling-off configuration [78], see Section 3.4.

A dangling terminus is extended with the error probability of an unbiased, non-templated extension, $\epsilon_{\text{non}} = 0.75$, and the corresponding extension rate is reduced by the stalling factor $\sigma_{\text{non}} = 250$ to k_0/σ_{non} [167, 78]. If the terminus is closed, i.e., with probability $1 - p(d, l')$, the stalling factor σ_l and the error probability ϵ_l depend only on the number l of mismatches at the extension site. The associated copying rate is

$$k_l = k_0/\sigma_l \, , \tag{3.1}$$

where $\sigma_0 = 1$ and the values for $l > 0$ are given in Fig. 3.2B. Experimentally, the basal extension rate, the stalling factors, and the error probabilities also depend on the exact sequence context, i.e, the templating and the incoming nucleotide as well as their neighbors [116, 77, 117, 71, 168, 121, 161]. However, averaging over results obtained for many random sequences with sequence-dependent parameters was found to be essentially equivalent to using sequence-averaged parameters instead [78]. This justifies the practical simplification that our model makes by distinguishing only between matching and non-matching base pairs. The average error probability ϵ_0 for extending a closed terminus with no mismatch is 0.08 for DNA and 0.17 for RNA [119, 80]. After a first mismatch, the error probability increases more than sevenfold for DNA and roughly threefold for RNA [78]. Systematic measurements of the copying accuracy following error clusters of size two are missing, but the existing data suggest that the error probability remains unchanged [78]. Typically, an initial mismatch stalls the copying speed by one to two orders of magnitude. A second mismatch then slows down the extension by another factor of six [77, 78].

This mathematical model corresponds to a Markov process, in which the stochastic template-directed polymerization dynamics depend only on the current state of the terminus. We analyze these dynamics with simulations based on the Gillespie algorithm [169] and with analytical approximations described below. For an overview of all relevant parameters, variables and observables see Table 3.1.

Parameter	
L	template length
k_0	basal extension rate
t_0	basal extension timescale ($1/k_0$)
σ_l	stalling factor after error cluster of size l
ϵ_l	error probability after error cluster of size l
k_{on}	primer-template association rate, set to $10^7\ \text{s}^{-1}$
c_{prim}	primer concentration
Variable	
l	size of error cluster including the terminus
k_l	extension rate following an error cluster of size l
d	distance to next upstream error cluster
l'	size of next upstream error cluster
τ	time window for copying or cycle duration
Observable	
t_{perf}	average copying time of an error-free product
$p(d, l')$	probability of unbound terminus
$f_e(\tau)$	error fraction
$Y(\tau)$	yield of completed strands
τ^*	optimal cycle duration
E_{waste}	wasted energy per completed copy
$t_{\text{cop}}(\tau)$	average time for the first error-free copy to appear
t^*_{cop}	minimum of $t_{\text{cop}}(\tau)$ with respect to τ

Table 3.1.: Overview of our parameters, variables and observables.

3.2.4. Kinetic separation of different error classes

Using the model of Fig. 3.2B, we simulate non-enzymatic template-directed polymerization with template length $L = 20$, tracking the number of full-length copies and their copying errors over time. We extract the time to completion for each full-length copy, to obtain the statistical distributions of completion times with the corresponding mean and median values for different error classes with a specified number of errors (see Fig. 3.3A). Here and below, all shown results are for DNA parameters whereas the corresponding plots for RNA parameters are shown in the Appendix A. As expected, copies containing no or few errors are completed much faster than highly erroneous copies. Error-free full-length copies display a near-Gaussian distribution of completion times (see Fig. A.1 in the Appendix) peaked at the mean completion time for perfect copies, $t_{\text{perf}} = L\, t_0$ (see inset of Fig. 3.3A).

Almost all copies, even the majority of the worst ones, reach full length within a completion time of $L\sigma_3 t_0$. Copies with few errors show asymmetric distributions of completion times (see Fig. A.1 in the Appendix), with tails at small times that overlap with the error-free distribution (see inset of Fig. 3.3A), which has implications for the limits of kinetic error discrimination (see below). We first turn to a trade-off that is inherent to kinetic error discrimination: Shortening the time window τ increases the fidelity of the obtained full-length copies, but decreases the yield.

3.2.5. Fidelity-yield trade-off

We measure the fidelity of the copying process via the error fraction $f_e(\tau)$, defined as the average fraction of wrongly incorporated nucleotides in full-length products when template-directed polymerization is stopped after time τ. The yield $Y(\tau)$ of the copying process is the fraction of templates for which copying has completed. Both, the error fraction and the yield increase with time and saturate when $\tau > L\sigma_3 t_0$ (circle symbols in Fig. 3.3B and Fig. 3.3C. The yield approaches 100%, since our model does not contain any side reactions such as template cleavage by hydrolysis. However, the error fraction concomitantly reaches values larger than 0.25. This is clearly too high to conserve, e.g., the function of a ribozyme, even if the ribozyme is relatively robust against mutations [170]. In the other extreme of very short times τ, the error fraction is dramatically reduced to less than 1%, but the yield becomes essentially zero.

The vertical dashed lines in Fig. 3.3B and Fig. 3.3C mark the mean completion time t_{perf} of perfect copies. In this regime, the yield grows strongest while the error fraction still remains low. Hence, for a single copying cycle, values of $\tau \approx t_{\mathrm{perf}}$ represent the best compromise. However, another interesting feature of the fidelity curve in Fig. 3.3C is that $f_e(\tau)$ apparently approaches a nonzero lower limit as $\tau \to 0$, which also appears consistent with the overlapping completion time distributions (inset of Fig. 3.3A). What factors determine this limit on how much kinetic error discrimination can improve the copying accuracy?

3.2.6. Kinetic error discrimination is limited

To understand the error discrimination for short times, we turn to an analytically solvable simplified model, which accurately describes the behavior in this regime. The underlying approximations rely on the observation that copies containing a cluster of multiple consecutive errors have a negligible probability to be completed at short times τ. Hence, we ignore the dependence of the model parameters on the cluster size l by using a single stalling factor σ and a constant error probability ϵ_l for all $l > 0$. Furthermore, we neglect the effect of preceding error clusters, i.e., we set

$p(d, l') = 0$, such that the copying rate is restored to k_0 immediately after one correct incorporation. We derive the analytical expressions for the resulting error fraction and yield in Section 3.4.4.

The analytical solutions for $\sigma = \sigma_1$ (solid lines in Fig. 3.3B and Fig. 3.3C) confirm that the simplified model is equivalent to our full model for small times τ. They deviate when the first error clusters emerge: In the simplified model, (i) the yield grows more rapidly and saturates earlier, since error clusters do not increase stalling, and (ii) the error fraction is smaller, since the error probability does not grow after the first mismatch. For $\tau \to \infty$, the simplified model predicts an error fraction $\epsilon_0/(1 + \epsilon_0 - \epsilon_1)$ in the limit of long templates (but with L fixed). This limit is independent of the stalling factor σ, since all strands reach full length.

Importantly, the simplified model lets us determine the lower limit f_e^{low} of the error fraction by including only copying processes with one or no error (see Section 3.4.4),

$$f_e^{\mathrm{low}} = \frac{\epsilon_0}{L} \frac{1 + \frac{(L-1)(1-\epsilon_1)}{\sigma(1-\epsilon_0)}}{1 + \epsilon_0 \frac{(L-1)(1-\epsilon_1)}{\sigma(1-\epsilon_0)}} \,. \tag{3.2}$$

In the strong stalling regime ($\sigma \gg L$) this reduces to $f_e^{\mathrm{low}} = \epsilon_0/L$, reflecting the absence of a kinetic penalty for errors in the last copying step. Fig. 3.3D compares the scaling of the lower bound (solid line) with $(L-1)/\sigma$ to the corresponding full analytical expression (symbols) in the limit $\tau \to 0$ for different values of L. For $L < \sigma$, the curves collapse onto one line and are well approximated by the lower bound, while they start to separate beyond this regime. Interestingly, long strands contain less errors than short ones for a fixed value of $(L-1)/\sigma$.

Returning to the full model, we ask how much the error fraction can be lowered by reducing τ. We consider the fold-reduction in the error fraction, $f_e(\infty)/f_e(\tau)$, evaluated both for $\tau \ll t_{\mathrm{perf}}$ and $\tau = t_{\mathrm{perf}}$ at different template lengths (see Fig. 3.3E). The first case merely illustrates the maximal possible fold-reduction at the cost of a vanishingly small yield. In contrast, the second case illustrates what is realistically attainable with a yield that is sizeable at small lengths, but decreases exponentially with L (see Fig. 3.3E).

Why does the accuracy increase with length when $\tau \ll t_{\mathrm{perf}}$? The finite error fraction is mostly due to isolated errors, since error clusters would strongly stall the copying process and hence prevent the copy from reaching full length. However, isolated errors are rare since an initial mismatch is likely to trigger an error cascade. The longer the strand, the higher the probability for an error-cluster at some point. Long strands with isolated mismatches are thus more unlikely than short strands.

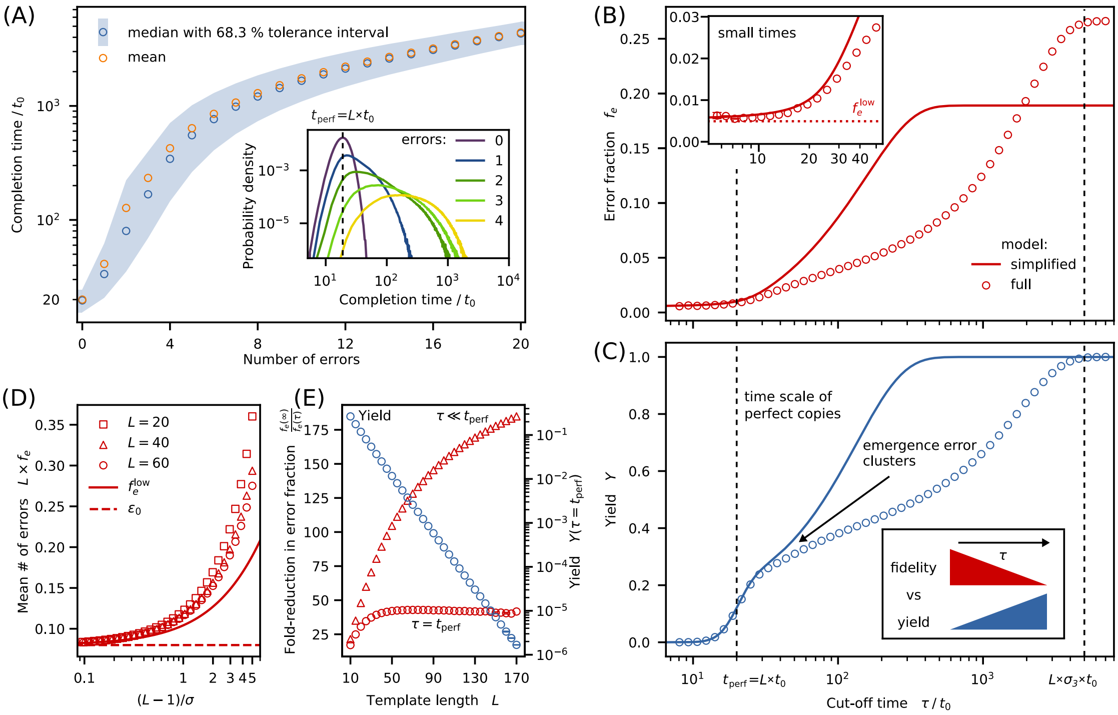


Figure 3.3.: (A) Kinetic separation of different error classes. Typical copying times increase rapidly with the number of errors, as shown here for a template of length $L = 20$. Main panel: Mean and median completion times with centered 68.3 % confidence interval. Inset: Distributions of the completion time for zero to four errors (normalization such that the sum of areas below the curves for all possible error numbers is one). The zero-error distribution peaks at $t_{\mathrm{perf}} = L\, t_0$, but overlaps with the distributions for one or more errors. (B) Fraction of errors in full-length copies, $f_e(\tau)$, and (C) yield $Y(\tau)$ as a function of the copying time window τ. The increase in copying fidelity for smaller τ is at the expense of the yield. Lines show the analytical solution of the simplified model, whereas data points are obtained from stochastic simulations of the full model (error bars shown only for statistical errors $> 1\%$ of the mean). The simplified model approximates the full model well when τ is not much larger than the mean completion time of perfect copies. (D) In the short time limit ($\tau \to 0$), the absolute number of errors within completed copies depends only on $(L-1)/\sigma$. For strong stalling ($\sigma > L$), the mean error number Lf_e (symbols) is well approximated by the lower bound Lf_e^{low} (solid line). For $\sigma \gg L$, Lf_e^{low} reduces to the error probability ϵ_0. For a fixed value of $(L-1)/\sigma > 1$, longer strands contain fewer errors than short ones. (E) The error fraction f_e can be decreased significantly by reducing τ compared to the error fraction obtained for $\tau \to \infty$. The reduction of the error fraction achieved by choosing $\tau = t_{\mathrm{perf}}$ goes hand in hand with a reduction of the yield Y.

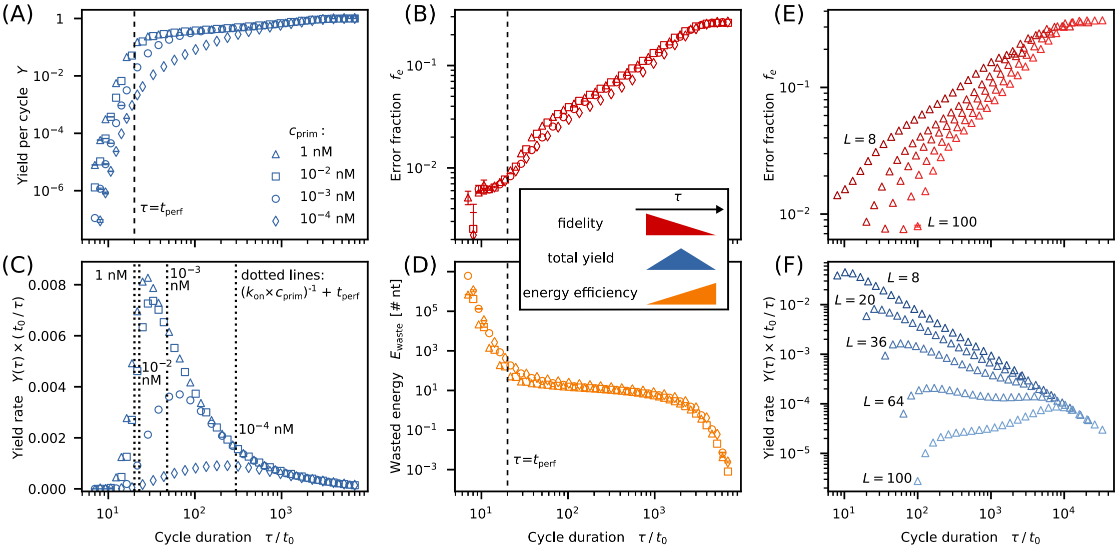


Figure 3.4.: Fidelity-yield-energy trade-off in cyclic environments (kinetic error filtering). The (A) yield per cycle and (B) mean error fraction both decrease as τ is lowered (the vertical dashed line marks t_{perf}). (C) However, the yield rate $Y(\tau)/\tau$ displays a maximum at intermediate τ values (vertical dotted lines indicate the timescales $(k_{\mathrm{on}}\, c_{\mathrm{prim}})^{-1} + t_{\mathrm{perf}}$ for comparison), such that fidelity and yield can be increased simultaneously over a wide range of τ values. (D) Stronger error filtering is also coupled to an increased energy waste, which is proportional to the number of nucleotides contained in uncompleted copies. (E) Error fraction $f_e(\tau)$ and (F) yield rate for different template lengths L (at fixed $c_{\mathrm{prim}} = 1$ nM).

3.2.7. Quantitative model for the kinetic error filtering scenario

We now turn to the full scenario for kinetic error filtering (see Fig. 3.2A) and follow one template over a long observation time in a periodically changing environment. It is clear from the above analysis that short cycle times τ will lead to high accuracy, while the yield $Y(\tau)$ per cycle will be poor. However, the overall system output is determined by the yield rate, i.e., the yield per unit time, $Y(\tau)/\tau$. We assume that the duration of the temperature peak in the full scenario is much shorter than τ, and that the peak temperature is high enough to separate templates from both partial and completed copies. In addition to the template-directed polymerization process, we have to account for template-primer binding (see Fig. 3.2A). Experiments [171, 172, 173] suggest an association rate k_{on} of about $10^7\,\mathrm{s}^{-1}\,\mathrm{M}^{-1}$. With an extension time of $t_0 = 1\,\mathrm{h}$, we have $k_{\mathrm{on}} = 3.6 \times 10^{10}/(t_0\,\mathrm{M})$. In our stochastic simulations, the time for each association event is drawn from an exponential distribution with mean $1/k_{\mathrm{on}} c_{\mathrm{prim}}$, where c_{prim} is the primer concentration.

Since kinetic proofreading consumes free energy to increase fidelity, we also seek to analyze the unproductive free energy consumption of kinetic error filtering. Every time a covalent bond is formed a leaving group is consumed. However, the assembly of copies that remain incomplete at the end of a cycle is unproductive. Thus, the wasted free energy per completed copy is

$$E_{\text{waste}} = \frac{N_{\bar{c}}}{N_c} \langle l_{\bar{c}} \rangle \, \Delta G_{\text{lg}} \, , \tag{3.3}$$

where N_c is the number of completed copies produced during the observation time t, $N_{\bar{c}}$ the number of incomplete copies, $\langle l_{\bar{c}} \rangle$ the average length of incomplete copies, and ΔG_{lg} the activation free energy per leaving group.

3.2.8. Cyclic environments mitigate the fidelity-yield trade-off

We revisit the trade-off between fidelity and yield in our full scenario for kinetic error filtering. We explore the behavior over all possible cycle durations τ, since natural non-equilibrium environments display a broad range of timescales over which their physico-chemical conditions vary (e.g., due to convective cycles, as discussed above). Per cycle, the yield (see Fig. 3.4A) and the error fraction (see Fig. 3.4B) for a template of length $L = 20$ display essentially the same τ-dependence as observed before in Fig. 3.3B and Fig. 3.3C, except for an additional dependence on the primer concentration, which affects the timescale of template-primer binding. However, the more relevant quantity now is the yield rate $Y(\tau)/\tau$. Remarkably, the yield rate displays a peak as a function of τ see Fig. 3.4C). The peak becomes more pronounced with increasing primer concentration. At our largest concentration, $c_{\text{prim}} = 1\,\text{nM}$, where the association time of the primer-template complex is negligible, the yield rate peaks at a cycle time close to t_{perf} (see Fig. 3.4C).

The peak in Fig. 3.4C implies that the fidelity-yield trade-off disappears over a certain range of cycle times: The fidelity and yield rate increase simultaneously as the cycle period τ is reduced from large times, until a value τ^* is reached where the yield rate is optimal. The τ-range of this simultaneous increase is largest for $c_{\text{prim}} = 1\,\text{nM}$, whereas the effect becomes weaker for smaller concentrations.

The simultaneous increase of fidelity and yield rate comes at a free energy cost: The wasted free energy per completed copy, E_{waste}, increases monotonically with decreasing τ (see Fig. 3.4D). For $\tau \approx t_{\text{perf}}$, hundred or more activated nucleotides are wasted per full-length copy, depending on the primer concentration. In constrast, for large τ almost no activated nucleotides are wasted.

How does the behavior in cyclic environments depend on the template length? Fig. 3.4E and Fig. 3.4F display the the error fraction $f_e(\tau)$ and the yield rate $Y(\tau)/\tau$ for different lengths L at the same primer concentration ($c_{\text{prim}} = 1$ nM). With increasing L,

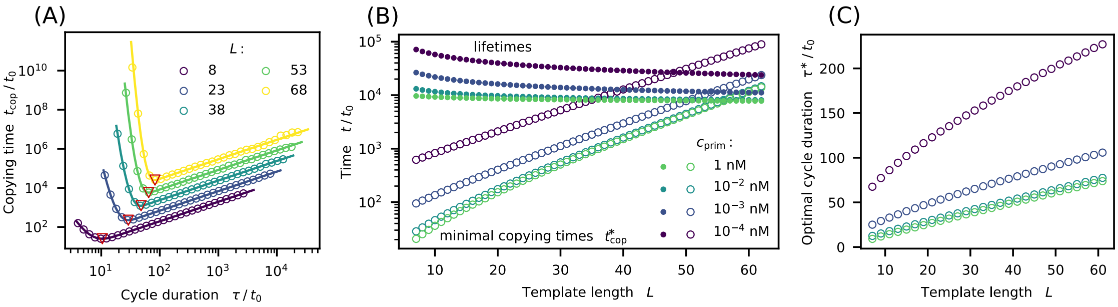


Figure 3.5.: Kinetic error filtering at the minimal copying time optimum. (A) Copying time t_{cop} as a function of the cycle duration τ for different template lengths at $c_{\mathrm{prim}} = 1$ nM. All curves display a distinct minimum (red triangles) at an optimal cycle duration τ^*. Symbols show data from stochastic simulations, while lines are obtained from (3.37). (B) The minimal copying time t^*_{cop} increases roughly exponentially with the template length, while the average template lifetimes decrease with L. These two timescales become equal for template lengths around 50 bases, depending on the primer concentration c_{prim}. (C) Dependence of the optimal cycle duration τ^* on L.

the yield maximum moves to longer cycle durations, becomes less pronounced, and eventually disappears. Concomitantly, a second maximum emerges at larger cycle periods, and hence larger error fractions. A significant increase in fidelity at high yield is only possible for lengths at which the left maximum still exists (see Fig. A.4).

3.2.9. Error-free copying

Is the fidelity-yield-boost in cyclic non-equilibrium environments sufficient to facilitate the emergence of molecular evolution? To not immediately lose a newly discovered functional sequence, a primitive copying process must at least form one accurate copy before the sequence is destroyed, e.g. by hydrolysis [71]. How long does it take to obtain the first error-free copy? To address this question, we compute the average copying time t_{cop}, defined as the average time for the first perfect copy to appear. We then compare t_{cop} to typical lifetimes of template sequences.

Using the exact results for our simplified model (see Section 3.4.4), we derive an analytical expression for t_{cop} that is also valid for the full model (see Section 3.5). For $\tau \gg (k_{\mathrm{on}}\, c_{\mathrm{prim}})^{-1}$, this expression simplifies to

$$t_{\mathrm{cop}}(\tau, L) = \frac{\tau}{(1-\epsilon_0)^L} \frac{1}{1 - \Gamma(L, \tau/t_0)/\Gamma(L)}, \tag{3.4}$$

where $\Gamma(L, \tau)$ and $\Gamma(L)$ are the incomplete and regular gamma function, respecitvely. Fig. 3.5A shows t_{cop} as a function of the cycle duration τ for different lengths at fixed

$c_{\text{prim}} = 1$ nM. All curves exhibit a distinct minimum at an optimal cycle duration τ^*. The copying time t^*_{cop} at the optimal cycle duration increases roughly exponentially with the template length Fig. 3.5B and Section 3.5), while τ^* grows roughly linearly with L (see Fig. 3.5C). As long as $c_{\text{prim}} \geq 10^{-2}$ nM, the primer concentration does not significantly affect t^*_{cop}. In this regime, and assuming $t_0 = 1$ h, a first perfect copy of a 20-mer would typically arise within a few days, whereas about 20 weeks would be required for a 50-mer.

How does the length-dependent optimal copying time $t^*_{\text{cop}}(L)$ compare to the lifetime of the template? For temperatures below 37° C, DNA hydrolysis is hardly measurable, but the lifetimes decrease rapidly with temperature [174, 175]. In environments with temperature peaks that separate copies from templates, the high temperature phases limit the template lifetime [28]. To estimate lifetimes, we apply the same temperature profile and environmental conditions as in the experiment of Ref. [28] and use the predictive formula for degradation rates given in Ref. [174], see Section 3.4 for details. The resulting lifetimes depend on L (see Fig. 3.5B), since the number of hydrolysis sites increases with L, while the number of temperature cycles decreases with L due to the growing optimal cycle duration (see Fig. 3.5C) . The length where the copying time $t^*_{\text{cop}}(L)$ matches the estimated lifetime is around $L \sim 50$, with only a weak dependence on the primer concentration (see Fig. 3.5B). Hence, kinetic error filtering can give rise to at least one error-free copy of DNA ~50-mers during their lifetime. With RNA parameters, this length threshold is at ~25-mers (see Fig. A.7).

3.3. Discussion

The mechanistic basis of kinetic error discrimination in non-enzymatic copying of nucleic acid sequences is experimentally well documented: initial errors stall the copying process and increase the error probability for subsequent nucleotides [77, 78]. We showed that these molecular effects give rise to a strong kinetic discrimination against errors in full-length copies, if the time window for template-directed polymerization is sufficiently short (see Fig. 3.3A). When this kinetic error discrimination mechanism is embedded in a length-selective cyclic environment, a kinetic error filtering scenario emerges (see Fig. 3.2A) with several interesting features: (i) Kinetic error filtering does not require any sophisticated enzymes, and could thus act as a prebiotic precursor to kinetic proofreading in template-directed polymerization. (ii) Kinetic error discrimination displays an intrinsic fidelity-yield trade-off (see Fig. 3.3B and Fig. 3.3C). This is in contrast to kinetic proofreading, which displays an intrinsic speed-accuracy trade-off [136]. However, the cyclic environment of the kinetic error filtering scenario creates a regime, in which the fidelity-yield trade-off is broken, such that reducing the cycle time

τ simultaneously increases both, fidelity (see Fig. 3.4B) and yield (see Fig. 3.4C), at the cost of chemical energy (see Fig. 3.4D). Energy efficiency is likely not a primary concern for early copying scenarios (but might become increasingly important as prebiotic living systems become more sophisticated and compete with each other). (iii) The cycle time τ can also be chosen to minimize the average time t_{cop} required to produce the first exact copy of a template (see Fig. 3.5A), rather than to maximize the yield (see Fig. 3.4C and Fig. 3.4F). Importantly, kinetic error filtering could sufficiently reduce t_{cop} to faithfully copy up to $\sim$50-mer templates within their lifetime (see Fig. 3.5B).

Kinetic error filtering could spontaneously arise in hydrothermal systems: Convective cycles produce periodic variations of the temperature and other physico-chemical properties of the local environment, creating limited time windows for copying, combined with enhanced loss and degradation rates for short strands [28]. Since the geometries of natural hydrothermal systems vary over a broad range, one may expect a correspondingly broad range of convective cycle times, such that different systems will naturally sample the τ values required for different template lengths and optimization criteria. Length selection could result, e.g., from an interplay between convection and thermophoresis [27, 162], from accumulation on mineral surfaces [163], or retention within lipid vesicles [164].

We note that the kinetic error filtering scenario studied here is inherently different from a previously described effect of post-mismatch stalling on the error threshold in a mutation-selection model [77]. The latter model describes the competition between replicators in an environment with constant carrying capacity, a scenario that may arise at a later evolutionary stage than considered here. On a mathematical level, kinetic error filtering relies on fluctuations and rare events in an explictly time-varying environment. In contrast, the effect reported in Ref. [77] results from a shift in the balance between the opposing average forces of mutation and selection.

For the stage of prebiotic evolution considered here, a key issue is the maintenace of a functional nucleotide sequence that may arise by chance. The sequence might display a catalytic activity that exerts a positive feedback onto its own synthesis, or a negative feedback onto its own degradation. Initially, its catalytic activity is likely not strong enough to significantly boost its replication. However, such a fledgling ribozyme (or DNAzyme) could further evolve, if a weak replication process supports its maintenance against degradation [176]. The relevant threshold for maintenance is that the sequence gives rise to at least one error-free copy before it is degraded, e.g. by hydrolysis [71]. Our analysis suggests that kinetic error filtering can reach this threshold for DNA sequences of up to $\sim$50 bases and RNA sequences of up to $\sim$25 bases (Fig. 3.5B and Fig. A.7). Short ribozymes and DNAzymes already display remarkable catalytic abilities, with DNAzymes not (clearly) inferior to ribozymes [64, 9]. In enzyme-free RNA copying, an important fraction of errors is due to G:U wobble paring [32, 119]. It is unclear whether the ribonucleotides available under prebiotic conditions correspond to the canonical ones found in extant living systems [177, 178, 179]. Alternative ribonucleotides not prone to wobble pairing could enable higher copying fidelities. For instance, replacing U with 2-thio-U significantly increases the fidelity in template-directed polymerization [180], such that RNA-based systems might reach similar error probabilities as DNA-based systems. Taken together, it appears that kinetic error filtering could push the enzyme-free copying of nucleic acid sequences via template-directed polymerization across an important threshold, facilitating the emergence and maintenance of sequences with catalytic functions.

3.4. Mathematical details

3.4.1. Recovery from error clusters

The extension rate k_l defined in (3.1) depends on the length l of the immediate error cluster at the extension site through the stalling factor σ_l. The error probability ϵ_l for the next incorporated nucleotide also depends on l. Moreover, effective extension rate and error probability also depend on the distance d to the preceding error cluster and its size l' (see Fig. 3.2B) and we write

$$k_{l,d,l'} \equiv k(l,d,l') \text{ and } \epsilon_{l,d,l'} \equiv \epsilon(l,d,l'). \tag{3.5}$$

If a matching nucleotide is incorporated after a single mismatch, the extension rate recovers to an almost normal level according to experimental observations [78]. After two matches following the isolated error, no effect on the copying process was measured anymore. We therefore assume that a preceding error cluster of length $l' = 1$ only affects the dynamics if $d = 1$. In this case, extension rate and error probability are given by $k(l,1,1) = 0.67 \times k_l$ and $\epsilon(l,1,1) = 1.25 \times \epsilon_l$ for $l \leq 2$.

In contrast, clusters of several mismatches influence the copying dynamics on a larger scale. The number of correctly incorporated nucleotides that are needed to restore the unperturbed extension dynamics increases with the size of the preceding error cluster. To include this effect, we use an extension of the binding model described in Ref. [78]. Matching nucleotides at the terminus following an error cluster are in a dangling configuration with probability p, which is a function of the distance d to and the size l' of the preceding error cluster, i.e.,

$$p = p(d,l'). \tag{3.6}$$

In Ref. [78], RNA folding software [109] is used to estimate p. It is shown, that the

d	1	2	3	> 3
$p(d, l' \geq 3)$	0.9	0.5	0.1	0
$p(d, l' = 2)$	0.5	0.1	0	0

Table 3.2.: A matching nucleotide at the primer's 3′-end is in an unbound configuration with probability $p(d, l')$ where d and l' denote the distance to and the size of the preceding error cluster.

perturbation decays quickly with the number of correctly incorporated nucleotides. For $d > 3$, the probability to observe the terminus in a bound state is approximately one regardless of the value of l' [78]. Within our model, the next copying steps after an error cluster are assumed to proceed normally with σ_l and ϵ_l at probability $1 - p(d, l')$. At probability $p(d, l')$ the copying process continues in a non-templated fashion with stalling factor and error probability given by σ_{non} and ϵ_{non}. Probabilities for bound and unbound configurations are then accounted for in a coarse-grained fashion: Effective extension rate $k(l, d, l')$ and error probability $\epsilon(l, d, l')$ are obtained by taking the average over the bound and dangling configuration, i.e.,

$$\begin{aligned} k(l, d, l' \geq 2) &= \left[1 - p(d, l')\right] k_l + p(d, l') \frac{k_0}{\sigma_{\text{non}}}. \\ \epsilon(l, d, l' \geq 2) &= \left[1 - p(d, l')\right] \epsilon_l + p(d, l')\, \epsilon_{\text{non}}. \end{aligned} \tag{3.7}$$

Numerical values for $p(d, l')$ used in the simulation are in line with Ref. [78] and are summarized in Tab. 3.2.

3.4.2. Analysis of fold-reduction in error fraction

The value for the error fraction in the limit $\tau \ll t_{\text{perf}}$ in Fig. 3.3E and was obtained from the simplified analytical model. However, we assume that the analytical and the full stochastic model give similar results in the short time limit. Obtaining a data set suited for statistical analysis for $\tau < 0.5\, t_{\text{perf}}$ from stochastic simulations was not possible since the yield was too poor for long templates.

3.4.3. Estimate for lifetime of DNA and RNA strands

According to Ref. [174] the rate of hydrolysis for a single-stranded RNA oligomer of L bases at a temperature T can be predicted by the following empirical formula:

$$\begin{aligned} k_{\text{degrad}} &= 247.4\,(L-1)k_{\text{bg}}10^{0.983(p\text{H}-6)}10^{-0.24(3.16-[\text{K}^+])} \\ &\quad \times \left[\text{Mg}^{2+}\right]^{0.80}\left[\text{K}^+\right]^{-0.419}10^{0.07(T-23)}, \end{aligned} \tag{3.8}$$

where $k_{\text{bg}} = 1.3 \times 10^{-9}\,\text{min}^{-1}$ is the background rate determined at pH 6, $[\text{K}^+] = 3.16\,\text{M}$ and at 23° C. The temperature T in the last term on the right-hand side of (3.8) is expressed as a multiple of 1° C. To predict the lifetime of DNA and RNA strands, we assume environmental conditions similar to the experimental study performed with RNA polymers of length $L = 60$ in Ref. [28], i.e., $\left[\text{Mg}^{2+}\right] = 0.05\,\text{M}$, $[\text{K}^+] = 0.05\,\text{M}$ and pH 8.3. Moreover, we assume the same profile for the temperature peaks separating copies from templates as in Ref. [28], i.e., $T = 68°$ C for $\tau_{\text{hot}} = 5.56 \times 10^{-4}\,\text{h}$. It is known that the stability of DNA strands against hydrolysis is much higher compared to RNA strands [174]. Therefore, we use the results based on (3.8) as a lower bound for the lifetime of DNA oligomers. For temperatures below 37° C, hydrolysis of DNA strands is hardly measurable [174, 175]. Hence, in the DNA scenario, we assume that hydrolysis only occurs during the temperature peaks. The average degradation rate over one optimal temperature cycle is then given by $\frac{\tau_{\text{hot}}}{\tau^*}k_{\text{degrad}}$. In contrast to DNA strands, RNA strands are also prone to hydrolysis at lower temperatures. To estimate the lifetime of an RNA oligomer, one therefore has to compute an average rate of hydrolysis, taking into account the degradation rate during the cold phase $k^{\text{cold}}_{\text{degrad}}$ and the degradation rate during the temperature peaks k_{degrad}. To obtain the correct average, $k^{\text{cold}}_{\text{degrad}}$ and k_{degrad} have to be weighted with the durations of the cold phase and the hot phase, respectively, i.e., $k^{\text{cold}}_{\text{degrad}} + \frac{\tau_{\text{hot}}}{\tau^*}k_{\text{degrad}}$. (Note that the duration of the cold phase is approximately equal to the duration of the cycle.)

Our estimates for the lifetimes are conservative, since hydrolysis within a double strand is rare compared to hydrolysis of a single strand. The double-strand configuration prevents the attacking 2’-OH from attacking the phosphodiester bond [175]. During the copying process, the template strand is partially protected by the extended primer. Therefore, the actual degradation rate might be smaller.

3.4.4. Analytical solutions for the simplified model

To study the lower limit of the error fraction, we introduced a simplified copying model that is analytically tractable. Here, we analyze the model in detail and derive expressions for its yield and error fraction. The simplified model partially neglects the polymerization history by taking only the last incorporation into account. If the nucleotide at the 3′-end of the (partially) extended primer is mismatched, the copying process is stalled by a constant factor regardless of the number of preceding errors, i.e. $\sigma_{l=1} = \sigma_{l>1} = \sigma$. The error probability also remains constant beyond the first mutation, ie. $\epsilon_{l=1} = \epsilon_{l>1}$. The simplified model further assumes that the unperturbed dynamics are restored immediately after the first correct monomer is integrated, and that the end of the partially extended primer is always bound to the template strands, i.e., $p(d, l') = 0 \;\forall\; d, l'$. Hence, error clusters do not affect the extension dynamics beyond a distance of one.

The master equation [140] describing the dynamics can be obtained by considering the probability $p_{m,n}(t)$ to have polymerized m steps with n mutations, but none of them in the last step, and the corresponding probability $q_{m,n}(t)$ to have one of the n mutations in the last step. With the basic extension rate k_0, the time-evolution of $p_{m,n}(t)$ must satisfy

$$\partial_t p_{m,n} = k_0(1-\epsilon_0)p_{m-1,n} + \frac{k_0}{\sigma}(1-\epsilon_1)q_{m-1,n} - k_0 p_{m,n} \tag{3.9}$$

The terms on the right-hand side account for the correct extension of a strand of length $m-1$ following a match (first term) or a mismatch (second term), and for the extension of a strand of length m with either a match or a mismatch following a match (third term). Similarly, $q_{m,n}(t)$ must satisfy

$$\partial_t q_{m,n} = k_0\epsilon_0 p_{m-1,n-1} + \frac{k_0}{\sigma}\epsilon_1 q_{m-1,n-1} - \frac{k_0}{\sigma} q_{m,n} \tag{3.10}$$

Here, the terms on the right-hand side describe a strand of length $m-1$ that is extended by a mismatch following a match (first term) or following a mismatch (second term), and a strand of length m that is extended with a match or mismatch following a mismatch (third term). Laplace transformation of these equations, such that $\tilde{p}_{m,n}(z) = \mathcal{L}\left\{p_{m,n}(t)\right\}$, leads to

$$z\tilde{p}_{m,n} = k_0(1-\epsilon_0)\tilde{p}_{m-1,n} + \frac{k_0}{\sigma}(1-\epsilon_1)\tilde{q}_{m-1,n} - k_0\tilde{p}_{m,n} \tag{3.11}$$

and

$$z_t\tilde{q}_{m,n} = k_0\epsilon_0\tilde{p}_{m-1,n-1} + \frac{k_0}{\sigma}\epsilon_1\tilde{q}_{m-1,n-1} - \frac{k_0}{\sigma}\tilde{q}_{m,n}. \tag{3.12}$$

Some algebra suffices to show that the solution is given by

$$\tilde{p}_{m,n}(z) = \sum_{i=1}^{n} \binom{m-n}{i}\binom{n-1}{i-1}\left(\frac{\epsilon_0(1-\epsilon_1)}{\epsilon_1(1-\epsilon_0)}\right)^i \frac{k_0^m\epsilon_1^n(1-\epsilon_0)^{m-n}}{(k_0+z)^{m-n}(k_0+\sigma z)^n} \tag{3.13}$$

$$\tilde{q}_{m,n}(z) = \sum_{i=1}^{n} \binom{m-n}{i-1}\binom{n-1}{i-1}\left(\frac{\epsilon_0(1-\epsilon_1)}{\epsilon_1(1-\epsilon_0)}\right)^i \frac{k_0^m \epsilon_1^n (1-\epsilon_0)^{m-n+1}\sigma}{(1-\epsilon_1)(k_0+z)^{m-n}(k_0+\sigma z)^n}. \tag{3.14}$$

Backtransforming, we obtain

$$\begin{aligned} p_{m,n}(t) = &\sum_{i=1}^{n} \binom{m-n}{i}\binom{n-1}{i-1}\left(\frac{\epsilon_0(1-\epsilon_1)}{\epsilon_1(1-\epsilon_0)}\right)^i k_0^m \epsilon_1^n (1-\epsilon_0)^{m-n} \\ &\times \frac{t^{m-1}e^{-k_0 t}}{\sigma^n (m-1)!} {}_1F_1(n, m, k_0 t(1-1/\sigma)). \end{aligned} \tag{3.15}$$

$$\begin{aligned} q_{m,n}(t) = &\sum_{i=1}^{n} \binom{m-n}{i-1}\binom{n-1}{i-1}\left(\frac{\epsilon_0(1-\epsilon_1)}{\epsilon_1(1-\epsilon_0)}\right)^i k_0^m \epsilon_1^n \frac{(1-\epsilon_0)^{m-n+1}}{1-\epsilon_1} \\ &\times \frac{t^{m-1}e^{-k_0 t}}{\sigma^{n-1} (m-1)!} {}_1F_1(n, m, k_0 t(1-1/\sigma)) \end{aligned} \tag{3.16}$$

where ${}_1F_1(a, b, x)$ is the confluent hypergeometric function [181] which can be written as

$${}_1F_1(a, b, x) = \sum_{n=0}^{\infty} \frac{a^{(n)} x^n}{b^{(n)} n!} \tag{3.17}$$

Here, $a^{(n)}$ and $b^{(n)}$ are rising factorials.

The relevant observable now is $\mathcal{P}_n(\tau)$, which denotes the probability to observe a complete polymerization product with n mutations when copying a template of length L after time τ. $\mathcal{P}_n(\tau)$ is given by

$$\begin{aligned} \mathcal{P}_n(\tau) = &\int_0^\tau \mathrm{d}t \left[k_0(1-\epsilon_0)p_{m-1,n}(t) + \frac{k_0}{\sigma}(1-\epsilon_1)q_{m-1,n}(t)\right] \\ &+ \int_0^\tau \mathrm{d}t \left[k_0\epsilon_0 p_{m-1,n-1}(t) + \frac{k_0}{\sigma}\epsilon_1 q_{m-1,n-1}(t)\right]. \end{aligned} \tag{3.18}$$

Introducing $\Phi_{L,n}(\tau)$ as

$$\Phi_{L,n}(\tau) = \int_0^\tau \mathrm{d}t \frac{k_0^L t^{L-1} e^{-k_0 t} {}_1F_1(n, L, k_0 t(1-1/\sigma))}{(L-1)!\sigma^n}, \tag{3.19}$$

and using that

$$\binom{a+1}{b+1} = \binom{a}{b} + \binom{a}{b+1}, \tag{3.20}$$

we can rewrite (3.18) and obtain

$$\begin{aligned}\mathcal{P}_n(\tau) = \epsilon_1^n\,(1-\epsilon_0)^{L-n}\Bigg[&\sum_{i=1}^{n}\binom{L-n}{i}\binom{n-1}{i-1}\left(\frac{\epsilon_0(1-\epsilon_1)}{\epsilon_1(1-\epsilon_0)}\right)^i\Phi_{L,n}(\tau)\\&+\frac{\epsilon_0}{\epsilon_1}\sum_{i=0}^{i-1}\binom{L-n}{i}\binom{n-1}{i}\left(\frac{\epsilon_0(1-\epsilon_1)}{\epsilon_1(1-\epsilon_0)}\right)^i\Phi_{L,n-1}(\tau)\Bigg].\end{aligned}\tag{3.21}$$

With (3.21) error fraction and yield for copies of length L as a function of τ are then given by

$$\begin{aligned}f_e(\tau) &= \frac{1}{Y(\tau)L}\sum_{i=0}^{L} n\,\mathcal{P}_n(\tau),\\ Y(\tau) &= \sum_{i=0}^{L}\mathcal{P}_n(\tau).\end{aligned}\tag{3.22}$$

(3.21) and (3.22) have to be evaluated by numerical integration. In the numerical integration, we use the calibrated Laplace approximation ${}_1\tilde{F}_1$ for the confluent hypergeometric function to insure numerical stability [182]. The alibrated Laplace approximation ${}_1\tilde{F}_1$ for the confluent hypergeometric function is given by

$${}_1\tilde{F}_1(a,b,x) = b^{b-\frac{1}{2}}\left(\frac{y^2}{a}+\frac{(1-y)^2}{b-a}\right)^{-\frac{1}{2}}\left(\frac{y}{a}\right)^a\left(\frac{1-y}{b-a}\right)^{b-a}e^{xy}\tag{3.23}$$

where

$$y = \frac{2a}{b-x\sqrt{(x-b)^2+4ax}}.\tag{3.24}$$

In Fig. 3.6 $f_e(\tau)$ and $Y(\tau)$ are plotted for $\epsilon_0 = 0.08$, $\epsilon_1 = 0.69$ and $\sigma = 25$ and compared to data from a corresponding stochastic simulation.

In the limit $\tau \to 0$, an approximativ but compact analytical expression for the error fraction $f_e(\tau)$ can be derived. In this limit, (3.19) can be approximated as

$$\Phi_{L,n}(\tau) \approx \frac{(\tau k_0)^L}{S^n(L-1)!}.\tag{3.25}$$

To arrive at (3.25) the Taylor expansion of the integrand of (3.19) is truncated at lowest order. From (3.25) we also see that the yield goes to zero in this limit. For $\tau \to 0$ mostly copies containing no or only one mutation contribute to the yield. A lower bound for the error fraction in the short time limit can therefore be obtained as

$$f_e^{\text{low}}(\tau \to 0) \approx \frac{\mathcal{P}_1(\tau)}{[\mathcal{P}_0(\tau) + \mathcal{P}_1(\tau)]\, L}. \tag{3.26}$$

Plugging in (3.25) into (3.21) then transforms (3.26) to

$$f_e^{\text{low}}(\tau \to 0) \approx \frac{\epsilon_0\,(1-\epsilon_0)^{L-2}\,(L-1)(1-\epsilon_1) + \sigma\epsilon_0(1-\epsilon_0)^{L-1}}{\left[\sigma\,(1-\epsilon_0)^{L} + \epsilon_0\,(1-\epsilon_0)^{L-2}\,(L-1)(1-\epsilon_1) + \sigma\epsilon_0(1-\epsilon_0)^{L-1}\right] L} \tag{3.27}$$

or to the more compact form used in the main text

$$f_e^{\text{low}}(\tau \to 0) \approx \frac{1}{L}\frac{\frac{\epsilon_0(L-1)(1-\epsilon_1)}{\sigma(1-\epsilon_0)} + \epsilon_0}{1 + \frac{\epsilon_0(L-1)(1-\epsilon_1)}{\sigma(1-\epsilon_0)}}. \tag{3.28}$$

For the opposite limit $\tau \to \infty$ and for large L we can estimate the error fraction, which turns out to be an upper bound. In this limit, (3.19) reduces to

$$\Phi_{L,n}(\tau \to \infty) \approx 1. \tag{3.29}$$

For long copies, it is justified to neglect the second term on the right-hand side of (3.21) which corresponds to a copying trajectory with a misincorporation in the final step. Using that

$$\binom{a}{b} = \frac{a}{b}\binom{a-1}{b-1} \tag{3.30}$$

(3.21) then becomes

$$\mathcal{P}_n(\tau \to \infty) \approx \epsilon_1^n\,(1-\epsilon_0)^{L-n} \sum_{i=1}^{n} \binom{L-n}{i} \frac{n}{i} \binom{n}{i} \left(\frac{\epsilon_0(1-\epsilon_1)}{\epsilon_1(1-\epsilon_0)}\right)^i \tag{3.31}$$

For $L \gg 1$ this approximate expression for $\mathcal{P}_n$ is sharply peaked. Neglecting the factor $\frac{n}{i}$ therefore only changes the overall scaling, which we will account for by the appropriate normalization later on. Replacing the binomials by their Gaussian approximations in

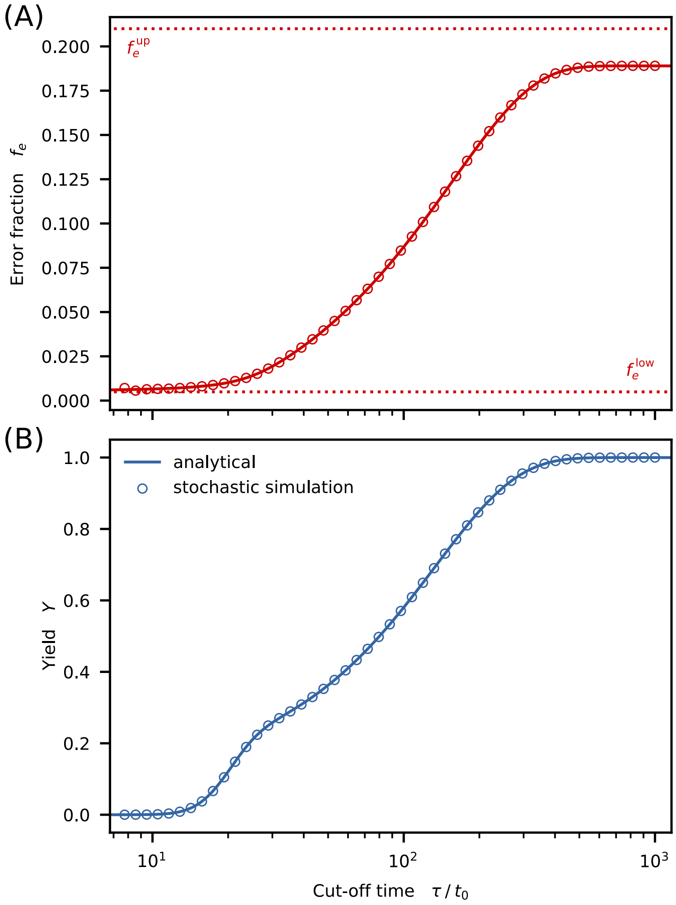


Figure 3.6.: (A) Error fraction $f_e(\tau)$ and (B) yield $Y(\tau)$ as functions of τ for $\epsilon_0 = 0.08$, $\epsilon_1 = 0.69$ and $\sigma = 25$ (see (3.22)). Dotted lines indicate lower and upper bound for the error fraction according to (3.27) and (3.33). Circles: stochastic simulation (one data set).

(3.31) then yields

$$
\begin{aligned}
\mathcal{A}_n &= \int_{-\infty}^{\infty} \mathrm{d}i \frac{\exp\left[-\frac{(i-\epsilon_0(L-n))^2}{2(L-n)\epsilon_0(1-\epsilon_0)}\right]}{\sqrt{2\pi(L-n)\epsilon_0(1-\epsilon_0)}} \frac{\exp\left[-\frac{(i-(1-\epsilon_1)n)^2}{2n\epsilon_1(1-\epsilon_1)}\right]}{\sqrt{2\pi n\epsilon_1(1-\epsilon_1)}} \\
&\approx \frac{\exp\left[-\frac{(n-\frac{\epsilon_0}{1+\epsilon_0-\epsilon_1}L)^2}{2L\left(\frac{\epsilon_0(1-\epsilon_0)}{(1+\epsilon_0-\epsilon_1)^2}-\frac{n}{L}\frac{\epsilon_0(1-\epsilon_0)-\epsilon_1(1-\epsilon_1)}{(1+\epsilon_0-\epsilon_1)^2}\right)}\right]}{\sqrt{2\pi L\left[\epsilon_0(1-\epsilon_0)-\frac{n}{L}\epsilon_0(1-\epsilon_0)-\epsilon_1(1-\epsilon_1)\right]}}
\end{aligned}
\tag{3.32}
$$

If we now replace $\frac{n}{L} \to \hat{\epsilon}$ with some irrelevant but constant value in the denominators of the variance terms, we can give an expression for the error fraction:

$$f_e^{\text{up}}(\tau \to \infty) = \frac{1}{L} \frac{\sum_{n=0}^{L} n\mathcal{A}_n}{\sum_{n=0}^{L} \mathcal{A}_n} \approx \frac{1}{L} \frac{\int_{-\infty}^{\infty} n\mathcal{A}_n}{\int_{-\infty}^{\infty} \mathcal{A}_n} = \frac{\epsilon_0}{1 + \epsilon_1 - \epsilon_1} \tag{3.33}$$

If $L \gg 1$ templates almost have no chance to complete without any error. An initial misincorporation, in turn, is likely to trigger an error cascade. However, if templates are relatively short, there is a fair chance of not having an initial mismatch at all and therefore not running into an error cascade. Hence, we expect (3.33) to be an upper bound for templates of finite length L.

3.5. Analysis of the copying time

In the main text, $t_{\text{cop}}(\tau, L)$ was introduced as the average waiting time for the first error-free copy to occur as a function of the cycle duration τ and the template length L. The differences in the dynamics between the simplified model (see Section 3.4.4) and the full model only become apparent after the first mismatch got incorporated. For an error-free copying process leading to a full-length product, the dynamics are identical in both models. Hence, we can build on the analytic results obtained in Section 3.4.4 to derive a formula for $t_{\text{cop}}(\tau, L)$.

According to (3.21) the probability $\mathcal{P}_0(\tau)$ to observe a complete polymerization product without mutations after time τ is given by

$$\mathcal{P}_0(\tau) = (1 - \epsilon_0)^L \, \Phi_{L,0}(\tau). \tag{3.34}$$

with

$$\Phi_{L,0}(\tau) = \int_0^\tau \mathrm{d}t \, \frac{t^{L-1} e^{-t} {}_1F_1(0, L, t(1 - 1/\sigma))}{(L-1)!}. \tag{3.35}$$

Note that in (3.35) all timescale are expressed in units of the basal extension timescale $t_0 = 1\,\text{h}$ as in the main text. We will use this convention throughout this section. Using that the confluent hypergeometric function is identical to one if the first argument is zero [181], i.e., $F_1(0, L, t(1 - 1/\sigma)) \equiv 1$, we obtain

$$\mathcal{P}_0(\tau) = \frac{(1 - \epsilon_0)^L}{(L-1)!} \int_0^\tau \mathrm{d}t \, t^{L-1} e^{-t} = \frac{(1 - \epsilon_0)^L}{\Gamma(L)} \left[\Gamma(L) - \Gamma(L, \tau)\right], \tag{3.36}$$

where $\Gamma(L)$ and $\Gamma(L, \tau)$ are the gamma and upper incomplete gamma function.

In the cycling scenario τ corresponds to the cycle duration. If we assume, that the average association time $\langle t_a \rangle = (k_{\text{on}} \times c_{\text{prim}})^{-1}$ is short in comparison to the cycle duration τ, $\mathcal{P}_0(\tau)$ represents the probability to obtain an error-free product within one cycle. Then, the average number of cycles, one has to wait until the first error-free full copy is observed, is $1/\mathcal{P}_0(\tau)$ and the copying time $t_{\mathrm{cop}}(\tau, L)$ is given as

$$t_{\mathrm{cop}}(\tau, L) = \tau \frac{1}{\mathcal{P}_0(\tau)} = \frac{\Gamma(L)}{[\Gamma(L) - \Gamma(L,\tau)]\,(1-\epsilon_0)^L}. \tag{3.37}$$

The situation is more complex if the primer concentration is small such that the association time t_{a} is not negligible anymore. In this case, the effective time window for the copying process within one cycle is given by $\tau - t_{\mathrm{a}}$. t_{a} in turn is distributed exponentially. The probability of observing an error-free product at the end of the cycle, therefore, takes the form

$$\mathcal{P}_0^{\mathrm{eff}}(\tau) = \int_0^\tau \mathrm{d}t_{\mathrm{a}}\, \mathcal{P}_0(\tau - t_{\mathrm{a}}) \exp\left(\frac{-t_{\mathrm{a}}}{k_{\mathrm{on}}\, c_{\mathrm{prim}}}\right), \tag{3.38}$$

where the exponential on the right-hand side corresponds to the association time distribution. Carrying out the integral, taking the inverse, and multiplying with the cycle duration τ then leads to

$$t_{\mathrm{cop}}(\tau, L) = \frac{\frac{\Gamma(L)}{(1-\epsilon_0)^L}\left(\frac{a-1}{a}\right)^L e^{\frac{\tau}{a}}}{\left(\frac{-1+a}{a}\right)^L \left(\tau^L E\left(1-L, \frac{(-1+a)\tau}{a}\right) + e^{\frac{\tau}{a}}\right) [\Gamma(L) - \Gamma(L,\tau)]\Big) - \Gamma(L)} \tag{3.39}$$

where we have used the abbreviation $a = k_{\mathrm{on}}\, c_{\mathrm{prim}}$ and where $E(x, y)$ is the exponential integral function which is defined as

$$E(x, y) = \int_1^\infty \mathrm{d}t \frac{e^{-yt}}{t^x}. \tag{3.40}$$

In Fig. 3.7 $t_{\mathrm{cop}}(\tau, L)$ according to (3.39) is plotted for $\epsilon_0 = 0.08$ and $c_{\mathrm{prim}} = 10^{-4}\,\mathrm{nM}$ for different values of L (straight lines) together with data points obtained from a corresponding stochastic simulation (circles). All curves show a distinct minimum. In general, the exact position of the minimum has to be determined numerically.

However, for large primer concentrations, such that the association time becomes negligible, we can derive approximative formulas for the position τ^* and the value t^*_{cop} of the minimum of the copying time according to (3.37). From $\frac{\mathrm{d}}{\mathrm{d}\tau} t_{\mathrm{cop}}(\tau) = 0$, we obtain

$$\Gamma(L) - \Gamma(L, \tau) = \gamma(L, \tau) = \tau^L e^{-\tau}, \tag{3.41}$$

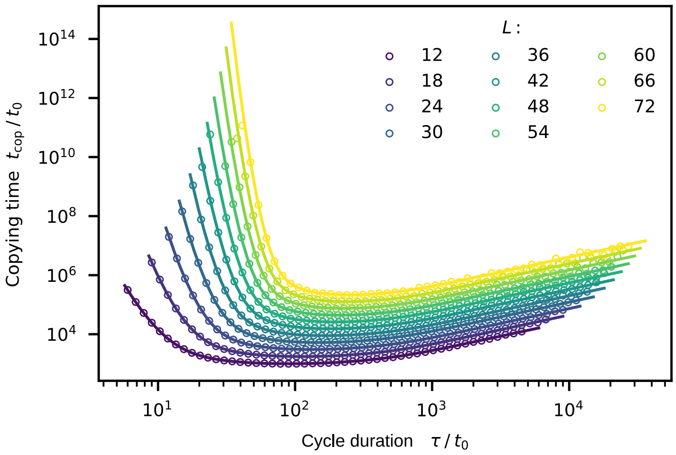


Figure 3.7.: Copying time t_{cop} as a function of the cycle duration τ for $c_{\text{prim}} = 10^{-4}$ nM. t_{cop} is the average time that passes until the first error-free copy of the template strand is produced. All curves show a distinct minimum. As the length L increases, the minimum moves to the right. Straight lines are obtained from (3.39), circles are obtained from simulation (one data set).

where $\gamma(L,\tau)$ is the lower incomplete gamma function defined as

$$\gamma(L,\tau) = \int_0^\tau t^{L-1} e^{-t} \mathrm{d}t. \tag{3.42}$$

To be able to solve for τ, we first need an approximate expression for the value of $\gamma(L,\tau)$. The integrand $I(L,t)$ in (3.42) can be written as:

$$I(L,t) = t^{L-1} e^{-t} = \exp\left\{\ln(t)(L-1) - t\right\} = \exp\left\{f(L,t)\right\}. \tag{3.43}$$

The value of t maximizing $f(L,t)$ also maximizes $I(L,t)$. From $\frac{\mathrm{d}}{\mathrm{d}t}(\ln(t)(L-1) - t) = 0$ this value is determined to be $t = L-1$. If we assume an upper integration boundary $\tau > L$ on the right-hand side of (3.42) we can perform a saddle point approximation to get an estimate for $\gamma(L,\tau)$. The assumption that $\tau > L$ is based on the observation that the yield per cycle drops quickly for $\tau < L$ (see Fig. 3.4C). According to the saddle point approximation, $\gamma(L,\tau)$ can be estimated as

$$\begin{aligned} \gamma(L,\tau) &\approx \exp\left\{f(L,L-1)\right\} \sqrt{\frac{2\pi}{f''(L,L-1)}} \\ &= \exp\left\{\ln(L-1)(L-1) - (L-1) + \frac{1}{2}\ln(L-1) + \frac{1}{2}\ln(2\pi)\right\}. \end{aligned} \tag{3.44}$$

For L sufficiently large, we can neglect the last two terms in the exponential on the right-hand side and further replace $(L-1)$ by L. Plugging this simplified result into (3.41) we obtain an equation determining the minimum position:

$$\exp\left\{\ln(L)(L) - L\right\} = \exp\left\{\ln(\tau)(L) - \tau\right\}. \tag{3.45}$$

The position of the minimum is roughly given by $\tau^* \approx L$. In fact, $\tau^* \gtrsim L$ since we neglected terms increasing τ^* (consistent with the saddle point approximation). Plugging this result into (3.37) we obtain an approximative formula for the minimal copying time t^*_{cop}:

$$t^*_{\text{cop}}(L) \approx \frac{L\,\Gamma(L)}{\gamma(L, L+\delta L)}(1-\epsilon_0)^{-L}. \tag{3.46}$$

We wrote $L + \delta L$ instead of L for the second argument of the lower incomplete gamma function in the denominator to emphasize that the upper integration boundary is actually slightly larger than L. Using Stirling's approximation, i.e., $\Gamma(L) = \sqrt{2\pi}L^L e^{-L}$ and the saddle point estimate for $\gamma(L, L+\delta L)$ for L sufficiently large, we end up with simple expression for the minimal copying time as a function of L, i.e.,

$$t^*_{\text{cop}}(L) \sim L(1-\epsilon_0)^{-L}. \tag{3.47}$$

Hence, the scaling of $t^*_{\text{cop}}(L)$ is dominated by the exponential factor $\exp\left\{-\ln(1-\epsilon_0)\right\}$. In a plot with a logarithmic y-axis (decadic logarithm) this corresponds to a slope of ~ 0.036 for $\epsilon_0 = 0.08$. This result fits the slope observed in Fig. 3.5B.

4. Simulating large oligonucleotide reaction networks[*]

In Chapter 5 and Chapter 6 we are investigating self-assembly scenarios emerging from large pools containing strands of various lengths. In contrast to the primer extension scenario discussed in Chapter 3, strands do not have distinct roles anymore. In these mixed scenarios all strands simultaneously serve as a template, primer, and substrate for the extension. Within such a pool, arbitrary complexes containing multiple strands assemble and disassemble continuously by hybridization and dehybridization. Some of these complexes may give rise to the ligation of two strands. In order for ligation to occur, the two strands have to be hybridized adjacently on a third strand. Moreover, complexes can enter or exit the system, and single-stranded segments may break by hydrolysis. For simplicity, we either assume a system composed of one self-complementary nucleotide (Chapter 5) or two complementary nucleotides denoted by X and Y for the sake of generality (Chapter 6). The underlying simulation method is the same for both the self-complementary and the binary system. This section aims to develop this simulation method in detail for future reference.

4.1. State space of occupation numbers

To formally describe the current state of the system, we have to introduce some nomenclature. A molecule containing L nucleotides linked covalently is called a *strand* of length L. A single nucleotide is a strand with length $L = 1$. Moreover, strands are assumed to be rigid and therefore cannot fold onto themselves. In addition, every strand is directed. An entity formed by several hybridized strands is referred to as a *complex*. In principle, all staggered conformations that can arise from the present strands are allowed regardless of the number of strands and potential *mismatches*, i.e., non-complementary nucleotide pairs. However, branched hybridization structures

[*] The chapter is adapted from the Supplementary Material (SM) belonging to the article: "Self-Assembly of Informational Polymers by Templated Ligation", by Joachim H. Rosenberger, Tobias Göppel, Patrick W. Kudella, Dieter Braun, Ulrich Gerland, and Bernhard Altaner published in Phys. Rev. X 11, 031055 (2020), under CC BY 4.0 License. See Ref. [183]. The author of this thesis is a shared first author of the paper.

and other non-linear complexes involving loops are excluded as implied by the rigid strand assumption. A single strand is also called a complex. Moreover, we say that two identical complexes belong to the same *species*. The number of identical complexes is called the *occupation number* or *copy number* of the species. Furthermore, we call a complex that contains two or three strands a *duplex* or a *triplex*, respectively. The overlapping horizontal region between two strands is referred to as a *hybridization site*, while the vertical interface between two strands hybridized adjacently on a third strand is called a *ligation site*.

If we assume a well-mixed scenario, the current state n of the system is solely described by the present species $S_i \in \mathcal{S}$ and their occupation numbers $N_i \in \mathbb{N}$. Here, $\mathcal{S}$ is the (countable) set of all possible species. In the case of a closed system without in- and outflux, the set of possible species is limited by the total number of nucleotides present in the system. The current state n can be written as a row vector where the i-th element corresponds to the copy number of the species $S_i \in \mathcal{S}$, i.e.,

$$n = (N_1, N_2, N_3, ...). \tag{4.1}$$

If a certain species is not present, its occupation number is zero (this is the case for most species, which are possible in principle.). The total number of complexes which are present in state n is trivially given by

$$N_n^{\text{tot}} = \sum_i N_i. \tag{4.2}$$

Moreover, we also define the set of all the species that are present in the current state n, i.e., that do have copy numbers different from zero as

$$\mathcal{S}_n = \{S_i : S_i \in \mathcal{S} \text{ and } N_i > 0\}. \tag{4.3}$$

With that, the total number of species in the current state n can be written as

$$S_n^{\text{tot}} = |\mathcal{S}_n|. \tag{4.4}$$

In a closed system, the number of possible states grows rapidly with the total number of nucleotides present in the system. In an open system the number of possible states is infinite. All possible states of the system n span the state space $\mathcal{N}$.

Every reaction, including the in- and outflux of complexes, increases or decreases the occupation number of one or two species and thus changes the state of system. We denote the set of all reactions which are possible in a given state n by $\mathcal{M}_n$. The number of possible reactions in the given state n is written as $M_n = |\mathcal{M}_n|$. Every reaction $\mu \in \mathcal{M}_n$ belongs to one of the allowed reaction types which are hybridization, dehybridization, ligation, cleavage (hydrolysis), influx, and outflux. While hybridization is a bimolecular reaction, all other reactions are monomolecular.

Moreover, we can uniquely map every reaction $\mu \in \mathcal{M}_n$ onto a tuple that specifies the reaction, i.e., the type of the reaction, the involved species, and the channel via which the reaction will potentially proceed. For a hybridization this tuple is given by (on, i, j, c). The first entry "on" specifies the type of the reaction. The second and the third entry i and j specify the two species S_i and S_j, which are involved in the reaction. In general, two species S_i and S_j can hybridize in multiple ways via different reaction channels c. The last entry in the tuple therefore specifies the reaction channel c. For a ligation the tuple is given by (lig, i, c). The first entry "lig" tells us that the reaction is of the ligation type. The second entry determines the species S_i of the complex in which the ligation can occur. In principle, several ligations are possible in a complex containing multiple strands and multiple ligation sites. Therefore, the third entry is used to specify the reaction channel c. In the same way the tuple (cut, i, c) specifies a cleavage reaction. Here the reaction channel c defines the position where a complex belonging to species S_i will potentially break. For in- and outflux the tuples only have two entries and are given by (in, i) and (out, i). In this case only the in- or outgoing species S_i has to be specified in the second entry.

To each reaction $\mu \in \mathcal{M}_n$ we can assign an elementary rate r_μ, which is also called reaction parameter (see [184]). The elementary rate r_μ is the rate at which an *individual* complex (monomolecular reactions) or a *pair of individual* complexes (bimolecular reactions) belonging to the species specified in the corresponding tuple react according to other parameters given in that tuple. All elementary rates have the unit of the inverse reference time t_0^{-1}. Using the mapping of μ onto tuples specifying the reactions, the elementary rates r_μ can also be written as $r_{\text{on}}(i, j, c)$, $r_{\text{off}}(i, c)$, $r_{\text{lig}}(i, c)$ and $r_{\text{out}}(i)$.

The reactions described above connect the different states in the state space $\mathcal{N}$. Moreover, they define the transition rates of a *Markov chain*.

In general, if the state $n' \in \mathcal{N}$ can be reached from the state $n \in \mathcal{N}$ via an elementary reaction, the corresponding transition rate is positive, i.e., $w(n \to n') > 0$; otherwise, it is zero, i.e., $w(n \to n') = 0$. In our case, if the system can transition from state $n \in \mathcal{N}$ to state $n' \in \mathcal{N}$ via the elementary reaction μ, we can write the corresponding transition rate as

$$w(n \to n') = h_\mu r_\mu, \tag{4.5}$$

where h_μ is a combinatorial factor accounting for the number of identical complexes or the number of identical pairs of complexes in which the reaction μ can occur. To simplify the notation we introduce the total rate r_μ^{tot} of the reaction μ as

$$r_\mu^{\text{tot}} = h_\mu r_\mu. \tag{4.6}$$

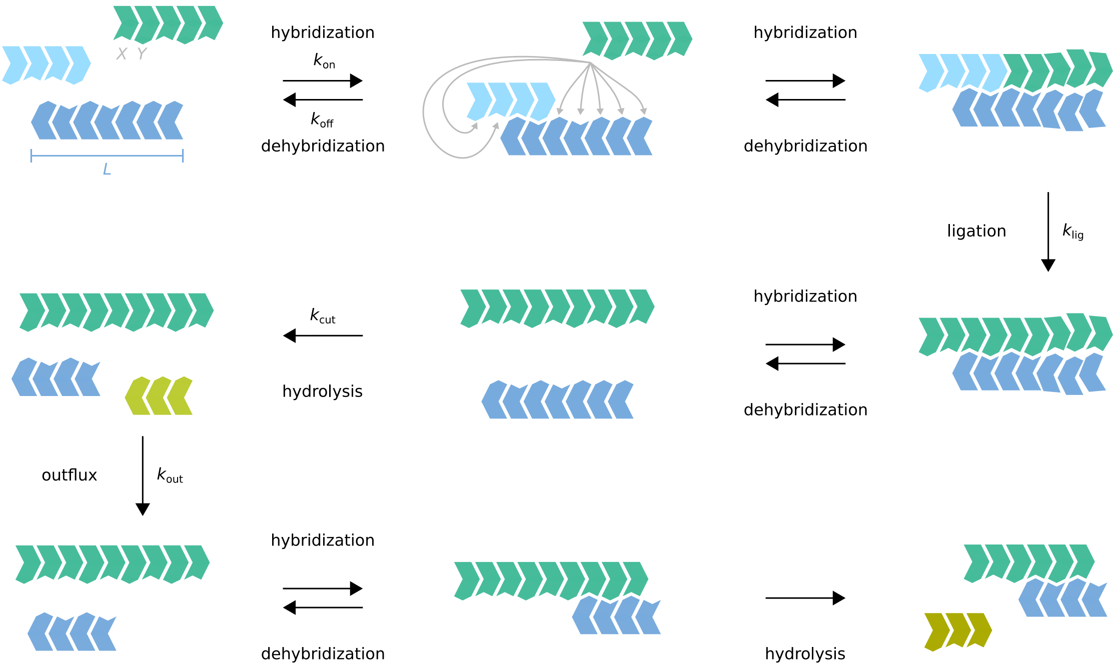


Figure 4.1.: Schematic illustration of the self-assembly emerging from pools of strands with various lengths. The elementary processes are hybridization, dehybridization, ligation on the template, cleavage via hydrolysis and the loss of complexes due to an outflux. In Chapter 6 we assume a binary system comprising two types of nucleotides denoted by X and Y and symbolized by different shapes. In Chapter 5 we are working in a simplified scenario where only one self complementary nucleotide exists. The color code used in this figure and the following figures is as follows: nucleotides that are linked covalently have the same color,i.e., every strand within one complex has its own color. Two strands with different colors adjacently hybridized on a third strand may get joined covalently by a ligation. Note, different colors do not stand for different sequences or strand lengths. Identical strands may have different colors.

With that, Eq. (4.5) becomes

$$w(n \to n') = r_{\mu}^{\mathrm{tot}}. \tag{4.7}$$

We will derive the combinatorial factors in detail later on. The combinatorial factors also become important when we link the elementary rates r_{μ} to the corresponding chemical rate constants k_{μ}, which one would use in a rate-equation approach based on the law of mass action.

The time evolution of the Markov chain can be simulated stochastically by means of *Gillespie algorithm* [169]. In the next section, we, therefore, present the basic ideas of the Gillespie algorithm. In the next but one section, we will describe the relation between the chemical rate constants that would be used in a rate-equation approach and the elementary rates appearing in our Markovian description of the system.

4.2. Foundations of the Gillespie algorithm

In order to compute the time evolution on the Markov chain stochastically, we employ the well-known Gillespie algorithm [169]. Given that the Markov chain is in state $n \in \mathcal{N}$ at time t, the transition rates $w(n \to n')$ for all possible transitions to all state $n' \in \mathcal{N}$ connected to the current state n via an elementary reaction are needed. The probability $p(n \to n^*)$ that the transition $n \to n^*$ is the next transition that will occur is given by the ratio of $w(n \to n^*)$ and the sum over all possible transition rates $\sum_{n'} w(n \to n')$, i.e.,

$$p(n \to n^*) = \frac{w(n \to n^*)}{\sum_{n'} w(n \to n')}. \tag{4.8}$$

If we assume that the transition $n \to n^*$ proceeds via the reaction μ, we can also write $p(\mu)$ instead of $p(n \to n^*)$. Introducing the total transition rate

$$r^{\text{tot}} = \sum_{\alpha=1}^{M_n} r_\alpha^{\text{tot}}, \tag{4.9}$$

the probability $p(\mu)$ for the next reaction being reaction μ then reads

$$p(\mu) = \frac{r_\mu^{\text{tot}}}{r^{\text{tot}}}. \tag{4.10}$$

To "choose" the next reaction on a computer, a uniform random number u_1 lying in the interval $[0, r^{\text{tot}}[$ is generated. The reaction μ that first satisfies the equation

$$\sum_{\alpha=1}^{\mu} r_\alpha^{\text{tot}} > u_1 \times \sum_{\alpha=1}^{M_n} r_\alpha^{\text{tot}} \tag{4.11}$$

is the reaction which is selected. Carried out in a naive way by forming the cumulative sum, the computational complexity of the selection process scales linearly with the number of possible reactions, i.e., $\mathcal{O}(M_n)$. However, the scaling of the CCof the selection process can be improved significantly by using a binary search tree such that a computational complexity of $\mathcal{O}(\ln(M_n)$ is achieved [185].

The time τ one has to wait until the next transition to any state $n' \in \mathcal{N}$ occurs is distributed exponentially and depends on the total transition rate r^{tot}. The probability $p(\tau)$ that the next transition will happen within the time interval $[\tau, \tau + d\tau[$ is given by

$$p(\tau) = r^{\text{tot}} e^{-r^{\text{tot}}\tau} d\tau. \tag{4.12}$$

Exponentially distributed waiting times can be generated easily from uniformly distributed random variables u_2 lying in the interval $[0, r^{\text{tot}}[$. Using the inverse transform sampling theorem we obtain

$$\tau = -\frac{\log\{u_2\}}{r^{\text{tot}}}. \tag{4.13}$$

Whenever a reaction is chosen, the Markov chain transitions from state n to the corresponding new state n^* and the system time is updated $t \to t + \tau$.

After the system transitioned to a new state n', several reactions that were possible in the previous state n might not be allowed anymore since the species involved in these reactions might have disappeared. All the reactions that are obsolete in the new state have to be identified and deleted from the computer memory. In the new state, some of the copy numbers associated with different species have changed and new species might be present. As a result, new reactions are possible. These new reactions have to be identified, created, and stored in the computer memory to update the set of possible reactions, i.e.,

$$\mathcal{M}_n \to \mathcal{M}_{n'}. \tag{4.14}$$

The computational complexity of this crucial update step may become very large if many different species with low occupation numbers are present and limit the performance of the algorithm. In Section 4.4 we present our solution to this complexity problem.

4.3. Relation between chemical rate constants and elementary rates

For the purpose of illustration, we first consider a system where the only allowed reaction types are hybridization and dehybridization. All other reactions types, i.e., ligation, cleavage, influx, and outflux, are excluded. Moreover, we assume that the system is not subjected to any external driving.

In this reduced system, complexes belonging to the species S_i and S_j can hybridize via the reaction channel c to form a complex of species S_k. The corresponding chemical rate constant reads $k_{\text{on}}(i, j, c)$. Complexes belonging to species S_k can disassemble again via the reaction channel c' to form two complexes belonging to the species S_i and S_j. For the inverse process, the corresponding chemical rate constant reads $k_{\text{off}}(k, c')$. The ratio of the rate constants $k_{\text{off}}(k, c')$ and $k_{\text{on}}(i, j, c)$ is related to the change in Gibbs free energy that is associated with the reaction via the equation [186]

$$\ln\left\{\frac{k_{\text{off}}(k, c')}{c^\circ k_{\text{on}}(i, j, c)}\right\} = \beta \Delta G^\circ_{\text{hyb}}(i, j, c), \tag{4.15}$$

where $c^\circ = 1\,\mathrm{mol/l}$ is the standard concentration. Furthermore, once the system reaches a chemical equilibrium, the concentrations of the different species that assemble and disassemble continuously by hybridization and dehybridization are solely determined by the initial concentrations and values, if the Gibbs free energies associated with the reactions $\Delta G^\circ_{\mathrm{hyb}}(i,j,c)$. Thermodynamic consistency now requires that we choose the elementary rates such that the Markovian approach would lead to the same chemical equilibrium as a chemical rate equation approach using the above chemical rate constants $k_{\mathrm{on}}(i,j,c)$ and $k_{\mathrm{off}}(k,c')$ would do. For that reason, we now seek to relate the chemical rate constants k_μ to the elementary rates r_μ. Here we have again used the index μ to simplify the notion. Recall that every index μ can be mapped onto a tuple precisely specifying the reaction. In the following, we will switch between these two notation wherever it is appropriate. To this end, we have to derive the combinatorial factors h_μ that appeared in Section 4.1. In our derivation we closely follow the original work by Gillespie [184]. For illustration purposes we also determine the combinatorial factors for the other reaction types that are allowed in Chapter 5.

To gain some intuition let us first look at the small chemical reaction network sketched in Fig. 4.2. Moreover, let us assume that the molecules in this network are not RNA or DNA oligomers but "objects" that do have a simple structure such that there is only one reaction channel for two interacting molecules. In this simplified scenario, the combinatorial factors are simply obtained by counting the number of individual molecules belonging to one species or by counting the number of distinct pairs formed by individual molecules belonging to one or two different species. For bimolecular reactions involving two different molecules belonging to two different species $S_i \neq S_j$, the combinatorial factor is given by

$$h_\mu = N_i N_j, \tag{4.16}$$

while for reactions involving two molecules belonging to the same species S_i the combinatorial factor is given by

$$h_\mu = N_i(N_i - 1)/2. \tag{4.17}$$

For monomolecular reactions of a molecules of species S_i the combinatorial factor is

$$h_\mu = N_i. \tag{4.18}$$

Using Eq. (4.17) the total elementary rate for a bimolecular reaction involving two molecules of the same species μ defined in Eq. (4.6) in Section 4.1 becomes

$$r_\mu^{\text{tot}} = \frac{N_i(N_i - 1)}{2} r_\mu. \tag{4.19}$$

If the number of molecules N_i is large, we approximately write $r_\mu^{\text{tot}} \approx N_i^2 r_\mu / 2$. Let us now consider an ensemble of systems, all prepared in the same state. In this ensemble, the quantity $\delta t \langle N_i^2 \rangle r_\mu / 2$ corresponds to the average occurrence of the reaction μ in the time interval δt. Neglecting fluctuations we can write

$$\left\langle N_i^2 \right\rangle = \left\langle (\langle N_i \rangle + \delta N_i)^2 \right\rangle = \langle N_i \rangle^2 + \left\langle \delta N_i^2 \right\rangle \approx \langle N_i \rangle^2 . \tag{4.20}$$

Let us assume that the product of the bimolecular reaction is a complex belonging to the species S_k and that complexes of that species S_k can not be produced by any other reaction. Then, the average change of the occupation number of species S_k can be described by a simple differential equation that reads

$$\langle \dot{N_k} \rangle = \frac{r_\mu}{2} \left\langle N_i^2 \right\rangle \approx \frac{r_\mu}{2} \langle N_i \rangle^2 . \tag{4.21}$$

We now rewrite the above equation using molder concentrations, i.e.,

$$c_i = \langle N_i \rangle / (N_A V), \tag{4.22}$$

where $N_A = 6.022 \times 10^{23} \text{mol}^{-1}$ is the Avogadro constant to obtain the chemical rate equation

$$\dot{c_k} = \frac{r_\mu}{2} V N_A c_i^2. \tag{4.23}$$

In this chemical rate equation, the expression in front of the concentration term on the right-hand side corresponds to the chemical rate constant. Hence, we can read off the relation

$$k_\mu = \frac{r_\mu}{2} V N_A \tag{4.24}$$

for the chemical rate constant k_μ and the elementary rate, i.e., reaction parameter r_μ for bimolecular reactions where the reaction molecules belong to the same species.

An analogous derivation for bimolecular reactions μ' involving two molecules belonging to different species $S_i \neq S_j$ yields

$$k_{\mu'} = V N_A r_{\mu'}, \tag{4.25}$$

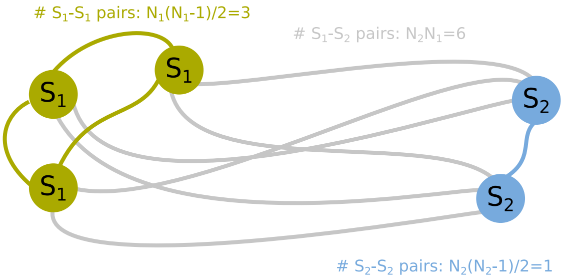


Figure 4.2.: Combinatorial factors for the reaction of two simple chemical species: The number of possible reacting pairs within a species S_i is given by $h_{(on,i,i)} = N_i(N_i - 1)/2$. For reactions involving two different species $S_i \neq S_j$ the combinatorial factors are given equal to $h_{(on,i,j)} = N_i N_j$. Green lines: Possible reactions within molecules belonging to species S_1. Blue lines: Possible reactions within molecules belonging to species S_2. Grey lines: Possible reactions involving molecules belonging to different species S_i and S_j with $S_i \neq S_j$.

while for monomolecular reactions the relation simply becomes

$$k_\mu = r_\mu. \tag{4.26}$$

However, the systems we are interested in are composed of molecules that are not "simple". Our "molecules" are single strands and complexes composed of multiple strands. In general, multiple reaction channels are possible between two interacting complexes. Some of these reaction channels may lead to identical reaction products. For that reason, the relations presented above have to be modified. We will derive these modifications in the following.

4.3.1. Hybridization

Usually, two complexes can hybridize via multiple reaction channels. Some of these reaction channels can lead to the same product complex. This is the case if one complex or both complexes are symmetric under a 180° rotation in the plane (see Fig. 4.3(a)). We denote the number of rotationally symmetric complexes involved in the reaction by $m_{\text{on}} \in \{0, 1, 2\}$. With that, the number of reaction channels leading to the same final complex is then given by $2^{m_{\text{on}}}$. Hence, the combinatorial factors h_μ for bimolecular reactions stated above have to be replaced by $h_\mu \to 2^{m_{\text{on}}} h_\mu$ (see Fig. 4.3(b) for and example with $m_{\text{on}} = 1$). As a result, the relation between the elementary rate and the chemical rate constant for a hybridization μ correspondoing to the tuple (on,i,j,c) is given by

$$k_{\text{on}}(i, j, c) = 2^{m_{\text{on}} - \delta_{ij}} r_{\text{on}}(i, j, c) V N_A. \tag{4.27}$$

The possible values that the prefactor $2^{m_{\text{on}} - \delta_{ij}}$ can take are summarized below:

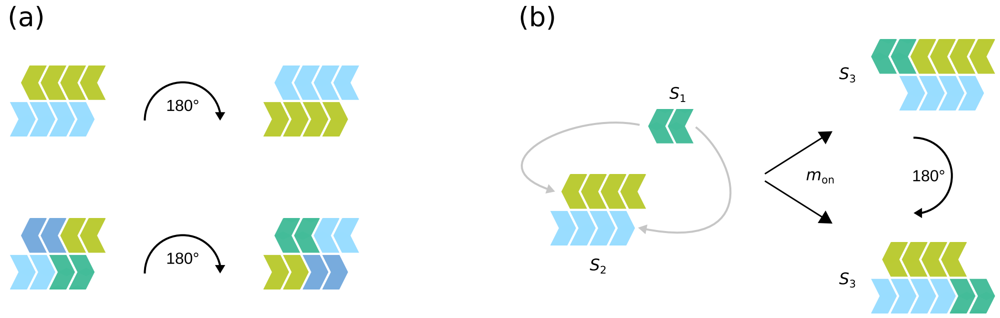


Figure 4.3.: For simplicity, we show strands and complexes that only contain one type of nucleotide which is self-complementary in this figure and in the following figures. However, all the states made also apply to the binary system. Recall the color code: nucleotides that are linked covalently have the same color ,i.e., every strand within one complex has a distinct color. Different colors do not correspond to different sequences or strand properties. Identical strands with the same length and the same sequence may have different colors. (a) Two examples for complexes that are symmetric under a 180° rotation in the plane. (b) The double stranded complex belonging to species S_1 is rotationally symmetric. Both hybridization channels lead to the same product complex belonging to species S_3.

	$m_{\text{on}} = 0$	$m_{\text{on}} = 1$	$m_{\text{on}} = 2$
$S_i = S_j$	1/2	1	2
$S_i \neq S_j$	1	2	4

4.3.2. Dehybridization

A nongeneric situation arises for rotationally symmetric complexes comprising six strands or more (see Fig. 4.4(a)). For every hybridization channel in such complexes, another hybridization channel exists, that leads to exact the same product complexes. The only exception is the channel corresponding to the hybridization site in the geometric center of the complex. The combinatorial factor for the dehybridization reaction disassembling the hybridization site in the geometric center simply is $h_\mu = N_i$. For all other dehybridizations the combinatorial factor is given by $h_\mu = 2N_i$. Generally, the relation between the elementary rate r_μ and the rate constant k_μ can be written as

$$k_{\text{off}}(i,c) = 2^{m_{\text{off}}(1-\delta_{jk})} r_{\text{off}}(i,c), \tag{4.28}$$

Moreover, the indices j and k of the Kronecker symbol correspond to the indexes labeling the species the product complexes belong to. For the dehybridization that disassembles the hybridization site in the geometric center we have $j = k$.

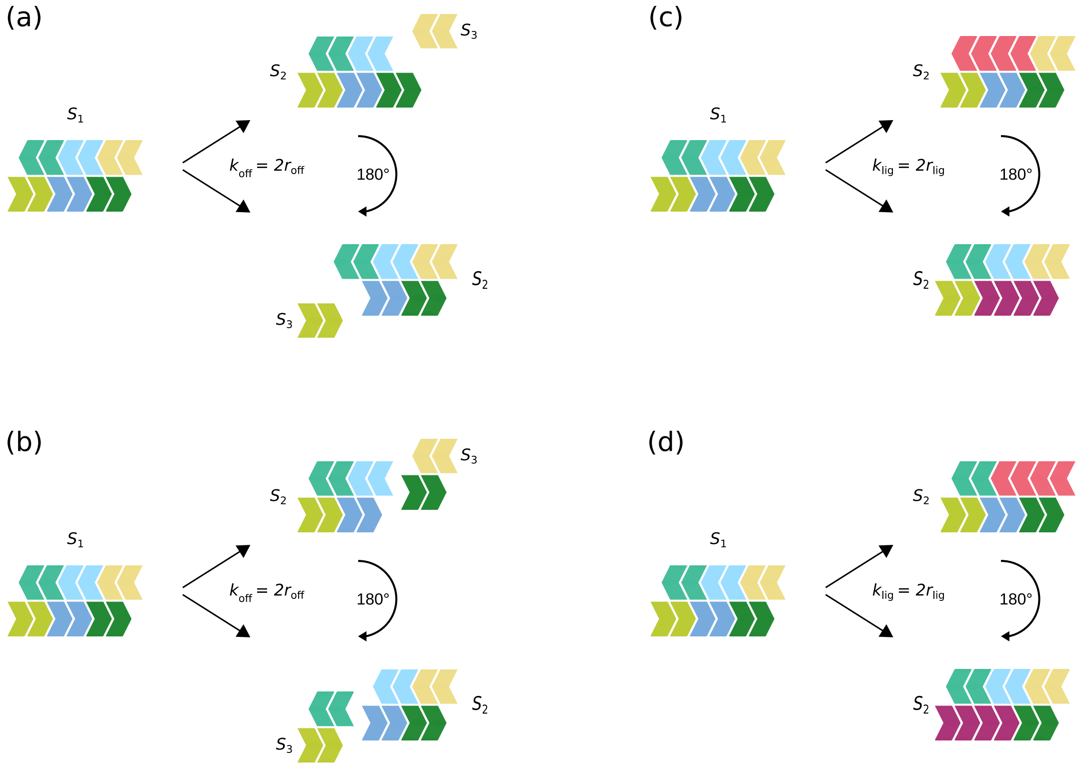

(a)
S3
S2
S1
$k_{\text{off}} = 2r_{\text{off}}$
180°
S2
S3
(c)
S2
S1
$k_{\text{lig}} = 2r_{\text{lig}}$
180°
S2
(b)
S3
S2
S1
$k_{\text{off}} = 2r_{\text{off}}$
180°
S2
S3
(d)
S2
S1
$k_{\text{lig}} = 2r_{\text{lig}}$
180°
S2

Figure 4.4.: (a) and (b) Examples for different elementary dehybridization reactions within one initial complex that lead to identical product complexes. This is the case for all dehybridizations of a rotationally symmetric complex consisting of more than four strands, $n \geq 4$, except the center dehybridization. (c) and (d) Examples for different elementary dehybridization reactions within one initial complex that lead to identical product complexes. (a)-(d) Note that the initial complex is rotationally symmetric.

4.3.3. Ligation

The reasoning to obtain the combinatorial factors for a ligation reaction μ corresponding to the tuple (lig, i, c) is analogus to the derivation of the combinatorial factors for dehybridization that we have just seen. The relation between the elementary rate and the rate constant can be written as

$$k_{\text{lig}}(i,c) = 2^{m_{\text{lig}}} r_{\text{lig}}(i,c), \tag{4.29}$$

where $m_{\text{lig}} = 1$ if the initial complex shows a rotational symmetry and $m_{\text{lig}} = 0$ otherwise, see Fig. 4.4(b).

4.3.4. Outflux

In Chapter 5 we assume a constant outflux rate which does not depend on the structure of considered complexes. Therefore, symmetry considerations do not play any role, and we simply have

$$k_{\text{out}}(i, c) = r_{\text{out}}(i, c). \tag{4.30}$$

4.4. Breaking down the hybridization process

As stated earlier, every reaction corresponds to a transition from the current state n to a new state n'. In particular, this is true for a hybridization reaction in which two individual complexes disappear, and one new larger complex is created. We will now have a closer look at this reaction.

In Fig. 4.5 a small system composed of complexes belonging to three different species is sketched. The lines connecting the three species represent possible hybridization reactions between the species. If the copy number of a species is larger than one, complexes belonging to this species can also interact with each other. However, these intra-species reactions are not shown to avoid an overloading of the figure. One of the reactions is highlighted. We assume that the highlighted reaction is the reaction that will occur next. After the reaction has taken place, some of the reactions that were possible in the old state n are no longer allowed in the new state n'. As a consequence, some of the connecting lines have disappeared. If the system would not only contain three distinct species but many different ones, there would be a huge number of outgoing lines for every species. As already mentioned, deleting the obsolete lines and "drawing" the new ones would be an operation with a large computational complexity.

One way to decrease the computational complexity is to separate the transport process that brings the complexes into proximity such that they can react in principle from the subsequent specific interaction process. The picture that we have in mind is the following: In order to react to a new complex, the two initial complexes belonging to species $S_i, S_j \in \mathcal{S}$ first have to collide. The elementary rate at which a collision between two complexes belonging to the species S_i and S_j occurs shall be denoted by $r_{\text{coll}}(i, j)$. If a collision occurs, there is a certain "acceptance" probability $p_{\text{a}}(i, j)$ that the two complexes will actually react directly after the collision. With probability $1 - p_{\text{a}}(i, j)$ nothing happens, and the two complexes move away again. However, if a reaction occurs, a hybridization channel has to be chosen. The probability to chose the reaction channel c is denoted by $p(c)$. The process of choosing one particular channel and then forming a new complex via this particular channel can be interpreted as an intermediate state with a vanishingly small lifetime. In this collision picture, the elementary rate for the hybridization $r_{\text{on}}(i,j,c)$ can be written as

$$r_{\text{on}}(i,j,c) = r_{\text{coll}}(i,j)p_{\text{a}}(i,j)p_{\text{c}}(i,j,c). \tag{4.31}$$

We can further coarse grain the acceptance and channel probability into one hybridization probability, i.e.,

$$p_{\text{hyb}}(i,j,c) = p_{\text{a}}(i,j)p_{\text{c}}(i,j,c). \tag{4.32}$$

The collision-based approach for the hybridization is sketched in Fig. 4.5. The number of connecting lines in this reaction network is significantly lower than in the original reaction network. Hence, the computational complexity of the update step $\mathcal{M}_n \to \mathcal{M}_{n'}$ has also become lower. Later on, we will introduce some assumptions that will further decrease the computational complexity. However, let us first take a closer look at the collision process in the next section.

4.5. A closer look at the collision process

All elementary rates appearing in our model solely depend on the structural properties of the reacting complexes. Our model thus neglects any spatial heterogeneity and assumes a homogeneous distribution of the reacting complexes. This assumption also implies that the local equilibrium distribution of the velocities is restored sufficiently quick after every reaction. These two assumptions are only valid if nonreactive elastic encounters are much more frequent than inelastic encounters, i.e., collisions leading to hybridizations. A system where the two assumptions are valid is often called a *well-mixed* system [184]. One could now come to the conclusion that in a well-mixed system, the acceptance probability p_a introduced in Section 4.4 has to be small such that the majority of all collisions does not lead to the formation of a new complex. However, two assumptions above can also be "made" valid for systems in which the acceptance probability p_a is not small. This can be achieved by introducing some non-reactive species that frequently hit the actual complexes. Yet, these non-reactive species do not have to be modeled explicitly. Instead, they can be assumed to act in the background. For RNA or DNA-like systems, those nonreactive species would naturally correspond to the solvent molecules.

In a well-mixed scenario, the elementary collision rate for two complexes belonging to the species S_i and S_j corresponds to the ratio of the average collision volume per unit time $\langle dV_{\text{coll}}(i,j)/dt \rangle$ to the system volume V according to Ref. [184], i.e.,

$$r_{\text{coll}}(i,j) = \frac{1}{V}\left\langle \frac{dV_{\text{coll}}(i,j)}{dt} \right\rangle. \tag{4.33}$$

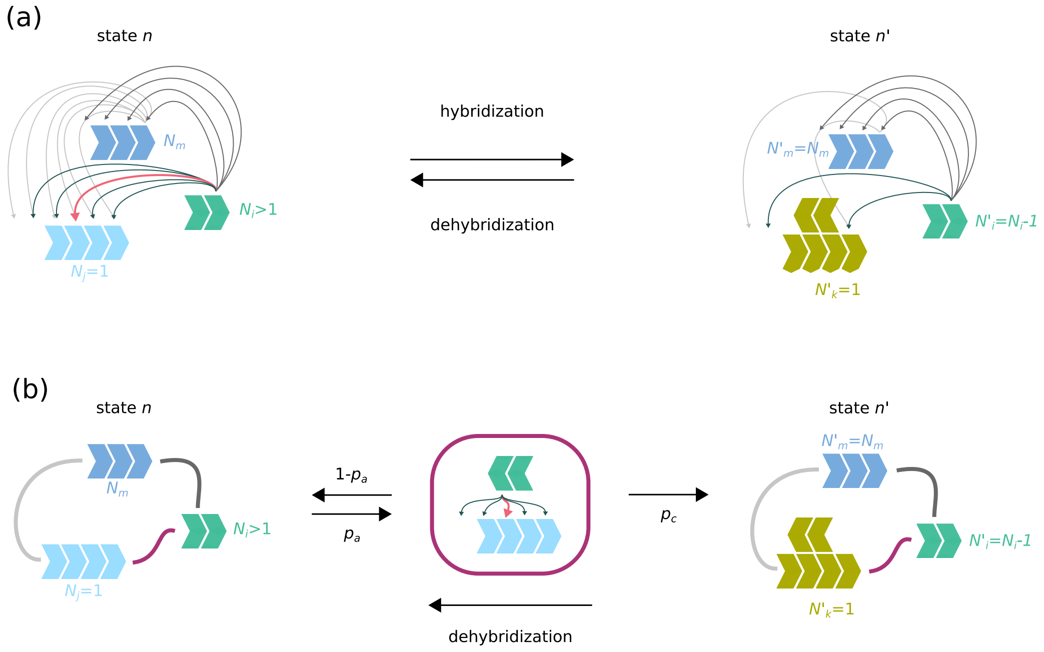


Figure 4.5.: Hybridization on the Markov chain on the space of copy numbers $\mathcal{N}$. (a) Hybridization reactions are treated as individual reactions. The system contains complexes belonging to three distinct species S_m, S_i and S_j with copy numbers $N_m, N_i > 1$ and $N_j = 1$. Each thin line connecting the instances of the species corresponds to a possible hybridization reaction. The elementary rates for these reactions are $r_{\text{on}}(i, j, c)$ and have to be weighted by the combinatorial factors accounting for the occupation numbers. The reaction that will occur next is highlighted. (b) Hybridization is interpreted as a two step reaction: A hybridization occurs after a collision via the transition through a transient state with a vanishingly small lifetime. The probability that two complexes form a transition state upon a collision is given by p_a. The probability of exiting the transition state via the hybridization channel c is p_c.

The term $\langle dV_{\text{coll}}(i, j)/dt\rangle$ is a function of the temperature and further depends on the properties of the solvent and the details of the underlying transport process. However, the scaling of the elementary collision rate with V^{-1} is a generic feature resulting from the assumption of spatial homogeneity. Moreover, if we substitute Eq. (4.33) into Eq. (4.31) and then plug in the result into Eq. (4.24) and Eq. (4.25), we can also convince ourselves that the rate constants for bimolecular reactions are indeed independent of the reaction volume (as they should be).

Before we end this section, we introduce the effective collisions rate $r_0(i, j)$ as

$$r_0(i, j) := N_A c^\circ \langle dV_{\text{coll}}(i, j)/dt\rangle, \tag{4.34}$$

where N_A is the Avogadro constant and $c^\circ = \text{mol/l}$ is the standard concentration.

Using this definition, we can rewrite Eq. (4.33), i.e.,

$$r_{\text{coll}}(i,j) = \frac{1}{VN_A c^{\circ}} r_0(i,j). \tag{4.35}$$

Eq. (4.35) will turn out to be useful in the next section where we are deriving the free energy associated with a hybridization site.

4.6. Gibbs Free Energies

In this section we will derive the Gibbs free energy associated with complexes and hybridization sites. The expression for Gibbs free energy depends on the specific choice of the elementary rates for hybridization and dehybridization reactions, i.e., on the corresponding chemical rate constants. In order to obtain a physically meaningful expression for the Gibbs free energy, care has to be taken when choosing the elementary rates. We will show that our specific choice for the elementary rates is thermodynamically consistent and in agreement with standard energy models for DNA or RNA [76, 107] (see also Section 2.1.3 in Chapter 2) complexes. In particular, we will show that the Gibbs free energy of a complex composed of multiple strands is independent of the specific assembly trajectory.

The stability of a hybridization site and the corresponding dehyridization rate depends on the number and nature of the interacting nucleotides at the hybridization site [133, 141, 142, 76, 107]. For the sake of simplicity, we assume a system where there is only one type of nucleotide and that this type of nucleotide is self-complementary. Moreover, we assume that the dehybridization rate decreases exponentially with the number of paired nucleotides that form the hybridization site. The same simplifications will be applied in Chapter 5 where we will also explain that such a simplified description emerges naturally in mean-field descriptions like the “random sequence approximation” [187]. In Chapter 6 we will drop this simplifying assumption and study systems where two different types of nucleotides are present. These two types of nucleotides are complementary. Moreover, in Chapter 6, the dehybridization rate does not decrease exponentially with the length of the hybridization site anymore, but with the number of nearest neighbor blocks contained in the hybridization site. However, most of the reasoning presented in the following will remain valid. The subtle differences will be discussed in Chapter 6.

4.6.1. Free energy of a hybridization site

In Section 4.5 we used a collision-based approach to derive the following expression for the elementary rate of a dehybridization reaction involving two complexes belonging to the species S_i and S_j:

$$r_{\text{on}}(i,j,c) = \frac{1}{VN_A c^{\circ}} r_0(i,j) p_{\text{a}}(i,j) p_{\text{c}}(i,j,c). \tag{4.36}$$

If we use this expression, we can rewrite Eq. (4.27), and obtain a new equation for the corresponding chemical rate constant:

$$k_{\text{on}}(i,j,c) = 2^{m_{\text{on}}-\delta_{ij}} p_{\text{a}} p_{\text{c}} r_0(i,j) \frac{1}{c^{\circ}}. \tag{4.37}$$

Recall, $m_{\text{on}} \in 0,1,2$ is the number of rotationally symmetric complexes involved in the reaction. Moreover, the rate constant $k_{\text{coll}}(i,j)$ corresponding to the elementary collision rate $r_{\text{coll}}(i,j)$ is given by

$$k_{\text{coll}}(i,j) := 2^{-\delta_{ij}} r_0(i,j) \frac{1}{c^{\circ}}. \tag{4.38}$$

The new complex resulting from the hybridization of the initial complexes belonging to the species S_i and S_j via channel c belongs to species S_k. We now choose the elementary rate $r_{\text{off}}(k,c)$ for the corresponding dehybridization, i.e, the inverse reaction to be

$$r_{\text{off}}(k,c) = p_c r_0(i,j) e^{\gamma l}. \tag{4.39}$$

Here, l is the length of the hybridization site, i.e., the number of nucleotides forming the hybridization site. Moreover, $\gamma < 0$ is a parameter that correponds to the (negative) binding energy per unit length in units of $k_{\text{B}}T$. We will discuss this parameter γ in detail in Chapter 5. According to Eq. 4.28, the corresponding chemical rate constant becomes

$$k_{\text{off}}(k,c) = p_{\text{c}} r_0(i,j) 2^{m_{\text{off}}(1-\delta_{ij})} e^{\gamma l}. \tag{4.40}$$

From the above expressions for the chemical rate constant we can deduce an expression for the Gibggs free energy $G_{\text{hyb}}(c)$ associated with the hybridization reaction via the hybridization channel c according to Eq. (4.15). We obtain:

$$\beta \Delta G_{\text{hyb}}(c) = \ln\left(\frac{k_{\text{off}}(k,c)}{c^{\circ} k_{\text{on}}(i,j,c)}\right) = \gamma l + \mu \ln(2) - \ln(p_{\text{a}}), \tag{4.41}$$

where we have introduced the term

$$\mu = \delta_{ij} - m_{\text{on}} + m_{\text{off}}(1 - \delta_{ij}). \tag{4.42}$$

Eq. (4.42) can be simplified if we know to the symmetry properties of species S_i, S_j and S_k to which the reaction and resulting complexes belong to. The following scenarios are possible:

1. S_i, S_j and S_k are rationally symmetric (see Fig. 4.6(a)). In this case we necessarily have $S_i = S_j$. Hence, we obtain

$$\mu = 1 - 2 + 1(1 - 1) = -1. \tag{4.43}$$

2. S_i and S_j are rationally symmetric, and S_k is not. In this case we necessarily have $S_i \neq S_j$. With that, we get

$$\mu = 0 - 2 + 0(1 - 0) = -2. \tag{4.44}$$

3. Either S_i or S_j are rationally symmetric, not both. This implies that $S_i \neq S_j$. In this case S_k can not show any symmetry. $S_i \neq S_j$.With that, we get

$$\mu = 0 - 1 + 0(1 - 0) = -1. \tag{4.45}$$

4. S_k is rationally symmetric, and both S_i and S_j are not. In this case we necessarily have $S_i = S_j$. This leads to

$$\mu = 1 - 0 + 1(1 - 1) = 1. \tag{4.46}$$

5. None of the involved species S_i, S_j and S_k is rationally symmetric implying that $S_i \neq S_j$ (see Fig. 4.6(b)) such that

$$\mu = 0 - 0 + 0(1 - 0) = 0. \tag{4.47}$$

With that, the second term on the right-hand side of Eq. (4.41) can be interpreted as follows: $\mu \ln(2)$ is a symmetry correction term, which yields either an energetic reward of $-\ln(2)$ or $-2\ln(2)$, if the hybridization caused a reduction in the number of symmetric complexes involved in the reaction by one or two, or an energetic penalty of $+\ln(2)$ if the number of symmetric complexes is increased by one. The term is zero if there is no change in the number of symmetric complexes. The $+\ln(2)$ term corresponds to the symmetry correction of $+0.43\,\text{kcal/mol}$ for the formation of a duplex

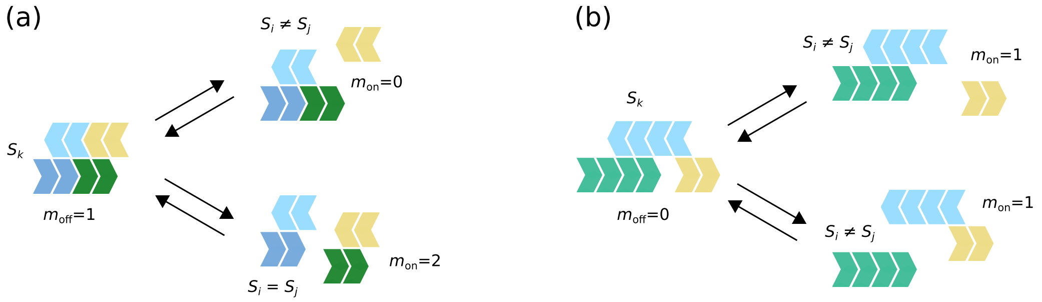


Figure 4.6.: m_{off} indicates whether the complex in which the dehyridization occurs is rotationally symmetric ($m_{\text{off}} = 1$) or not ($m_{\text{off}} = 0$). The parameter $m_{\text{on}} \in 0, 1, 2$ counts the number of symmetric complexes involved in the hybridization process. (a) Example for the dehybridization of a rotationally symmetric complex ($m_{\text{off}} = 1$). In this example, the dehybridization either leads to two non-symmetric complexes belonging to different species $S_i \neq S_j$ ($m_{\text{on}} = 0$) or to two rotationally symmetric complexes belonging to to same species, i.e., $S_i = S_j$. In this case we have $m_{\text{on}} = 2$. (b) Dehybridization of a non-symmetric complex: The dehybridization leads either to two non-symmetric complexes ($m_{\text{on}} = 0$) or one non-symmetric complex and one complex with a rotational symmetry ($m_{\text{on}} = 1$).

from two identical single strands appearing in the nearest neighbor data bases for RNA and DNA [76, 107] (see also Section 2.1.3 in Chapter 2).

The first term on the right-hand side of Eq. (4.41), $\gamma l = \beta \Delta G_b^\circ$, is the standard binding free energy (in units of $k_{-1}T$) associated with the hybridization site which results from the base pairing.

The third term on the right-hand side of Eq. (4.41) is the logarithm of the probability that a hybridization actually occurs upon collision. It can be thought of as the probability for the formation of a first base pair p_{first} times the probability that the formation of the first base pair leads to zipping of the single-stranded segments onto each other p_{zip}, (see [188]). We can therefore write $p_{\text{a}} = p_{\text{1bp}} p_{\text{zip}}$. Hence, the term $\nu = -\ln(p_a) \geq 0$ has the interpretation of an energy penalty associated with the formation of the first base pairs upon hybridization.

4.6.2. Total Gibbs free energy of a complex

To obtain the total Gibbs free energy $\Delta G_{\text{tot}}^\circ$ of a complex C containing multiple strands, we take the sum over all the Gibbs free energy ΔG_{hyb} that emerge along the assembly trajectory of the complex. In our simple picture the binding energy associated with a hybridization site only depends on its length. In particular, it is independent of the binding energy of other hybridization sites also present in the complex. Hence, the contributions from the binding energies simply sum up to $\gamma \sum_{h \in C} l_h$, where the index h denotes the different hybridization sites in the complex C. For a complex containing

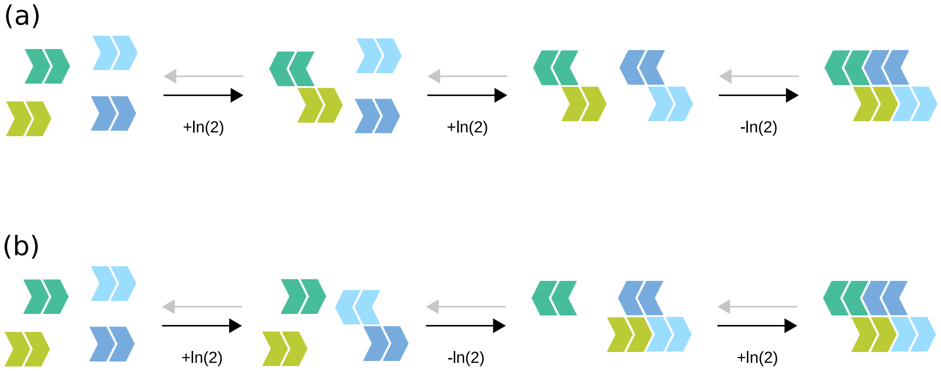


Figure 4.7.: The assembly trajectory (a) and (b) both lead to the same final complex. The symmetry terms $\mu \ln(2)$ occurring in the different steps are indicated along the assembly trajectory. The hybridization reactions along the two trajectories are associated with different Gibbs free energies since the numbers of rotationally symmetric complexes after the second hybridization are different. However, the Gibbs free energies of the final complex are identical since both trajectories accumulate the same symmetry terms.

N_{strands} strands, $N_{\text{strands}} - 1$ hybridizations must have occurred along the assembly trajectory. For every hybridization the energetic association penalty ν accrues. Hence, the total energetic association penalty is given by $(N_{\text{strands}} - 1)\nu$. As illustrated in Fig. 4.7, the symmetry correction terms accruing along an arbitrary assembly trajectory always sum up to $+\ln(2)$ if the final complex is rationally symmetric. If the final complex does not show a rotational symmetry, the symmetry correction terms always sum up to zero. Hence, for the total symmetry correction term we write $\rho \ln(2)$, where $\rho = 1$ if the complex is rotationally symmetric and $\rho = 0$ otherwise. The total Gibbs free energy $\Delta G^{\circ}_{\text{tot}}$ then reads

$$\beta \Delta G^{\circ}_{\text{tot}} = \beta \sum_{c=1}^{n-1} \Delta G^{\circ}_{\text{hyb}}(c) = \sigma \ln(2) + (N_{\text{strands}} - 1)\nu + \gamma \sum_{h \in C} l_h. \tag{4.48}$$

From Eq. (4.48) we see that the total Gibbs free energy $\Delta G^{\circ}_{\text{tot}}$ only depends on the structure of the complex and not on the preceding assembly process. Different assembly trajectories leading to the same final complex will result in identical total Gibbs free energies as required by thermodynamic consistency.

4.7. Specific choice of the kinetic parameters and implementation details

In the previous sections we derived general expressions for the elementary rates based on a collision approach. These expressions led to thermodynamically consistent Gibbs free energies for hybridization sites and complexes. However, the general expressions for the elementary rates are not unique and allow for different choices within some constraints. We will precise the specific choices we made in our model in the following paragraphs. However, before we can specify the elementary rates, we have to comment on some aspects of our implementation.

A naive implementation would tread individual complexes explicitly. If N_i complexes of species S_i exist, the implementation would hold N_i distinct instances of that species in the computer memory. The number of monomolecular reactions would then scale with the number of individual complexes, while the number of bimolecular reactions would approximately scale with the number of pairs that could be formed from individual complexes. The total number of reactions that would have to be stored in the computer memory would become very large in this case. A more sophisticated implementation only keeps one instance of every present species in the computer memory together with the copy number. The rates for the reactions involving these instances then have to be weighted accordingly with the copy numbers. These weights correspond to the combinatorial factors that we derived for the mapping of the elementary rates onto chemical rate constants in Section 4.3. If a monomolecular reaction involving one instance of a species gets chosen, this instance "sends out" one individual complex, which will actually perform the reaction. The copy number of the instance that "sent out" the complex is reduced by one. If the updated copy number is zero, the instance has to be deleted from the computer memory. If the new complex(es) resulting from the reaction does (do) not belong to any species that is (are) already present, one (two) new instance of new species with copy number one has (have) to be created in the computer memory. Otherwise, if the species is (are) already present, the copy of the corresponding instance has (have) to be raised by one. In the next step, the rates of the existing reactions have to be updated according to the changes in the copy numbers. If one (two) new species emerged in the reaction, new reactions between the new species and the already existing species have to be created. The procedure for a bimolecular reaction, i.e., a collision is similar, except that two complexes have to be sent out for the reaction. These two complexes can either come from two different instances if $S_i \neq S_j$ or from the same instance if $S_i = S_j$. In the latter case the copy number of the instance has to be decreased by two.

Our implementation is of the second, more sophisticated type. In the following, we will show that the number of collision reactions that have to be stored in the computer memory can be reduced even more if clever yet thermodynamically consistent choices of the kinetic parameters are made.

4.7.1. Constant elementary collision rate

The elementary collision rate $r_{\mathrm{coll}}(i, j)$ introduced for two individual compelexes via Eq. (4.31) in Section 4.4 explicitly depends on the species S_i and S_j of the two complexes. With that, the resulting rate for the collision between two instances of the species S_i and S_j (see above) in our implementation rate constants depends explicitly on the species S_i and S_j.

Imagine that due to a reaction a species S_k, which had not been present in the system before the reaction occurred, emerges. In this case a new instance corresponding to that new species S_k has to be created. To update the set of possible reactions we would now have to compute the elementary rates for every possible collision between this instances and all other instances corresponding to the species $S_i \in \mathcal{S}_n$ which are present in the new system state n. The rate constants and the corresponding reactions would then have to be stored somewhere in the computer memory. The computational complexity of this update step scales with the number of different species $|\mathcal{S}_n|$. If one species, i.e., one instance, disappears, the computational complexity of the update step is of the same order. Also reactions that do not lead to the creation of new instances but only change the copy numbers of existing instances require an update step scaling with $|\mathcal{S}_n|$. The computational complexity can be decreased drastically if we assume that the rate constant for the collision is constant, i.e., identical for all pairs of species $S_i, S_j \in \mathcal{S}_n$. Then, all collisions can be summarized to one single global collision rate $r_{\mathrm{coll}}^{\mathrm{tot}}$, which only depends on the absolute number N_n^{tot} of complexes contained in the current system state n, i.e.,

$$r_{\mathrm{coll}}^{\mathrm{tot}} = \frac{N_n^{\mathrm{tot}}(N_n^{\mathrm{tot}} - 1)}{2} r_{\mathrm{coll}}, \tag{4.49}$$

where r_{coll} is the constant elementary collision rate for two individual complexes.

When the Gillespie algorithm has selected the global-collision as the next reaction to be performed, two species, i.e., instances, have to be chosen randomly according to their copy numbers. This selection process can be sped up using a binary search tree. The chosen instances then "send out" individual complexes for the actual collision reaction.

According to Eq (4.35) the elementary collision rate for two individual complexes is given by

$$r_{\text{coll}} = \frac{1}{VN_A c^\circ} r_0. \tag{4.50}$$

The inverse of the elementary collision rate $t_0 = r_0^{-1}$ sets the time scale for the collision events. In Chapter 5 and Chapter 6 we will use t_0 as our fundamental time unit and express all time scales as a multiple of this fundamental time unit.

The assumption of the constant elementary rate for the collision process is further discussed in Section 4.7.5.

4.7.2. Hybridization upon collision and dehybridization

Two colliding complexes belonging to species S_i and S_j can form multiple different product complexes via different hybridization channels. The number of possible hybridization channels is denoted by Θ. We also refer to this quantity as the channel factor. Θ is a function of S_i and S_j, i.e., $\Theta = \Theta(i,j)$. For simplicity, we assume that that the acceptance probability p_a is always one if $\Theta > 0$ and zero otherwise, i.e.,

$$p_{\text{a}} = \begin{cases} 0, & \Theta(i,j) = 0 \\ 1, & \Theta(i,j) > 0. \end{cases} \tag{4.51}$$

For instance, this implies that two fully hybridized duplexes without any single-stranded overhangs will not react upon the collision, i.e., will not change their conformation during the collision process. Another implication is that the initiation penalty for a "productive" encounter is always zero within this simplification, i.e., $\nu = -\ln(p_a) = 0$. Having a second look at Eq. (4.36), we see that a constant acceptance probability $p_a < 1$, which is independent of the channel factor, would have the same effect as a rescaling of the reaction volume while keeping the number of complexes fixed. Moreover, we assume that all hybridization channels are equally likely such that probability to "pick" a particular channel c becomes

$$p_{\text{c}} = \frac{1}{\Theta(i,j)}. \tag{4.52}$$

Consequently the rate constant for the hybridization $r_{\text{on}}(i,j,c)$ is given by

$$r_{\text{on}}(i,j,c) = \frac{1}{\Theta(i,j)} \frac{1}{VN_A c^\circ} \frac{1}{t_0}, \tag{4.53}$$

while the corresponding chemical rate constant reads

$$k_{\text{on}}(i,j,c) = 2^{m_{\text{on}} - \delta_{ij}} \frac{1}{\Theta(i,j)} \frac{1}{c^\circ t_0}. \tag{4.54}$$

The elementary rate and the chemical rate constant for the inverse reaction, i.e., the hybridization, then become

$$r_{\text{off}}(k,c) = \frac{1}{\Theta(i,j)} e^{\gamma l} \frac{1}{t_0}, \tag{4.55}$$

and

$$k_{\text{off}}(k,c) = \frac{1}{\Theta(i,j)} 2^{m_{\text{off}}(1-\delta_{ij})} e^{\gamma l} \frac{1}{t_0}. \tag{4.56}$$

We will interpret and discuss the kinetics resulting from this parametrization in detail in Section 4.7.5. Moreover, we will convince ourselves that our parametrization is thermodynamically consistent.

4.7.3. Outflux

We also assume the elementary ouflux rate to be independent of the properties of the instances of the species and set it to a constant value, i.e., $r_{\text{out}}(i) = r_{\text{out}}$. In analogy to the single collision instance for the collision reaction we can introduce a single instance for the outflux reaction occurring with rate

$$r_{\text{out}}^{\text{tot}} = N_n^{\text{tot}} r_{\text{out}}. \tag{4.57}$$

Again, if according to the Gillespie algorithm the next event to occur is the outflux of a complex, we randomly choose one instance according to its copy number which will then “send out” one complex, which will then be removed from the system.

Ligation

In Chapter 5 the elementary rate of ligation is assumed to be constant, $r_{\text{lig}}(i,c) = r_{\text{lig}}$. In Chapter 6, where we treat sequences explicitly, the ligation rate depends on the nucleotides in the vicinity of the ligation site.

In our implementation each ligation reaction that is possible within one instance of a species is stored “inside” that instance. Hence, we do not coarse-grain the ligation reactions associated with different instances of species to one single global ligation reaction, including all instances.

Cleavage

In Chapter 5 the cleavage reaction does not exist, while we assume a constant rate for the cleavage of bonds within single-stranded segments in Chapter 6. Every instance of a species “contains” a summarizing cleavage reaction. The rate of this total cleavage reaction for one instance corresponds to the number of bonds within all single-stranded segments times the constant cleavage rate for one bond (times the copy number of the instance). If the summarizing cleavage reaction associated with one instance is selected by the Gillepsie algorithm as the next reaction to occur, the instance "sends out" an individual complex for the reaction. The bond that will get cleaved in this complex is chosen randomly.

4.7.4. Scaling of the kinetic parameters of a stationary system

Consider the reaction fluxes ϕ in terms of the concentration vector $\vec{c}$ and the rate constants:

$$\phi_{\text{on}}(\vec{c}, k_{\text{on}}, k_{\text{off}}, k_{\text{lig}}, k_{\text{out}}) \propto c_i c_j k_{\text{on}}, \tag{4.58}$$

$$\phi_{\text{off}}(\vec{c}, k_{\text{on}}, k_{\text{off}}, k_{\text{lig}}, k_{\text{out}}) \propto c_k k_{\text{off}}, \tag{4.59}$$

$$\phi_{\text{lig}}(\vec{c}, k_{\text{on}}, k_{\text{off}}, k_{\text{lig}}, k_{\text{out}}) \propto c_k k_{\text{lig}}, \tag{4.60}$$

$$\phi_{\text{out}}(\vec{c}, k_{\text{on}}, k_{\text{off}}, k_{\text{lig}}, k_{\text{out}}) \propto c_k k_{\text{out}}. \tag{4.61}$$

A transformation of the form

$$\vec{c} \to \vec{c}' = \alpha_c \vec{c}, \tag{4.62}$$

$$k_{\text{on}} \to k'_{\text{on}} = \alpha_{\text{on}} k_{\text{on}}, \tag{4.63}$$

$$k_{\text{off}} \to k'_{\text{off}} = \alpha_{\text{off}} k_{\text{off}}, \tag{4.64}$$

$$k_{\text{lig}} \to k'_{\text{lig}} = \alpha_{\text{lig}} k_{\text{lig}}, \tag{4.65}$$

$$k_{\text{out}} \to k'_{\text{out}} = \alpha_{\text{out}} k_{\text{out}} \tag{4.66}$$

that scales all reaction fluxes by the same amount changes only the intrinsic time-scale of the dynamics. In particular, it leaves the stationary distributions invariant.

By forming the three independent ratios of the reaction fluxes, one finds that all transformations that lead to the same ratios

$$\alpha_1 := \frac{\alpha_{\text{off}}}{\alpha_{\text{lig}}}, \tag{4.67}$$

$$\alpha_2 := \frac{\alpha_{\text{off}}}{\alpha_{\text{out}}}, \tag{4.68}$$

$$\alpha_3 := \frac{\alpha_{\text{off}}}{\alpha_{\text{on}} \alpha_{\text{c}}}, \tag{4.69}$$

are equivalent with respect to the stationary distribution. In particular, it shows that scaling the hybridization rate has the same effect as scaling the concentration.

4.7.5. Thermodynamically consistent kinetics for nearest-neighbor models

For a thermodynamically consistent model the parametrization of the (de-)hybridization kinetics is constraint by the binding energy. In particular, thermodynamic consistency implies that that the free energy $\Delta G^\circ_{\mathrm{tot}}$ (see Eq. (4.48)) of any complex is independent of the assembly trajectory. Consequently, using an energy model that is in the spirit of common nearest neighbor models, *cf.* Eq. (4.70), introduces constraints on the available parameterizations.

Thermodynamic consistency with nearest-neighbour models (see Section 2.1.3 in Chapter 2) requires that the energy for a hybridization channel, $\Delta G^\circ_{\mathrm{hyb}}(c)$, is independent of θ. $\Delta G^\circ_{\mathrm{hyb}}$ is determined by the choice of the rate constants k_{on} and k_{off} via $\beta \Delta G^\circ_{\mathrm{hyb}} = \frac{k_{\mathrm{off}}}{c^\circ k_{\mathrm{on}}}$. The rate constants and rates can be mapped onto each other via Eq. (4.27) and Eq. (4.28). Choosing a collision based Ansatz for the on-rate (Eq. (4.31)) $r_{\mathrm{on}} = (V N_A c^\circ)^{-1} r_0 p_{\mathrm{a}} p_{\mathrm{c}}$, and writing the off-rate as $r_{\mathrm{off}} = a e^{\gamma l}$, $\Delta G^\circ_{\mathrm{hyb}}$ becomes

$$\Delta G^\circ_{\mathrm{hyb}} = \gamma l + \mu \ln(2) + \ln\left(\frac{a}{p_a p_c r_0}\right). \tag{4.70}$$

To get a path independent $\Delta G^\circ_{\mathrm{tot}}$, the acceptance probability p_{a}, the channel probability p_{c} and the effective collision rate must be chosen such that the third term of Eq. (4.70) becomes constant, and hence independent of the properties of the complexes. $\nu = -\ln(p_a)$ is interpreted as the initiation energy. In the simplest form of the kinetics, the channel probability, p_c must sum to unity for all of the θ different channels. The simplest choice is $p_c = 1/\theta$. Hence, Eq. (4.70) becomes

$$\Delta G^\circ_{\mathrm{hyb}} = \gamma l + \mu \ln(2) + \nu + \ln\left(\frac{a\theta}{r_0}\right). \tag{4.71}$$

To have a thermodynamically consistent energy model, we have to eliminate the factor θ from the binding energy. Otherwise, the total energy of a complex, Eq. (4.48), would depend explicitly on the properties of all involved strands.

There are two possible canonical choices of a and r_0 that can be made to set the last term to zero[†]: (a) $a = r_0/\theta$ or (b) $a = 1$ and $r_0 = r_0^* \theta$, where r_0^* is a constant. Both choices lead to a consistent energy model ut may be considered microscopically unsatisfactory: (a) This is the choice we used for our model. At first glance, it seems odd that the off-rate depends on the number of channels θ. However, this problem is intrinsic to using such a kinetics in nearest-neighbor type models. (b) The second choice has the advantage that the θ-dependence is contained in the hybridization rate and absent from dehybridization. It has the microscopic advantage that the hybridization

[†] a remaining constant could be absorbed into ν

depends on θ. However, this is also physically questionable from the perspective of microscopic, diffusive dynamics (see below).

Whatever the exact choice, as long as it is not exponential in the strand length microscopic factors like θ only contribute subexponentially to the length-scales derived from the competition of time scales.

The primary reason for our choice (a) is computational. Having a collision rate independent of the nature of the complexes allows us to sample colliding pairs without considering their possible hybridizations *a priori*. Any other choice would increase the complexity because of the additional computation that needs to be performed for all pairs of species. After each hybridization we would have to update the whole reaction network which scales as $\sim \langle\theta\rangle\,(N_n^{\text{tot}}-1)$, where $\langle\theta\rangle$ is the average number of hybridization channels. As explained above our implementation is fast because by assuming a constant collision rate, we can reduce the collision events to a single total reaction. Only after the colliding pair is drawn we chose a channel and then update the possible hybridization and dehybridization reactions just in time. Moreover, a closer look at the kinetics (b) reveals that it is not necessarily more physical, as long as the underlying transport process is not specified. For example, $r_0 = r_0^*\theta$ neglects the decrease in mobility with strand length as would be the case for regular diffusion. To illustrate this aspect more thoroughly let us consider a simple model of two diffusing complexes with diffusion coefficients D_1, D_2. We assign to the complexes the hydrodynamic radii R_1, R_2 and assume that the complexes undergo a reaction as soon as their distance becomes smaller than $R_1 + R_2$. The collision rate is then obtained via the Smoluchowski rate coefficient [189, 190, 191]

$$r_{\text{coll}} = \frac{1}{V}4\pi(R_1+R_2)(D_1+D_2). \tag{4.72}$$

Using that the hydrodynamic radius is inversely proportional to the diffusion coefficient [192], Eq. (4.72) becomes

$$r_{\text{coll}} \sim \frac{1}{V}4\pi\frac{(D_1+D_2)^2}{D_1D_2}. \tag{4.73}$$

Experimentally, the following relation between diffusion coefficient and length for single (double) stranded DNA has been found:

$$D \sim L^{-\nu}, \tag{4.74}$$

where $\nu = 0.45$ ($\nu = 0.67$) [193, 194]. Thus the collision rate is proportional to

$$r_{\text{coll}} \sim \frac{(L_1^\nu + L_2^\nu)^2}{L_1^\nu L_2^\nu} =: f(L_1, L_2, \nu). \tag{4.75}$$

This expression scales differently with L_1 and L_2 than $\theta = L_1 + L_2 - 1$, compare Fig. 4.8(left) with Fig. 4.8(right).

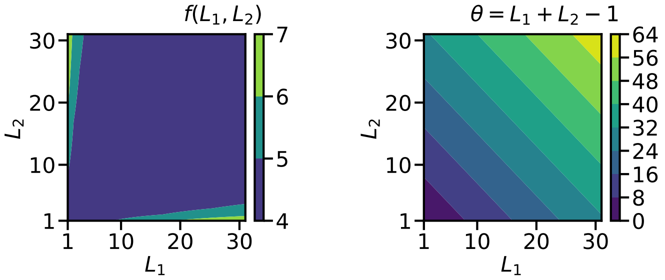


Figure 4.8.: (left) $r_{\text{coll}} \sim f(L_1, L_2)$ for $\nu = 0.45$. (right) number of channels $\theta(L_1, L_2)$. $f(L_1, L_2)$ and $\theta(L_1, L_2)$ have different scaling.

One potential solution to this apparent dilemma would be the introduction of an intermediate state. This state would be characterized by two molecules that have collided, but have not yet formed a hybridization complex. From this intermediate state, molecules can then either hybridize along a channel or go back into solution. In accordance with microscopic reversibility the dehybridization has to pass through this intermediate state, which effectively allows the strands of a complex to reassemble. In that case, the factors weighing the hybridization into different channels could be completely arbitrary since the corresponding probabilities would not need to sum to unity, but only determine the life-time of this intermediate state. Acceptance probabilities with a general parameter dependence can be formulated independently from an (optional) “initiation energy” in a nearest-neighbor model. However, such a simulation would require many more (unknown) parameters and a different implementation.

4.7.6. Approximations to improve the performance of the implementation

We close the discussion of our simulation method by presenting two “hacks” that allowed us to speed up the simulation process considerably.

In Chapter 5 we discuss a system that is diffusively coupled to a large reservoir containing monomers and dimers only. The coupling to the reservoir is implemented by keeping the number of single-stranded monomers and dimers in the reaction volume constant. The timescales for dehybridizations in complexes composed of monomers and dimers only are short compared to the time scales for ligation and outflux processes. We can therefore assume that the single-stranded monomers, dimers, and complexes containing only monomers and dimers reach a binding equilibrium quickly. Moreover, we can assume that this binding equilibrium is not affected much by the other processes going on in the reaction volume since monomers and dimers are by far the most abounded species. In other words, we can assume that there is a background formed by single-stranded monomers, dimers, and simple complexes comprising only monomers and dimers that remains approximately constant. We call the species forming this background *background species*. Collisions between, and dehybridization within background species only lead to small fluctuations in the background. However, since the background species typically have the highest copy numbers, collision and dehybridization involving only background species are the most frequent reactions. Sampling all these "background" reactions is computationally costly and will, as already mentioned, only lead to small fluctuations in the background. A simple trick to overcome this sampling problem is to assume a constant background, separate this background from the species that do not belong to one of the background species and collisions and dehybridization within this background. To illustrate this idea we look at the time scale between two collision events in a straightforward implementation where the background is not separated from the other species. Let us assume that there are in total N_{tot} complexes in the system. Out of these N_{tot} complexes, N_1 complexes belong to one of the background species and N_2 do not. As already said, in a typical situation, we have $N_1 \gg N_2$. The time scale t_{coll} between two collision events is given by the inverse total collsion rate, i.e.,

$$t_{\text{coll}} \approx \left[\frac{N_{\text{tot}}^2}{2}\right]^{-1} = \left[\frac{(N_1 + N_2)^2}{2}\right]^{-1} = \left[\frac{\left(N_1^2 + 2N_2 + N_2^2\right)^2}{2}\right]^{-1} . \tag{4.76}$$

The N_2^2-term on the right-hand side of Eq. (4.76) is the dominant contribution and leads to a high collision frequency. If we would now exclude all collisions involving only complexes belonging to background species, we could "ban" the N_2^2-term from Eq. (4.76). As a result, the frequency of collisions between individual complexes would become much smaller. Hence, we would have to sample way fewer collisions to advance the system time by the same amount of time without changing the physics much. In our implementation used to generate the results for Chapter 5 , we realize the idea that we just described: We do not consider collisions between complexes from the background in our implementation and we only account for collisions involving at most one complex belonging to the background. The constant background is created before the actual simulation starts. To do so, we leta system evolve where where the number of single-stranded monomers and dimers is fixed and where the only reactions that are allowed are collisions, i.e., hybridizations and dehybridizations. Once a binding equilibrium is reached, i.e., when the running time average of occupation numbers of the different species remains approximately constant, we start to take a large number of “snapshots” separated by a given time interval. The constant background is then obtained by forming the ensemble average over all the snapshots. Note that due to the averaging process some species in the background may have non-integer copy numbers. However, these non-integer copy numbers are not problematic since they ultimately only correspond to the weight at which these species are drawn to react by the algorithm.

In Chapter 6 we are considering a closed system with a fixed total mass. In this case we cannot assume a background formed by “simple” complexes and single-stranded monomers and dimers that remains constant over time. However, a similar “hack” to speed up the implementation is possible. The idea is as follows: Although their copy numbers vary over time, monomers will always be the most abundant species. If two monomers collide and hybridize to a “double-strand”, this “double-strand” will dissociate quickly due to its very low binding stability. As a result most of the monomers are free in solution. The tiny fraction of monomers that are bound to other monomers can be neglected. At this point it is important to stress that we do not neglect the binding of monomers to oligomers with $L > 1$. In such configurations monomers can potentially ligate to another strand and contribute to the growth process. We only neglect “double strands” formed by two monomers. During the short time interval that such a “double strand” exists, it cannot react further or be involved in a growth process. Everything it can do is to dissociate again. Neglecting these “double strands” formed of two monomers will therefore not change the global dynamics of the system. However, this approximation allows us to discard all collision reactions that involve exactly two monomers. These kinds of collision events would be the most frequent ones due to the high copy numbers of the monomer species. Avoiding the sampling of these collision events allows us to advance the system time in larger steps and to improve the performance of the simulation.

5. Growth regimes in polymer self-assembly by templated ligation[*]

The emergence of evermore complex entities from prebiotic building blocks is a key aspect of origins of life research. The RNA-world hypothesis posits that RNA oligomers known as ribozymes acted as the first self-replicating entities. However, the mechanisms governing the self-assembly of complex informational polymers from the shortest prebiotic building blocks were unclear. In the first assembly scenario discussed in Chapter 3, we made the assumption of having one distinct template strand with one specific primer statically bound at one specific end surrounded by mononucleotides only. We now drop this simplifying assumption and assume a pool composed of strands of various lengths without distinct roles. Now, all strands serve as a template, primer, and substrate for the extension at the same time. One open issue in the self-assembly via tamplated polymerization and ligation concerns the relation between concentration and oligonucleotide length, usually assumed to be exponentially decreasing. Here, we show that a competition of timescales in the self-assembly of informational polymers by templated ligation generically leads to nonmonotonic strand-length distributions with two distinct length-scales. The first length scale characterizes the onset of a strongly nonequilibrium regime and is visible as a local minimum. Dynamically, this regime is governed by extension cascades, where the elongation of a "primer" with a short building block is more likely than its dehybridization. The second length scale appears as a local concentration maximum and reflects a balance between degradation and dehybridization of completely hybridized double strands in a heterocatalytic extension-reassembly process. Analytical arguments and extensive numerical simulations within a sequence-independent model allowed us to predict and control these emergent length scales. Non-monotonic strand-length distributions confirming our theory were obtained in thermocycler experiments using random DNA sequences from a binary alphabet. This chapter emphasizes the role of structure-forming processes already for the earliest

[*] The chapter is adapted from the article: "Self-Assembly of Informational Polymers by Templated Ligation", by Joachim H. Rosenberger, Tobias Göppel, Patrick W. Kudella, Dieter Braun, Ulrich Gerland, and Bernhard Altaner published in Phys. Rev. X 11, 031055 (2020), under CC BY 4.0 License. See Ref. [183]. The author of this thesis is a shared first author of the paper. The experiments described in the paper, as well as in this chapter, have been performed by Patrick W. Kudella in the laboratory of Dieter Braun.

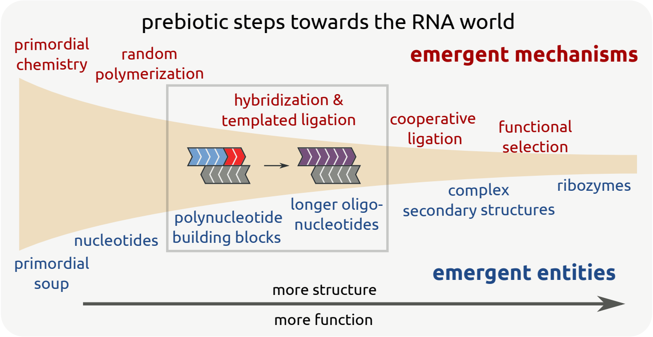


Figure 5.1.: Prebiotic evolution is a multi-step process that creates new entities exhibiting emergent mechanisms of interaction. This chapter outlines the emergence of structured oligonucleotides from the smallest building blocks.

stages of prebiotic evolution. The accumulation of strands with a typical length reveals one possible starting point for higher-order self-organization events that ultimately lead to a self-replicating, evolving system.

5.1. Introduction

A key question in research on the origins of life is how structure and biochemical complexity could emerge from unstructured conditions on early Earth. One of the most well-known hypotheses in this context is that of an "RNA world" [31, 32, 33, 34, 35, 36]. In this scenario, RNA oligomers acted as both the carrier of information and "ribozymes", *i.e.*, catalytic molecules allowing for the replication of this information and other metabolic functions. Yet, the RNA-world hypothesis does not address the question of how RNA strands that are complex enough to act as functional ribozymes came into being [80, 58, 72, 73, 60]. In the light of evolutionary principles, a multistep scenario of self-organization seems plausible, *cf.* Fig. 5.1. However, the intermediate steps on the way toward functional polynucleotides are still not well understood.

The smallest ribozymes known today are 30–100 nt long [64, 53, 65]. A common view is that the reliable self-assembly of replicating RNA molecules required specific sequences of 20–30 nt in length [63, 79, 42, 40]. It has been shown that selective (Watson-Crick) base pairing can lead to a vast reduction of complexity in sequence space, a phenomenon called cooperative ligation [59, 195]. Moreover, a recent hypothesis suggests that a replicating catalytic network would emerge as a "virtual circular genome", which self-assembles from an initial distribution of short oligonucleotides [79].

The efficiency and viability of such catalytic networks strongly depend on the relative concentrations of oligonucleotides of different lengths. Commonly exponential length distributions are assumed [196, 79]. While a decay in length is natural, the exact shape of length distributions emerging from short building blocks is not known [79]. Models and experiments have reported different observations [60, 187, 59].

Importantly, for low concentrations, the concatenation of oligonucleotides is dominated by a process known as *templated ligation* [60, 78, 197, 167]. Unlike *random ligation*, where two oligonucleotides directly combine into a longer strand, templated ligation involves a third strand, *cf.* Fig. 5.1. The third strand, also called the "template", enables the covalent bonding of two other strands adjacently hybridized on the template. Thus, the self-assembly of oligonucleotides cannot be captured by standard polymerization theory, where exponential length distributions are well understood [198, 199].

Probing the length distribution arising from templated ligation is challenging [79]. When forming covalent bonds between oligonucleotides, an energy gap needs to be overcome. In enzyme-free situations, this energy can be provided by an activation chemistry often involving imidazole or EDC [68, 130, 72, 73, 115]. The yields are usually small, experiments require a long time, and results can be obscured by side products [62, 63, 80, 81]. As a consequence, experimental research has focused on so-called "primer extension", which regards the extension of a short ($\sim$10 nt) oligonucleotide "primer" on a longer template, rather than self-assembly of oligonucleotides from small building blocks [68, 121, 69, 70, 116, 77, 71, 168].

An alternative for providing the energy for bond formation is using a ligase, which can be either an RNA-based ribozyme or a modern protein. The latter is not prebiotically plausible. However, these enzymes drastically increase yields and reaction speeds. While ligases require oligomers of lengths between six and ten nucleotides, enzymatic systems can still serve as conceptual models to explore the principles of self-assembly and ligation-based early replication [59, 58].

In order to study the self-assembly from smallest building blocks, we employed a computational and analytical approach based on a minimal "bottom-up" model. A transition between two dynamical regimes featuring differently decaying distributions has been reported recently [187]. These results were obtained within a coarse-grained, deterministic model, which does not capture the entire complexity associated with templated ligation.

A general yet simple theory identifying the generic properties of the self-assembly from shortest oligomer building blocks has been missing. The goal of this study is to close this gap. To this end, we investigated a model that captures the elementary mechanism of self-assembly: the hybridization of strands to form arbitrary complexes on which templated ligation can occur. To focus on the assembly process alone, the dependence on oligonucleotide sequences was neglected: The binding energy of a hybridization site is proportional to its length and characterized by a single parameter γ, reflecting a typical binding energy *per nucleotide*, which emerges naturally in mean-field descriptions like the "random sequence approximation" [187].

Our main result is that the competition of timescales between (length-dependent) dehybridization, extension, and a degradation or observation timescale generically leads to a *nonmonotonic strand-length* distribution. We show that different dynamic processes govern different regions in the space of strand lengths. The boundaries between these regions are given by a local minimum at a length $L_{\min}$ and a local maximum at $L_{\max} > L_{\min}$, which can be approximated by two analytical length scales $L^* \sim L_{\min}$ and $L^\dagger \sim L_{\max}$. This accumulation of strands at the typical length scale $L_{\max}$ constitutes a novel structure-forming process. Many of the microscopic details only enter the theory via a single parameter that characterizes the effective rate of extension. This allowed us to apply our theory to experiments, where a nonmonotonic length distribution emerges from the enzymatic ligation of a random pool of DNA sequences in a thermocycler.

5.2. Comparison to other models in the literature

Previous theoretical work largely studied templated ligation by effective models. The description of the state space had been reduced to strand lengths, without taking into account the hybridization complexes explicitly [200, 187, 201, 197, 202, 203, 204, 205, 195, 206, 207]. In such a coarse-grained picture, (de)hybridization and templated ligation are not elementary reactions but are combined into an effective extension reaction. To specify the corresponding rate, the intricacies of the assembly process are neglected and *a priori* assumptions regarding the relevant configurations are made [203, 200, 202, 201, 197, 195]. Many models neglect the dependence of the binding energy on the number of paired nucleotides [203, 200, 202, 201, 197, 195, 206, 205]. Others consider a length-dependent dehybridization rate only up to some cutoff length such that the time scale of ligation is always much larger than the time scale of the dehybridization kinetics [187]. A study addressing the full complexity of the assembly was limited by small system sizes [208].

5.3. Model

Figure 5.2(a) sketches the model dynamics. Short oligomers enter a reaction volume V, where they hybridize to form (partially) double-stranded complexes. If oligomers aggregate in suitable configurations, they may undergo templated ligation. Eventually, all complexes leave the reaction vessel at a constant rate, mimicking a flow reactor.

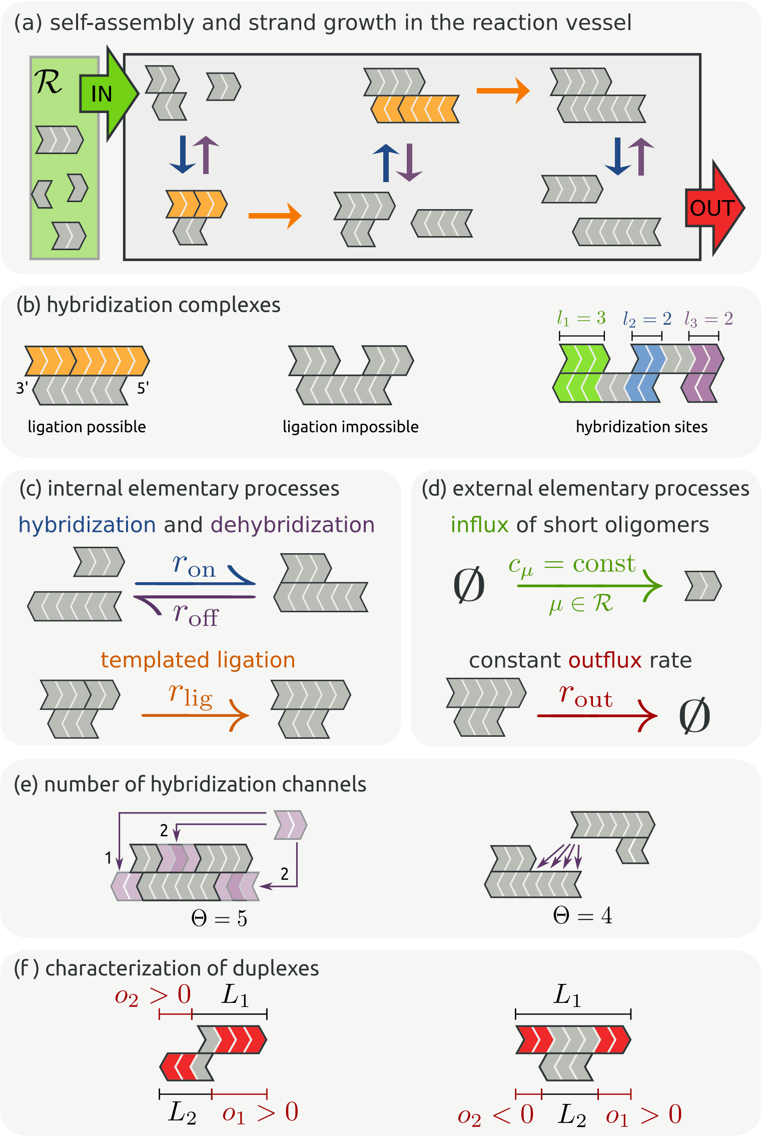


Figure 5.2.: (a) Short strands entering the reaction vessel from the reservoir are the initial building blocks of the system. Inside the vessel, strands form various complexes via hybridization and dehybridization. Subsequent ligation leads to longer strands eventually leaving the system. (b) Examples of higher-order complexes with multiple hybridiziation sites. (c) The internal elementary processes are hybridization, dehybridization and templated ligation with corresponding rates r_{on}, r_{off} and r_{lig}. (d) The external elementary processes couple the system to its environment. Short strands of length $L = \mu$ for $\mu \in \mathcal{R}$ are chemostated via the coupling to an external reservoir of initial building blocks at fixed concentrations c_μ. All complexes leave the system at a constant rate r_{out}. (e) When two complexes collide, they can form Θ different hybridization complexes. (f) Duplexes $D := (L_1, L_2, o_1)$ are uniquely characterized by the strand lengths $L_1, L_2 \in \mathbb{N}$ and one of the overhangs $o_i \in \mathbb{Z}$ of strand S_i, $i \in \{1, 2\}$, at its $3'$ end. Overhangs o_i can be negative, as for the case of o_2 in the right-hand example.

5.3.1. Strands and complexes

The basic element of our dynamics is a directed oligomer called a strand, which consists of covalently linked nucleotides. All linear conformations that can arise from a set of strands are allowed, see Fig. 5.2(b). Only self-folding and branched hybridization structures are excluded. While these effects might become important for longer strands, they can be neglected when dealing with the self-assembly from short strands. The overlapping region between two strands is referred to as a hybridization site. Single strands are called m-mers. We explicitly refer to monomers, dimers, trimers, and tetramers for $m = 1, 2, 3, 4$.

5.3.2. Elementary processes and parameters

The internal elementary reactions are hybridization, dehybridization and templated ligation, see Fig. 5.2(c). Hybridization and dehybridization are assumed to be elementary and reversible reactions with rates r_{on} and r_{off}. Thermodynamic consistency [209, 210] connects r_{on} and r_{off} to the free energy ΔG_b° of a hybridization site:

$$\frac{r_{\text{off}}}{r_{\text{on}}} = (VN_A c^\circ) e^{\beta \Delta G_b^\circ}, \tag{5.1}$$

where $\beta =^{(k_{\text{B}}T)-1}$, k_{B} is Boltzmann's constant and T denotes the absolute temperature, N_A is the Avogadro constant and $c^\circ = 1\,\text{mol/l}$ is the standard concentration.

When two strands of length L_1 and L_2 are hybridized adjacently on a third strand, they may ligate to a new strand of length $L_1 + L_2$. The ligation rate r_{lig} is assumed to be independent of L_1, L_2, the directionality of the strands and microscopic details. The uniform ligation rate can be interpreted as an effective average. A more detailed model could reflect that short oligomers predominately ligate to $3'$ ends [79, 211] due to the underlying chemistry [212, 213] or include stalling effects [77, 214, 215]. Since template-free ligation is a much slower process than templated ligation [78], it is neglected. Moreover, two external reactions connect the system to its environment, *cf.* Fig. 5.2(d): (i) A coupling to a reservoir fixes the concentrations c_m of m-mers with $m \in \mathcal{R}$. (ii) Each complex exits the system at a constant rate r_{out}.

5.3.3. Thermodynamics and kinetics of hybridization

The binding energy ΔG_b° of a hybridization is assumed to be directly proportional to the length chlerof the binding site l, see Fig. 5.2(b),

$$\beta \Delta G_b^\circ(l) = \gamma l, \tag{5.2}$$

where $\gamma < 0$ is a parameter that gives the (negative) binding energy per unit length in units of $k_\mathrm{B}T$.

Equation (5.1) thermodynamically constrains the ratio of r_on and r_off. An additional kinetic parameter is needed for a full parametrization of the rates. Here, we use a constant rate of collision between two complexes $r_\mathrm{coll} = (VN_A c^\circ t_0)^{-1}$, where $t_0 = (r_0)^{-1}$ is a microscopic, intensive, collision timescale, see Section 4. All times are measured in units of t_0. In general, two colliding complexes can form multiple configurations via Θ distinct hybridization channels, see Fig. 5.2(e). The probability of choosing each of these channels is assumed to be equal,

$$p_\mathrm{hyb} = 1/\Theta,\ \Theta > 0. \tag{5.3}$$

Hence, the hybridization rate via a given channel is

$$r_\mathrm{on} = r_\mathrm{coll}\, p_\mathrm{hyb}. \tag{5.4}$$

For the dehybridization rate we obtain from Eq. (5.1)

$$r_\mathrm{off} = \frac{1}{\Theta} e^{\gamma l}. \tag{5.5}$$

In reality, the collision rate depends on the properties of the colliding complexes, the solvent and temperature. A parametrization where the binding energy γl is attributed to the dehybridization rate r_off is a common kinetic assumption, and has been confirmed experimentally [133, 141, 142]. The kinetic assumptions Eqs. (5.4) and (5.5) reduce the computational complexity, while still maintaining the sampling of all configurations thermodynamically consistent, see Section 4.

In addition to this standard model, we also consider a "bounded" variant, where the dehybridization rate cannot become smaller than a minimal rate r_cut, such that $r_\mathrm{off} = r_\mathrm{cut}$ if $e^{\gamma l}/\Theta < r_\mathrm{cut}$. The bounded model can be thought of as an effective implementation of a system subjected to an external mechanism causing dehybridization of *all* complexes with a timescale of $\tau \sim (r_\mathrm{cut})^{-1}$. Such a situation can be realized by the thermal cycling in a "thermal trap" situated in a hydrothermal vent or be the consequence of other naturally occurring cycles [187, 125, 124, 29, 162, 216].

5.3.4. Standard parameters

Our primary focus is a scenario where the building blocks entering from the reservoir are dimers only. If not indicated otherwise, $c_2 = 2$ mM. The volume V is chosen such that 10^4 single-stranded dimers are present. This dimer-only scenario is the simplest model allowing for templated ligation and makes analytical considerations easier. If

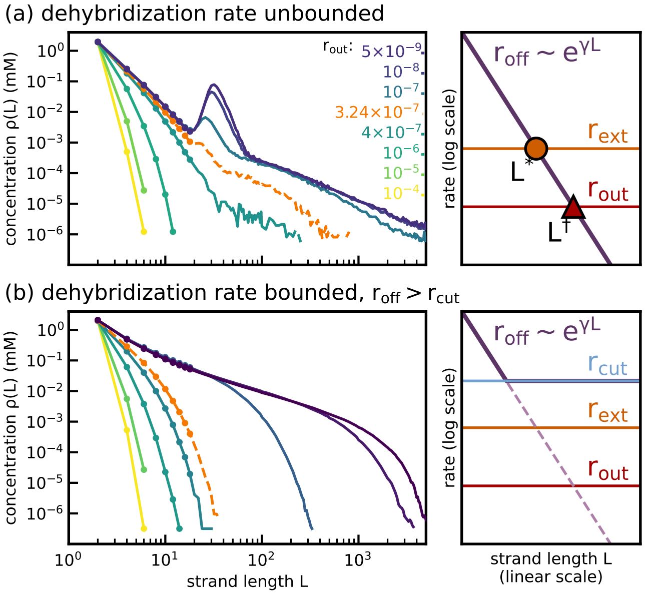


Figure 5.3.: Stationary length distributions (left) and competition of time scales (right) for the standard model (a) and its bounded variant (b) for different values of the outflux rate r_{out}. In the bounded model, dehybridization cannot become smaller than $r_{\text{cut}} = 0.05$. Dehybridization is thus faster than ligation ($r_{\text{lig}} = 2.5 \times 10^{-3}$) for all lengths. In both models, the length distributions develop long tails when decreasing the outflux rate r_{out}. The orange (dashed) curves correspond to a system where the outflux rate takes the crossover value $r_{\text{out}} = 3.24 \times 10^{-7}$, *cf.* Eq. (5.6). For outflux rates below the transition value, the unbounded model exhibits a nonmonotonic length distribution with a local minimum and maximum at L_{min} and L_{max}.

not otherwise stated, $\gamma = -0.5$, $r_{\text{lig}} = e^{-6}$ and $r_{\text{out}} = 5 \times 10^{-9}$. In the bounded model r_{cut} is a further parameter.

5.4. Simulation results and analysis

The main observable in this chapter is the length distribution of strands $\rho(L)$. It expresses the concentration of strands of length L, irrespective of the complexes they belong to.

5.4.1. Self-enhancing catalysis leads to long-tailed distributions

Self-assembly via templated ligation is a self-enhancing mode of growth, where long strands facilitate their own formation. This process competes with degradation. For large outflux rates, strands remain inside the reaction volume only for short times and participate in few or even no templated ligations. The resulting stationary length distribution is therefore expected to be short tailed.

In contrast, for a small outflux rate, strands spend more time inside the system and thus have a higher chance to serve as a template or to get ligated, leading to the formation of longer strands. These longer strands again allow for larger hybridization sites and are better templates. Consequently we expect the existence of a crossover value for the outflux rate $r_{\text{out}} = r^{\text{c}}_{\text{out}}$, where the formation of longer strands is dominantly self-enhancing. In the following section, we derive the value of the crossover rate

$$r^{\text{c}}_{\text{out}} = 2(c_2)^2 \left(e^{-4\gamma} + 2e^{-3\gamma}\right) r_{\text{lig}}, \tag{5.6}$$

under the assumption that (i) short-tailed distributions are dominated by the smallest building blocks and (ii) timescales of the dehybridization of these building blocks are small compared to the timescale of ligation.

We probed the stationary distribution for various values of the outflux rate r_{out}. Simulation results for the standard model are shown in Fig. 5.3(a). Figure 5.3(b) gives the analogous results for the bounded model.

Since the derivation of Eq. (5.6) does not rely on the dynamics of long strands affected by the cutoff, we expect the same transition from short- to long-tailed distributions in both scenarios. For sufficiently large outflux rates the resulting short-tailed length distributions look quantitatively similar. The curves for the crossover outflux rate $r^{\text{c}}_{\text{out}} = 3.24 \times 10^{-7}$ obtained from Eq. (5.6) are indicated as dashed (orange) lines. The long-tailed distributions for small outflux rates differ significantly: In the standard model, Fig. 5.3(a), a local minimum and maximum emerge. In contrast, the long-tailed distributions in the bounded model, Fig. 5.3(b), decay monotonically.

This behavior is rationalized in the right-hand column of Fig. 5.3, where we sketch the dependence of the (effective) rates of the processes affecting the strand length. The crucial effective growth process is the extension reaction, *i.e.*, hybridization of a third strand followed by ligation. The effective rate is denoted by r_{ext}. In the unbounded model, the dehybridization rate r_{off} intersects the horizontal lines corresponding to constant extension and outflux rates at two distinct length scales L^* and $L^\dagger$. This already hints at the two emergent length scales L_{min} and L_{max} in the length distribution. This intersection does not occur for the bounded model.

An analogous argument to Eq. (5.6) for the transition from long to short tails was made in Ref. [187]. There, the authors studied assembly in a model, where strands break by cleavage. The crucial difference between their work and our unbound model is that in their model ligation is always the slowest process.

5.4.2. Estimation of the outflux rate at the transition from short- to long-tailed distributions

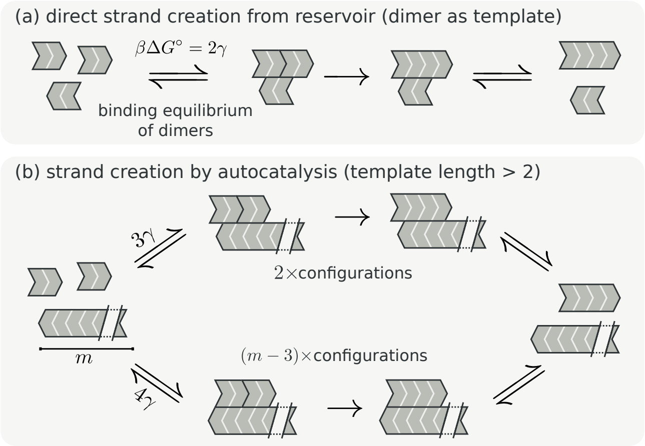

(a) direct strand creation from reservoir (dimer as template)
$\beta\Delta G^\circ = 2\gamma$
binding equilibrium of dimers
(b) strand creation by autocatalysis (template length > 2)
3γ
2×configurations
m
$(m-3)$×configurations
4γ

Figure 5.4.: (a) Formation of a tetramer from the dimer background. A total overlap of two leads to a total binding energy of $\beta\Delta G^\circ = 2\gamma$. (b) Templated ligation of dimers on an m-mer. There are two overhanging configurations with $\beta\Delta G^\circ = 3\gamma$ and $m-3$ configurations with $\beta\Delta G^\circ = 4\gamma$.

The transition from short-tailed to long-tailed distributions occurs when the direct production of long strands from reservoir strands is balanced by the production involving long strands as templates, *cf.* Fig. 5.4. In the following, the corresponding crossover value of the outflux rate r_{out} in the dimer-only model, Eq. (5.6), is derived.

Consider the total concentration $\rho_>$ of strands with a length larger than two, *i.e.*, strands not provided by the reservoir. In a steady state we have

$$0 = \partial_t \rho_> = \phi - \rho_> r_{\text{out}}, \tag{5.7}$$

where ϕ is the concentration flux indicating processes by which $\rho_>$ grows, namely the formation of tetramers from dimers. Notice that the formation of strands with $L \geq 4$ does not change $\rho_>$. In general, this templated ligation can happen in all triplex configurations with two dimers that are adjacently hybridized. Ignoring higher-order complexes, we assume that the dominant contribution to the production of longer strands arises from a ligation reaction happening at triplexes consisting of two dimers and a templating strand of length $L \geq 2$, see Fig. 5.4.

As the hybridization dynamics of dimers are fast, we assume a biding equilibrium. This means that the ratio of the concentration of a triplex and its constituents is determined by its binding energy.

With the elementary rates for hybridization and dehybridization defined in Section 5.3, the binding energy of a complex C is given by

$$\beta \Delta G^{\circ}_{\mathrm{tot}}(C) = \gamma \sum_{i \in C} l_i + \sigma \ln(2), \tag{5.8}$$

where we sum over all hybridization sites. The term $\sigma \ln(2)$ is a "symmetry penalty" that occurs if the complex is rotationally symmetric ($\sigma = 1$) and is zero ($\sigma = 0$) otherwise (see Section 4.6).

Using Eq. (5.8), the ligation flux for triplexes consisting of dimers only is $\phi_2 = (c_2)^3 e^{-2\gamma} r_{\mathrm{lig}}$, see Fig. 5.4 (a). In contrast, the ligation corresponding to templates of length $m > 2$ is

$$\phi_m = (c_2)^2 \left[2e^{-3\gamma} + (m-3)e^{-4\gamma}) \right] c_m r_{\mathrm{lig}}, \tag{5.9}$$

where we took into account the different configurations of the relevant triplexes: see Fig. 5.4(b).

We separate the ligation flux into two components, $\phi = \phi_2 + \phi_>$. The first term, ϕ_2, only involves the building blocks provided by the reservoir. In contrast, the second term, $\phi_> := \sum_{m>2} \phi_m$, involves longer strands. The transition occurs when the latter dominates the former.

Assuming that the length distribution is dominated by single strands, we approximate $\rho_> \approx \sum_{m>2} c_m$, to obtain an expression (lower bound) for $\phi_>$ as

$$\phi_>(\rho_>) \approx (c_2)^2 \left[2e^{-3\gamma} + e^{-4\gamma} \right] r_{\mathrm{lig}} \rho_> .$$

In the stationary situation the balance equation (5.7) is

$$0 \approx \phi_2 + \phi^{\mathrm{c}}_>(\rho_>) - \rho_> r_{\mathrm{out}}, \tag{5.10}$$

which is solved by the crossover value $\rho_> = \rho^{\mathrm{c}}_>$. In this approximation, autocatalysis starts to dominate the production of longer strands from the background when $\phi^{\mathrm{c}}_> = \phi(\rho^{\mathrm{c}}_>) > \phi_2$. In terms of the outflux rate this means that autocatalysis dominates if

$$r_{\mathrm{out}} < r^{\mathrm{c}}_{\mathrm{out}} = 2(c_2)^2 \left[e^{-4\gamma} + 2e^{-3\gamma} \right] r_{\mathrm{lig}}. \tag{5.11}$$

5.4.3. Competition of timescales enables extension cascades and persisting complexes

We now focus on the standard model without effective thermal cycling and with a sufficiently low outflux rate. It already became clear that the nonmonotonic behavior stems from complexes for which dehybridization is not necessarily the fastest process. If the binding energy of a duplex is close to zero, it dehybridizes quickly. In contrast, if the binding energy has a large absolute value, the duplex is stable and the extension with a third strand becomes probable. The extended complex is even more stable and another extension becomes even more probable. We call this phenomenon an *extension cascade*.

Disregarding dehybridization and outflux for now, an extension cascade only stops when no further extension is possible. In our model this is only the case for a fully hyridized duplex consisting of two maximally overlapping strands of the same length. These duplexes persist for long times. The fate of such a long-lived complex is determined by either dehybridization or outflux.

Structure of complexes

We partition complexes into different classes by distinguishing between single strands, duplexes and higher-order complexes, *cf.* Fig. 5.5(a). We further subdivide duplexes according to "parity": Fully hyridized duplexes have zero parity. In contrast, duplexes with even or odd overhangs have even or odd parity. Note that in the dimer-only model, mixed parities are excluded, because all strand lengths are even.

Extension cascades only reach a terminal fully hyridized duplex when they start from even duplexes. Duplexes with odd parity will undergo quasi-infinite extension cascades. Figures 5.5(a,b) show the partitioned length distribution. Short strands are mostly single stranded. In contrast, the concentration peak is dominated by fully hyridized duplexes. The effect of quasi-infinite extension cascades is visible in the tail. Higher-order complexes are less abundant and do not contribute significantly to the shape of the distribution.

The minimum at $L = L_{\text{min}}$ is due to the increase of the concentration of fully hyridized duplexes at a characteristic length scale $L^* \lesssim L_{\text{min}}$, which we will derive below. Figure 5.5(c) shows that L^* is the typical length-scale on which duplexes become stable enough for extension cascades to start.

Kinetics of duplexes

Since the dehybridization rate depends on the length of the hybridization site, it connects time scales to length scales. As such, the characteristic scales L^* and $L^\dagger$

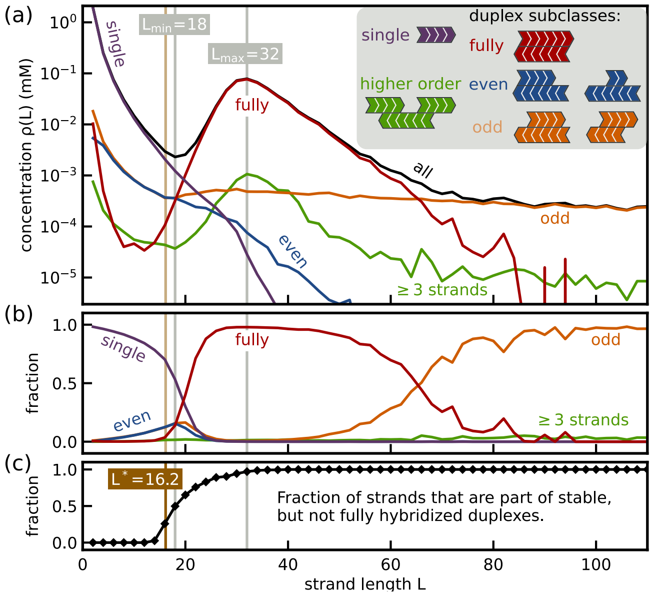


Figure 5.5.: (a) Partitioning the contributions of the different subgroups to the strand-length distribution reveals the dominant configurations: Short strands are mostly single stranded. Strands with lengths around the peaks are in the persistent fully hyridized zero-parity configuration. In the dimer-only model odd duplexes never reach a fully hyridized state and cause the long tail of the distribution. (b) The probability of different complex types conditioned on strand length. (c) The probability that a duplex with nonzero parity is stable conditioned on strand length. Around $L = L^*$ (*cf.* Eq. 5.18) this probability increases rapidly.

also divide the length distribution into different dynamical regimes. Since the length distribution is dominated by single strands and duplexes, we now consider the kinetics of duplexes in detail.

A duplex consisting of strands S_1 and S_2 with lengths L_1 and L_2 is fully characterized by the 3-tuple $D := (L_1, L_2, o_1)$. The number o_1 is the (positive or negative) overhang of strand S_1 on its $3'$ end; see Fig. 5.2(f). When the two strands collide, they can form $\Theta = L_1 + L_2 - 1$ different duplexes. Applying this to Eq. (5.5), the dehybridization rate becomes

$$r_{\text{off}}^{\text{dupl}}(D) = \frac{1}{L_1 + L_2 - 1} e^{\gamma l(D)}. \tag{5.12}$$

First, we formally derive the onset of extension cascades at L^*: Hybridization of a short m-mer can occur on one of the two nonzero overhangs o_i, $i \in (1, 2)$, of the duplex D and results in a triplex T_i. If the m-mer is subsequently ligated to its neighboring strand, we call the combined process an *extension*. In that case, the length of the hybridization site of the m-mer with the duplex is $z_i = \min(|o_i|, m)$. To calculate an effective rate for this process, we assume that the dynamics of the m-mer hybridization is fast compared to the ligation rate and to the dehybridization rate of the duplex D. Consequently, the concentrations of the duplex D and the triplex T_i can be assumed to be at a binding equilibrium and we obtain

$$c_{T_i} = c_D c_m e^{-\gamma z_i}. \tag{5.13}$$

With this, we define the effective extension rate with an m-mer as the ratio of the rate of ligations from that triplex and the duplex concentration, *i.e.*, $r_{\text{ext},m} = r_{\text{lig}} c_T / c_D$. Using Eq. (5.13) and taking into account that there are generally two ligation sites ($i = 1, 2$), the extension rate reads

$$r_{\text{ext},m}(D) = r_{\text{lig}} c_m \sum_{\substack{i \in \{1,2\} \\ o_i \neq 0}} e^{-\gamma z_i}. \tag{5.14}$$

Hence, the extension rate with a short m-mer is

$$r_{\text{ext}}(D) = \sum_m r_{\text{ext},m}(D). \tag{5.15}$$

The ratio of r_{ext} and r_{off} gives the condition for the onset of extension cascades for the duplex D, $1 < r_{\text{ext}}(D)/r_{\text{off}}^{\text{dupl}}(D)$. As dimers are the most abundant species, we approximate $r_{\text{ext}}(D) \gtrsim r_{\text{ext},2}(D)$, yielding the lower bound:

$$1 < \frac{r_{\text{ext},2}(D)}{r_{\text{off}}^{\text{dupl}}(D)}. \tag{5.16}$$

To be more systematic, we now consider a system containing only strand lengths smaller or equal to some fixed value L_0. We then determine the minimal L_0 such that duplexes appear which can undergo extension cascades. Using Eqs. (5.12) and (5.14) we write the ratio in Eq. (5.16) as

$$\frac{r_{\text{ext},2}(D)}{r_{\text{off}}^{\text{dupl}}(D)} = (L_1 + L_2 - 1) c_2 r_{\text{lig}} \sum_{\substack{i \in \{1,2\} \\ o_i \neq 0}} e^{-\gamma(l + z_i)}. \tag{5.17}$$

This ratio is largest for symmetric duplexes with $L_1 = L_2 = L_0$ where $l(D) + z_i = L_0$. The two duplex configurations maximizing the ratio are thus the odd duplex $D_{\pm 1} = (L_0, L_0, \pm 1)$ and the even duplex $D_{\pm 2} = (L_0, L_0, \pm 2)$. The smallest L_0, for which extension cascades are possible, defined as L^*, are found by solving

$$1 = 2(2L^* - 1)c_2 r_{\text{lig}} e^{-\gamma L^*}, \tag{5.18}$$

which yields $L^* \approx 16.2$. As the shortest building blocks are dimers, $L^*_\bullet = \lceil L^* \rceil$ is calculated by ceiling L^* to the next even integer, *i.e.*, $L^*_\bullet = 18$.

For strong binding, *i.e.*, $\gamma < -1$, the subexponential length dependence which enters via the channel number Θ can be neglected. To leading order one then has

$$L^* \approx \ln\left(c_2 \frac{r_{\text{lig}}}{r_0}\right)\gamma^{-1}, \tag{5.19}$$

where we made the dependence of the microscopic kinetic parameter r_0 explicit.

The distinct peak in the strand-length distribution is caused by fully hyridized duplexes $(L, L, 0)$ being end points of extension cascades. These duplexes persist until they dehybridize or leave the system. This gives rise to two different fates depending on their length. For L smaller than some critical value $L^\dagger$, $r_{\text{off}}^{\text{dupl}}(L, L, 0) > r_{\text{out}}$, duplex production in the stationary state is mostly balanced by dehybridization. For long duplexes with $L > L^\dagger$, we have $r_{\text{off}}^{\text{dupl}}(L, L, 0) < r_{\text{out}}$ and hence the stationary concentration is mostly determined by a balance of their production with the outflux. The outflux rate r_{out} is independent of L, whereas r_{off} decreases exponentially with L. We thus expect the existence of two different regimes where the stationary concentration of the fully hyridized duplexes exhibits a different dependence on L. We can find the length where the dehybridization rate becomes smaller than the outflux rate by

$$r_{\text{off}}(L^\dagger, \gamma) = \frac{e^{\gamma L^\dagger}}{2L^\dagger - 1} = r_{\text{out}}. \tag{5.20}$$

Solving this equation numerically for the standard parameters, we obtain $L^\dagger = 30.07$. Ceiling to the next even integer yields $L^\dagger_\blacktriangle = \lceil L^\dagger \rceil = 32$.

As above, we may ignore the logarithmic kinetic dependence on the length for strong binding and obtain

$$L^\dagger \approx \ln\left(\frac{r_{\text{out}}}{r_0}\right)\gamma^{-1}. \tag{5.21}$$

From Fig. 5.5(a) we see that $L^\dagger$ coincides with the position of the maximum L_{max}, whereas L^* serves as a proxy for the position of the minimum L_{min}.

In Section 5.4.5 we perform an extensive screening of the parameter space demonstrating that the analytical estimates Eqs. (5.18) and (5.20) are generally valid. Moreover, the transient behavior of the length distribution in a closed system is discussed in Section 5.4.6. There, the global transient observation time τ_{obs} limits the maximal lifetime of any complex and thus plays the same role as the global degradation timescale r_{off}^{-1} in an open system.

5.4.4. Monomer-dimer mixtures

So far, we have studied systems using dimers as initial building blocks. While this made our analytical considerations easier, only strands of even length appeared in the system. These strands enabled infinite extension cascades starting from duplexes with odd parity. As a result, the length distribution had a heavy tail, see Fig. 5.5(a).

Figure 5.6 shows the length distribution for a reservoir containing monomers and dimers at a total building block concentration of $c_{\text{tot}} = c_1 + c_2 = 2$ mM. The fraction of monomers $f_{\text{m}} := c_1/c_{\text{tot}}$ is set to 70%. Now, infinite extension cascades are suppressed and the long tail collapses. The partitioning of complexes into various substructures shows that fully hyridized duplexes again dominate the tail of the distribution. As above, duplexes with finite overlap are distinguished by the parity of their overhangs, with the addition of mixed parity duplexes, having different parities at the different sites.

The general understanding of the characteristic features of the length distribution developed above remains valid. Repeating the calculations leading to Eq. (5.18) for the onset of extension cascades, with the combined extension rate for both monomers and dimers leads to the same equation, with the dimer concentration c_2 replaced by the total concentration c_{tot}, see Section 5.4.7. The position of the maximum does not depend on the building block concentration, *cf.* Eq. (5.20). Hence, the peak in Fig. 5.6 is roughly at the same position as in the dimer-only system. We confirm the validity of this result by probing different monomer fractions f_{m} in Section 5.4.7.

5.4.5. Exploration of parameter space

To verify that our results of Section 5.4.3 are indeed generic, we performed an extensive parameter sweep. In each row of Fig. 5.7, a single parameter is varied while all the other parameters are fixed at their standard values. The left-handed column in Fig. 5.7 shows simulated stationary length distributions. The right-handed column presents the analytical expressions for the (ceiled) values of L^* and $L^\dagger$ with the characteristic lengths L_{min} and L_{max} from the simulation result. A colored curve in the left-handed panel corresponds to the accordingly colored marker in the right-handed panel. The tails of the distributions are smoothed using a standard running-average smoothing algorithm.

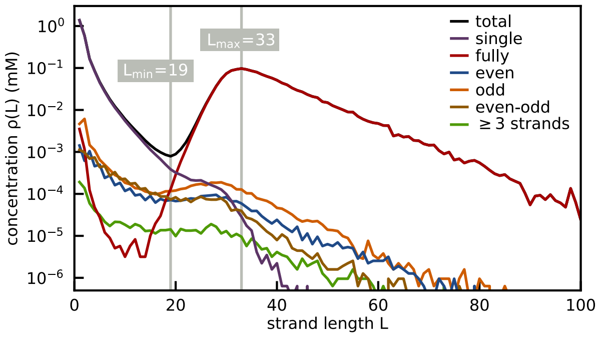


Figure 5.6.: Partitioned length distribution for a system coupled to a reservoir containing monomers and dimers at a monomer fraction of $f_m = 0.7$. In contrast to Fig. 5.5, virtually all strands with $L > L^*$ belong to a fully hyridized duplex.

Figure 5.7(a) shows the result for a variation of the outflux rate r_{out}. The transition from a short- to a long-tailed length distribution was already discussed in Section 5.4.1. As the outflux rate should not influence the onset of extension cascades, we expect the minimum to remain constant, which the simulation confirms. Increasing the outflux rate shifts L_{max} to lower lengths in a logarithmic way in accordance with Eq. (5.21).

In Fig. 5.7(b) we vary the binding energy γ. We observe that increasing the binding energy displaces the characteristic peak toward shorter strands. The behavior of both curves is roughly inverse proportional: $L \propto -\gamma^{-1}$.

Next, we vary the bare ligation rate r_{lig}; see Fig. 5.7(c). The position of the maximum remains unchanged, since the transition determining the fate of a fully hyridized state is not affected by the ligation rate, see Eq. (5.20). In accordance with Eq. (5.18), decreasing r_{lig} logarithmically shifts the onset of extension cascade and the position of the minimum to larger lengths. For the smallest ligation rate plotted, we cross the transition towards short-tailed distributions described in Section 5.4.1, and the characteristic peak in the length distribution disappears.

Figure 5.7(d) shows the effect of varying the dimer concentration c_2. Since reducing c_2 logarithmically reduces the effective rate of extension with a dimer, higher concentrations enable extension cascades already for duplexes consisting of shorter strands, shifting the minimum to the left. Again, the position of the peak remains constant. For the smallest concentration shown we cross the transition toward a short-tailed distribution.

In summary, the phenomenological positions of the minimum L_{min} and the peak L_{max} are well described by the expressions for L^* and $L^\dagger$, Eqs. (5.18) and (5.20).

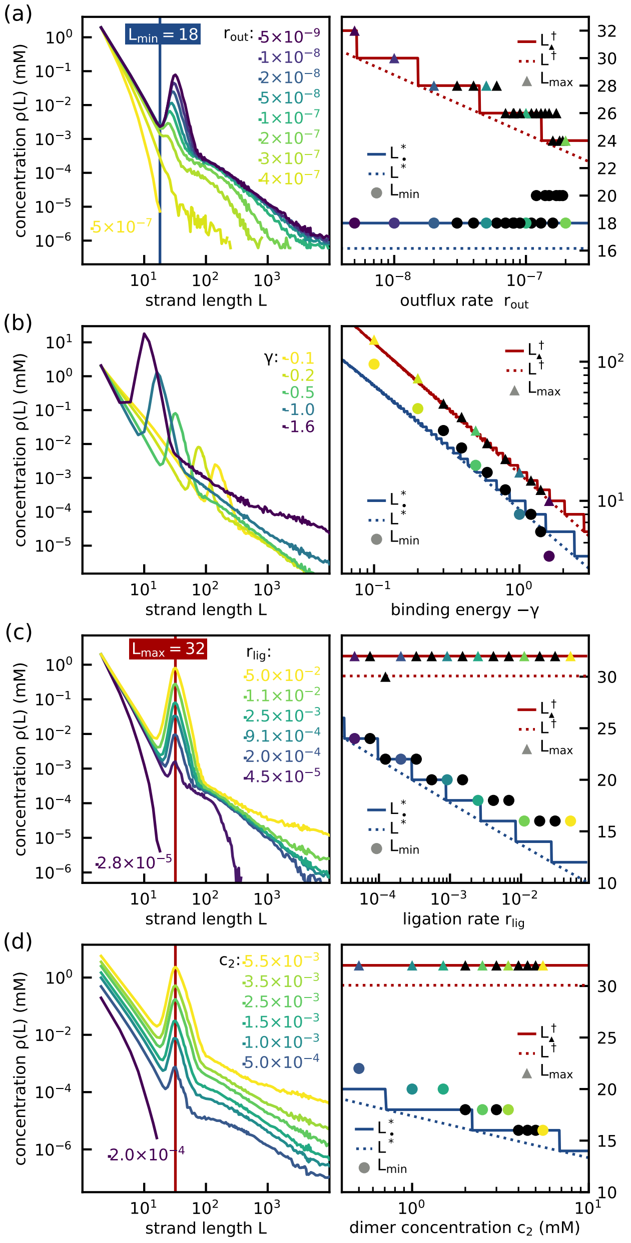


Figure 5.7.: Probing the parameter space of the dimer-only model. Left-handed column: stationary length distributions. Right-handed column: comparison of the observed values L_{min} and L_{max} and the predictions for L^* and $L^\dagger$ via Eqs. (5.18) and (5.20). Variable parameters are (a) the outflux rate r_{out}, (b) the dimensionless binding energy per nucleotide γ, (c) the bare ligation rate r_{lig} and (d) the concentration of chemostated single-stranded dimers c_2.

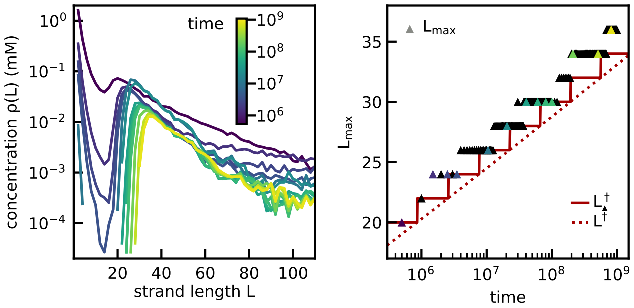


Figure 5.8.: Transient strand distributions: Left: temporal development of the length distribution in a closed system. Over time the concentration of short strands decreases and the minimum develops into depleted region. Right: the position of the maximum L_{max} shifts logarithmically with time toward longer lengths.

5.4.6. Transient behavior in closed systems

Next, we investigate a closed systems without influx or outflux. We prescribed the concentration of initial building blocks and let the system evolve transiently. Because of the irreversibility of the ligation reaction, closed systems are not ergodic: Short building blocks will deplete and the final configuration contains only two very long strands. However, this stationary state will never be reached on practical timescales.

Thus we consider a transient state at intermediate times. We focus on the situation where long strands have already formed, and extension cascades are possible, with still a sufficient amount of short building blocks available. Then, the system behaves similar to the steady state in an open system with small outflux rates.

As in the stationary case, we observe a minimum and maximum in the length distribution. Figure 5.8(a) shows the length distribution for the standard choice of parameters for various values of the transient observation time $t = \tau_{\mathrm{obs}}$. Figure 5.8(b) shows that the position of the maximum increases logarithmically with the observation time.

In order to get an intuition for this behavior, we again use an argument involving the competition of time scales. As in the open systems, strands longer than L^* will dominantly occur in fully hyridized configurations. In contrast, the second time scale is not determined by a global outflux rate and fully hyridized duplexes eventually dehybridize with a length-dependent rate $r_{\mathrm{off}}(L)$.

Yet, dehybridization of duplexes of length L only plays a role for observation times longer than $\tau_{\mathrm{obs}} \sim r_{\mathrm{off}}(L)^{-1}$.

We thus expect the global transient observation time τ_{obs} to play the same role as the time scale r_{off}^{-1} in an open system. The length scale $L = L^\dagger$ that determines the peak in a

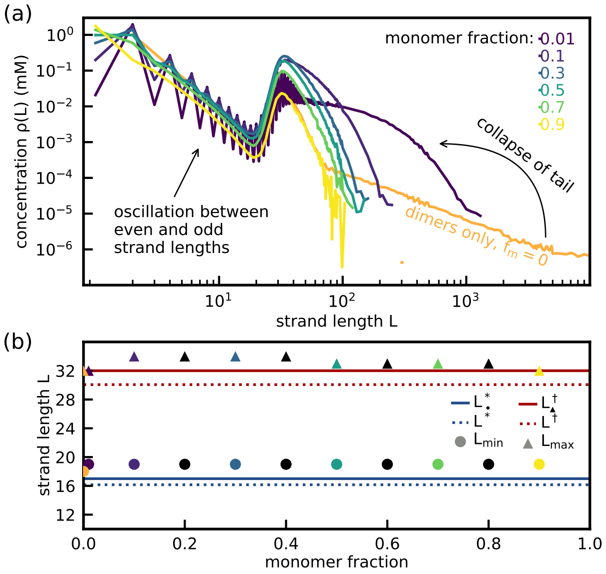


Figure 5.9.: (a) Length distributions for different monomer-dimer mixtures. The monomer fraction f_m is varied between zero and 90 % at a total concentration $c_\mathrm{tot} = 2$ mM. For low f_m the concentration between even and odd strands oscillates heavily for short strands. The long tail that is present for $f_\mathrm{m} = 0$ (orange curve, only even strand lengths shown) collapses even for very small f_m. (b) L^* in the monomer-dimer system is calculated via Eq. (5.25), which is the same as the formula for the dimer-only system upon substituting the dimer concentration with the total concentration c_tot.

closed system can then be obtained by replacing r_off with τ_obs^{-1} in Eq. (5.20) or (5.21). In that case, the position of the peak should increase logarithmically with time, consistent with the results shown in Fig 5.8(b).

5.4.7. Exploring monomer-dimer mixtures

Figure 5.9(a) shows the length distribution for a reservoir where the total initial building block concentration $c_\mathrm{tot} = c_1 + c_2 = 2\,\mathrm{mM}$ is constant. We then vary the monomer fraction $f_\mathrm{m} := \frac{c_1}{c_\mathrm{tot}}$ from zero to 90%. The orange curve is the dimer-only system at standard parameters, showing the long tail caused by the infinite extension cascades. For any finite monomer concentration, infinite extension cascades are suppressed and the long tail collapses.

The length distributions for finite monomer fractions look qualitatively similar. The larger f_m, the less nucleotide mass is added by the influx, and the lower the concentrations. The lower the monomer concentration, the more of the bias toward strands of even lengths is retained. The bias dominates for short strands, leading to the zigzag pattern visible in Fig. 5.6(a). For long strands, the bias vanishes. In accordance with Eq. (5.20), the position of the maximum is unchanged, as it does not depend on the building block concentration.

The minimum position, *i.e.*, the typical length L^* for the onset of extension cascades, is derived analogously to the dimer-only model using the condition $1 < r_{\text{ext}}(D)/r_{\text{off}}^{\text{dupl}}(D)$, *cf.* Section 5.4.3. Instead of considering the extension with a dimer only, one needs to include the extension with a monomer. The extension rate is

$$\begin{aligned} r_{\text{ext}}(D) &\approx r_{\text{ext},1} + r_{\text{ext},2} \\ &= r_{\text{lig}} \sum_{\substack{i\in\{1,2\} \\ o_i \neq 0}} \left[c_2 e^{-\gamma[\min(|o_i|,2)]} + c_1 e^{-\gamma[\min(|o_i|,1)]} \right]. \end{aligned} \tag{5.22}$$

The criterion for extension cascades then reads

$$\begin{aligned} 1 \leq &(L_1 + L_2 - 1) r_{\text{lig}} \times \\ &\sum_{\substack{i\in\{1,2\} \\ o_i \neq 0}} \left[c_2 e^{-\gamma[l+\min(|o_i|,2)]} + c_1 e^{-\gamma[l+\min(|o_i|,1)]} \right]. \end{aligned} \tag{5.23}$$

The right-hand side of the Eq. (5.23) is maximal for the odd duplex configuration $D_{\pm 1} = (L_0, L_0, \pm 1)$, for which $l + \min(|o_i|, 2) = l + \min(|o_i|, 1) = L_0$, which leads to

$$1 \leq 2(2L_0 - 1) r_{\text{lig}} (c_2 + c_1) e^{-\gamma L_0}. \tag{5.24}$$

Consequently, L^* for monomer-dimer mixtures obeys

$$1 = 2(2L^* - 1) r_{\text{lig}} c_{\text{tot}} e^{-\gamma L^*}, \tag{5.25}$$

where $c_{\text{tot}} = c_1 + c_2$ is the total concentration of building blocks.

Equation (5.25) is the same formula as for the dimer only-system, except that the dimer concentration c_2 is substituted by the total concentration of building blocks c_{tot}. In accordance with formula Eq. (5.25), we observe that the position of the minimum is constant $L_{\text{min}} = 19$ under variation of the monomer fraction while keeping the total concentration fixed at $c_{\text{tot}} = 2$ mM, see Fig. 5.9(b).

5.4.8. Growth of complexes

The length scale $L^\dagger$ relates the dehybridization to the outflux timescale. However, its role in the dynamics is not straightforward. We now show that $L^\dagger$ is a typical scale where self-enhancing processes leading to the growth of strands and complexes break down.

Trajectories of stable duplexes

In what follows, we investigate the trajectories of extension cascades starting from stable duplexes until they leave the system in a fully hyridized configuration. We sample trajectories from the steady state of the monomer-dimer system with a monomer fraction of $f_\mathrm{m} = 70\,\%$, see Fig. 5.6. Our sampling algorithm is consistent with the events that occur in a steady state and is explained in Appendix B.1.

An initial stable duplex consists of a long strand of size L_long and a short strand of size $L_\mathrm{short} \leq L_\mathrm{long}$. It has an overlap l_initial and a length $C_\mathrm{initial} = L_\mathrm{long} + L_\mathrm{short} - l_\mathrm{initial}$ *cf.* Fig. 5.10(a). These stable duplexes are the starting point for extension cascades and eventually become fully hyridized duplexes of length $C_\mathrm{final} \geq C_\mathrm{initial}$. If the length of the initial complex is the same as that of the final complex, $C_\mathrm{final} = C_\mathrm{initial}$, no extension occurs beyond the length of the original duplex and we speak of *pure primer extension*. In contrast, if $C_\mathrm{final} > C_\mathrm{initial}$, processes occurred that extended the length of the initial duplex and we speak of *duplex extension*. A detailed look at extension cascades involving duplex extension can be found in Section 5.4.8.

Fig. 5.10(b) shows the joint probability distribution $p(C_\mathrm{initial}, l_\mathrm{initial})$. It is maximal for $C_\mathrm{initial} \sim L^\dagger$ and $l_\mathrm{initial} \sim L^*$. The accumulation of probability at this point characterizes a *typical initial configuration*, but does not determine the length of the individual strands.

Figure 5.10(c) shows $p(L_\mathrm{long}, L_\mathrm{short})$. We see that it is restricted to the lower triangle defined by $L^* \lesssim L \lesssim L^\dagger$. The boundaries of this region reflect our analysis above: Strands shorter than L^* typically do not bind strongly enough to start extension cascades. In contrast, strands longer than $L^\dagger$ are mostly double stranded and thus not available to form the initial duplexes. The approximately uniform behavior of the distribution in that region indicates that no particular combination of strand lengths is preferred.

Next, we consider the length C_final of the fully hyridized duplex that marks the end of an extension cascade. Figure 5.10(d) shows the joint probability distribution $p(C_\mathrm{final}, C_\mathrm{initial})$. The diagonal line $C_\mathrm{initial} = C_\mathrm{final}$ indicates pure primer extension and has a total weight of ($\sim 17\,\%$). The point $C_\mathrm{final} = C_\mathrm{initial} = 31 \sim L^\dagger$ has the maximal individual weight ($\sim 2.5\,\%$).

Finally, Fig. 5.10(e) shows $p(C_\mathrm{final}, L_\mathrm{long})$. Purely autocatalytic processes, where the long strand facilitates the formation of itself, are contained on the diagonal line $L_\mathrm{long} = C_\mathrm{final}$. These autocatalytic trajectories constitute a fraction of about 2.5 % of all extension cascades.

Autocatalytic trajectories ($L_\mathrm{long} = C_\mathrm{final}$) are a subset of trajectories with pure primer

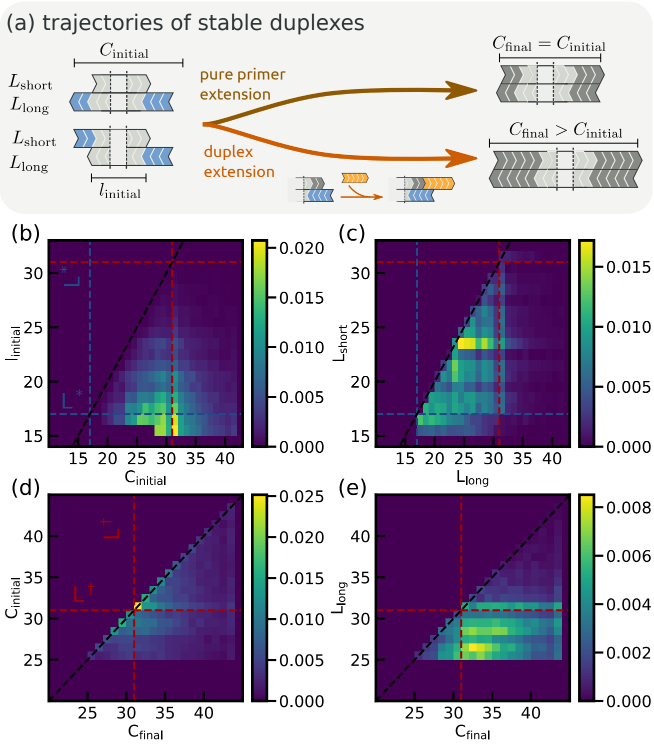


Figure 5.10.: (a) Trajectories start with an initial stable duplex characterized by its strand lengths L_{long} and L_{short} together with the initial overlap l_{initial} and complex length C_{initial}. Trajectories not creating new single-stranded regions beyond the initial complex are referred to as "pure primer extension". In contrast, duplex extension leads to a final complex with $C_{\mathrm{final}} > C_{\mathrm{initial}}$. (b–e) Trajectory statistics can be understood from various joint probability distributions, where $L^* \sim 17$ and $L^\dagger \sim 31$. See main text for details

extension ($C_{\mathrm{initial}} = C_{\mathrm{final}}$). Most extension cascades, however, lead to fully hyridized duplexes that are longer than either of the two strands of the initial complex. In order to emphasize the cooperative effects, we refer to this more general process as heterocatalytic growth.

Catalytic growth and reassembly

Auto- and hetero-catalytic *cycles* are formed by combining extension cascades with a dehybridization of the final duplex and eventual reassembly. Fig. 5.11(a) illustrates an autocatalytic cycle, while Fig. 5.11(b) shows heterocatalytic growth. Note that in general heterocatalytic processes involve duplex extensions.

In the following we investigate how such catalytic cycles shape the strongly nonequi-

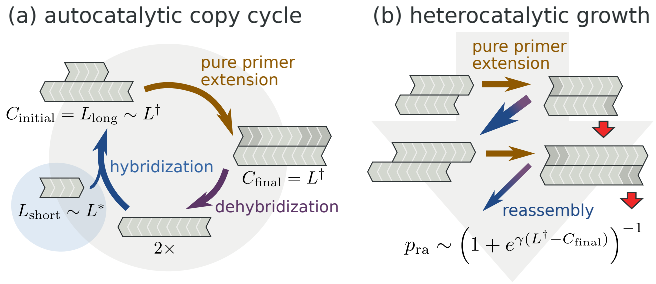

(a) autocatalytic copy cycle
(b) heterocatalytic growth
pure primer extension
$C_{\text{initial}} = L_{\text{long}} \sim L^{\dagger}$
hybridization
$C_{\text{final}} = L^{\dagger}$
$L_{\text{short}} \sim L^{*}$
dehybridization
2×
pure primer extension
reassembly
$p_{\text{ra}} \sim \left(1 + e^{\gamma(L^{\dagger} - C_{\text{final}})}\right)^{-1}$

Figure 5.11.: Heterocatalytic (a) and autocatalytic (b) processes for the growth of strands. In the strongly nonequilibrium regime, extension cascades cover the available overhang of stable duplex and form longer fully hyridized strands. These long strands can then dehybridize and reassemble, thus creating new overhangs to be covered by extension cascades. The reassembly probability p_{ra} is determined by the balance between dehybridization and outflux and decays to zero fast for $L \gtrsim L^{\dagger}$.

librium regime $L^{*} \leq L \leq L^{\dagger}$ of the strand-length distribution: After a fully hyridized duplex is reached at the end of an extension cascade it will dehybridize or leave the system. If it dehybridizes, it may hybridize to another single strand and create a new stable duplex with a new overhang. By this *reassembly*, long strands catalyze the formation of other long strands. The reassembly probability p_{ra} is mostly determined by the competition between outflux and dehybridization. A sigmoidal dependence on the length C_{final} follows:

$$p_{\text{ra}} \sim \frac{r_{\text{off}}}{r_{\text{off}} + r_{\text{out}}} \sim \left(1 + e^{\gamma(L^{\dagger} - C_{\text{final}})}\right)^{-1}. \tag{5.26}$$

The reassembly probablity p_{ra} decays exponentially with the C_{final}. Thus, the production rate of longer strands $C_{\text{final}} \sim L^{\dagger}$ is drastically reduced.

In summary, the strongly nonequilibrium catalytic strand growth is constrained to strand length between L^{*} and $L^{\dagger}$. It is this self-enhancing dynamical behavior, which leads to the increased production of strands with lengths $L^{*} < L < L^{\dagger}$. This effect directly yields a region in the strand length distribution, where concentration increases with length. To the right of the peak at $L \sim L^{\dagger}$, catalytic cycles producing longer strands are too slow in order to compete with the outflux rate r_{out}. Similarly, in the analogous transient situation, the observation time τ_{obs} is too short to allow the catalytic production of strands beyond a certain length.

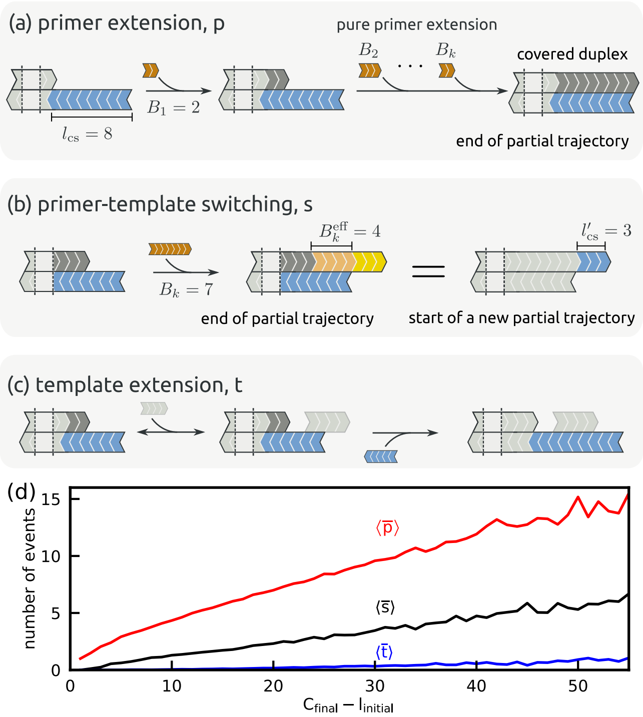


Figure 5.12.: Pure primer extension (a) versus duplex extension (b),(c). The overhang at the beginning of a (partial) trajectory is called a copy site (blue) with length l_{cs}. (b) In primer-template switching events a building block extends the primer beyond the original copy site. The original copy site is fully covered and a new copy site is formed. The roles of primer and template are exchanged. (c) Copy sites can grow independently of the original primer by template extension with the help of a helper strand. (d) The number of extension events occurring during the covering of the total copy site $C_{final} - l_{initial}$ split according to different types of extensions.

Beyond pure primer extension

Figure 5.10(a) already depicted an example of an extension step leading to the growth of a complex beyond its initial length. In the following, we take a more detailed look at this phenomenon.

Growth happens essentially independently at each end of a duplex. It thus makes sense to take the perspective of a single end, since it allows us to distinguish the two strands by their roles: We call the strand whose end is overhanging the template, whereas the other strand is called the primer. Moreover, we refer to the length of the overhang at the start of a trajectory as its initial copy site length l_{cs}.

The obvious mechanism that leads to duplex extension is depicted in in Fig. 5.12(b).

It occurs when the original primer is extended with a strand that is longer than the (remaining) length of the copy site. After this extension, the roles of primer and template are reversed and a new copy site is created. We thus denote this process as *primer-template switching*.

A complex undergoing an extension cascade is not always a simple duplex. Ligation reactions can also occur away from the stable hybridization site. We say that *template extension* occurs, if another strand facilitates the extension of the template strand, see Fig. 5.12(c). From the perspective of the stable hybridization site, the length of its associated copy site l_{cs} has increased.

Figure. 5.12(d) shows the number of extension events along a trajectory as a function of the total single-stranded length that is covered during the trajectory, $C_{\mathrm{final}} - l_{\mathrm{initial}}$. The standard primer-extension steps (p, red curve) are most common. In contrast, template extension (t, blue curve) is rare. For large $C_{\mathrm{final}} - l_{\mathrm{initial}}$, the number of events behaves strictly linear and primer-template switching (s, black curve) is approximately three times less likely than primer extension. For small values of $C_{\mathrm{final}} - l_{\mathrm{initial}}$, the relative fraction of primer-template switching increases, since a short available overhang increases the chance of primer-template switching.

5.5. Experimental system

To test our theory experimentally, we used DNA strands of length $L_{\mathrm{bb}} = 12$ as basic building blocks in a closed volume. The strands have random sequences drawn from a binary alphabet of A (adenine) and T (thymine). As discussed above and in Section 5.4.6, in closed systems the global transient observation time τ_{obs} plays the same role as r_{off}^{-1} in an open system.

Enzyme-free templated ligation is slow and not compatible with experimental timescales [59]. Ligases speed up the assembly process, but require the formation of complexes involving at least three strands with $L \gtrapprox 12$ and the ligase. The probability of finding such complexes decreases with temperature. Further, the ligase activity itself is temperature dependent, resulting in a nontrivial temperature dependence of the effective extension rate. In isothermal systems, one may encounter a stalemate situation. For high temperatures, the extension rate is small since the formation of the required complexes is thermally suppressed. In contrast, for low temperatures, the dehybridization rate is small and the system is effectively frozen. This stalemate can be resolved using temperature cycles [59, 58]: During the cold phase, the extension rate is initially high until virtually all possible ligations in existing complexes have occurred. Hence, the hot phase is required to create new ligatable complexes. However, temperatures in the hot phase must still be such that the binding energy γ remains

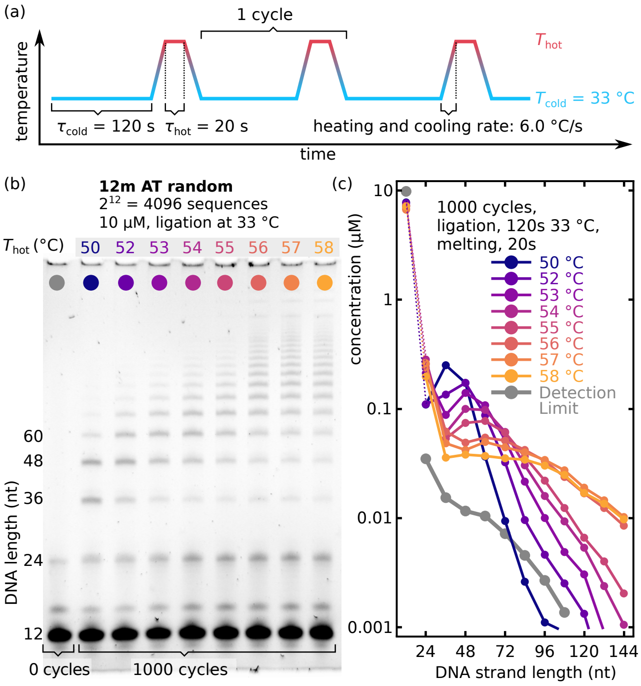


Figure 5.13.: Product concentration analysis for a 12 nt random sequence AT-only pool. (a) Experimental temperature profile. Ligation occurs for 120 s at 33 °C after which the sample is heated to the variable hot reassembly temperature T_{hot} for 20 s. (b) Image of a polyacrylamide gel electrophoresis.The first lane on the left shows the "baseline" sample, which is similar to the other lanes but was not subjected to temperature cycling. The other lanes have the same ligation conditions but different temperatures for dissociation. (c) Quantitative results for the length distribution. Non-monotonic length distributions with a maximum and minimum were observed.

negative. Only then is the binding energy still proportional to the overlap length, such that the competition of timescales gives rise to a nonmonotonic length distribution.

5.5.1. Experimental method and results

Our experiment was performed using a TAQ DNA ligase from *NEB* and a *ThermoFisher* ProFlex PCR system to generate the temperature profile shown in Fig. 5.13(a). This setup is similar to the setup used in [58]. The analysis of the length distributions is done by running the sample in a polyacrylamide gel electrophoresis (PAGE), post staining the DNA with intercalating SYBR gold dye and taking fluorescent images of the gel in a *BioRad* ChemiDoc MP. Concentration quantification is performed with a custom software extracting the lane intensity from gel CCD images (see Appendix B.2). The bands visible at lengths of 16 nt and 24 nt for all lanes in Fig. 5.13(b) are artifacts from the ligation buffer and DNA synthesis, respectively.

We analyzed the length distribution for various observation times τ_{obs} for different isothermal conditions and cycling scenarios, where temperature alternated between $T_{\text{cold}} = 33\,^{\circ}\text{C}$ at variable temperature T_{hot}, *cf.* Fig. 5.13(a). Isothermal experiments resulted in no product formation within 60 and 116.5 h, *cf.* Appendix B.2. For low temperatures, even short duplexes with strands of length L_{bb} cannot separate. For high temperatures, the extension is suppressed because no stable ligatable complexes are formed.

Cyclic conditions led to different product distributions, shown in Fig. 5.13(b). The length distribution decayed quickly for $T_{\text{hot}} = 50\,^{\circ}\text{C}$, while it decayed slowly for $T_{\text{hot}} = 58\,^{\circ}\text{C}$. All length distributions showed a nonmonotonic behavior exhibiting a local minimum L_{min} between 24 and 48 nt and a maximum L_{max} between 36 and 72 nt. For higher dissociation temperatures the peak was found to be flatter and wider. The shape of the distribution changed significantly in a limited range for T_{hot}.

5.5.2. Effective theory

In order to understand the behavior of the experimental system when varying the temperature T_{hot} in the hot phase, we consider the thermodynamics of the standard Gibbs free energy ΔG°. It enters our theory as the central temperature-dependent binding parameter $\Delta G^{\circ}/(k_{-1}T)$. To leading order in T, the Gibbs energy can be written as $\Delta G^{\circ} = \Delta H^{\circ} - T\Delta S^{\circ}$, where the standard enthalpy ΔH° and standard entropy ΔS° are temperature-independent microscopic parameters [107, 76] (see also Section 2.1.3 in Chapter 2).

The most significant effects occur when the binding energy changes sign at the critical temperature $T_{\text{c}} = \Delta H^{\circ}/\Delta S^{\circ}$. Assuming that ΔH° and ΔS° scale linearly with the length of the hybridization site, T_{c} is independent of length. For our experiment, we estimated the expected value of T_{c} between 60 °C and 75 °C (see Appendix B.2).

At positive binding energy, *i.e.*, for $T_{\text{hot}} > T_{\text{c}}$, strands of all lengths dissociate quickly in the hot phase. We are then in a situation akin to the bounded model discussed in Section 5.4.1. In order to observe a nonmonotonic distribution, the dehybridization (and thus the reassembly) rate in the hot phase must decay exponentially with strand length, which requires $T_{\text{hot}} < T_{\text{c}}$. For our effective theory we employ a linear expansion of the binding parameter γ below the critical temperature T_{c}:

$$\gamma(T) = \frac{\Delta G_1^{\circ}}{k_{-1}T} = \frac{T - T_{\text{c}}}{\xi}. \tag{5.27}$$

In this formula, ΔG_1° is a typical binding energy per nucleotide. The parameter ξ has

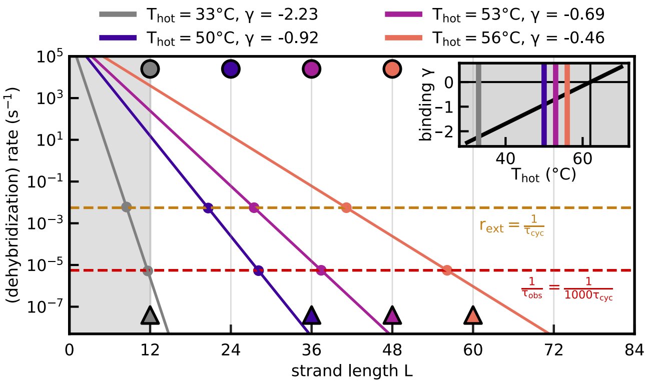


Figure 5.14.: Temperature dependence of the emerging length scales in the effective theory. Solid lines: dehybridization rate $r_{\text{off}} = r_0 e^{\gamma L}$ for various values of T_{hot}. The inset shows the linear function $\gamma(T)$, Eq. (5.27), with $T_c = 62\,°\text{C}$ and $\xi = 13$ K. Horizontal dashed lines denote the effective extension rate τ_{cyc}^{-1} and the inverse observation time τ_{obs}^{-1}. As in Fig. 5.3(a), the competition of timescales determines the scales L^* and $L^\dagger$ as the intersection between the solid and dashed lines. They are mapped to the observable length scales L_{min} (circles) and L_{max} (triangles) by ceiling the intersection point to the next multiple of $L_{\text{bb}} = 12$. By approaching T_c, the binding energy and thus the slope become smaller in magnitude, and the intersection points move to larger lengths.

units of temperature and characterizes the (inverse) slope of $\gamma(T)$ around T_c. From $\Delta G_1^\circ = \Delta H_1^\circ - T\Delta S_1^\circ$ it follows that $\xi = -k_{-1}T_c^2/\Delta H_1^\circ \approx 30$ K for typical enthalpies and entropies (see Appendix B.2).

Unlike T_c, the parameter ξ is inversely proportional to ΔH_1° and thus depends on strand length. Our simple model is based on effective binding energies of (self-complementary) nucleotides. It is therefore questionable whether the value of $\xi \approx 30$ K obtained from standard libraries for matching nucleotides is appropriate here. Since hybridization sites also contain mismatches, the correct value of ξ describing the experimental behavior is likely smaller. Under the assumption that a nucleotide has a probability of 1/2 to encounter its complement, an adjusted value of $\xi \approx 15$ K seems reasonable. Finally, for a full parametrization of the dehybridization rate $r_{\text{off}}(L; T_{\text{hot}}) \sim r_0 \exp[\gamma(T_{\text{hot}})\, L]$, we need to specify the collision rate r_0. While the exact value of this rate depends on microscopic details, experimental evidence suggest that a value of $r_0 = 10^6\ \text{s}^{-1}$ is reasonable [173, 172, 171].

Figure 5.14 shows the length-dependence of the dehybridization rate r_{off} for various values of T_{hot} below $T_c = 62\,°\text{C}$ and $\xi = 13$ K. One should recall that the extension rate r_{ext} is the effective rate at which a duplex binds to a third strand and subsequently ligates. As explained initially, extensions are likely to happen only in the cold phase. Because of frustrated dehybridization, we expect a single extension per duplex per cycle. Consequently, the extension rate determining L^* is given by $r_{\mathrm{ext}} \sim \tau_{\mathrm{cyc}}^{-1}$. In transient systems without outflux, the inverse observation time $\tau_{\mathrm{obs}}^{-1} = N_{\mathrm{cyc}}\tau_{\mathrm{cyc}}$ replaces r_{out} in determining $L^\dagger$.

For $\tau_{\mathrm{cycle}} = 180$ s and $N_{\mathrm{cyc}} = 1000$, we obtain the two horizontal lines in Fig. 5.14. The intersections with the length-dependent dehybridization rate determines the scales L^* and $L^\dagger$ as a function of T_{hot}. The big dots and triangles denote the values $L^*_\bullet$ and $L^\dagger_\blacktriangle$ obtained by ceiling to the next integer multiple of L_{bb}.

We observe that the scales L^* and $L^\dagger$ shown in Fig. 5.14 agree well with the experimental observations for $L_{\min}$ and $L_{\max}$ shown in Fig. 5.13. The exact values of L^* and $L^\dagger$ depend on the exact values for the parameters r_0, ξ and T_{c}, for which we used reasonable estimates. As such, they should not be confused with rigorous predictions. Nonetheless, both the order of magnitude and the qualitative dependence of experimental results on T_{hot} are fully captured by our effective theory.

However, we expect the effective theory to break down close to the critical temperature T_{c}. For small γ, the contributions to the dehybridization rate due to microscopic details play a more prominent role. Further, when approaching T_{c}, the characteristic features of the distribution shift to larger lengths. Then, both the experimental timescales and the overall oligonucleotide mass become limiting and the depletion of building blocks starts to play a role. Moreover, the quantitative evaluation of the gel plots is more difficult for long strands, *cf.* Fig. 5.13(a). For this case, other effects, like self-folding of strands, may be an important mechanism that is absent from the theory, *cf.* Ref. [58].

5.6. Discussion

5.6.1. Summary

Since major transitions in evolution appear to have occurred when smaller entities were coming together to form larger ones [217], a multi step scenario toward increased complexity also seems natural in a prebiotic context. While the importance of templated ligation in this scenario is clear [218], the assembly dynamics emerging from this interaction of short building blocks were not fully understood previously.

Using a minimal bottom-up model, we showed that a nonmonotonic length distribution arises from the competition of three timescales, or equivalently, the corresponding rates:

1. The dehybridization rate r_{off} which decreases exponentially upon increasing strand length L, with the decay determined by the binding parameter per nucleotide γ.

2. An effective extension rate r_{ext} of a duplex, which is determined by the ligation rate r_{lig}, γ and the concentrations of building blocks.
3. A global timescale determined by the outflux rate r_{out} or an observational time τ_{obs}.

The competition between r_{ext} and r_{out} determines whether we see long-tailed distributions at all: If r_{out} is larger than r_{ext}, ligations are unlikely. The competition between r_{off} and r_{ext} leads to the emergence of extension cascades at a typical length scale L^*: As soon as strands in a hybridization complex have a length such that $r_{\mathrm{ext}} > r_{\mathrm{off}}$, they undergo extension cascades resulting in persistent configurations, which cannot extend further. The fate of such a configuration is determined by the competition between r_{off} and r_{out}: Fully hyridized duplexes shorter than $L^\dagger$ dehybridize before leaving the system. The single strands released in this way subsequently act as templates in newly formed primer-template complexes and thus catalyze further strand growth.

The combination of extension cascades and reassembly represents (auto or hetero)catalytic cycles producing longer strands from shorter building blocks. In the strand-length distribution, this strongly nonequilibrium regime is visible as an increase in concentration with length. Extension cascades are fast. Therefore, the dehybridization time of the fully hy-ridized duplex at the end of the cascade determines the completion-time scale of these cycles. For strand lengths where this timescale becomes comparable to transient or global degradation times, these catalytic cycles have no time to complete, and the length distribution decays.

The validity of this scenario was revealed using a state-of-the-art simulation. To our knowledge, no comparable simulation is available to this date. As our experiments demonstrate, the emerging length scales can be tuned by changing environmental parameters such as the melting temperature T_{hot} without changing the chemistry. On early Earth, strands of a characteristic scale $L^\dagger$ emerging from the self-assembly could act as building blocks of a higher level of organization. Moreover, length-dependent accumulation of such strands might trigger novel effects like phase transitions [219, 220, 162, 29].

5.6.2. Advantages and shortcomings of our model

Our minimal model allowed us to reveal universal features of the self-assembly process and to derive analytical expressions for the emerging length scales. Yet, there are several aspects that our model does not capture. It does not allow for secondary structures like hairpins and other "nonproductive" configurations [79, 58]. While these (potentially functional) structures will likely be important in later stages of evolution, in the current scenario they probably have the same role as fully hyridized duplexes.

The strongest simplification in this study is the negligence of any explicit sequence dependence for the binding energy, *i.e.*, the use of self-complementary nucleotides. However, an effective self-complementary description of hybridization arises naturally in a mean-field “random sequence approximation” [187]. It assumes that differences in binding energies between the (complementary and noncomplementary) nucleotide pairs are small with respect to the average binding energy γ per nucleotide.

While this scenario constitutes the extreme of vanishing sequence selectivity, a comparable situation arises in the other limit of perfect selectivity. There, only fully complementary strands bind, and a similar description is achieved using a combinatorial factor for each nucleotide, which can be incorporated by a reduction of γ or using effective concentrations (see also Section 4.7.4). During extension cascades, the incorporation of mismatches is suppressed, since building blocks matching the template bind stronger and thus lead to higher extension rates. Additional stalling effects for nonmatching short building blocks likely enhance this effect. [77, 214, 215]. Moreover, since ligation reactions are irreversible in our model, it corresponds to the high dissipation limit of sequence copying, which generally increases fidelity, *cf.* Ref. [221]. Consequently, we expect that in a fully sequence-dependent model, where mismatches are penalized, the maximum in the length distribution is still present and caused by fully hyridized duplexes with comparatively few errors.

5.6.3. Outlook

Our study provides a first step toward understanding the emergence of structure in a kinetically and thermodynamically consistent bottom-up approach. Extending our algorithm to include sequence-dependent parameters can be a starting point for future studies.

In any explicitly sequence-dependent model, the timescales involved in the extension-reassembly process would include the sequence of the strand in addition to its length [59]. In particular, such a model extension would allow for a more direct study of evolutionary processes in sequence space. As discussed above, sequence selectivity and thus replication may arise during extension cascades. The combined heterocatalytic and autocatalytic nature of the assembly process emphasizes the importance of cooperation, *cf.* Section 5.4.8. Therefore, model extensions could also provide a testing ground for abstract frameworks involving catalytic networks [222, 195, 223].

6. Thermodynamic and kinetic sequence selection in enzyme-free polymer self-assembly*

The RNA world is one of the principal hypotheses to explain the emergence of living systems on the prebiotic Earth. It states that RNA oligonucleotides acted as both the carriers of genetic information as well as catalytic molecules, promoting their own replication. However, the RNA world does not explain the origin of the functional RNA molecules, called ribozymes, in the first place. How did the transition from the pre-RNA to the RNA world occur? A starting point to answer this question is to unveil the dynamics in sequence space on the lowest level, i.e., where mononucleotide and short oligonucleotides come together and collectively evolve into larger molecules. In the second self-assembly scenario discussed in Chapter 5, the sequences of oligonucleotides were treated in a mean-field picture. In this chapter, we now make the sequence dependence of the self-assembly process explicit and study the self-assembly of polymers from a random initial pool of short binary building blocks via templated ligation. Templated ligation requires two strands that are hybridized adjacently on a third strand. The stability of such a configuration crucially depends on the sequence context and, therefore, significantly influences the reaction probability. Moreover, non-complementary nucleotide pairs in the vicinity of the ligation site stall the formation of new covalent bonds. These thermodynamic and kinetic aspects are explicitly considered in our stochastic approach, building on a nearest-neighbor energy model. Based on this model, we investigate the system-level dynamics inside a non-equilibrium RNA reactor enabling a fast chemical activation of the termini of interacting oligomers. Moreover, the RNA reactor subjects the oligomer pool to periodic temperature changes inducing the repeated reshuffling of the system. The binding stability of strands typically grows with the number of complementary nucleotides forming the hybridization site. While shorter strands unbind spontaneously during the cold phase, larger complexes only disassemble during the temperature peaks. Inside

*The chapter is adapted from the manuscript: "Thermodynamic and kinetic sequence selection in enzyme-free polymer self-assembly inside a non-equilibrium RNA reactor", by Tobias Göppel, Joachim H. Roesenberger, Bernhard Altaner and Ulrich Gerland, which intended to be submitted for publication in *Life* in February 2022. The author of this thesis is the only first author of the manuscript.

the RNA reactor, strand growth is balanced by cleavage via hydrolysis such that the oligomer pool eventually reaches a non-equilibrium stationary state characterized by its length and sequence distribution. How do motif-dependent energy and stalling parameters affect the sequence composition of the pool of long strands? As a critical requirement for self-enhancing sequence selection, we identify kinetic stalling due to non-complementary base pairs at the ligation site. Kinetic stalling enables cascades of self-amplification that result in a strong reduction of occupied states in sequence space. Moreover, we discuss the significance of the symmetry breaking for the transition from a pre-RNA to an RNA World and the origins of life.

6.1. Introduction

Modern biology separates between two roles in information transfer: DNA and RNA store the genetic information, whereas proteins carry out the encoded functions and, in particular, replicate the genetic information [30]. The blueprints for the proteins are in turn stored in the DNA and RNA. Inevitably, the question arises, which came first, the polymer carrying the instructions for the proteins or the proteins assembling the polymers? As RNA can not only store the genetic information but also fold into catalytically active structures [39, 40, 224, 41, 225], it stands at the core of one of the most prominent hypotheses for the emergence of living systems – the so called *RNA world* [31, 32, 33, 34, 35, 36]. It proposes that RNA oligonucleotides acted as both the carriers of information as well as functional molecules, promoting their own replication. Hence, the RNA world provides an elegant solution to the dilemma raised above. Yet, the RNA world does not explain the origin of the catalytic RNA molecules, called ribozymes [11]. While recent experimental work revealed potential prebiotic pathways to synthesize nucleotides [55, 56, 57], the mechanisms assembling these building blocks into functional molecules remain still unclear [42, 61, 59, 28, 62, 68, 63]. According to Refs. [64, 53, 65], the smallest ribozymes known today are approximately 30 to 100 nucleotides long. More complex ribozymes that could, e.g., assist replication are likely to have a minimum length of more than 150 nucleotides [46, 66, 67]. For polymers of a length between 30 to 150, a total of 10^{18} to 10^{90} distinct sequences is possible. However, the subset of catalytically active sequences is generally believed to be marginal [58]. Therefore, the spontaneous emergence of a functional sequence from a random pool of nucleotides seems highly unlikely, and the question arises: How did the transition from the *pre-RNA* to the RNA world occur?

To answer this question, one must first unveil the dynamics in sequence space on the lowest level, i.e., when mononucleotides and short oligonucleotides inside a reaction volume collectively self-assemble into longer strands. [2, 12]. The self-assembly is governed by a multi-step process called *templated ligation* [218, 60, 78, 197, 167]. In this process, two strands that are hybridized adjacently on a third *template* strand become joined covalently. In contrast to random ligation concatenating two arbitrary strands, the process of templated ligation is sequence-selective in two-fold ways. First, the (de)hybridization dynamics on the template strand is sequence-selective: Complementary nucleotide pairs at the hybridization sites increase the stability of the complex of strands and accordingly also increase the probability for a new covalent bond to be formed. Moreover, complementary hybridization sites of the same length comprising different sequence motifs can have different binding stabilities. The stability of a hybridization site depends on the hybridization energy. The hybridization energy is mainly determined by the stacking interactions of neighboring nucleotide pairs [226]. Changing their order or flipping one of the pairs generally leads to new stacking interactions, changes the energy and thus alters the complex's stability. Hence, certain sequence motifs might be favored over others thermodynamically [227, 107, 76, 113]. Second, the ligation step is motif-selective: Non-complementary nucleotide pairs in the vicinity of the ligation site stall the formation of a new covalent bond. As a result, the formation of new strands from shorter fragments that do not match the template strand is also suppressed kinetically [228, 77, 78, 119].

Probing the enzyme-free self-assembly of long strands from a random pool of mononucleotides and short fragments experimentally is challenging. Typically the experiments require long times while the reaction yields remain low and undesired side products obscure the results. Moreover, it remains an unsolved technical problem how to track the evolution of the whole sequence pool simultaneously [79, 80, 81, 62, 63]. Due to these constraints, non-enzymatic self-assembly experiments either employed initial oligonucleotides with precisely designed sequences limiting the product space [72, 73, 130, 74, 75, 131] or focused on *primer extension* scenarios. In the latter scenario, a defined primer that is statically bound to a defined longer template strand gets extended by mononucleotides and short oligomers [68, 121, 69, 70, 116, 77, 71, 168].

Moreover, two conceptual, experimental studies investigated the emergence of progressively longer strands from DNA-oligomers using DNA ligases to accelerate the assembly dynamics and to obtain better yields [58, 59]. In the first study [58], all possible 12-mers that can be formed from a binary alphabet of A and T are present initially. The assembly dynamics give rise to structured sequence pools characterized by a reduced sequence entropy compared to a random pool. The emerging longer strands are either characterized by a large A or T content since mixed strands are more prone to self-inhibition due to hairpin formation. In the second study [59], three pairs of carefully designed complementary sequences composed of 20 nucleotides were used as basic building blocks. The authors demonstrated that certain subsets of sequence motifs composed of two basic building blocks form cooperative networks. Due to the cooperative dynamics, some of the motifs occur far more frequently in longer strands than others. Since the initial building blocks are already quite long in both studies, the binding energies of bound strands are large such that small differences in the stacking energies associated with adjacent nucleotide pairs do not matter. In order to achieve a separation of strands, both studies employed temperature cycling. However, subtle differences in the stacking energies might trigger sequence selection already on the level of the shortest oligomers, i.e., dimers and trimers, for which dissociation occurs spontaneously subtle than being induced externally. Since a sequence bias introduced early might feedback onto itself, it could have a substantial impact on the pool of longer strands at later stages. In summary, an experimental study exploring growth dynamics into longer polymers starting from a pool of small building blocks is still missing. [79].

Investigating the collective growth from small building blocks theoretically or by means of computer simulations in a model including the essential features of self-assembly, i.e., sequence-dependent (de)hybridization and ligation dynamics, is also challenging: First, the number of possible complex configurations grows exponentially fast as strands become longer. Second, there is an intrinsic separation of time scales between the fast dissociation of short and the slow dissociation of long hybridization sites and the slow ligation step.

To date, no theoretical study on self-assembly via templated ligation accounted for the motif-dependent thermodynamic and kinetic aspects of hybridization and bond formation (see Section 6.4.2). Therefore the following questions remained open:

1. What are the emerging dynamics in sequence space as strands grow longer?
2. Which are critical parameters that enable self-enhancing sequence selection?
3. How do motif-dependent thermodynamic and kinetic parameters affect the selection process?

In Chapter 5 we used the simulation method developed in Chapter 4 to investigate a model that partially reflects the complexity of the self-assembly process. In this first study, we treated the sequence dependence of the (de)hybridization dynamics in a mean-field picture where the dissociation rate decreases exponentially with the length of the hybridization site. We identified several growth regimes arising from the competition of timescales for dissociation and extension. Moreover, we showed that depending on external control parameters, the strand-length distribution in the stationary state can exhibit a non-monotonous shape characterized by a distinct strand length. In this chapter, we extended the model to cope with the augmented complexity, i.e., treat sequences explicitly. The starting point of the present study is a closed reaction volume that is initialized symmetrically with mononucleotides and a few dimers. Within the reaction volume, oligomers grow and degrade via templated ligation and hydrolysis. Eventually, the dynamics of the pool converge to a non-equilibrium stationary state characterized by its length and sequence distribution. To address the above questions, we consider different model variants. We start with a simple, purely energetically controlled scenario, where the additional kinetic stalling is absent, and the stacking energies for all complementary neighboring nucleotide pairs are identical. This scenario distinguishes solely between complementary and non-complementary pairings. We then increase the complexity step-wise, making the model variants more and more realistic. In the second scenario, we add a varying motif-depend bias favoring specific stacking interactions. In the third scenario, we turn back to the energetically unbiased system, this time with varying additional kinetic stalling. Finally, the fourth scenario combines a varying energetic bias with different stalling parameters. Our main finding is that thermodynamic discrimination between correctly and incorrectly paired nucleotides within hybridized strands is not sufficient by itself to promote self-enhanced sequence selection that drives the pools significantly away from the random state. Distinct patterns in sequence space only arises if non-complementary strand termini at the ligation site further slow down the ligation step significantly, i.e., if strong kinetic stalling is applied. In this case, a small thermodynamic bias for certain sequence-motifs trigger a self-enhancing dynamics such that the thermodynamically favored sequence motif dominates the stationary state.

6.2. Model

6.2.1. Strands and complexes

We consider a binary system composed of two complementary nucleotides X and Y. A molecule containing L nucleotides linked covalently is called a *strand* of length L (see Fig. 6.1a). A single nucleotide is a strand of $L = 1$. As in Chapter 5, strands are directed and point from the $-$ to the $+$ end (called 5′ and 3′ end for real polynucleotides). Moreover, strands are assumed to be rigid and hence can not fold onto themselves. An entity formed by several hybridized strands is referred to as a *complex*. All staggered conformations that can arise from a set of single strands are allowed inside the RNA reactor regardless of the number of strands and *mismatches*, i.e., non-complementary nucleotide pairs (see Fig. 6.1b, c and Fig. 6.5 in Section 6.5.2). However, branched hybridization structures and other non-linear complexes involving loops are excluded as implied by the rigid strand assumption. Furthermore, we call a complex that contains two or three strands a *duplex* or a *triplex*, respectively. The overlapping horizontal region between two strands is referred to as a *hybridization site*. Moreover, the vertical interface between two strands hybridized adjacently on a third strand is called a *ligation site*.

We assume that the non-equilibrium RNA reactor enables a fast chemical (re)activation of the termini of all strands present in the system. Hence, we do not model the activation step explicitly.

6.2.2. Elementary reactions

Strands and complexes form new complexes via *hybridization*, *dehybridization*, *templated ligation* and *hydrolysis* (see Fig. 6.1). All reactions are assumed to be elementary and occur with sequence- and structure-dependent rates k_{on}, k_{off}, k_{lig}, and k_{cut}. Assuming constant environmental conditions, k_{on} and k_{off} are related to the *hybridization energy* ΔG_{hyb} associated to a hybridization site via the thermodynamic consistency requirement [209, 210, 229]:

$$\frac{k_{\mathrm{off}}}{k_{\mathrm{on}}} = (VN_A c^{\circ}) e^{\beta \Delta G_{\mathrm{hyb}}}, \tag{6.1}$$

where $\beta = (k_{\mathrm{B}}T)^{-1}$, k_{B} is Boltzmann's constant and T denotes the (absolute) temperature, V and N_A are the reaction volume and Avogadro constant and $c^{\circ} = 1\,\mathrm{mol/l}$ is the reference concentration. From now on, we will express all concentrations as a multiple of the reference concentration. Moreover, in the following, we use the dimensionless hybridization energy

$$\Gamma = \beta \Delta G_{\mathrm{hyb}}. \tag{6.2}$$

Γ is obtained by summing over motif-depended stacking energies of nearest-neighbor blocks [76, 107, 113] (see Section 6.2.4). Γ thus reflects the number of complementary and non-complementary nucleotide pairs and their arrangement. Generally, mismatches increase the hybridization energy, therefore, reducing the stability of a complex.

The rate k_{lig} at which two strands that are located next to each other on a third strand ligate, depends on the paired nucleotides in the vicinity of the ligation site. Mismatches lead to *kinetic stalling*, i.e., a reduction of the ligation speed [228, 77, 78, 119]. We model the kinetic stalling using the *kinetic stalling factors* $\Phi_{\pm} \leq 1$. The stalling factors $\Phi_{\pm}$ are functions of the *complementarities* $\kappa_{\pm i} \in \{1, 0\}$ of the paired nucleotides in the vicinity of the ligation site. The value 1 indicates a complementary pair, whereas the value 0 indicates a non-complementary pair of nucleotides. Φ_{-} takes the complementarities κ_{-1}, κ_{-2} of the two nucleotides in the $-$ direction of the ligation site into account, while Φ_{+} is an equivalent expression for the two nucleotides in the $+$ direction (see Fig. 6.1c, d and Section 6.2.5 for more details). The two stalling factors are then multiplied with the *basal ligation rate* λ. With that, the formal definition of k_{lig} reads

$$k_{\mathrm{lig}} = \lambda\, \Phi_{-}\left(\kappa_{-1}, \kappa_{-2}\right)\, \Phi_{+}\left(\kappa_{+1}, \kappa_{+2}\right). \tag{6.3}$$

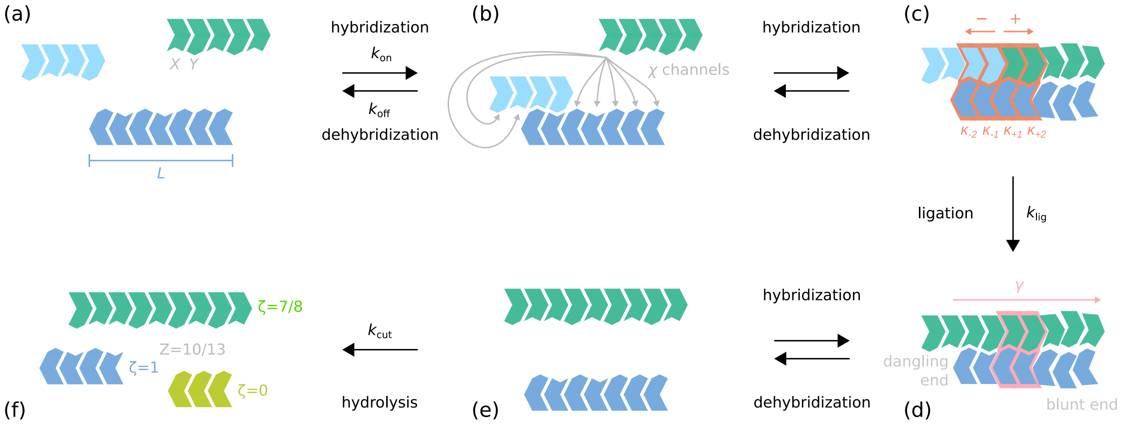


Figure 6.1.: Schematic illustration of the dynamics inside the RNA reactor. The elementary processes are hybridization, dehybridization, templated ligation, and hydrolysis with corresponding elementary rates k_{on}, k_{off} and k_{lig}, and k_{cut}. The elementary rates k_{on}, k_{off}, k_{lig} are functions of the sequence context. (a) Strands have a binary sequence and are directed. L denotes their length. (b) When two molecules collide, they can form χ different hybridization complexes. (c) Hybridization sites within complexes (horizontal interfaces) can contain mismatches. Two strands ($-$ and $+$ strand) located adjacently on another strand may get joined covalently via templated ligation. The speed of the ligation reaction depends on the complementarity κ of the nucleotide pairs at the ± 1 and ± 2 position (red box). Non-complementary pairings lead to kinetic stalling. (d) The stability of a hybridization site is governed by the hybridization energy ΔG_{hyb}. ΔG_{hyb} is obtained by summing over stacking energies γ associated with nearest-neighbor blocks (purple box) and considering terminal nucleotide pairs. ΔG_{hyb} and γ depend on the structural and sequential context. Mismatches weaken the binding. (e) Covalent bonds within single strands or single-stranded segments may get cleaved via hydrolysis at a constant rate. (f) In Section 6.3.1 we introduce the zebraness ζ of a strand level and system-level zebraness Z. The zebraness ζ of a strand ζ is the fraction of zebra motifs, i.e., alternating binary motifs, whereas system-level zebraness Z measures how zebra-like, i.e., alternating or homogeneous, the pool is as a whole.

Since random ligation of two strands in the absence of a template is weak compared to templated ligation [60, 78, 197, 167], we neglect it in our model.

While covalent bonds within double-stranded parts of complexes are assumed to be stable against hydrolysis, bonds within single-stranded sections get cleaved [175, 79, 230, 174] (see Figs. 6.1e and Figs. 6.1f). The corresponding rate is sequence-independent, i.e.,

$$k_{\text{cut}} = \text{const.} \tag{6.4}$$

With that, the overall degradation rate for a single strand of length L is $(L-1)k_{\text{cut}}$, for example. In real systems, k_{cut} varies by several orders of magnitude as a function of environmental parameters and crucially depends on the polymer's backbone chemistry [174, 175, 230, 231, 232].

6.2.3. Kinetics of hybridization and dehybridization

Since Eq. (6.1) only constrains the ratio of k_{on} and k_{off}, an additional kinetic parameter is required to fix the kinetics of the model. However, the chosen parametrization has only a minor effect on the global kinetics, given that ligation and hydrolysis are rare compared to hybridization and dehybridization (see Section 6.2.7). Our approach uses the constant rate of collision between two complexes $k_{\text{coll}} = (VN_A c^\circ t_0)^{-1}$, where t_0 is the collision time scale. In the following, we express all time in units of collision time scale t_0.

In general, two colliding complexes can form multiple hybridization configurations via χ distinct hybridization channels (see Fig. 6.1b). If $\chi > 0$, the probability of choosing one particular channel reads

$$p_{\text{hyb}} = 1/\chi. \tag{6.5}$$

Hence, the rate for a hybridization via a given channel is

$$k_{\text{on}} = k_{\text{coll}}\, p_{\text{hyb}}, \tag{6.6}$$

whereas the dehybridization rate becomes

$$k_{\text{off}} = \frac{1}{\chi} e^{\Gamma}. \tag{6.7}$$

If $\chi = 0$, no hybridization can occur. This is the case if one of the colliding complexes is a duplex without any overhang. A parametrization attributing the hybridization energy Γ to the dehybridization rate is common in theoretical approaches and has been confirmed experimentally [133, 141, 142]. Our choice of k_{on} and k_{off}, i.e., Eqs. (6.6) and (6.7) is rationalized in more detail in Ref. [183] which studies self-assembly in a sequence-

system	RNA		DNA	
nucleotides	A,U	C,G	A,T	C,G
$\overline{\gamma_{\text{com}}}$	-1.74	-5.00	-1.40	-3.26
δ_γ	-0.46	0.60	0.42	-0.65

Table 6.1.: $\overline{\gamma_{\text{com}}}$: mean stacking energies for complementary nearest neighbor blocks in binary RNA and DNA systems at a reference temperature of 37°C in units of $k_{\text{B}}T$ [76, 107]. δ_γ: difference between alternating and homogeneous blocks (see Eq. (6.12)). Note that the sign of δ_γ depends on whether A and U or T or G and C are considered for the binary system.

independent model with hybridization energies proportional to the overlap lengths. The kinetic assumptions Eqs. (6.6) and (6.7) reduce the computational complexity considerably, while still sampling complexes in a thermodynamically consistent way (see Section 6.5.3).

6.2.4. Hybridization energy

Detailed models to compute the free energy of given RNA and DNA secondary structures exist [107, 76, 113] (see also Section 2.1.3 in Chapter 2). These models build on the so-called stacking interactions of neighboring nucleotide pairs at the hybridization sites [112, 113]. Every nearest-neighbor interaction, i.e., every block of two adjacent nucleotide pairs, is associated with a motif-dependent stacking energy. These stacking energies additively contribute to the total free energy. Additional contributions to the total free energy take into account non-linearities of secondary structures such as loops, branching points, and particular end configurations.

Our coarse-grained model which excludes non-linear complex structures conserves the essential feature of the detailed nearest-neighbor models. The central element of our energy model is the stacking interaction of two neighboring nucleotides pairs P_i and P_{i+1} with

$$P_i \text{ and } P_{i+1} \in \left\{ \begin{matrix} X \\ \cdot \\ Y \end{matrix}, \begin{matrix} Y \\ \cdot \\ X \end{matrix}, \begin{matrix} X \\ \\ X \end{matrix}, \begin{matrix} Y \\ \\ Y \end{matrix} \right\}, \tag{6.8}$$

where dots symbolize hydrogen bonds between complementary nucleotides. To every block of adjacent nucleotide pairs $[P_iP_{i+1}]$ we assign a dimensionless stacking energy $\gamma\left([P_i\, P_{i+1}]\right)$. (Note that the last two pairs are non-complementary. Therefore, the nucleotide are not connected via a dot.) The hybridization energy is then given by the sum over all stacking energies and contributions ϵ_- and ϵ_+ accounting for the terminal nucleotide pairs at the $-$ and the $+$ end of double-stranded segment (see Fig. 6.1d), i.e.,

$$\Gamma = \sum_{i \,\in\, \text{blocks}} \gamma_i + \epsilon_- + \epsilon_+. \tag{6.9}$$

The contributions $\epsilon_\mp$ for the $\mp$ end also depend on the structural and sequential context. If the $\mp$ terminal nucleotide pair forms a *dangling end*, i.e., is preceded or followed by an unpaired nucleotide, we have $\epsilon_\mp \neq 0$. If the terminal nucleotide pair is part of a ligation site, there also is a contribution $\epsilon_\mp \neq 0$. If otherwise, it corresponds to *blunt end* of a complex, we have $\epsilon_\mp = 0$ (see Section 6.5.1 for details).

For simplicity, we assume symmetric stacking energies, i.e., $\gamma\left([P_i\,P_{i+1}]\right) = \gamma\left([P_{i+1}\,P_i]\right)$. Moreover, complementary nearest-neighbor blocks are either *alternating* if

$$[P_i\,P_{i+1}] \in \left\{ \begin{bmatrix} X-Y \\ \cdot \quad \cdot \\ Y-X \end{bmatrix}, \begin{bmatrix} Y-X \\ \cdot \quad \cdot \\ X-Y \end{bmatrix} \right\}, \tag{6.10}$$

or *homogeneous* if

$$[P_i\,P_{i+1}] \in \left\{ \begin{bmatrix} X-X \\ \cdot \quad \cdot \\ Y-Y \end{bmatrix}, \begin{bmatrix} Y-Y \\ \cdot \quad \cdot \\ X-X \end{bmatrix} \right\}. \tag{6.11}$$

Here, the $-$ symbol stands for a covalent bond. We denote stacking energies assigned to alternating and homogeneous blocks by γ_{alt} and γ_{hom}. Motivated by the observation that $\gamma_{\text{alt}} \neq \gamma_{\text{hom}}$ in DNA and RNA systems (see Tab. 6.1) [76, 107], we treat the energy difference

$$\delta_\gamma = \gamma_{\text{alt}} - \gamma_{\text{hom}}. \tag{6.12}$$

as a variable parameter. Without loss of generality, we assume $\delta_\gamma \leq 0$ for our model. Moreover, we assume constant stacking energies γ_{1nc} and γ_{2nc} for nearest neighbor blocks containing one or two non-complementary nucleotide pairs. Since blocks containing mismatches weaken the binding, their stacking contributions are positive. The contribution for a block with two mismatches is larger than for a block with only one mismatch. In summary, the block-wise contributions obey the hierarchy

$$\gamma_{\text{alt}} \leq \overline{\gamma_{\text{com}}} \leq \gamma_{\text{hom}} < 0 < \gamma_{\text{1nc}} < \gamma_{\text{2nc}}, \tag{6.13}$$

where $\overline{\gamma_{\text{com}}}$ is the average energy value of complementary blocks (see Tab. 6.2), i.e.,

$$\overline{\gamma_{\text{com}}} = \left(\gamma_{\text{alt}} + \gamma_{\text{hom}}\right)/2. \tag{6.14}$$

6.2.5. Kinetic stalling

Our kinetic stalling model draws on recent experimental findings [77, 119, 78]. Mismatches directly at the ligation site affect the ligation speed more substantially than distant ones. If the nucleotide pair at the ±1 position is non-complementary ($\kappa_{\pm 1} = 0$), a mismatch at the ±2 position ($\kappa_{\pm 2} = 0$) amplifies the stalling effect. Otherwise, a mismatch at the ±2 position has no effect, i.e.,

$$\Phi_{\pm}\left(\kappa_{\pm 1}, \kappa_{\pm 2}\right) = \begin{cases} 1 & \text{for } \kappa_{\pm 1} = 1 \wedge \kappa_{\pm 2} \in \{0, 1\} \\ \sigma_1 & \text{for } \kappa_{\pm 1} = 0 \wedge \kappa_{\pm 2} = 1 \\ \sigma_1 \sigma_2 & \text{for } \kappa_{\pm 1} = 0 \wedge \kappa_{\pm 2} = 0 \end{cases}, \tag{6.15}$$

where $\sigma_1 \leq \sigma_2$. If the hybridization site in the $+$ or $-$ direction contains only one nucleotide pair (see Fig. 6.1c), we use Eq. (6.15) with $\kappa_{\pm 2} = 1$. The formation of new covalent bonds is a non-equilibrium process: Every ligation requires energy which needs to be provided by the environment in the form of an activation chemistry [115, 81, 212]. The strength of the stalling effect depends on that underlying activation chemistry as well as the type of nucleotides being used [78, 77, 119], therefore, we treat σ_1 and σ_2 as variable parameters (see Tab. 6.2).

6.2.6. Effective cyclic environment

According to the energy model defined in Eq. (6.9), hybridization energies for long, primarily complementary hybridization sites become arbitrarily negative. Hence, the corresponding dehybridization rates converge to zero exponentially. As a result, strands can be bound in duplexes without single-stranded overhangs over long times. This effect is called template inhibition and leads to freezing of the dynamics [233, 79, 80]. To overcome this problem, we assume cyclic variations of the physico-chemical conditions (temperature, pH, or salt concentrations) inside the RNA reactor such that all hybridized strands separate within the period time τ [28, 123]. Aforesaid oscillatory conditions arise for example due to convection flows induced by temperature gradients or micro scale water cycles at heated gas–liquid interface. Both scenarios arise naturally in rock fissure in the vicinity of hydrothermal vents [124, 162, 234, 29, 125]. They are modeled effectively by introducing a lower bound for the dehybridization rate [187, 183], i.e., modifying Eq. (6.7) such that

$$k_{\text{off}} = \max\left\{\frac{1}{\chi} e^{\Gamma}, k_{\text{low}}\right\}, \tag{6.16}$$

process	parameter	value
hybridization	k_{coll}	1
	c_{tot}	$0.01\,c^{\circ}$
dehybridization	$\overline{\gamma_{\text{com}}}, \gamma_{\text{1nc}}, \gamma_{\text{2nc}}, \delta_{\gamma}$	$-1.25, 0.375, 0.75, [-0.3, 0]$
	l_{low}	7
ligation	l_{lig}	10
	σ_1, σ_2	$[0,1]$, $[0.1,1]$
hydrolysis	l_{cut}	18.5

Table 6.2.: Summary of parameters used in Section 6.3.

where $\tau = k_{\text{low}}^{-1}$. With that, the (dis)assembly dynamics of long complementary complexes do not obey the thermodynamic consistency requirement Eq. (6.1) anymore. Nonetheless, the kinetics are still plausible: The constant collision rate is a reasonable approximation for a collision process with a diffusion coefficient decaying with length compensated by a cross section growing with length (see Chapter 4).

6.2.7. Parametrization of rates

We can parametrize every rate constant k_* introduced so far by a dimensionless length l_* such that

$$k_* = e^{\overline{\gamma_{\text{com}}}\, l_*}. \tag{6.17}$$

This presentation will prove convenient in the later analysis of the results as it connects time scales to length scales. For example, $l_{\text{low}} = 7$ tells us that entirely complementary hybridization sites composed of more than seven nucleotides dissociate as quickly as altogether complementary hybridization sites comprising exactly seven nucleotides. Parameters used in the following are summarized in Tab. 6.2. Moreover, $l_{\text{lig}} = 10$ signifies that the timescale of a dehybridization for a hybridization site counting more than ten nucleotides would be slower than the bare ligation timescale if the lower bound with l_{low} would not have been introduced.

6.3. Results

6.3.1. Boundary conditions and observables

We aim to investigate the model dynamics starting on the lowest level, i.e., where mononucleotide and a few short oligonucleotides collectively evolve into larger entities. Will the sequences of longer strands be random, or will they show patterns? We chose the arguably simplest setting for our study, which is a closed reaction volume that does not exchange complexes with the environment. In such a setting, we expect the dynamics to settle to a stationary state eventually. We initialize the reaction volume symmetrically with 5000 nucleotides distributed over 4920 mononucleotides and 40 dimers. Moreover, we adjust the reaction volume such that the total nucleotide concentration is given by $c_{\text{tot}} = 0.01\,c^{\circ}$. The ratio of the initial monomer to dimer concentration is $c_1^{\text{init}} : c_2^{\text{init}} = 123 : 1$.

Our focus is on the evolution of the length distribution and the dynamics in sequence space. The length distribution c_L expresses the concentration of strands of length L, irrespective of whether they are part of a complex or not. We denote the mean length by $\overline{L}$. To describe the dynamics in sequence space, we aim for a simple observable with an intuitive and straightforward interpretation. Therefore, we introduce the *zebraness* as a characterization of a strand's sequence. The zebraness $\zeta(S)$ of a strand S of length L_S is the number of alternating "zebra" submotifs $X - Y$ or $Y - X$ within its sequence divided by the number of binary motifs $L_S - 1$ (see Fig. 6.1f). With that, a random sequence S_r is expected to have $\zeta(S_r) = 0.5$ on average. Moreover, the system-level zebraness Z characterizes how zebra-like the ensemble of strands is. It is given by

$$Z = \frac{\sum_S \zeta(S)\,(L_S - 1)}{\sum_S (L_S - 1)}, \tag{6.18}$$

where, the summation is performed over all individual strands with $L > 1$. A system containing strands with homogenous strands only would have $Z = 0$, whereas, for a system exclusively composed of strands with alternating sequences, we would have $Z = 1$.

All plots show ensemble averages which are taken over 20 independent realizations of the dynamics.

6.3.2. Overview of key findings

Before we analyze the four different scenarios outlined in the introduction in detail, we briefly overview our key results. First, we study the simplest variant of our model where both kinetic stalling and energetic bias are absent, i.e., $\delta_\gamma = 0$ and $\sigma_1 = \sigma_2 = 1$.

This scenario only distinguishes between nearest-neighbor blocks containing zero, one, or two mismatches. Alternating and homogenous blocks have identical energetic properties. While non-complementary pairings decrease the complex's stability, erroneous pairings at the ligation site do not reduce the bare ligation rate. The first model variant does not give rise to motif selection; the composition of the sequence pool remains entirely random, i.e., $Z = 0.5$.

In the second scenario, we increase the model complexity and introduce an energetic bias $\delta_\gamma < 0$ for alternating blocks while still assuming a non-discriminative ligation. The energetic bias favors the hybridization of strands with zebra-like sequences. This time, a weak zebra pattern $Z > 0.5$ is induced transiently during the initial growth phase. However, the pattern vanishes almost completely as the system approaches the steady-state (see Fig. 6.2).

The dynamics in sequence space change drastically if kinetic stalling with $\sigma_1, \sigma_2 < 1$ is applied. If non-complementary nucleotide pairs at the ligation site slow down the formation of a covalent bond, distinct patterns in sequence space can emerge. In the third scenario, we investigate the correlation between the strength of the kinetic stalling effect and the reduction of possible states in sequence space assuming identical energetic properties for alternating and homogeneous blocks, i.e., $\delta_\gamma = 0$. Within this setting, we observe a spontaneous symmetry breaking in sequence space. Independent realizations of the dynamics evolve to stationary state, dominated by either zebra motifs with $Z < 0.5$ or homogeneous motifs with $Z > 0.5$ (see Fig. 6.3). Moreover, we see that a dominant pattern emerges such that $Z \to 0$ or 1 if the stalling effect is strong enough.

In the fourth scenario, we show that a slight energetic bias $\delta_\gamma < 0$, can become self-amplifying if kinetic stalling is present (see Figs. 6.4). Depending on the strength of the kinetic stalling, the system converges to either a partial or pure zebra state characterized by either $Z > 0.5$ or $Z \to 1$.

6.3.3. No energetic bias and no kinetic stalling: $\delta_\gamma = 0$ and $\sigma_1 = \sigma_2 = 1$

This section aims to answer whether the energetic discrimination of matches and mismatches alone is sufficient to give rise to spontaneous symmetry breaking in sequence space such that $Z \neq 0.5$. To this end, we study the simplest variant of our model with neither energetic bias nor kinetic stalling, i.e., $\delta_\gamma = 0$ and $\sigma_1 = \sigma_2 = 1$.

Initially, the growth dynamics of the mean length $\overline{L}$ is slow until $t \approx 8.8 \times 10^9$ (see dark blue curve in Fig. 6.2a). At this point, the mean length $\overline{L}$ starts to increase rapidly. We refer to this time point as the *onset* of growth and denote it by $\widehat{t}$. After the steep increase, $\overline{L}$ reaches a plateau value. The inset shows the steady-state length distribution displaying a double-exponential shape.

The ensemble average of the zebraness Z initially fluctuates and then converges to

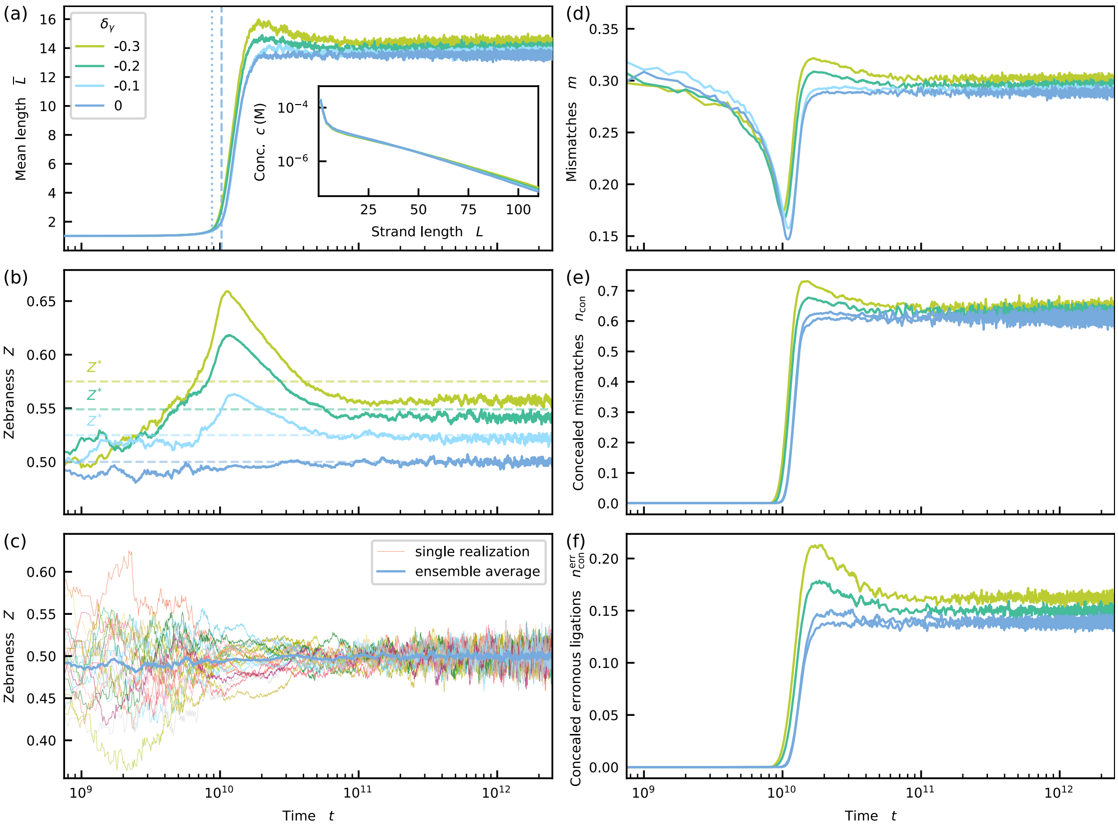


Figure 6.2.: Mean length $\overline{L}$ and system-level zebraness Z as functions of time for $\sigma_1 = \sigma_2 = 1$ and various δ_γ. (a) A sharp increase of $\overline{L}$ appears at $\widehat{t}$. For $\delta_\gamma = 0$, the dashed line corresponds to $\widehat{t}$ resulting from the formal definition, whereas the dotted line is the prediction obtained from Eq. (6.19). For $\delta_\gamma < 0$, $\overline{L}$ reaches a maximum before decaying gradually to the stationary value. The inset shows the steady-state length distributions. (b) If there is no energetic bias, i.e., $\delta_\gamma = 0$, no distinct patterns emerge in sequence space, and hence $Z = 0.5$ (see also Section 6.5.4). If an energetic bias $\delta_\gamma < 0$ is applied, Z grows initially and then decays when $\overline{L} \approx 7$. The final value is slightly above the random state $Z = 0.5$ and below the simple thermodynamic estimate Z^* (see Eq. (6.22)). (c) Single realizations of the dynamics for $\delta_\gamma = 0$ behave similar to ensemble average. Strong fluctuations for small times stem from low numbers of strands with $L > 1$. (d) The fraction of mismatches m first increases and then decreases as the mean length becomes longer. (e) The fraction of concealed mismatches, i.e., mismatches not affected by energetic discrimination grows simultaneous with the mean length. (e) Over time, concealed erroneous ligations become frequent and destroy the initial sequence bias.

$Z = 0.5$ (see Fig. 6.2b). Looking at single trajectories (see Fig. 6.2c) reveals a behavior similar to the ensemble average. The initial values of $Z \lessgtr 0.5$ on the single trajectory-level are due to small numbers of strands with $L > 1$. A value of $Z = 0.5$ hints towards an entirely random sequence pool but does not exclude motif correlations on larger scales. However, analyzing distributions of longer motifs reveals that the final sequence composition is indeed random (see Section 6.5.4).

The evolution of the mean length shows some interesting features. After a lag phase, its increase becomes exponential at $t = \widehat{t}$ (see Fig. C.1 in the Appendix). Formally, we define the onset of growth $\widehat{t}$ by intersecting the tangents to the $\overline{L}$–curve at $t = 0$ and the point where the increase is strongest (dashed line in Fig. 6.2a, for details see Section 6.5.5 and Fig. C.2). We observe that $\widehat{t}$ coincides with the moment in time at which *higher-order ligations*, i.e., ligations involving at most one monomer become more abundant than ligations joining two monomers to a dimer (see Fig. 6.7 in Section 6.5.5). Moreover, we can predict the onset with a relative error smaller than 15% by the following formula derived in Section 6.5.5 (dotted line in Fig. 6.2).

$$\widehat{t} \approx \log \left[\frac{c_{\text{tot}}^2 k_{\{1,1|2\}} - k_{\text{cut}}}{c_{\text{tot}} c_2^{\text{init}} k_{\{1,2|2\}}} \right] \Big/ \left(c_{\text{tot}}^2 k_{\{1,1|2\}} - k_{\text{cut}} \right) \tag{6.19}$$

Here, c_2^{init} is the initial dimer concentration and $k_{\{1,1|2\}}$ and $k_{\{1,2|2\}}$ are the effective rate constants at which new dimers or trimers are formed from monomers or monomers and dimers. $k_{\{1,1|2\}}$ is given by

$$k_{\{1,1|2\}} = \frac{k_{\text{lig}}}{K_{\{1,1|2\}}}, \tag{6.20}$$

where $K_{\{1,1|2\}}$ is the effective dissociation constant averaged over all triplexes involving two monomers and one dimer. An analogous expression exists for $k_{\{1,2|2\}}$. Repeating the computer experiment with k_{cut} and k_{lig} different from the standard values given in Tab. 6.2 confirmed the validity of Eq. (6.19) (see Fig. C.3 and Fig. C.4 in the Appendix).

Moreover, altering k_{cut} and k_{lig} while keeping the other parameters fixed revealed that the mean length $\overline{L}$ in the stationary state does not explicitly depend on these two variables but only on their ratio, i.e.,

$$\overline{L} = \overline{L} \left(k_{\text{lig}} / k_{\text{cut}} \right) \quad \text{for} \quad t \to \infty. \tag{6.21}$$

The same dependence of the mean length on the ratio $k_{\text{lig}}/k_{\text{cut}}$ can be derived analytically for a random ligation model [235].

6.3.4. Energetic bias in the absence of kinetic stalling: $\delta_\gamma < 0$ and $\sigma_1 = \sigma_2 = 1$

In the previous section, we saw that the energetic discrimination between complementary and non-complementary nucleotide pairs alone is insufficient for the spontaneous emergence of patterns in sequence space. Therefore, we now ask whether an energetic bias $\delta_\gamma < 0$ favoring the binding of zebra motifs can induce zebra patterns that become self-amplifying, while kinetic stalling is still absent ($\sigma_1 = \sigma_2 = 1$).

The energetic bias $\delta_\gamma < 0$ affects the dynamics of the mean length only marginally (see light blue and green curves in Fig. 6.2a). The onset of growth $\widehat{t}$ is estimated by a formula similar to Eq. (6.19) with an accuracy of $\sim 15\%$ (see Section 6.5.5 and Figs. C.5–C.7 in the Appendix). The steep increase after the lag phase is followed by a gradual descent to the steady state value, slightly below the maximum. Moreover, the steady-state length distribution is similar to the scenario without energetic bias.

In sequence space, a simple thermodynamic estimate Z^* for the final zebraness can be made based on a two-state system with an energy difference δ_γ

$$Z^* = \frac{1}{1+e^{\delta_\gamma}}. \tag{6.22}$$

Since this estimate neglects any correlation and feedback effects, one could naively expect that the zebranass resulting from the simulated model dynamics reaches a value larger than Z^*. Indeed, the observable initially grows beyond the estimate Z^*. However, the growth stops when the mean length reaches a value of $\overline{L} \approx 7$. At that point, the zebraness starts to decay and converges to a stationary value simultaneously with $\overline{L}$ (see Fig. 6.2b). In the stationary state, the zebraness Z is only slightly above the result for random sequences $Z = 0.5$, and below the simple thermodynamic estimate Z^*.

6.3.5. Loss of energetic discrimination prevents sequence selection

Why are the initially emerging zebra patterns triggered by the energetic bias $\delta_\gamma < 0$ in Fig. 6.2b neither amplified nor maintained? In the following, we analyze the growth processes in detail to give an intuitive explanation.

Strand growth requires the formation of complexes comprising at least three strands. The more negative the hybridization energies of the hybridization sites in these complexes are, the more stable the configurations become and the higher the probability for a ligation gets. Non-complementary nucleotide pairs increase the hybridization energy and, therefore, weaken the binding. To analyze the effect of these mismatches, we define the overall fraction of mismatches m as

$$m = \frac{N_{\text{non}}}{N_{\text{pairs}}}, \tag{6.23}$$

where N_{pairs} and N_{non} are the absolute numbers of nucleotide pairs and non-complementary nucleotide pairs in all present complexes. Initially, the fraction of mismatches m decreases since complexes mostly containing complementary nucleotides that emerge during the early growth persist longer and hence contribute more substantially to the average. However, the fraction of mismatches starts to increase when the mean length becomes larger (see Fig. 6.2d). This increase of the mismatch fraction arises from the loss of thermodynamic discrimination induced by the cut-off k_{low} in the dehybridization rate, i.e., the effective temperature cycles (see Section 6.2.6). Although the hybridization energy may become arbitrarily negative for large hybridization sites, the dehybridization rate can not become smaller than the the lower bound k_{low}. The length scale associated with k_{low} is $l_{\text{low}} = 7$ (see Section 6.2.7). This implies that an entirely complementary hybridization site comprising more than seven pairs has the same stability as a mismatch-free hybridization site composed of exactly seven pairs. Moreover, mismatches in extended hybridization sites might have no effect on the rate for unbinding because the hybridization site still contains a high number of matches. If many matches are present, the hybridization energies are strongly negative such that the lower threshold still determines the rate for dehybridization. This effect enables *concealed mismatches*. Concealed mismatches are mismatches that do not lower the dehybridization rate k_{off} of a hybridization site. Replacing a concealed mismatches with a complementary pair would not decrease k_{off} further since it is already given by the cut-off, i.e., $k_{\text{off}} = k_{\text{low}}$. The longer the strands become during the first growth phase, the more concealed mismatches emerge. With the absolute numbers of mismatches and concealed mismatches in all present complexes N_{non} and N_{con}, we now introduce the fraction of concealed mismatches n_{con} as

$$n_{\text{con}} = \frac{N_{\text{con}}}{N_{\text{non}}}, \tag{6.24}$$

The evolution of n_{con} shown in Fig. 6.2e reveals that most of the occurring mismatches are concealed, once $\overline{L}$ has become approximately twice as large as l_{low}. Concealed mismatches also occur at the strand termini at ligation sites and may lead to the formation of new binary motifs which are not complementary to the templating motif at the ligation site. We call such a ligation involving at least one concealed mismatch a *concealed erroneous ligation*. Dividing the number of concealed erroneous ligations N_{err} per time by the overall number of ligations N_{lig} per time gives the fraction of concealed erroneous ligations $n_{\mathrm{con}}^{\mathrm{err}}$, i.e.,

$$n_{\mathrm{con}}^{\mathrm{err}} = \frac{N_{\mathrm{err}}}{N_{\mathrm{lig}}}. \tag{6.25}$$

Every erroneous concealed ligation mitigates the present bias in sequence space and leads to randomness. Erroneous concealed ligation are the reason why the initial sequence patterns decay almost to the random level $Z = 0.5$. However, not all hybridization sites, particularly the shorter ones, have a dehybridization rate determined by the lower bound. As the initial bias for binary zebra motifs on the system level decreases and sequences become more random, non-concealed mismatches in shorter hybridization sites become more likely. Hence, short hybridization sites not yet affected by the lower bound for the unbinding rate become less stable on average and contribute less to the growth process. This explains why the small maxima in the mean length and the other observables shown in Figs Fig. 6.2a and Fig. 6.2d-f disappear as the bias for binary zebra motifs fades away.

6.3.6. Kinetic stalling in the absence of energetic bias: $\delta_\gamma = 0$ and $\sigma_1, \sigma_2 < 1$

Concealed erroneous ligations prevent sequence selection during the self-assembly process in the model variant without kinetic stalling. In the presence of kinetic stalling, concealed erroneous ligations will be suppressed. In this section, we thus investigate the more realistic model variant where kinetic stalling is applied and vary the strength of the kinetic stalling σ_1 from 0 to 0.1 while fixing $\sigma_2 = 0.1$.

Initially, the dynamics of the mean length $\overline{L}$ is qualitatively similar to the systems without kinetic stalling studied before (see Fig. 6.2a and Fig. 6.3a). However, the onset of growth $\widehat{t}$ appears later. The time point of the onset $\widehat{t}$ can be predicted by a formula analogous to Eq. (6.19), which considers the kinetic stalling effect, with an error $< 15\%$. For $\sigma_1 = 0.05$, the values for $\widehat{t}$ from the prediction and the formal definition are highlighted by the dotted and dashed lines in Fig. 6.3a (for details see Section 6.5.5 and Figs. C.8–C.12 in the Appendix). On larger timescales, the model including kinetic stalling deviates from the earlier model. After the steep increase, $\overline{L}$ does not directly settle to a steady-state. Instead, it grows gradually and converges to a constant value eventually. Visualizations with a linear x- and a logarithmic or linear y-axis reveal that the initial increase after the lag phase is approximately exponential, while the increase during the second growth phase is approximately linear (see Figs. C.13-C.15 in the Appendix). For $\sigma_1 = 0, 0.05$ or 0.067, similar stationary mean lengths are reached. However, the relaxation time increases with σ_1, such that it takes more than ten times longer for a system with $\sigma_1 = 0.067$ (see inset of Fig. 6.3a) to converge to the stationary state than for a system with infinite stalling. For $\sigma_1 \leq 0.067$, the steady-state value of the mean length is more than twice as large as for $\sigma_1 = \sigma_2 = 1$ scenario. For $\sigma_1 = 0.1$, the increase of the mean length during the second growth phase is small during the time window of observation. From Fig. 6.3a we can not deduce whether the mean length already approached a stationary value, or whether it will keep on growing until a similar value as for $\sigma_1 = 0.67, 0.05, 0$ is reached. If a stationary value were already reached, it would be significantly smaller than for $\sigma_1 = 0.67, 0.05, 0$. For $\sigma_1 = 0.1$ (as well as for $\sigma_1 = 1$) the simulation times are large and prevented us from analyzing at longer time scales. However, visualizing the curve for $\sigma_1 = 0.1$ in a coordinate system with a linear x-axis might suggest that the system indeed already converged to a stationary state (see Fig. C.13 in the Appendix). We will discuss the behavior for $\sigma_1 = 0.1$ in more detail in Section 6.3.9 and provide further evidence why the behavior, in this case, might be qualitatively different from the behavior for $\sigma_1 = 0.67, 0.05, 0$. Moreover, the length distributions in the stationary state look qualitatively similar to the ones seen earlier (see Fig. C.14 in the Appendix).

Is the novel behavior of the mean length shown in Fig. 6.3a related to a novel motif-selective dynamics in sequence space? We now investigate the evolution of the strands' sequences. Since the initial pool of sequences is symmetric and since neither zebra nor homogeneous binary motifs are preferred energetically, we do not expect a preference for a single realization to go to either a zebra ($Z > 0.5$) or non-zebra ($Z < 0.5$) state. Hence, the system-level zebraness Z is not appropriate to describe an ensemble of realizations. As a meaningful observable to quantify the sequences bias on the ensemble level, we, therefore, introduce the *patterness* Π as

$$\Pi = \max\{Z, 1 - Z\}\,. \tag{6.26}$$

During the first growth phase of $\overline{L}$, a bias $\Pi > 0.5$ is established for all values of σ_1. The dominance of the bias correlates with the strength of the kinetic stalling. During the slow second growth phase, Π gradually increases and reaches a stationary value simultaneously with $\overline{L}$ (see Fig. 6.3b and Fig. C.11). For $\sigma_1 = 0, 0.05$ or 0.067, we observe a value of $\Pi \approx 1$ in the stationary state. Hence, on the realization level, the final pool contains either (almost) entirely alternating or homogeneous sequences. We classify such states in sequence space as *pure*. For $\sigma_1 = 0.1$, the final sequence composition within the observation time window is also is non-random. However, the patterns are not pure, i.e., $0.5 < \Pi < 1$. We refer to these states as *partially mixed* states. Whether the system has already converged to a stationary state with $\Pi < 1$ or will further evolve to a pure state as for $\sigma_1 = 0.67, 0.05, 0$ remains unclear at this point (see above). Fig. 6.3c displays the evolution of the zebraness of all realizations forming the steady-state for $\sigma_1 = 0.05$. On average, one half of the realizations evolves towards the $Z = 1$ state, while the other half evolves towards the $Z = 0$ state. Hence, the symmetry of the initial state is broken spontaneously: either zebra or homogeneous motifs are selected. (See Figs. C.8–C.12 for equivalent visualizations of single realizations for other σ_1 values.)

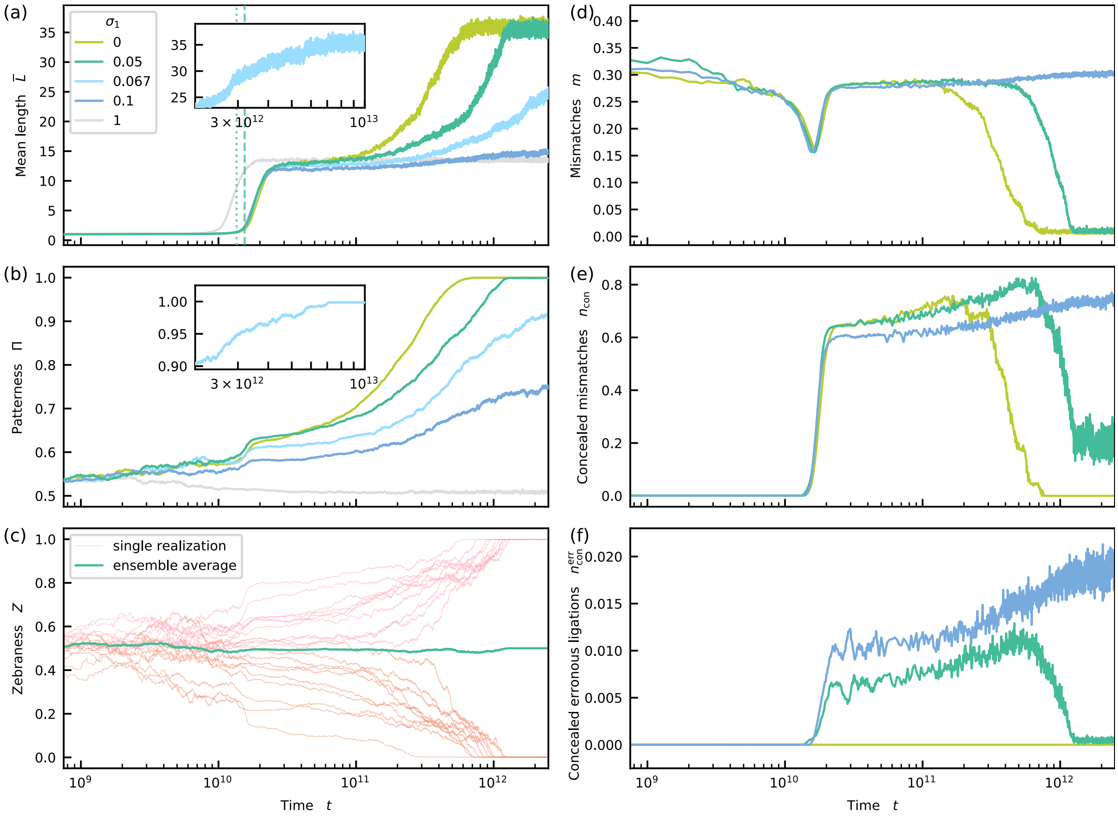


Figure 6.3.: Evolution of the mean length $\overline{L}$ and system-level zebraness Z in a kinetic stalling scenario without energetic bias, i.e., $\delta_\gamma = 0$. For reference, we also show the $\sigma_1 = \sigma_2 = 1$ curves (gray). (a) For $\sigma_1 = \sigma_2 < 1$, $\overline{L}$ shows two distinct growth phases. While the first one is rapid, the second one is slow. Relaxation to the steady-state for $\sigma_1 = 0.067$ appears much later (see inset). For $\delta_\gamma = 0.05$, the dashed line corresponds to $\widehat{t}$ resulting from the formal definition, whereas the dotted line is the prediction. (b) The bias for alternating or homogeneous patterns established in the first growth phase becomes amplified during the second growth phase. For strong stalling ($\sigma_1 \leq 0.067$), the final pool comprises either pure zebra or fully homogeneous sequences. (c) The symmetry of the random initial state is broken spontaneously. For $\sigma_1 = 0.05$, equal fractions of realizations evolve to the zebra or homogeneous state. (c)-(e) Dynamics of mismatches, concealed mismatches, and concealed erroneous ligations for $\sigma_1 = 0, 0.05, 0.1$. For details, see the main text.

6.3.7. Hydrolysis and stalling boost sequence selection

The previous section revealed a coupling between sequence selection and two distinct growth phases in the kinetic stalling scenario. We now interpret and explain this coupling. Here, we consider the cases $\sigma_1 = 0$ and $\sigma_1 = 0.05$ where all trajectories eventually converge to a pure state. The case where $\sigma_1 = 0.1$ is discussed in Section 6.3.9.

On the level of individual trajectories, fluctuations lead to a small bias in the motif composition even before the mean length starts to grow rapidly (see Fig. 6.3c). This early asymmetry in the distribution of alternating and homogeneous motifs governs the fate of the realization as seen from Fig. 6.3c.

During the first growth phase (see Fig. 6.3a), monomers and short strands self-assemble into longer strands. The first growth phase ends when most of the initial monomers are consumed. At the end of this initial growth phase, a significant bias towards alternating or homogeneous motifs is present. However, a considerable fraction of binary motifs does not yet reflect system-level bias, i.e., differs from the dominant binary motifs (which are either $X - Y$ and $Y - X$ or $X - X$ and $Y - Y$). Therefore, mismatches in bound strands are still frequent (see Fig. 6.3d). Moreover, since the average strand length is already significantly above l_{low}, most mismatches are concealed (see Fig. 6.3e).

When monomers and short strands do not dominate the pool anymore, hydrolysis becomes important. Every time a strand breaks an existing binary motif vanishes. During the second growth phase, fragments of broken strands are reassembled to longer strands via ligation on a template strand. (For an analysis of sequence patterns of strands of specific lengths, see Fig. C.17 in the Appendix.) If the kinetic stalling is strong, the ligation of two strands is (almost) impossible if a mismatch occurs at the ligation site and the fraction of concealed erroneous ligations is (close to) zero (see Fig. 6.3f). Hence, every ligation forms a new binary motif that (almost) always complements the templating motif at the ligation site. If the templating motif is zebra-like (homogeneous), the new motif is zebra-like (homogeneous), too. Over time, all binary motifs created during the initial growth phase, particularly those that do not reflect the system bias, get destroyed at a uniform rate. At the same time, binary motifs emerging during the second growth phase likely reflect the system bias. Consequently, the bias becomes self-enhancing (see Fig. 6.3b). Newly created binary motifs enhance the system bias even more, while all motifs that do not reflect the asymmetry in motif space become extinct eventually. As a result, the motif composition becomes more and more ordered and mismatches become rarer (see Fig. 6.3d). Remaining mismatches are now even more unlikely to affect the dehybridization rate since the rate is determined by the lower bound k_{low} for long and primarily complementary hybridization sites. Hence, the fraction of concealed mismatches increases slightly (see Fig. 6.3e). Moreover, if kinetic stalling is finite ($\sigma_1 > 0$), the fraction of concealed erroneous ligations also slightly increases (see Fig. 6.3f).

Since the energetic properties are symmetric, every realization chooses spontaneously to converge either to the zebra-like or homogeneous state. As strands become more similar, i.e., complementary in their motif composition, triplex configurations achieve higher stability on average. Recall that the dehybridization rate decays exponentially with the hybridization site energy (see Eq. (6.7)). This increase of stability enhances the probability of templated ligations. Therefore, the symmetry breaking in sequence space couples to an enhanced growth, which becomes apparent in the further increase of the mean length. For strong stalling, the second growth phase ends when the system has reached an (almost) entirely alternating or homogeneous state (see Figs. 6.3a and 6.3b). At that point, the fraction of mismatches is close to zero. The few mismatches that still occur are mostly due to strayed mononucleotides and short oligomers sitting on longer strands. Since these mononucleotides and short oligomers are far from being affected by the lower bound and unbind quickly, the fraction of concealed mismatches takes small values again.

6.3.8. Energetic bias in the presence of kinetic stalling: $\delta_\gamma < 0$ and $\sigma_1 = \sigma_2 < 1$

The previous section revealed that spontaneous motif selection occurs as a result of kinetic stalling. If no energetic bias is applied, the sequence pool converges to a stationary state which is either dominated by homogenous or alternating sequences. In addition, the energetic symmetry can be broken explicitly if a small energetic bias favoring zebra motifs is applied. To understand the emergent phenomena in this setting, we study two systems with strong and one with weak kinetic stalling for various energetic biases $\delta_\gamma < 0$.

First, we consider the case where $\sigma_1 = 0.05$. There, most parts of the description in the previous section ($\delta_\gamma = 0$) also apply here (see Figs. 6.4a and 6.4b). Again, we can predict the onset of growth $\widehat{t}$ with an error $< 15\%$ (for details see Section 6.5.5 and Figs. C.18–C.20 in the Appendix). The steady-state value of $\overline{L}$ depends weakly on the energetic bias. However, the final state is reached earlier if the bias is stronger. In sequence space, all trajectories end in a pure zebra state $Z = 1$. Hence, symmetry breaking in sequence space is now induced energetically as expected because of the explicit symmetry breaking in the energy landscape. Second, we investigate a scenario with $\sigma_1 = \sigma_2 = 0.1$. The mean length grows strongly in the beginning as before (see Fig. 6.4c). Again, we are able to predict the onset of growth $\widehat{t}$ with an error $< 10\%$ (for details see Section 6.5.5 and Figs. C.21–C.23 in the Appendix). The fast growth phase is followed by either a marginal increase ($\delta_\gamma = -0.1, -0.2$) or decrease ($\delta_\gamma = -0.3$) of

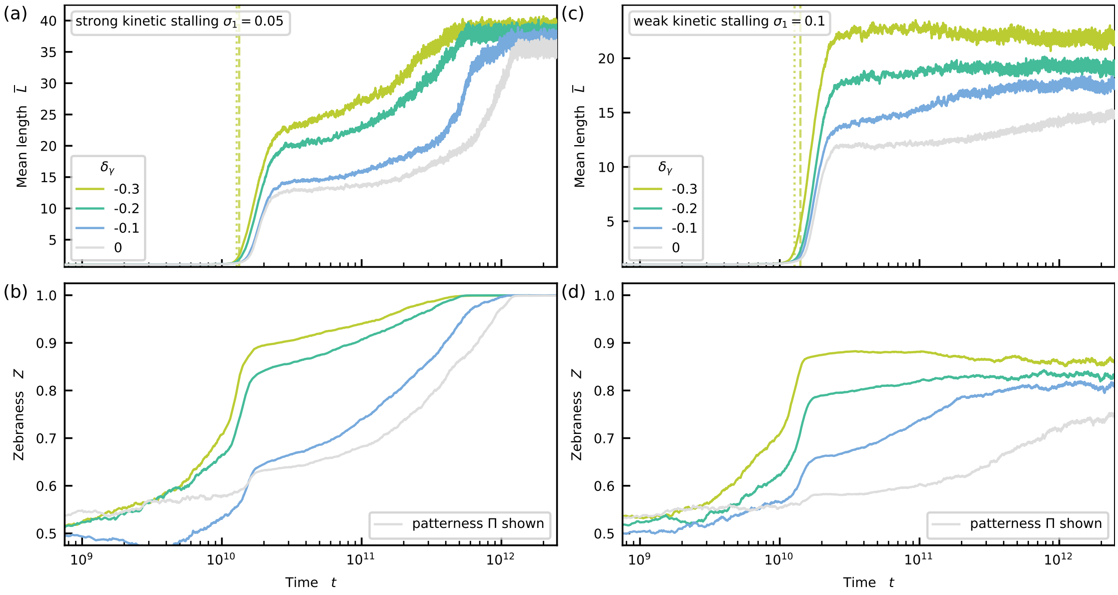


Figure 6.4.: Left column: scenario with $\sigma_1 = 0.05$ (strong kinetic stalling), right column: scenario with $\sigma_1 = 0.1$ (weak kinetic stalling). (a) The mean length $\overline{L}$ grows again in two steps. The stronger the bias, the earlier the relaxation to the stationary value. For $\delta_\gamma = -0.3$, the dashed line corresponds to $\widehat{t}$ resulting from the formal definition, whereas the dotted line is the prediction. (same for (c)). (b) Pronounced zebra pattern emerge during the first growth phase. The pattern become pure during the second growth phase. (c) A gradual increase or decay follows the rapid growth phase. The steady-state value of $\overline{L}$ correlates with the strength of the energetic bias. (d) While $\overline{L}$ slightly increases (decreases), Z also increases (decreases). The sequence pool in the stationary-state shows mixed patterns dominated by zebra motifs. The fraction of zebra motifs depends on the energetic bias. (a)-(d) For reference, we also show the patterns Π for $\delta_\gamma = 0$ curves (gray).

$\overline{L}$ to a stationary value correlating with the strength of the energetic bias. A strong zebra pattern $Z > 0.5$ is induced in sequence space during the initial increase of the mean length (see Fig. 6.4d). While $\overline{L}$ grows (decays) during the second phase, Z also grows (decays). Eventually, the sequence pool converges to a partially mixed stationary state with a significant majority of zebra motifs such that $0.5 < Z < 1$. The excess of zebra motif again correlates with the strength of the energetic bias. Moreover, from Figs. C.21–C.23 in the Appendix, it becomes clear that all single trajectories behave similarly to the ensemble mean, i.e., show steady-state values of the zebras above 0.5.

6.3.9. Weak versus strong kinetic stalling regimes

In Section 6.3.6, we speculated whether the system with $\sigma_1 = 0.1$ and without energetic bias (blue curve in Fig. 6.3) reaches a stationary state characterized by $\Pi < 1$ in contrast to the scenarios with $\sigma_1 = 0.067, 0.5, 0$, where $\Pi = 1$ and referred to the visualization Fig. C.13 with a linear x-axis. However, this visualization did not allow for a clear conclusion either. It could be that the alleged partially mixed stationary state is only transient and that a pure state is reached on much larger time scales. Though the findings from Section 6.3.8 suggest that the stationary state of the system with $\sigma_1 = 0.1$ and without energetic bias is indeed qualitatively different from the scenarios with $\sigma_1 \leq 0.067$ and without energetic and marked by $\Pi < 1$. The curves for $\sigma_1 = 0.1$ and various energetic biases clearly converge to stationary states with $\Pi < 1$ (see colored curves in Fig. 6.4c and Fig. 6.4d). If the stationary state is partially mixed in the presence of an energetic bias, it is not too far-fetched to assume that it is also partially mixed if the energetic bias is absent.

Naturally, the question arises, whether a critical value for σ_1 exists above which the system always reaches a state with $\Pi = 1$ for a given value of the energetic bias δ_γ. This first question directly leads to a second question, namely, what would be the nature of the corresponding non-equilibrium phase transition? We leave the answer to this question open for future research. However, finding an answer might be challenging since the relaxation time to the stationary state will probably diverge. At this point, we will content ourselves with hypothesizing that two different regimes might exist without drawing an exact border: For *strong* kinetic stalling, the system converges to a *pure* state, while for *weak* kinetic stalling, it converges to a *partially mixed* state.

In the light of the above hypothesis, we now analyze the dynamics of mismatches and concealed erroneous ligations in the energetically unbiased system case for $\sigma_1 = 0.1$ (see blue curve in Fig. 6.3d-Fig. 6.3f as for $\sigma_1 = 0.0$ and $\sigma_1 = 0.05$. The self-amplification of the dominant binary motif comes to a halt during the second growth phase (see Fig. 6.3a and Fig. 6.3a). Consequently, the fraction of concealed mismatches does not decrease again as for $\sigma_1 = 0.0$ and $\sigma_1 = 0.05$. Since most of the strands are long enough, most of the mismatches are concealed. Moreover, concealed erroneous ligations are not fully suppressed and still occur in the stationary state. Hence, the weak kinetic stalling scenario includes features of both dynamic regimes described in Section 6.3.5 and Section 6.3.7.

6.3.10. Validity of our model and application to primer extension

In the results section, we focused on self-assembly scenarios where all strands (apart from monomers) are equally important because there are no distinct template, primer, and substrate strands as in typical primer-extension situations. However, in Section 6.6, we show that our modeling of the kinetic stalling and the (de)hybridization kinetics in combination leads to copying dynamics in primer-extension situations consistent with the experimental literature. For completeness, we here give a brief overview. In primer extension experiments, a non-complementary nucleotide at the primer terminus slows down the extension process. Moreover, such a mismatch increases the probability of another misincorporation. Within our model, the slowdown of the extension process and the accumulation of misincorporation stem from two contributions. Kinetic stalling reduces the bare ligation rate. In addition, mismatches at the primer's end weaken the monomer binding, increase its dehybridization rate, and hence render the next extension even less probable. We summarize both contributions to a combined stalling effect. Moreover, a mismatch at the primer's terminus reduces the thermodynamic discrimination of a monomer hybridized next to it. Hence, the fraction of misincorporations increases.

6.4. Discussion

6.4.1. Summary

Our study investigated the self-assembly of prebiotic polymers from binary building blocks via templated ligation inside a non-equilibrium RNA reactor. Three main questions motivated our study.

1. What are the dynamics in sequence space as strands grow longer from a random pool?
2. Which critical parameters drive the system away from a random state towards a state characterized by distinct patterns in sequence space?
3. How do non-uniform thermodynamic and kinetic parameters affect these sequence patterns?

We identified kinetic stalling as a critical requirement for self-enhanced sequence selection. The final sequence space shows no structure if the underlying stacking energies are uniform without kinetic stalling. In contrast, spontaneous symmetry breaking occurs for strong kinetic stalling. The final pool contains either entirely homogenous or zebra-like sequences. In scenarios without kinetic stalling, any energetically induced sequence bias vanishes almost completely as strands grow. In contrast, in the presence of kinetic stalling, subtle differences in the stacking energies trigger cascades of self-amplification, leading to highly ordered sequence pools.

Our results hint towards the existence of two different stalling regimes. We hypothesize that for strong kinetic stalling, the system converges to a pure state with $\Pi = 1$, while for weak kinetic stalling, it converges to a partially mixed state characterized by $\Pi < 1$.

Initially, the mean length shows burst-like growth dynamics after a short lag phase. The onset of the rapid growth coincides with the time point where higher-order ligations become more abundant than ligations joining two monomers and can be predicted analytically.

6.4.2. Prior work, our model, and future extensions

In an earlier model for prebiotic self-assembly, strands only grow via random ligation [176]. There, self-folding and complex formation introduced a protection mechanism against hydrolysis for double-stranded segments. Moreover, the ensemble of strands was assumed to reach a binding equilibrium immediately after a random ligation occurred. In this model, protection against hydrolysis could extend the system's sequence memory. However, the effect was only transient, and all selected patterns vanished eventually.

Previous theoretical studies considering growth via templated ligation generally explored effective models that reduce the state space to (sub-)sequences without considering complex formation explicitly. [200, 187, 201, 197, 202, 203, 204, 205, 195, 206, 207]. Such approaches do not treat (de-)hybridization and ligation as elementary steps. Instead, the reactions are coarse-grained into one extension process. The specification of the corresponding rate, neglects the intricacies of the assembly mechanisms and requires a priori assumptions regarding the relevant configurations [203, 200, 202, 201, 197, 195]. Moreover, many models ignore that the hybridization energy is a function of the number and nature of the paired nucleotides and use constant (de-)hybridization rates. [203, 200, 202, 201, 197, 195, 206, 205]. Other studies treat the sequence dependence of (de-)hybridization employing mean-field techniques where sequence correlations are dismissed [187]. Such simplifications result in systems effectively containing only one type of self-complementary nucleotide as in Chapter 5 and any form of sequence selection is necessarily absent. In contrast, our stochastic approach explicitly takes the sequence-dependent thermodynamic and kinetic aspects of templated ligation into account. The amount of details in our model is novel. Yet, it is essential to study the emergence of patterns in sequence space in non-enzymatic growth. For that reason, despite its complexity, our model should still be considered as a minimal model.

Although being already quite complex, our model also made simplifying assumptions. Future studies need to increase the complexity even more and relax some of our assumptions. In particular, one has to consider non-linear complexes containing loops and multiple branches. Such configurations can give rise to self-folding, self-templating, template inhibition, and gelation [58, 220, 176]. All these features potentially influence the sequence dynamics. However, we expect that these effects only become important in the long time limit once the strands have reached sufficient size for the formation of secondary structures. Consequently, the discussion of the emergence of structured sequences on shorter timescales in the limit of strong kinetic stalling is expected to hold, even if secondary structures are taken into account. In addition, one has to extend the alphabet size from two to four. The question here is whether pure states containing only a minimal number of sub motifs exist. In the future, the model could also include additional reactions such as non-templated polymerization and ligation, and recombination [236, 176, 237, 205, 238, 239, 240] or length selective environments as employed in Chapter 3. The first two reactions probably play a role in the formation of the first short oligomers, whereas a flow-through system preferentially accumulating long strands can trigger a so-called *escalation of polymerization* [27].

6.4.3. Plausibility of a binary alphabet

Our study assumed a binary alphabet following previous theoretical work [207, 241, 60, 202, 203, 221, 242, 206]. While this assumption simplified the analysis, there is also evidence for a two-letter alphabet preceding the modern four-letter alphabet [243, 244, 245, 34, 121, 246]. The plausibility is also underlined by the fact that functional sequences composed of only two types exist [247, 248]. For the sake of generality, we referred to the two types of nucleotides appearing in our model as X and Y. This terminology was motivated by the idea that a pre-RNA, sometimes called *prebioitic XNA*, or alternative RNA nucleotides, may have existed before the modern RNA came into being. [249, 250, 178, 177, 251, 252, 10]. Various backbone chemistries [253, 254, 255, 256, 257, 179], non-canonical nucleotides [84, 85, 86, 87, 258, 259] and chemical modifications [260, 180, 261] are eligible, some of which, e.g., PNA and TNA are more plausible to emerge [262, 263, 264] under the conditions on the early Earth than RNA.

6.4.4. Significance for origins of life

What is the origin of the first ribozymes heralding the transition from the pre-RNA to the RNA world? Catalytic ribosomal activity requires long strands with distinct sequences, which are unlikely to emerge spontaneously from a random pool [58]. In Darwinian evolution, the assembly of low-level building blocks into higher-level entities triggered significant developments [217]. In the light of this evolutionary principle, a multi-step process towards greater complexity eventually resulting in functionality also seems natural in the chemical evolution on the prebiotic Earth. Here, we potentially unveiled one of the first steps following the emergence of early nucleotides. This step forms structured oligonucleotides that could serve as building block on the next higher level of self-organization down the road to functional ribozymes.

In our study, we considered four different model variants with and without kinetic stalling (see Sections. 6.3.3–6.3.8). Since kinetic stalling is probably inevitable in non-enzymatic templated ligation, the latter scenario appears purely academic at first glance [228, 77, 78, 119]. Yet, this model variant is essential to separate the effects and identify kinetic stalling as the crucial mechanism enabling self-enhancing sequence selection (see Sec. 6.3.5). Moreover, the strength of the stalling effect depends on the underlying activation and nucleotide chemistry, [78, 77, 115] and both weak and strong kinetic stalling scenarios, leading to potentially qualitatively different outcomes, are plausible (see Sections. 6.3.6, 6.3.8 and 6.3.9). Furthermore, in an early RNA-World scenario, a primitive ribozyme catalyzing ligations might have a poor ability to discriminate mismatching ends kinetically. In this case, the ribozyme would operate in a regime where thermodynamics mainly controls the discrimination between complementary and non-complementary nucleotides. This regime would be close to the variant without the kinetic stalling of our model.

Moreover, our study revealed that minor differences in the motif-dependent stacking energies significantly affect the dynamics in sequence space. The conceptual experimental studies in Refs. [58, 59] probably did not capture this aspect for the following reason: The experiments used DNA 12- or 20-mers as initial building blocks and a ligase to promote the formation of covalent bonds. The ligase requires overlaps of at least six nucleotides to work efficiently [265]. Moreover, the experiments were performed under temperature cycling. The applied temperature cycling was too fast for the long strands to reach a binding equilibrium. Therefore, the hybridization timescale of mostly complementary hybridization sites is always set by the duration of the cold phases [183]. Hence, subtle variations in the stacking energies are not visible. In contrast, the effective cycling in our model is slow enough for thermodynamics to govern the (de-)hybridization of short strands leading to an amplification of the stacking bias (see Section 6.3.8). The cut-off of the dehybridization rate only affects longer strands emerging from the pool which is already biased. It is worthwhile mentioning that the ligase used in Refs. [58, 59] only joins strands if their ends match the template strand [266]. Consequently, it operates in the infinite kinetic stalling regime.

In non-enzymatic scenarios involving kinetic stalling, the strands formed from the initial pool are already the result of a primary selection process. Selection is not imposed externally but stems from a self-organizing replication network [223]. We showed that the ability to form self-organizing and self-amplifying replication networks is inherent in template-directed growth and does not require higher-level mechanisms such as sequence-specific template inhibition as a result of self-folding, reported previously [58]. In the eighties, Kauffmann promoted the concept of the *autocatalytic set* – a set of molecules that mutually catalyze their formations [128] – as a chemical intermediate on the way to biological life [129]. Since that, autocatalytic sets have been the subject of many theoretical and experimental studies [218, 43, 267, 268, 269, 270, 271]. Once our system has reached a stationary state in the strong kinetic stalling regime, it shows the key features of such an autocatalytic set: strands with a specific pattern promote the formation of new strands of arbitrary length, showing the same pattern. Importantly, this concrete realization of an autocatalytic set emerges naturally from an unstructured initial pool without requiring any external (pre-) engineering. This insight could bridge the gap between strand formation and self-sustaining sequence replication and offers a path towards Darwinian evolution on early Earth.

6.5. Mathematical details

6.5.1. Terminal base pair energies

The energies ϵ_j (with the index j denoting to either $-$ or $+$ end) assigned to terminal nucleotide pairs are functions of the complex structure and the sequence context as well. If a terminal nucleotide pair coincides with the end of a complex, a so-called *blunt end*, there is no additional contribution, i.e., $\epsilon_j = 0$. In contrast, a terminal nucleotide pair followed or preceded by an unpaired nucleotide, a so-called *dangling end*, adds a non-zero contribution to hybridization energy, i.e., $\epsilon_j \neq 0$ [76, 107].

Again, a mismatch results in an energetic penalty $\epsilon_j > 0$, weakening the binding,

process	parameter	value
dehybridization	$\overline{\epsilon_{\text{com}}}, \epsilon_{1\text{nc}}, \delta_\epsilon$	$-0.625, 0.375, [-0.15, 0]$

Table 6.3.: Summary of energy parameters for dangling ends used in Section 6.3

while a complementary nucleotide pair leads to a reward $\epsilon_j < 0$ increasing stability. As for the stacking energies $\gamma\left([P_i\, P_{i+1}]\right)$, we distinguish between complementary *alternating* terminal configurations such as

$$\begin{bmatrix} & Y \\ & \cdot \\ Y- & X \end{bmatrix} \text{ or } \begin{bmatrix} Y- & X \\ \cdot & \\ X & \end{bmatrix}, \tag{6.27}$$

and complementary *homogeneous* terminal configurations as for example

$$\begin{bmatrix} & X \\ & \cdot \\ Y- & Y \end{bmatrix} \text{ or } \begin{bmatrix} Y & \\ \cdot & \\ X- & X \end{bmatrix}. \tag{6.28}$$

We denote the associated energies by ϵ_{alt} and ϵ_{hom}. As before, we treat the energy difference

$$\delta_\epsilon = \epsilon_{\text{alt}} - \epsilon_{\text{hom}}. \tag{6.29}$$

as a variable parameter. Moreover, for a dangling end configuration involving one mismatch, we assume a constant contribution ϵ_{1nc}. The terminal nucleotide pair contributions obey the inequality

$$\epsilon_{\text{alt}} \leq \overline{\epsilon_{\text{com}}} \leq \epsilon_{\text{hom}} < \epsilon_{\text{1nc}}, \tag{6.30}$$

with $\overline{\epsilon_{\text{com}}} = \left(\epsilon_{\text{alt}} + \epsilon_{\text{hom}}\right)/2$. Parameters values used in Section 6.3 are summarized in Tab. 6.3. For simplicity, we set the energy difference between alternating and homogeneous dangling end contributions to half of the value of the difference between full alternating and homogeneous blocks, i.e.,

$$\delta_\epsilon = \frac{1}{2}\delta_\gamma. \tag{6.31}$$

Next, we consider the contribution to the hybridization energy resulting from a terminal nucleotide pair, part of a ligation site. While, in principle, one can choose the contributions due to dangling and blunt ends freely (within a reasonable range), contributions coming from ligation sites are constrained. The reason is that one ligation site is involved in two hybridization sites.

To formulate this constraint, we introduce the total hybridization energy $\beta\Delta G_{\text{tot}}(C)$ of a complex C. $\beta\Delta G_{\text{tot}}(C)$ is obtained by first summing over the stacking energies of all *continuous* nearest neighbor blocks within all hybridization sites of the complex, i.e., all blocks of the form

$$\begin{bmatrix} X- & Y \\ \cdot & \cdot \\ Y- & X \end{bmatrix} \text{ or } \begin{bmatrix} X- & X \\ \cdot & \cdot \\ Y- & X \end{bmatrix} \text{ or } \ldots, \tag{6.32}$$

where the nucleotide pairs are linked by two covalent bonds (represented by the $-$ symbols). Next, we add the sum over all dangling end contributions. In the last step, we add an energy contribution for every ligation site, i.e., for every *noncontinuous* nearest neighbor block of the form

$$\begin{bmatrix} X & Y \\ \cdot & \cdot \\ Y- & X \end{bmatrix} \text{ or } \begin{bmatrix} X- & X \\ \cdot & \cdot \\ Y & X \end{bmatrix} \text{ or } \ldots, \tag{6.33}$$

where one covalent bond is missing. Since the covalent bond does not affect the stacking interaction, the energy contribution to the total hybridization energy $\beta \Delta G_{\text{tot}}(C)$ of noncontinuous blocks is the same as for the corresponding continuous blocks [76]. For γ as a function of the (non)continuous nearest neighbor blocks, we have, for example

$$\gamma\left(\begin{bmatrix} X- & Y \\ \cdot & \cdot \\ Y- & X \end{bmatrix}\right) = \gamma\left(\begin{bmatrix} X & Y \\ \cdot & \cdot \\ Y- & X \end{bmatrix}\right). \tag{6.34}$$

With that, $\beta \Delta G_{\text{tot}}(C)$ reads

$$\beta \Delta G_{\text{tot}}(C) = \sum_{i \in \substack{\text{continuous} \\ \text{blocks}}} \gamma_i + \sum_{d \in \substack{\text{dangling} \\ \text{ends}}} \epsilon_d + \sum_{l \in \substack{\text{ligation} \\ \text{sites}}} \gamma_l \tag{6.35}$$

Note that we would obtain an identical total hybridization energy if we would replace all non-continuous nearest neighbor blocks in the complex C with continuous blocks, i.e., if we would join the $+$ and $-$ strands at all ligation sites.

We now consider two complexes C_1 and C_2 reacting to a new complex C_3 containing one or two new ligation site. The energy difference between the states before and after the reaction has to correspond to the hybridization energy Γ_{new} associated with the newly formed hybridization site, i.e.,

$$\Gamma_{\text{new}} \overset{!}{=} \beta \Delta G_{\text{tot}}(C_3) - [\beta \Delta G_{\text{tot}}(C_1) + \beta \Delta G_{\text{tot}}(C_2)] \tag{6.36}$$

Γ_{new} could also be interpreted as the energy that is needed to reverse the reaction. The constraint formulated via Eq. (6.36) defines the energetic contribution of the newly formed ligation site(s) to Γ_{new} associated with the newly formed hybridization site.

In practice, the constraint Eq. (6.36) reduces to the difference between the stacking contribution(s) $\gamma_l^{(C_3)}$ associated with the newly formed non-continuous nearest neighbor block(s) in the new complex C_3 and the corresponding dangling end contribution(s) $\epsilon_d^{(C_1)}$ and $\epsilon_d^{(C_2)}$ of the old complexes C_1 and C_2.

Newly formed ligation sites also affect the hybridization energies of hybridization sites that were already existing in C_1 or C_2 before the reaction occurred. Hence, these hybridization energies have to be recalculated comparing the total hybridization energies of the new complex C_3 and the compounds that result from virtually dissolving the hybridization site. This procedure is explained in more detail in App. 6.5.2, where we explicitly derive hybridization energies in several exemplary complex configurations.

Note that in general

$$\beta\Delta G_{\mathrm{tot}}(C) \neq \sum_{h \in \substack{\text{hyb.} \\ \text{sites}}} \Gamma_h. \tag{6.37}$$

This inequation becomes an equation only if the complex C does not contain any ligation sites. Moreover, we can interpret $\beta\Delta G_{\mathrm{tot}}(C)$ as the total energy stored within the complex. It is equal to the energy difference between the fully disassembled state, where we have only single strands.

6.5.2. Examples for hybridization energies

To illustrate the calculation of the hybridization energy, we consider four complexes, sketched in Fig. 6.5.

In example Fig. 6.5a, the duplex C_1, and the single strand C_2 react to the triplex C_3. Before the reaction, the total hybridization energies of the complexes are

$$\beta\Delta G_{\mathrm{tot}}(C_1) = \epsilon_{\mathrm{hom}} + \gamma_{\mathrm{alt}} + \epsilon_{\mathrm{alt}}, \tag{6.38}$$

$$\beta\Delta G_{\mathrm{tot}}(C_2) = 0. \tag{6.39}$$

After the reaction, the total hybridization energy is

$$\begin{aligned}\beta\Delta G_{\mathrm{tot}}(C_3) = {} & \epsilon_{\mathrm{hom}} + \gamma_{\mathrm{alt}} + \epsilon_{\mathrm{alt}} \\ & + \epsilon_{\mathrm{alt}} + \gamma_{\mathrm{alt}} + 2\gamma_{\mathrm{hom}}\end{aligned} \tag{6.40}$$

As stated above, the hybridization energy Γ_{new} associated with the new hybridization site is the difference between the total hybridization energies before and after the reaction, i.e.,

$$\begin{aligned}\Gamma_{\mathrm{new}} &= \beta\Delta G_{\mathrm{tot}}(C_3) - [\beta\Delta G_{\mathrm{tot}}(C_1) + \beta\Delta G_{\mathrm{tot}}(C_2)] \\ &= \epsilon_{\mathrm{alt}} + \gamma_{\mathrm{alt}} + 2\gamma_{\mathrm{hom}}.\end{aligned} \tag{6.41}$$

Since the reaction did not lead to a new ligation site, Γ_{new} corresponds to the sum over all nearest-neighbor blocks plus the danging end contributions.

In example Fig. 6.5b, two single strands C_1 and C_2 form a new duplex C_3. The total hybridization energy before the reaction is zero. Hence, the hybridization energy Γ_{new}

of the new hybridization site equals the total hybridization energy $\beta\Delta G_{\text{tot}}(C_3)$ after the reaction, i.e.,

$$\begin{aligned}\Gamma_{\text{new}} &= \beta\Delta G_{\text{tot}}(C_3) - 0 \\ &= \epsilon_{\text{alt}} + \gamma_{\text{alt}} + 2\gamma_{1\text{nc}} + \gamma_{\text{hom}} + \epsilon_{\text{hom}}\end{aligned} \tag{6.42}$$

Again, since no ligation sites are involved, Γ_{new} is equivalent to the sum over stacking and end contributions.

In example Fig. 6.5c, the emerging complex C_3 features two new ligation sites. The total hybridization energies of the initial complexes C_1 and C_2 are

$$\beta\Delta G_{\text{tot}}(C_1) = \gamma_{\text{alt}} + \epsilon_{\text{alt}}, \tag{6.43}$$

$$\beta\Delta G_{\text{tot}}(C_2) = \gamma_{\text{alt}} + 2\epsilon_{\text{alt}}. \tag{6.44}$$

The total energy of complex C_3 is

$$\beta\Delta G_{\text{tot}}(C_3) = 5\gamma_{\text{alt}} + \epsilon_{\text{alt}}. \tag{6.45}$$

As before, the difference between the total hybridization energies before and after the reaction determines the hybridization site energy Γ_{new} of the new hybridization site,

$$\begin{aligned}\Gamma_{\text{new}} &= \beta\Delta G_{\text{tot}}(C_3) - [\beta\Delta G_{\text{tot}}(C_1) + \beta\Delta G_{\text{tot}}(C_2)] \\ &= 3\gamma_{\text{alt}} - 2\epsilon_{\text{alt}}.\end{aligned} \tag{6.46}$$

In the last example, Fig. 6.5d, a mononucleotide C_1 hybridizes to the duplex C_2 resulting in the triplex C_3. The hybridization energies before and after the reaction are

$$\beta\Delta G_{\text{tot}}(C_1) = 0, \tag{6.47}$$

$$\beta\Delta G_{\text{tot}}(C_2) = \epsilon_{\text{alt}} + \gamma_{\text{alt}} + 2\epsilon_{\text{hom}}, \tag{6.48}$$

$$\beta\Delta G_{\text{tot}}(C_3) = \epsilon_{\text{alt}} + 2\gamma_{\text{alt}} + 2\epsilon_{\text{hom}}. \tag{6.49}$$

Consequently, the hybridization energy Γ_{new} assigned to the mononucleotide is

$$\begin{aligned}\Gamma_{\text{new}} &= \beta\Delta G_{\text{tot}}(C_3) - [\beta\Delta G_{\text{tot}}(C_1) + \beta\Delta G_{\text{tot}}(C_2)] \\ &= \gamma_{\text{alt}}.\end{aligned} \tag{6.50}$$

In the examples Fig. 6.5c and Fig. 6.5d, we also have to recalculate the hybridization energy of the hybridization sites that were present before the reactions since they now also involve an energetic contribution from a ligation site. To this end, we virtually dissolve the hybridization site(s) that were already existing and then virtually reassemble these hybridization sites again. Reassembling the hybridization site(s), we apply the same procedure of calculating the hybridization energy as described before. The updated hybridization energy (energies) obtained that way now include(s) the correct contribution of the newly formed ligation site(s). Moreover, in the examples Fig. 6.5a, Fig. 6.5c and Fig. 6.5d, we have to reconsider the channel factors associated with already existing hybridization sites. To update these channel factors, we again virtually disassemble the complex into its compounds (see Fig. 6.5e). We then virtually reassemble the parts and thereby count the number of different configurations that would be possible. With the updated channel factors and hybridization energies, we recompute the dehybridization rates according to Eq. (6.7).

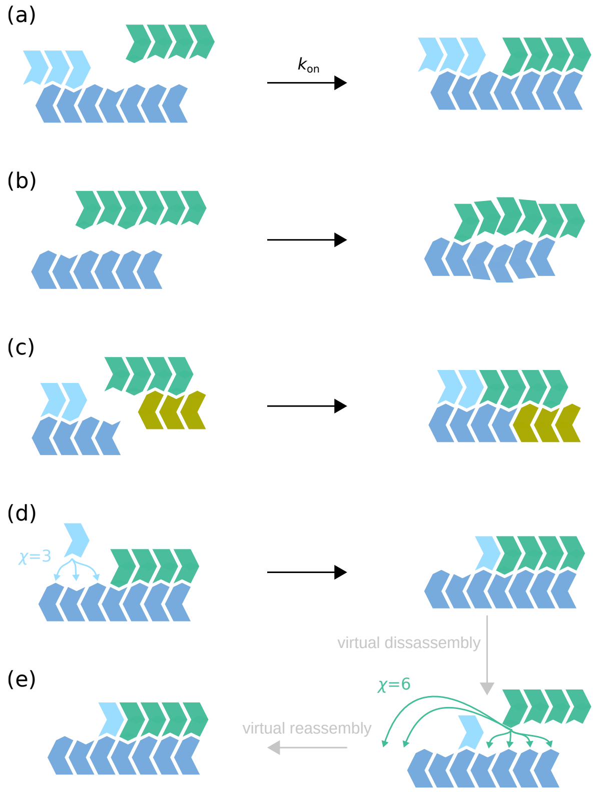


Figure 6.5.: (a) A duplex C_1 with two dangling ends and a single strand C_2 form a new triplex C_3 with a blunt end. C_3 has no ligation site. (b) Two single strands C_1 and C_2 react to a duplex C_3 displaying a mismatch and two dangling ends. (c) Two duplexes C_1 and C_2 hybridize to new complex C_3 featuring two new ligation sites. (d) A mononucleotide C_1 hybridizes onto a duplex C_2. The new triplex has a new ligation site. (e) We need to update the channel factor χ to renew the dehybridization rate k_{off} associated with the hybridization site that already existed before the binding of the monomer. To this end, we virtually dissolve this hybridization site and directly reassemble the complex and recount the possible reaction channels and obtain a new (integer) value for the channel factor χ (see text).

6.5.3. Thermodynamics of hybridization

With the elementary rates defined in Eqs. (6.6) and (6.7), the total free energy $\Delta\mathcal{G}_{\text{tot}}(C)$ of a complex C is found to be

$$\beta\Delta\mathcal{G}_{\text{tot}}(C) = \beta\Delta G_{\text{tot}}(C) + \rho\ln(2), \tag{6.51}$$

for constant environmental conditions (see Chapter 4). The first term on the right-hand side is the total hybridization energy defined in Eq. (6.35) and the second term is a *symmetry penalty* that occurs if the complex is rotationally symmetric ($\rho = 1$) and is zero ($\rho = 0$) otherwise. The free energy $\beta\Delta\mathcal{G}_{\text{tot}}(C)$ is linked to the dissociation constant K_D occurring in a mass-action approach where all concentrations are expressed in units of the standard concentration c° via [209]

$$\beta\Delta\mathcal{G}_{\text{tot}}(C) = \ln\left(K_D\right). \tag{6.52}$$

Thermodynamically, the symmetry penalty can be understood as a decrease in the (standard internal) entropy by a factor of $\ln(2)$ due to the rotational symmetry.

Kinetically, it is rationalized by looking at the interaction probability of two complexes belonging to the same species versus the interaction probability of two complexes representing different species. For equal concentrations, complexes representing different species interact twice as often as complexes belonging to the same species. While the complex resulting from a collision between distinguishable complexes is never symmetric, the interaction of identical molecules always leads to a complex with a rotational symmetry (see Chapter 4). Hence, the $\rho\ln(2)$-term arises naturally in any collision based kinetic model [235]. We emphasize that the symmetry penalty is due to a reduced product-formation rate rather than a decrease in stability of the complex.

Moreover, the symmetry penalty also appears in the standard databases for free energies of hybridized oligonucleotides [107, 76]. These databases also add a constant *initiation penalty* to the total free energy $\beta\Delta\mathcal{G}_{\text{tot}}(C)$. The initiation penalty is a constant multiplied by $(n-1)$, where n is the number of strands forming the complex. The initiation penalty accounts for the loss of system entropy due to the fusion of separate entities into one new complex. However, this penalty term can be set to zero through a rescaling of concentrations and therefore does not occur in our approach (see Chapter 4).

6.5.4. Distribution of longer motifs

In Section 6.3.3, we discussed a scenario without kinetic stalling ($\sigma_1 = \sigma_2 = 1$) and without energetic bias ($\delta_\gamma = 0$). There, the zebraness converged to $Z = 0.5$ in all individual realizations, pointing to an entirely random sequence pool (see Fig. 6.2c).

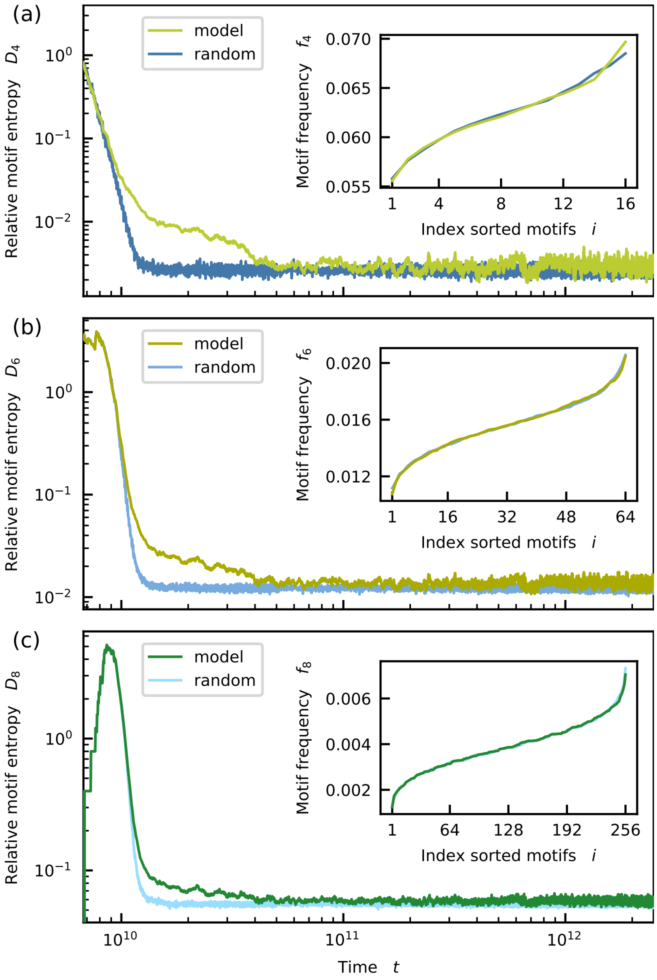


Figure 6.6.: Relative entropies of the distributions of (sub)motifs of size four (a), six (b), and eight (c) as a function of time. Green curves: data obtained from simulations of the full model dynamics. Blue curves: data generated by the corresponding random process. At every time point, the number of (sub) motifs generated by the random process equals the number of (sub)motifs found in the simulation output. For small times, correlations in sequence lead to an increased relative entropy in the model dynamics. For large times, model dynamics and random processes yield similar results. The insets show the normalized frequency of (sub)motifs sorted by abundance in for the last time point $t = 10^4$.

However, this result does not exclude a non-random distribution of longer (sub)motifs in the steady-state. To rule out correlations on larger scales, we analyze the distributions of (sub)motifs of lengths $n > 2$ in detail. Therefore, we compare the (sub)motif distributions $\mathcal{P}_n$ from the simulated model dynamics to the (sub)motif distributions $\mathcal{R}_n$ obtained from a true random process. To this end, we count the number of (sub)motifs of size n contained in all strand with $L \geq n$ from the simulation output at every time point. (A strand of length $L \geq n$ contains $L - n + 1$ (sub)motifs of size n.) For every time point we also generate an equal number of random motifs of size n. We then analyze the evolution of $\mathcal{P}_n$ and $\mathcal{R}_n$ by means of the relative entropy D_n (also called Kullback-Leibler divergence) with respect to the uniform distribution $\mathcal{U}_n$ [272]. For the distributions $\mathcal{P}_n$ and $\mathcal{U}_n$, the relative entropy $D_n(\mathcal{P}_n,\mathcal{U}_n)$ is given by

$$D_n(\mathcal{P}_n,\mathcal{U}_n) = \sum_{m \in M_n} \mathcal{P}_n(m) \log_2 \left[\frac{\mathcal{P}_n(m)}{\mathcal{U}_n(m)} \right], \tag{6.53}$$

where the sum runs over all possible motifs $m \in M_n$ of the given length n. Using that

$$\mathcal{U}_n(m) = \frac{1}{2^m}, \tag{6.54}$$

Eq. (6.53) simplifies to

$$D_n(\mathcal{P}_n,\mathcal{U}_n) = m + \sum_{m \in M_n} \mathcal{P}_n(m) \log_2 \left[\mathcal{P}_n(m) \right]. \tag{6.55}$$

The relative entropy measures how much the distribution $\mathcal{P}_n$ deviates from the distribution $\mathcal{U}_n$. The smaller $D_n(\mathcal{P}_n,\mathcal{U}_n)$, the more similar are $\mathcal{P}_n$ and $\mathcal{U}_n$. Since $\mathcal{U}_n$ is the uniform distribution, i.e. the distribution with the largest entropy, we have that $D_n(\mathcal{P}_n,\mathcal{U}_n) \geq 0$. Only a few long strands exist at early times and the occupation numbers for most motifs are zero. For this reason, we can not directly compare $\mathcal{P}_n$ and $\mathcal{R}_n$ by means of the relative entropy. The definition of relative entropy Eq. (6.53) requires that the relative frequency of any motif $m \in M_n$ is always finite for the second distribution. Therefore, $D_n(\mathcal{P}_n, \mathcal{R}_n)$ would be ill-defined as long as are not all motifs present.

Fig. 6.6a–Fig. 6.6c show the evolution of $D_n(\mathcal{P}_n,\mathcal{U}_n)$ (green lines) and $D_n(\mathcal{R}_n,\mathcal{U}_n)$ (blue lines) for $n = 4, 6, 8$. The high values of the relative entropies at early times stem from small numbers of (sub)motifs. As the number of (sub) motifs grows with time, $\mathcal{P}_n$ and $\mathcal{R}_n$ become more similar to the uniform distribution, and the relative entropies decay. However, the decay of $D_n(\mathcal{P}_n,\mathcal{U}_n)$ is slower, indicating that sequence correlations exist during early strand growth. Eventually, $D_n(\mathcal{P}_n,\mathcal{U}_n)$ and $D_n(\mathcal{R}_n,\mathcal{U}_n)$ converge to the same stationary value. This result implies that the motif distribution resulting from the model dynamics is entirely random. The insets in Fig. 6.6 show the relative frequencies f_n of the motifs sorted by abundance for the last time point. Simulation output and random process give similar results.

6.5.5. Onset of growth

In Sections. 6.3.3–6.3.8, we have seen that the mean length $\overline{L}$ grows rapidly after once the first new oligomers are formed from existing mononucletoides and dimers. Plotting the data with a logarithmic y-axis reveals that the growth dynamics is approximately exponential (see Fig. C.1 in the Appendix). Moreover, increasing the strength of kinetic stalling (the stacking bias) shifts the onset of the rapid growth phase to later (earlier) times. The goal of this section is to derive an approximative formula for the onset of growth $\widehat{t}$.

To this end, we first focus on the scenario with $\delta_\gamma = 0$ and $\sigma_1 = \sigma_2 = 1$ discussed in Section 6.3.3 (see Fig. 6.2a). Formally we define $\widehat{t}$ as the abscissa of the intersection of the tangents to $\overline{L}(t)$ at $t = 0$ and the point where the growth is strongest (see dotted lines in Fig. 6.7a).

To gain a microscopic picture of the growth dynamics, we look at the statistics of ligation events over time and distinguish between *first-order* and *higher-order ligations*. First-order ligations correspond to reactions where two monomers ligate to a dimer on a template of arbitrary length $L \geq 2$. In contrast, higher-order ligations involve at most one monomer and lead to new strands with $L \geq 3$. (We can further differentiate higher-order ligations into *second* and *third-order* ligations depending on whether the reaction comprises one or no monomer.) Fig. 6.7b reveals that $\widehat{t}$ coincides approximately with the time point t_{high} where higher-order ligations become more frequent than first-order ligations, i.e., for $t = t_{\text{high}}$ such that we have

$$r_h\left(t\right) = r_f\left(t\right) \tag{6.56}$$

where $r_h\left(t\right) = \frac{N_h(t)}{\nu \Delta t}$ and $r_f\left(t\right) = \frac{N_f(t)}{\nu \Delta t}$ are the numbers N_f and N_h of first- and higher-order ligations per time interval $\Delta t = 2.5 \times 10^8\, t_0$. Moreover, ν is a normalization constant. (In fact, Fig. 6.7b reveals that $\widehat{t}$ corresponds precisely to the time point, where third-order reactions become more abundant than first-order ligations.)

To obtain an analytic estimate t_{est} for t_{high}, we consider the dynamical mean-field rate-equation for the evolution of the dimer concentration $c_2(t)$. This equation neglects the explicit sequence dependence and coarse-grains (de)hybridization and ligation into one effective *extension*. It is given by

$$\begin{aligned}\dot{c_2} = {} & (c_1)^2 c_2 k_{\{1,1|2\}} + (c_1)^2 \sum_{L\geq 3} c_L k_{\{1,1|L\}} \\ & - c_1 (c_2)^2 k_{\{1,2|2\}} - c_2 \sum_{L_1 \geq 2} \sum_{L_2 \geq 2} c_{L_1} c_{L_2} k_{\{2,L_1|L_2\}} \\ & - c_2 k_{\mathrm{cut}} + \sum_{L\geq 3} 2 c_L k_{\mathrm{cut}}.\end{aligned} \tag{6.57}$$

The first term on the right-hand side of Eq. 6.57 describes the creation of a new dimer via a first-order ligation with a dimer serving as the template. $k_{\{1,1|2\}}$ is the rate constant for this effective extension process. We explain it below, together with the rate constants for the other extension processes. The second term also relates to the formation of a new dimer but using a template with $L \geq 3$. The reaction is again a first-order ligation. The third term is a loss term accounting for the ligation of a monomer to a dimer, on a dimeric template. This reaction is a higher-order ligation. The fourth term is again a loss term describing a higher-order ligation of a dimer to a strand with $L_1 \geq 2$ on a template with $L_2 \geq 2$. The second to last term accounts for the loss of dimers due to cleavage with rate constant k_{cut}. The last term represents the gain of dimers due to cleavage of strands of length $L \geq 3$. There, a dimer can break apart at either side of the longer strand. Eq. (6.57) is valid if (1) concentrations are small enough such that the total strand concentration is approximately equivalent to the concentration of single strands, (2) complexes composed of more than three strands are negligible and, (3) the time scales for ligation and dehybridization are separated such that $k_{\mathrm{lig}} \ll k_{\mathrm{off}}$. The assumptions (1)-(3) are satisfied (see Fig. C.24 in the Appendix and Tab. 6.2).

We derive the effective rate constant for an extension $k_{\{1,1|2\}}$ as follows. First, we compute the average Boltzmann factor over the set of all complexes comprising exactly two monomers and one dimer $\mathcal{C}_{\{1,1|2\}}$. The Boltzmann factor associated with a specific complex $C \in \mathcal{C}_{\{1,1|2\}}$ is determined by its total binding energy $\Delta\mathcal{G}_{\mathrm{tot}}(C)$ (see App. 6.5.3). The sequence-averaged Boltzmann factor then defines the sequence-averaged dissociation constant $K_{\{1,1|2\}}$ via

$$\frac{1}{K_{\{1,1|2\}}} = \sum_{C \in \mathcal{C}_{\{1,1|2\}}} \frac{e^{-\beta \Delta \mathcal{G}_{\mathrm{tot}}(C)}}{\left|\mathcal{C}_{\{1,1|2\}}\right|}. \tag{6.58}$$

Recall that all concentrations are expressed as a multiple of the reference concentration $c^{\circ} = 1\,\mathrm{mol/l}$. For that reason, the dissociation constant appears as a dimensionless quantity in Eq. (6.58). For complexes formed of exactly one dimer and two monomers, the total binding energy reduces to stacking energy associated with one nearest neighbor block such that

$$\begin{aligned}\frac{1}{K_{\{1,1|2\}}} = \frac{1}{16}\Bigg[&\exp\left\{-\gamma\left(\begin{bmatrix} X & X \\ & \\ X- & X \end{bmatrix}\right)\right\} + \exp\left\{-\gamma\left(\begin{bmatrix} Y & X \\ \cdot & \\ X- & X \end{bmatrix}\right)\right\} + \\ &\exp\left\{-\gamma\left(\begin{bmatrix} X & Y \\ & \cdot \\ X- & X \end{bmatrix}\right)\right\} + \exp\left\{-\gamma\left(\begin{bmatrix} Y & Y \\ \cdot & \cdot \\ X- & X \end{bmatrix}\right)\right\} + \ldots\Bigg].\end{aligned} \tag{6.59}$$

This equation further reduces to

$$\frac{1}{K_{\{1,1|2\}}} = \frac{1}{4}\left[e^{-\gamma_{2\text{nc}}} + 2e^{-\gamma_{1\text{nc}}} + e^{-\overline{\gamma_{\text{com}}}}\right]. \tag{6.60}$$

Second, we multiply the inverse of the average dissociation constant with the ligation rate k_{lig}. The result is

$$k_{\{1,1|2\}} = \frac{k_{\text{lig}}}{K_{\{1,1|2\}}}. \tag{6.61}$$

Coarse-gaining (de)hybridization and ligation into an effective extension this way is valid as long as the ligation timescale is much slower than the dehybridization time scales such that there is enough time for the (de)hybridization dynamics to equilibrate. By our choice of the ligation rate (see Section 6.2.6), this premise is clearly fulfilled for monomers and dimers.

The rate constant $k_{\{1,2|2\}}$ is obtained analogously. The other rate constants are less trivial. However, we will neglect them later on, anyway.

For $t < t_{\text{high}}$, mostly monomers and dimers populate the pool. We, therefore, assume that t_{high} roughly matches the time point where the loss terms related to higher-order ligations balance the gain terms in Eq. (6.57). Eq. (6.57) has no analytic solution. We thus have to make (crude) approximations to obtain the estimate t_{est} for t_{high}. First, we neglect all terms that involve strands of length $L > 2$ or do not include at least one monomer. The resulting simplified equation then reads:

$$\dot{c_2} = (c_1)^2 c_2 k_{\{1,1|2\}} - c_1 (c_2)^2 k_{\{1,2|2\}} - c_2 k_{\text{cut}}. \tag{6.62}$$

Eq. (6.62) is a cubic ordinary differential equation since $c_1 + 2c_2$ is a constant. Hence, the simplified equation still does not allow for a closed analytic solution for $c_2(t)$, and further (crude) approximations will be necessary.

For $t \to 0$ the dimer concentration grows exponentially (see dotted line in Fig. 6.7), i.e.,

$$c_2(t) \approx c_2(0) \exp\left[\left([c_1(0)]^2 k_{\{1,1|2\}} - k_{\text{cut}}\right) t\right]. \tag{6.63}$$

We now assume the monomer concentration to remain constant, i.e., $c_1(t) = c_1(0) \approx c_{\text{tot}}$. For $t \to \infty$, the dimer concentration then approaches a steady state concentration $\widetilde{c}_2$ which is given by

$$\widetilde{c}_2 = \frac{c_{\text{tot}}^2 k_{\{1,2|2\}} - k_{\text{cut}}}{c_{\text{tot}} k_{\{1,1|2\}}}. \tag{6.64}$$

With that, we estimate the time point for which the right-hand side of Eq. (6.62) vanishes by equating Eq. (6.63) and Eq. (6.64). We obtain

$$t_{\text{est}} = \frac{\log\left[c_{\text{tot}}^2 k_{\{1,1|2\}} - k_{\text{cut}}\right] - \log\left[c_{\text{tot}} c_2(0) k_{\{1,2|2\}}\right]}{c_{\text{tot}}^2 k_{\{1,1|2\}} - k_{\text{cut}}}. \tag{6.65}$$

t_{est} from Eq. (6.65) is an estimate for the time point t_{high} where higher-order ligations become more frequent than first-order ligations. Comparing this estimate to the exact value extracted from simulation data in Fig. 6.7b, we see that t_{est} is only slightly smaller than t_{high}. Hence, Eq. (6.65) yields a solid estimate for the transition from the first-order to the higher-order regime. Moreover, Fig. 6.7a shows that t_{est} matches the $\widehat{t}$ with less then 10% error. We conclude that Eq. (6.65) is a useful approximation for the onset of the rapid growth. In addition, t_{est} corresponds well to the time point where the dimer concentration c_2 reaches a maximum (see Fig. 6.7c) underlining the validity of our initial assumptions. In Fig. C.3 in the Appendix, we show that the prediction Eq. (6.65) is robust under parameter variations.

We now turn to the more general case involving energetic bias $\delta_\gamma < 0$ and kinetic stalling and $\sigma_1 = \sigma_2 < 1$. The energetic bias is automatically accounted for by averaging over all complex configurations. To include kinetic stalling, we introduce the effective sequence-averaged dissociation constant $\widetilde{K}_{\{1,1|2\}}$ in analogy to Eq. (6.58) as

$$\frac{1}{\widetilde{K}_{\{1,1|2\}}} = \sum_{C \in \mathcal{C}_{\{1,1|2\}}} \left[\frac{e^{-\Delta\mathcal{G}_{\text{tot}}(C)}}{\left|\mathcal{C}_{\{1,1|2\}}\right|} \mathcal{S}(C)\right], \tag{6.66}$$

where we weight every term in the sum on the right-hand side with the *overall effective stalling* factor

$$\mathcal{S}(C) = \Phi_-\left(\kappa_{-1}(C), 1\right) \Phi_+\left(\kappa_{+1}(C), 1\right). \tag{6.67}$$

$\mathcal{S}(C)$ considers both the + and the − monomer at the ligation site (see Eq. (6.15)). An analogous definition holds for $\widetilde{K}_{\{1,2|2\}}$. To compute $k_{\{1,1|2\}}$ and $k_{\{1,2|2\}}$ for the rate equation Eq. (6.62) we now use $\widetilde{K}_{\{1,1|2\}}$ and $\widetilde{K}_{\{1,2|2\}}$ (see Eq. (6.61)). Figs. C.5–C.12 and Figs. C.18–C.23 in the Appendix show that the generalized approach also yields a reasonable estimate for the onset of growth.

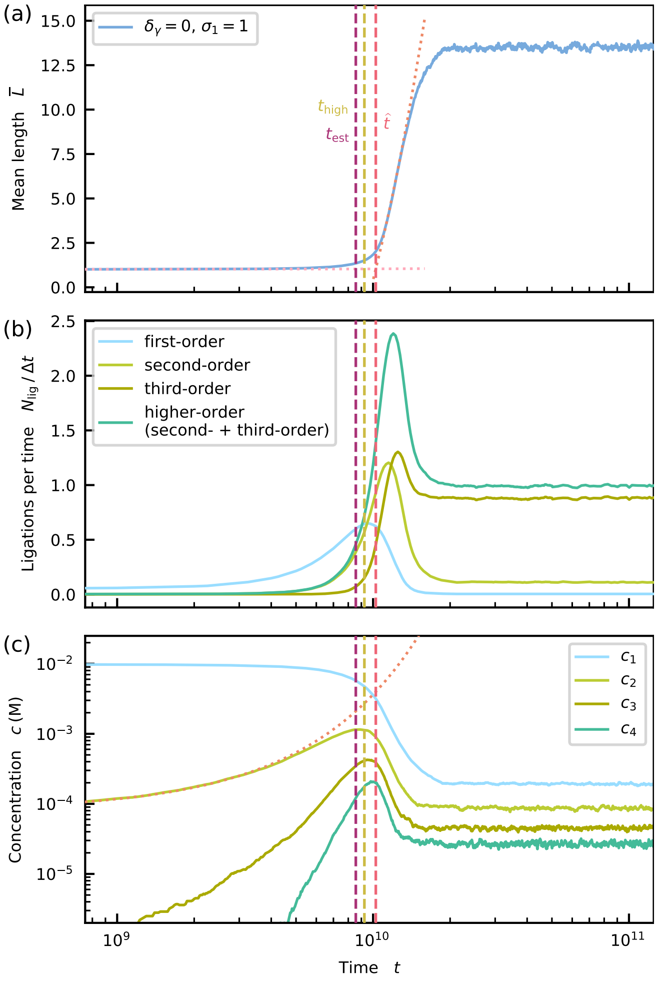


Figure 6.7.: (a) We formally define the onset of growth $\hat{t}$ (dashed line) by intersecting the tangents to $\overline{L}(t)$ at $t = 0$ and the point where the increase is steepest (dotted lines). (b) t_high is defined as the time point where higher-order ligations become more frequent than first-order ligations. Our estimate t_est obtained from Eq. (6.65) matches t_high well (compare dashed lines). Curves are normalized such that, on average, there is one ligation per time interval in the steady-state. (c) The initial exponential growth of the dimer concentration is described by Eq. (6.63) (dotted line). The dimer concentration has a maximum at t_high.

6.6. Application to primer extension

This section investigates a typical *primer extension* scenario, where a *primer* bound to one longer *template* becomes extended stepwise in the absence of thermal cycling. We assume that the primer-template complex is stable, i.e., does not dissociate, and that the solution surrounding this complex only contains mononucleotides. The first assumption implies that the primer is long enough, such that the dehybridization timescale is much larger than the (effective) extension timescale (see Section 6.2.6). Moreover, we focus on ligations involving the (partially extended) primer and neglect the formation of new dimers from mononucleotides on the template. It is well known experimentally that non-complementary nucleotides at the primer's end slow down the extension process and trigger the accumulation of misincorporation. These effects stem from the interplay of two contributions. Kinetic stalling reduces the bare ligation rate. In addition, mismatches at the primer's end weaken the monomer binding, increase its dehybridization rate, and render the next extension less probable. Moreover, non-complementarities at the primer's end reduce the thermodynamic discrimination of a hybridized monomers resulting in an increased fraction of misincorporations.

To develop a quantitative description based on our model for (de)hybridization and ligation (see Section 6.2), we compare the two situations sketched in Fig. 6.8. In Fig. 6.8a, no mismatches occur, and the extension (hybridization and subsequent ligation) proceeds in an unperturbed way. In Fig. 6.8b, the primer terminates with a mismatch. According to our model energy model (see Section 6.2.4 and App. 6.5.1), the hybridization site energy assigned to the mononucleotide in Fig. 6.8b is less negative than in Fig. 6.8a. The single nucleotide in Fig. 6.8b will therefore unbind faster. Hence, primer extension becomes less likely. We call this effect *thermodynamic stalling*. In addition, the bare ligation rate is multiplied by a factor $\Phi_{-}(\kappa_{-2} = 1, \kappa_{-1} = 0) = \sigma_1 \leq 1$ in the situation of Fig. 6.8b as described in Section 6.2.5. Therefore, primer extension becomes even more unlikely. In the main text, we referred to this contribution as *kinetic stalling*. Note that even in the scenario without kinetic stalling ($\sigma_1 = \sigma_2$=1) in Section 6.3.4, thermodynamic stalling due to enhanced dehybridization rates is always present.

We now quantify the thermodynamic stalling contribution. To this end, we compare the dehybridization rates $k_{\mathrm{off}}^{(c)}$ and $k_{\mathrm{off}}^{(n)}$ of the mononucleotides for the complementary and non-complementary primer termini. For simplicity, we assume an energetically unbiased scenario where $\delta_\gamma = \delta_\epsilon = 0$. The dehybridization rates depend exponentially on the hybridization energies $\Gamma^{(c)}$ and $\Gamma^{(n)}$. Applying the procedure to compute

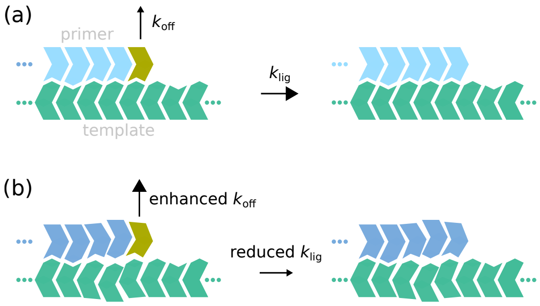


Figure 6.8.: A monomer hybridized adjacent to matching primer terminus (a) has a lower dehybridization rate than a monomer bound next to a non-complementary terminus (b), leading to thermodynamic stalling. Moreover, kinetic stalling reduces the ligation rate.

hybridization energies described in Section 6.5.2, $\Gamma^{(c)}$ and $\Gamma^{(n)}$ are given by

$$\Gamma^{(c)} = \overline{\gamma_{\mathrm{com}}} + \overline{\epsilon_{\mathrm{com}}} - \overline{\epsilon_{\mathrm{com}}} = \overline{\gamma_{\mathrm{com}}}, \tag{6.68}$$

$$\Gamma^{(n)} = \gamma_{1\mathrm{nc}} + \overline{\epsilon_{\mathrm{com}}} - \epsilon_{1\mathrm{nc}}. \tag{6.69}$$

The ratio of $k_{\mathrm{off}}^{(c)}$ and $k_{\mathrm{off}}^{(n)}$ now defines defines the thermodynamic stalling effect ϑ, i.e.,

$$\vartheta = \frac{k_{\mathrm{off}}^{(c)}}{k_{\mathrm{off}}^{(n)}} = \exp\left[\Gamma^{(c)} - \Gamma^{(n)}\right] \tag{6.70}$$

$$= \exp\left[\overline{\gamma_{\mathrm{com}}} + \epsilon_{1\mathrm{nc}} - (\overline{\epsilon_{\mathrm{com}}} + \gamma_{1\mathrm{nc}})\right] \tag{6.71}$$

Note that channel factors do not play a role here since they cancel out.

Multiplying the kinetic and the thermodynamic stalling factors yields the *combined stalling* factor. For the stacking parameters used in the main text, we obtain $\vartheta \approx 0.5$. For $\sigma_1 = 0.1$, the combined stalling factor has a value of 0.05. This result aligns with overall stalling factors ranging between 0.1 and 0.003 measured in non-enzymatic primer extension experiments [77, 78, 119].

Primer extension experiments typically also consider the error fraction ω, which is defined as the ratio of the rate for an erroneous extension and the overall extension rate. We now quantify the error fraction that results from our model. To this end, we introduce the dehybridization rates $k_{\mathrm{off,r}}^{(c)}$ and $k_{\mathrm{off,w}}^{(c)}$ for *right* and *wrong* mononucleotides hybridized adjacent to a complementary primer terminus. Assuming that monomer concentrations are sufficiently low such that there is no competition for the extension site and that equal amounts of both nucleotide types are present, the error fraction is

$$\omega^{(c)} = \left(1 + \frac{k^{(c)}_{\text{off,w}}}{k^{(c)}_{\text{off,r}}\,\sigma_1} \right)^{-1} . \tag{6.72}$$

Relating the hybridization rates to the hybridization energies as before, Eq. (6.72) becomes

$$\omega^{(c)} = \left(1 + \frac{1}{\sigma_1} \exp\left[\gamma_{1\text{nc}} + \epsilon_{1\text{nc}} - (\overline{\gamma_{\text{com}}} + \overline{\epsilon_{\text{com}}})\right] \right)^{-1} . \tag{6.73}$$

Using the same stacking parameters as above and $\sigma_1 = 0.1$, we obtain $\omega^{(n)} = 0.72\%$. This result agrees with the experimentally observed error fraction of 0.8% in a binary DNA system containing C and G monomers [77].

The error fraction $\omega^{(n)}$ for the extension of a primer ending with mismatch is derived analogously and reads

$$\omega^{(c)} = \left(1 + \frac{1}{\sigma_1} \exp\left[\gamma_{2\text{nc}} + \epsilon_{1\text{nc}} - (\gamma_{1\text{nc}} + \overline{\epsilon_{\text{com}}})\right] \right)^{-1} . \tag{6.74}$$

For $\sigma_1 = 0.1$ we find that $\omega^{(n)} \approx 3\omega^{(c)}$. This observation is consistent with the experimental finding that non-complementarities at the primer terminus significantly increase the error fraction in the subsequent (stalled) extension process [78].

7. Summary and outlook

In this thesis we investigated three different conceptual models for self-assembly and replication via template-directed polymerization and ligation. All three scenarios give rise to a rich growth dynamics and emphasize the importance of the two related processes in the context of the emergence of the first self-replicating and evolving systems. Template-directed polymerization and ligation are composite reactions requiring several steps. The timescales associated with the different steps can allow for large variations and depend on the non-equilibrium environmental conditions. Moreover, these time scales can compete with other important timescales, such as the timescale for degradation. All three scenarios have in common that certain emergent phenomena only occur if the different time scales are within certain limits. To a certain extent, the different time scales can be moved by a clever choice of environmental conditions.

First, we considered a scenario where the template strand gets replicated by the extension of a defined primer strand with single nucleotides. In contrast to "modern" kinetic proofreading enabling multiple rounds of error correction, only one error discrimination step per nucleotide occurs in non-enzymatic primer extension. Therefore, the fraction of misincorporations in enzyme-free primer extension is several orders of magnitude larger as compared to the error fraction observed for copying in cells. Consequently plausible scenarios for the spontaneous emergence of molecular evolution and transmission of information require an accuracy-enhancing mechanism. We propose that kinetic error filtering could be a potential prebiotic precursor to kinetic proofreading in template-directed polymerization. Kinetic error filtering is based on the experimental observation that initial incorporation errors stall the downstream copying process, and lead to kinetic discrimination of accurate and erroneous copies. Within a limited time window, only accurate copies reach full length while erroneous ones remain unfinished. Coupling kinetic discrimination to a length-selective environment, where short copies are removed preferentially, gives rise to kinetic error filtering. Kinetic error filtering results in a pool containing mostly accurate full-length copies. As a concrete realization for an environment enabling kinetic error filtering, we suggested a hydrothermal vent system. In such a system, the interplay of convection and thermophoresis leads to limited time windows for the copying process due to periodic thermally induced strand separation and enhanced loss and degradation rates for short strands. Since kinetic error discrimination prevents erroneous partial copies from reaching full length, it necessarily results in an accuracy-yield trade-off per replication cycle. However, we showed that the accuracy and the overall number of completed copies obtained from one template over a long but fixed time could be optimized simultaneously by adjusting the duration of the cycles. This simultaneous optimization is coupled to increased consumption of chemical energy, which has to be provided by the environment and gives rise to a trade-off between accuracy, overall copy number, and energy efficiency. While energy efficiency is not crucial in early replication scenarios, it might become increasingly important as prebiotic living systems become more sophisticated and compete with each other. Moreover, we proved analytically that an upper limit for the accuracy that can be achieved within one cycle by kinetic discrimination exists.

Our study is based on some important simplifying assumptions: (1) Our study assumed that the solution surrounding the primer-template complexes contains only single nucleotides. (2) We neglected template inhibition due to self-folding. (3) We assumed that unfinished potentially erroneous copying products do not rebind to the template strand. As a result of spontaneous template-free oligomerization, the "prebiotic soup" surrounding the primer-template complex should also contain significant quantities of dimers and trimers [68]. Published data [68] and preliminary data from the Richert group suggest that the speed, as well as the accuracy of the primer extension process, could be significantly enhanced in the presence of dimers. Moreover, the presence of longer oligomers as substrates for the extension process could also help to circumvent the inhibition problem arising from self-folding. As shown in Ref. [40] trimers can "invade" and unfold secondary structures and render the copying of template sequences possible, which would be blocked in relatively stable hairpin configurations. Neglecting product-template rebinding is only justified if we assume the product concentration to be much smaller than the primer concentration such that the time scale for the reassociation is smaller than the time scale for the association with a "fresh" primer and the subsequent copying process. This assumption implies that the template concentration has to be very small as well (at most in the nanomolar range [80]). However, as product and template strands become more abundant during the exponential replication process, this assumption breaks down. Then, full or erroneous partial copies could rebind to the "naked" template strand or replace a bound primer strand. In this way erroneous partial copies could reach full length over multiple cycles and increase the overall error fraction. However, in this product-template rebinding scenario as well, short oligomers present in the surrounding solution could have a beneficial effect [80]. If these short oligomers bind reversibly further downstream the extension site, they could prevent full and erroneous partial product strands from docking and displacing the partially extended primer. Future related studies should explicitly account for the different effects described above and estimate their impact on kinetic error filtering. Moreover, our model only distinguished between matching and non-matching nucleotides and used sequence-averaged parameters. Future studies could discard this mean-field picture and consider sequences explicitly. It is known that rates for (un)binding and ligation depend significantly on the exact sequence context [107, 76, 78, 117, 116]. The probability for self-folding also crucially depends on the template sequence. A model treating the sequence dependence in an explicit manner could also be used to search for the optimal template configuration.

In the second scenario, we studied self-assembly resulting from a large pool of strands of various lengths. In this scenario, strands no longer have clear role — each strand acts as a template, primer, or substrate strand. Our goal was to focus on the self-assembly process alone. Therefore, we treated the sequence dependence of hybridization and ligation in a mean-field picture. In this mean-field picture, the binding energy associated with a hybridization site is proportional to its length. The proportionality constant is the (negative) binding energy per nucleotide. Its value was chosen to be of the order of the thermal energy. Moreover, in this mean-field approach, the rate for the ligation process becomes constant. The model resulting from the mean-field description can be considered as a *null model* for self-assembly via templated ligation. The central finding of our study of this null model is that the stationary state (or typical state in the long time limit) is characterized by a non-monotonous strand-length distribution. Moreover, we found analytical expressions that predict the two characteristic lengths corresponding to the local minimum and the maximum of the strand-length distribution. These characteristic length scales associated with the maximum do not depend on the underlying ligation chemistry. We speculate that the strands of the characteristic length associated with the maximum could serve as the building blocks on the next higher level of self-organization ultimately leading to a self-replicating, evolving system.

The non-monotonous shape of the strand-length distribution arises from the competition between different timescales. The first time scale in this competition is the timescale of the dehybridization reaction. This timescale grows exponentially with the length of the hybridization site. The second time scale corresponds to the typical time needed to extend a strand within a given duplex with a short oligomer, i.e., a monomer or a dimer. This timescale is constant and depends on the bare ligation rate and the (fixed) concentrations of monomers and dimers. The third timescale is either the global observation time in case of a closed reaction volume, or the timescale, which is determined by the constant outflux rate for individual complexes in the case of an open system.

In short, the mechanism giving rise to the maximum in the strand-length distribution can be rationalized as follows: Let us image a duplex that has exactly one short single-stranded overhang. The "primer" sitting on the strand with the overhang could now become extended. In our model, an extension is not a single-step process. An extension occurs if a monomer or dimer hybridizes onto the "template" strand next to the primer strand. In this new configuration, a ligation reaction would be possible. However, since the rate for the bare ligation reaction is small, the monomer or dimer will most likely unbind and go back into the solution. Multiple binding and unbinding events of monomers and dimers would be necessary until a ligation finally takes place. If the strands that form a duplex are relatively short as well, the hybridization site within the duplex is necessarily short, too. Consequently, the duplex itself will also dissociate quickly. If both strands forming the duplex are shorter than a characteristic length, the timescale for the dissociation is smaller than the effective time scale for an extension process. However, suppose both strands are long enough, such that they can form a "stable" duplex with a long enough hybridization site. In case the desiccation timescale can become smaller than the effective extension time scale. Now, it is likely, that the primer becomes extended before it unbinds. Every extension following the first one increases the length of the hybridization site. As a result, the stability of the complex increases even more and more extension will probably occur. Thus, stable duplexes can be considered as the starting point for what we call an *extension cascade*. During an extension cascade, the role of primer and template can switch multiple times if, by chance, a long strand that protrudes over the overhang binds to the primer stand. An extension cascade comes to an end when a duplex either reaches a fully hybridized configuration without any overhang, or exits the system. Fully hybridized stable duplexes are "inert", i.e., can not grow any further. Hence, the lengths of the strands forming the duplex are preserved for a certain time. The resulting accumulation of stable duplexes causes an increase in the strand length distribution above a certain characteristic length. This characteristic length coincides with the minimal strand length that is required to form a complex, stable enough to be the starting point for an extension cascade. Eventually, a fully hybridized duplex will dissociate. The strands "freed" by the dissociation can grow further in via another extension cascade. In an open system, a freed strand eventually becomes "trapped" in a fully hybridized duplex with a length such that a dissociation becomes less likely than an outflux event. The length of fully hybridized duplexes, which the dissociation timescale becomes equal the inverse of the outflux rates, coincides with the position of the maximum. In a closed system, the global observation time plays the role of the timescale for dissociation. Since the dissociation time scale increases exponentially with the length of the hybridization site, the position of the maximum then shifts to longer lengths logarithmically in time.

To show that the non-monotonous strand-length distribution feature is a generic feature, we designed a proof-of-principle experiment using random DNA 12-mers to mimic the basic building blocks of lengths one and two in our theoretical model. This proof-of-principle experiment employing a thermocycler and a ligase enzyme to speed up the assembly process was performed in the Braun group and indeed reproduced the non-monotonous behavior.

In order to study the emergent phenomena described above, we first had to develop an appropriate simulation framework that is able to tread the hybridization, dehybridization, and ligation as separate reactions and that can create new complex containing multiple strands in arbitrary staggered conformations. To our best knowledge, no other simulation toolkit to investigate self-assembly of oligonucleotides in such a detailed fashion is currently available. The simulation is based on the Gillespie Algorithm and implemented in $C++$.

The emergence of extension cascades and "spontaneous" primer extension within a pool of strands of various lengths resembles the potential dynamics of the replication of a *virtual genome* described Ref. [79] by Zhou et al. The authors of this perspective article consider a virtual circular master sequence, the virtual circular genome. Virtual here means that the master sequence does not appear as a whole. The virtual master sequence is only present in a fragmented way in the form of many shorter strands. The sequences of these strands correspond to partially overlapping subsequences or reverse subsequences of the virtual circular master sequence. Taken together, all fragments fully cover the virtual circular genome multiple times as depicted in Fig. 7.1. The fragments are encapsulated in a reaction volume that is coupled to a reservoir constantly providing new monomers and also few dimers and trimers. The authors hypothesize that the fragments of the virtual genome could grow and replicate via the formation of initial stable complexes and subsequent extension cascade, i.e., by "spontaneous" primer extension. In order to "free" strands that are bound in "non-productive" configurations (their equivalent to fully hybridized duplexes in our model), cyclic variations of the environmental conditions are proposed. By replication of single fragments, the virtual genome could become replicated as a whole.

A variant of our simulation framework that explicitly takes into account the sequence information of strands (see below) could be used to realize a first implementation of the replicating dynamics of a virtual genome. In the pure primer extension scenario studied in Chapter 3, we used the two observables yield and error fraction to characterize the average number of copies obtained from one defined template strand as well as their "quality". These two observables would now have to be generalized to describe the replica of the virtual circular genome emerging in a system initialized with fragments of the master sequence. The replication of subsequences of the virtual circular genome competes with their degradation and the formation of new sequences which are not part of the original virtual genome. These new sequences can either arise due to errors during the copying process, or due to the ligation of two longer sequences extending the short template strand at least on one side. The second process can produce sequences that fall out of the virtual circular genome even though there is no mismatch on the short template and although all the involved fragments are contained in the genome. In

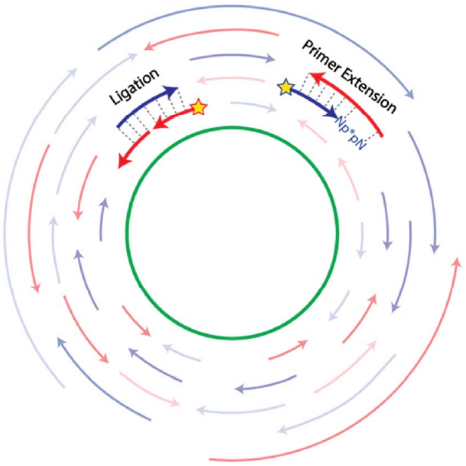


Figure 7.1.: The green circle in the center corresponds to the virtual circular master sequence, which is not present as a whole in the system. Red and blue arrows correspond to (reverse) subsequences that, taken together, cover the whole circular virtual genome multiple times. This figure is adapted from Fig. 3 in Ref. [79] published under CC BY 4.0 License.

the latter case, it would be interesting to find out under what conditions the information of the virtual master sequence could at least be retained in the system for as long as possible. Therefore a new observable measuring the preservation of information would be necessary.

In the third scenario, we made the sequence dependence of the self-assembly process from small building blocks via templated ligation explicit. For simplicity, we used a binary system where only two types of nucleotides, denoted by X and Y for generality, appear. X and Y are assumed to be complementary. The energy associated with a hybridization site is computed using an energy model of the nearest neighbor type. To every block of neighboring nucleotide pairs, a stacking energy is assigned. Each of these nearest-neighbor stacking energies contribute linearly to the overall energy of the hybridization site, which then determines the rate for dehybridization. The stacking energy associated with a nearest-neighbor block is a function of the motif context. Blocks comprising non-complementary pairs reduce the stability of a hybridization site, while complimentary blocks enhance the stability. The contributions of the latter are again non-uniform and depend on the arrangement of the matching nucleotides. This is motivated by the fact that homogeneous and alternating “zebra” blocks have different stacking energies in binary RNA or DNA systems. With that, our energy model conserves the essential feature of more detailed nearest-neighbor models for DNA and RNA. Moreover, the ligation step is motif-dependent as well. As in the primer extension scenario discussed in Chapter 3, non-complementary nucleotide pairs in the vicinity of the ligation site kinetically stall the formation of a new covalent bond. To circumvent template blocking and to allow for frequent "reshuffling" of complexes, we employed periodically changing environmental conditions that lead to the melting of complexes. These periodic changes were implemented effectively via a lower bound for the rate of dehybridization. In contrast to Chapter 5, the effective cycling parameters were chosen such that the resulting strand-length distribution is always monotonically decaying. The growth of strands is balanced by the cleavage of single-stranded segments of strands via hydrolysis. The cleavage rate was taken to be independent of the sequence context.

Up to date, no theoretical study on self-assembly via templated ligation considered the motif-dependent energetic and kinetic aspects of hybridization and bond formation in such a detailed manner.

In our conceptual study, the strength of the kinetic stalling effect as well as the difference between the stacking energies for homogeneous and alternating "zebra" blocks were treated as variable parameters. The goal of this project was to find out how these two parameters influence the dynamics in sequence space as strands grow longer. To this end, we initialized a closed reaction volume symmetrically with mostly monomers and a few dimers and let the system evolve until a stationary state was reached. As a simple observable to characterize patterns in sequence space, we introduced the system level *zebraness*, telling us which fraction of all binary motifs within all present strands is alternating, i.e., of the form $X - Y$ or $Y - X$.

Our main finding is that kinetic stalling constitutes a necessary requirement for pattern formation in sequence space. If mismatches in the vicinity of the ligation site are not penalized kinetically, i.e., do not stall the ligation step, no patterns in sequence space emerge in the case where there is no difference between the stacking energies for homogeneous and alternating blocks. If an energetic bias favoring alternating nearest-neighbor blocks is applied, a distinct bias for "zebra" motifs in sequence space is observed transiently. The initial sequence bias decreases almost completely as strands grow longer. As a source for the loss of the initial sequence bias, we identified *concealed mismatches*. Concealed mismatches are non-complementary nucleotide pairs which "escape" from thermodynamic discrimination. These types of mismatches can arise in hybridization sites that are long enough, i.e., contain enough complementary nucleotide pairs such that the corresponding dehybridization rate is set by the lower bound, i.e., by the effective cycle duration in any case. Turning a concealed mismatch into a match would not alter the stability of a complex. If a concealed occurs at the ligation site, and if the stalling effect is "switched off", the formation of a wrong binary motif working against the bias in sequence space on the system level is not suppressed.

The dynamics in sequence space is fundamentally different if kinetic stalling is applied such that mismatches at the ligation site always experience kinetic discrimination. In this case, if homogenous and alternating blocks have identical stacking energies, small fluctuations in sequence pool occurring early on trigger cascades of self-amplification that spontaneously break the symmetry in sequence space. If the stalling effect is strong enough, the reaction volume contains either entirely homogenous or zebra-like sequences in the stationary state. Hence, the zebraness either converges to one or zero. In contrast, if alternating blocks are slightly favored energetically, alternating binary motifs always dominate the sequence pool in the stationary state in the presence of a strong kinetic stalling effect.

In all scenarios, the mean length shows burst-like growth dynamics. The onset of the rapid growth coincides with the time point where *higher-order ligations*, i.e., ligations involving at most one monomer become more abundant than ligations joining two monomers and can be predicted analytically.

For the sake of generality, we referred to the two types of nucleotides appearing in our model as X and Y. Within the system considered in the third scenario, X and Y stand for complementary nucleotides in the Watson-Crick sense. Nucleotide pairs formed by two Xs or two Ys, respectively nucleotides are considered as mismatches and generally reduce the stability of a hybridization site, while pairs involving one nucleotide of each type are considered a match. A reinterpretation of the meaning of X and Y coming along with some small changes of the energy parameters nearest neighbor blocks would lead to a new conceptual model that could be used to study another important aspect of the origins of life, which is the emergence of homochirality. All important biomolecules appearing in living systems such as proteins and polynucleotides are homochiral, i.e., composed of subunits with an identical chirality (handedness) [101]. Yet, the prebiotic soup assumed to be racemic, i.e., of mixed chirality and it is unknown by which mechanism this asymmetry could occur [102, 103, 104]. The general laws of physics and chemistry, as we know them, do not depend on chirality, therefore, a mirror-symmetric biology could be possible in theory [105].

In the new conceptual model, X and Y would stand for two different chiral forms in which an abstract self-complementary nucleotide can appear. Experimental measurements suggest that pairings involing nucleotides of different chiralities weaken the stability of a hybridization site [273, 274, 275]. Therefore, the energy parameters would have to change such that nucleotide pairs involving two Xs or two Ys, respectively, would be favored energetically over mixed pairings. Moreover, it has been that some kind of stalling effect also appears if the nucleotides at the ligation site show different chiralities [276, 277]. The relevant quantity to characterize a single strand in this new model would be the *chiral enantiomeric excess*, which is one for a homochiral strand and zero for a strand containing as many Xs as Ys. For a closed system initialized symmetrically with nucleotides and short oligomers of both chiralities, or for an open system coupled to a mixed but symmetric pool of short building blocks, one could now ask under which conditions two sub-pools of (mostly) homochiral strands belonging to the X- or Y type emerge. In this context, it would also be interesting to extend the model by a new type of reaction: the deracemization of monomers. In this reaction, a monomer of the X becomes a monomer of the Y type and vice versa. This reaction could lead to a global homochiral final state in a closed system. We have the following in mind: Due to fluctuations, one of the two sub-pools shows a faster initial growth. Hence, either the pool of X or Y monomers are depleted faster. However, the deracemization reaction tends to equilibrate the two monomer pools. As a result, monomers corresponding to the slower-growing sub-pool of longer strands are more frequently converted. Eventually, the strong inherent non-linearity in the growth via template-directed ligation could lead to the extinction of one form of chirality. A suitable observable to characterize the dynamics in chirality space would be the *global chiral enantiomeric excess*, which is one for a fully homochiral state and zero of the racemic state.

Appendices

A. A kinetic error filtering mechanism for enzyme-free copying of nucleic acid sequences

A.1. Additional figures

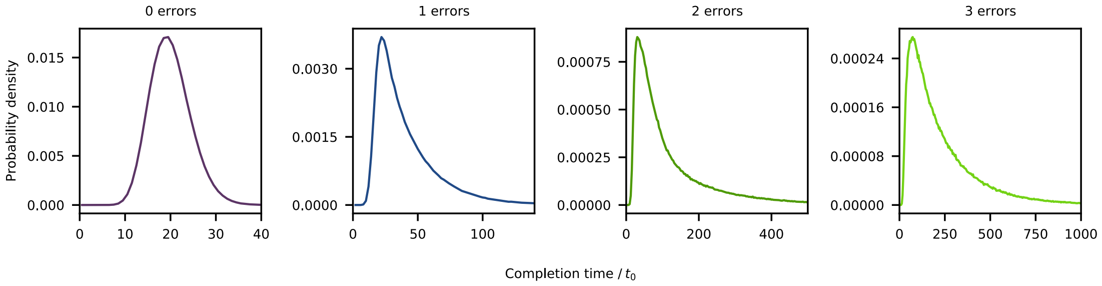


Figure A.1.: Distributions of the completion time for zero to three errors. Typical completion times increase rapidly with the number of errors. The zero-error distribution approximately has the shape of a Gaussian distribution with a distinct peak at $t_{\text{perf}} = Lt_0 = 20$.

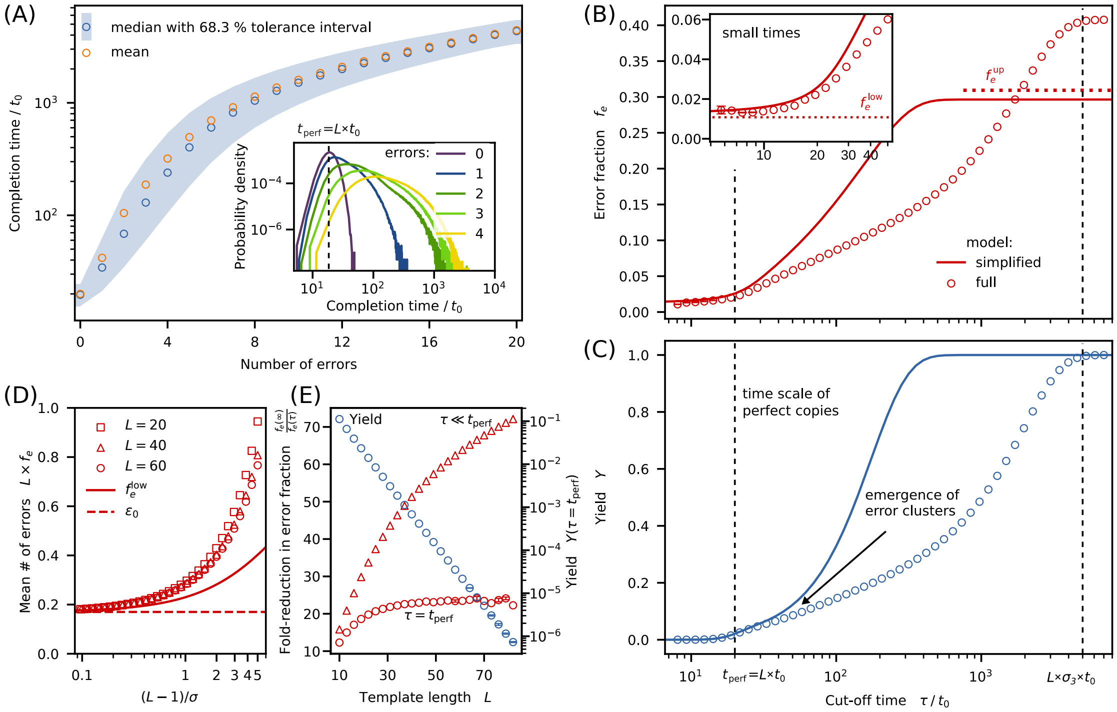


Figure A.2.: Same as Fig. 3.3 but with RNA parameters. (A) Comparing the probability densities for the RNA system to the DNA system shows that full-length products containing zero or only a few errors are more unlikely in the RNA system. (B) and (C) The error fraction for $\tau = \tau_{\text{comp}}$ is more than a factor of 1.5 larger in the RNA system. Compared to the DNA system, the increase at $\tau = L\,t_0$ is less pronounced in the RNA system. Dotted red lines indicate the lower and upper bounds for the error fraction derived for the simplified model. (D) The absolute number of error for a given configuration in the RNA system is always larger than in the DNA system. (E) The error-reduction effect is by a factor of roughly tow smaller compared to the DNA system. The decrease of the yield is steeper in the RNA system (roughly a factor of two in the logarithmic plot).

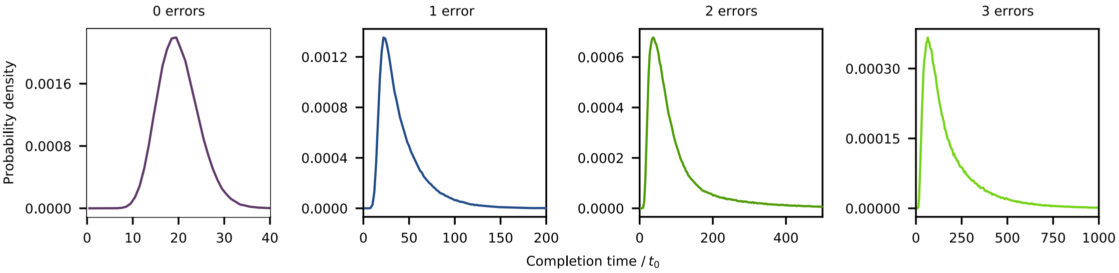


Figure A.3.: Same as for Fig. A.1 but with RNA parameters.

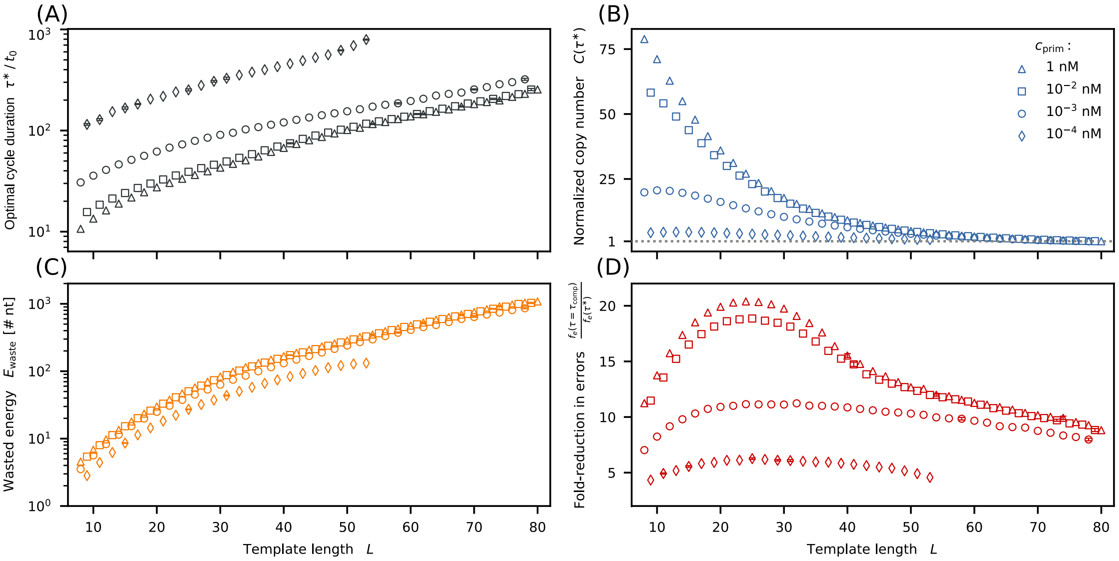


Figure A.4.: Kinetic error filtering at the yield rate optimum. For a more systematic analysis of the effect of the template length, we introduce two new quantities: the cycle duration τ_{comp}, for which 99% of all copies reach completion ($Y(\tau_{\mathrm{comp}}) = 0.99$), and the ratio $C(\tau)$ of the yield rates evaluated at τ and τ_{comp}, i.e., $C(\tau) = \left[Y(\tau) \times \tau_{\mathrm{comp}}\right] / \left[0.99 \times \tau\right]$. We refer to $C(\tau)$ as the normalized copy number, since it corresponds to the overall number of full copies obtained over a fixed time interval $\Delta t \gg \tau$ for a given cycle duration τ relative to the total number of full copies obtained in a scenario where the cycle duration is τ_{comp}. The maxima of the normalized copy numbers for different template lengths L are at the same positions as the maxima of the yield rates in Fig. 3.4. The numerical procedure to determine the potions of the maxima is explained in Section A.2. With that, we can define "simultaneous optimization": A simultaneous optimization of fidelity and yield is possible for those values of L for which the a left maximum of $C(\tau)$ exits and is larger than one. (A) Over the range of template lengths for which a simultaneous optimum exists, the optimal cycle duration τ^* increases with L. (B) The normalized copy number $C(\tau^*)$ decreases with L. As expected, smaller primer concentrations lead to smaller copy numbers. For low concentrations, as well as for larger L, the total yield is only slightly improved. (C) The wasted energy per completed copy evaluated at τ^* grows significantly as L is increased. (D) The fold-change in the error fraction, $f_e(\tau_{\mathrm{comp}})/f_e(\tau^*)$ is largest for template lengths between 20 and 30. However, the accuracy is significantly improved for all concentrations and lengths.

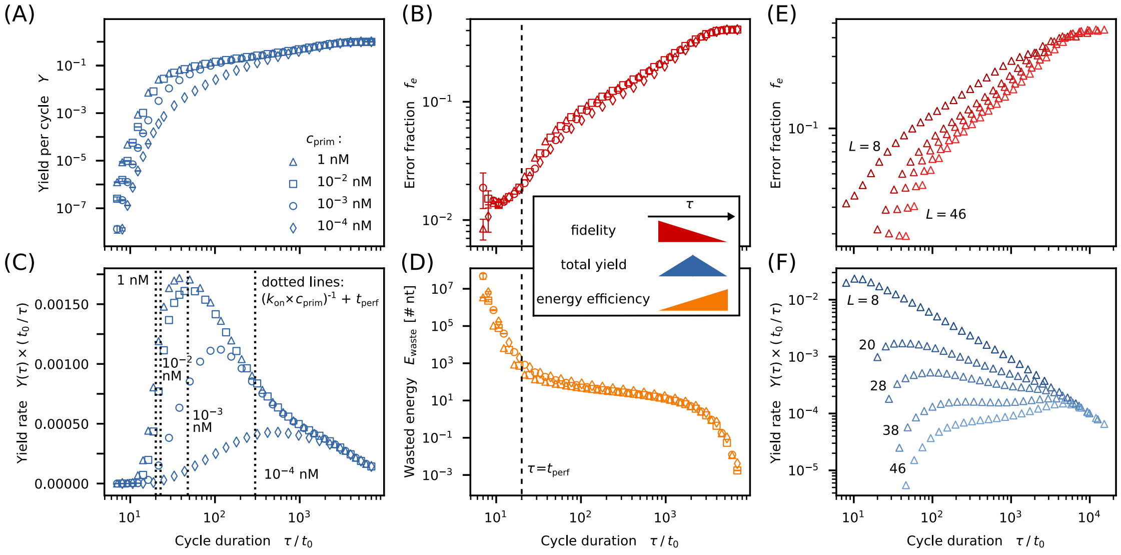


Figure A.5.: Same as for Fig. 3.4 but with RNA parameters.

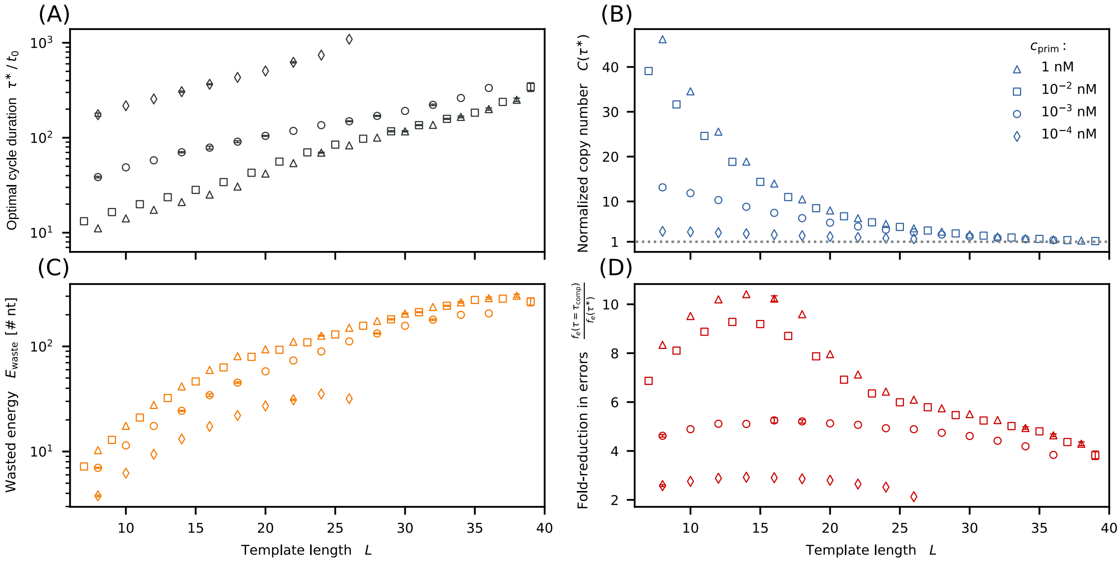


Figure A.6.: Same as for Fig. A.4 but with RNA parameters.

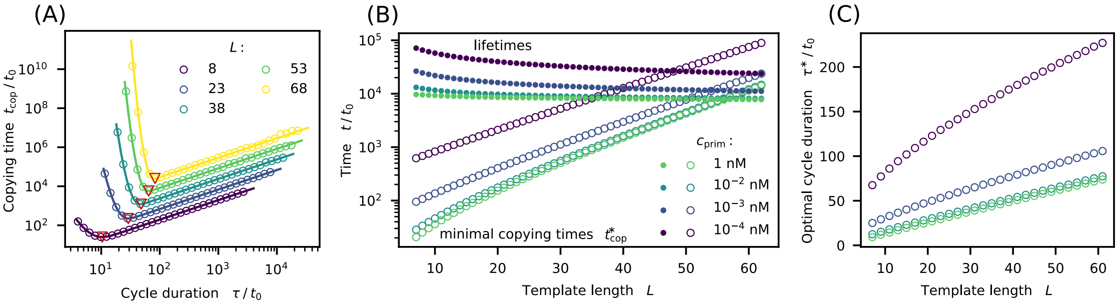


Figure A.7.: Same as for Fig. 3.5 but with RNA parameters.

A.2. Analysis of the Normalized Copy Number

In Fig. A.4 the maximum value of the normalized copy number $C(\tau)$ achieved for $\tau = \tau^*$ is plotted as a function of the template length L. To obtain the maximum value of C for a given L, the following procedure is carried out:

- First τ_{comp} is estimated. To this end, the yield per cycle $Y(\tau)$ is computed as an average value from a sufficiently large ensemble of trajectories for $\tau = \tau_i$ with $i \in \{1, 2, \ldots, 50\}$ and $\tau_1 = 0.5 \times L$ and $\tau_{50} = 1.5 \times L \times \sigma_{\text{non}}$ (see Fig. A.8, circles). The spacing $\tau_{i+1} - \tau_i$ increase exponentially with i. The ensemble size varies with the value of τ_i (see implementation for details).
- The index $i = j$ for which $Y(\tau_i) > 0.99$ for the first time is identified. The values of $\tau_{\text{below}} = \tau_{j-2}$ and $\tau_{\text{above}} = \tau_{j+1}$ are stored. τ_j serves as a first estimate for τ_{comp}.
- As a next step, we estimate the optimal cycle duration τ^* maximizing the yield rate. The yield rate $R(\tau_i) = Y(\tau_i)/\tau_i$ is computed from the same set of ensembles. The index $i = m$ for which $R(\tau_i)$ is maximized, is identified (see circles in Fig. A.10). The values of $\tau_{\text{left}} = \tau_{m-2}$ and $\tau_{\text{right}} = \tau_{m+2}$ are stored. τ_m serves as a first estimate for the optimal cycle duration τ^*.
- After the coarse search for τ_{comp} and τ^*, new simulations to generate data to determine more precise values of τ_{comp} and τ^* are carried out.
- For the refined search for τ_{comp} new ensembles of trajectories are simulated for τ_k with $k \in \{1, 2, \ldots, 100\}$ and $\tau_1 = \tau_{\text{below}}$ and $\tau_{100} = \tau_{\text{above}}$ (see Fig. A.8 and Fig. A.9, bullets). The spacing is chosen to be constant.
- Also for the refined search for τ^*, new ensembles of trajectories are simulated for τ'_k with $k \in \{1, 2, \ldots, 100\}$ and $\tau'_1 = \tau_{\text{left}}$ and $\tau'_{100} = \tau_{\text{right}}$ (bullets in Fig A.10). Again, the spacing is chosen to be constant for the refined search.
- The data set $\{(\tau_k, Y(\tau_k))\}$ with $k \in \{1, 2, \ldots, 100\}$ and $\tau_1 = \tau_{\text{below}}$ and $\tau_{100} = \tau_{\text{above}}$ generated for the refined search for τ_{comp} is now further analyzed.
- The index $k = l$ for which the running mean $\frac{1}{5}\sum_{n=-2}^{2} Y(\tau_{k+n}) > 0.99$ for the first time is determined. After that, a straight line is fitted to a the subset of data points for the refined search given by $\{(\tau_k, Y(\tau_k))\}$ with $k \in \{l-5, \ldots, l+5\}$ (see red line in Fig. A.8 and Fig. A.9). We now identify τ_{comp} with the abscissa of the intersection point of the straight line obtained from the fit and the horizontal line $Y = 0.99$ (red square and red horizontal line in Fig. A.8 and Fig. A.9).

- The error fraction f_e at the saturation time τ_{comp} is determined in a similar way: First, a linear function is fitted to the subset of data points corresponding to the refined search $\{(\tau_k, f_e(\tau_k))\}$ with $k \in \{l-5, \ldots, l+5\}$. After the fit parameters are determined, the linear function is evaluated at τ_{comp}. A similar procedure is carried out to obtain the wasted energy $E_{\text{waste}}(\tau_{\text{comp}})$.

- The data set $\left\{\left(\tau'_k, Y(\tau'_k)\right)\right\}$ with $k \in \{1, 2, \ldots, 100\}$ and $\tau'_1 = \tau_{\text{left}}$ and $\tau'_{100} = \tau_{\text{right}}$ generated for the refined search for the optimal cycle duration τ^* is now also further analyzed.

- To this end, the yield rate $R(\tau'_k)$ is mapped onto the normalized copy number $C(\tau'_k)$ via $C(\tau'_k) = R(\tau'_k) \times \frac{\tau_{\text{comp}}}{0.99}$.

- With that, the index $k = l$ for which the running mean $\frac{1}{5}\sum_{n=-2}^{2} C(\tau'_{k+n})$ is maximal is determined. After that, a parabola is fitted to a the subset of data points obtained from the refined search given by $\left\{\left(\tau'_k, C(\tau'_k)\right)\right\}$ with $k \in \{l-12, \ldots, l+12\}$, and its maximum is determined (red line and red triangle in Fig. A.10). Abscissa and ordinate of the maximum are identified with the τ^* and $C(\tau^*)$.

- The error fraction f_e at τ^* is determined as follows: A linear function is fitted to the subset of data points corresponding to the refined search $\left\{\left(\tau'_k, f_e(\tau'_k)\right)\right\}$ with $k \in \{l-12, \ldots, l+12\}$. Once the fit parameters are determined, the linear function is evaluated at τ^*. A similar procedure is carried out to obtain the wasted energy E_{waste} at τ^*.

With increasing L the position τ^* of the maximized normalized copy number $C(\tau^*)$ on the left-hand side (red triangles) shifts to the right. The value of $C(\tau^*)$ thereby decreases. For larger values of L the maximum on the left-hand side finally vanishes, while a second maximum on the right-hand side, located close to τ_{comp}, emerges. The same observation is made for smaller values of c_{prim} (not shown). The maximum on the right-hand side corresponds to a gain in overall yield and copying fidelity which is only marginal. Hence, we say that a simultaneous improvement of copy number and accuracy is only possible up to a certain length (red triangles in Fig. A.10). If no maximum on the left-hand side exists or if it is smaller than one the mechanism breaks down.

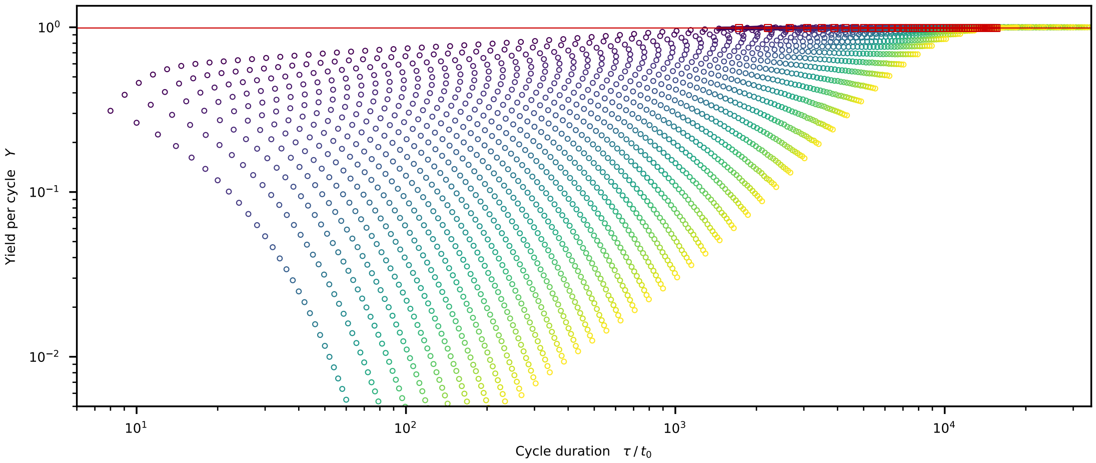


Figure A.8.: Different colors correspond to different values of the length L. Dark purple: $L = 8$, yellow: $L = 102$, L is increased by steps of two. Circles correspond to the coarse-grained search for $\tau = \tau_{\text{comp}}$ such that $Y(\tau > 0.99)$ for the first time coming from small values $Y(\tau)$ while bullets correspond the the refined search (see zoom in Fig. A.9). The red horizontal line marks the $Y = 0.99$ threshold. Red squares highlight the points that are identified with $\left(\tau_{\text{comp}}, Y(\tau_{\text{comp}})\right)$.

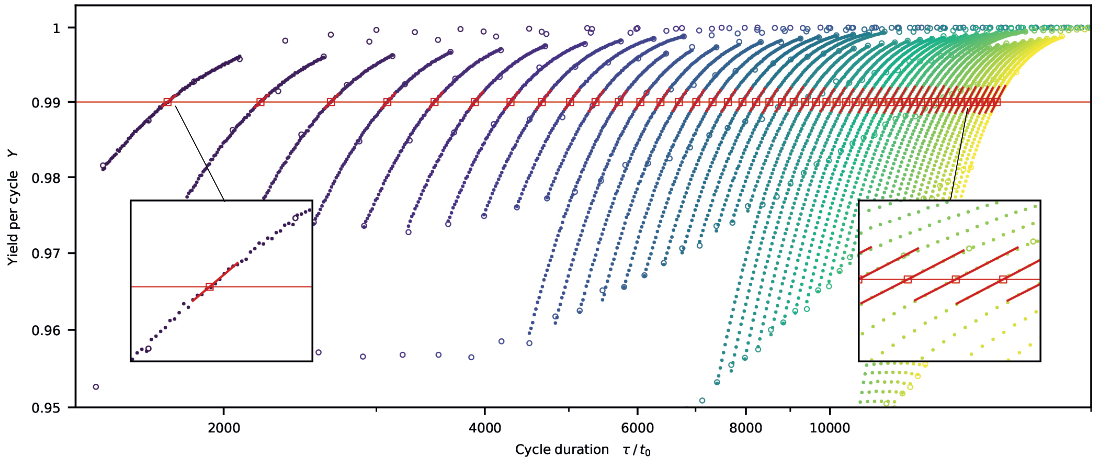


Figure A.9.: Different colors correspond to different values of the length L. Dark purple: $L = 8$, yellow: $L = 102$, L is increased by steps of two. Coarse-grained search: circles, refined search: bullets. Red lines correspond the the liner fits. The intersections points (red squares) of the linear fits and the horizontal line $Y = 0.99$ are identified with the points $(\tau_k, f_e(\tau_k))$.

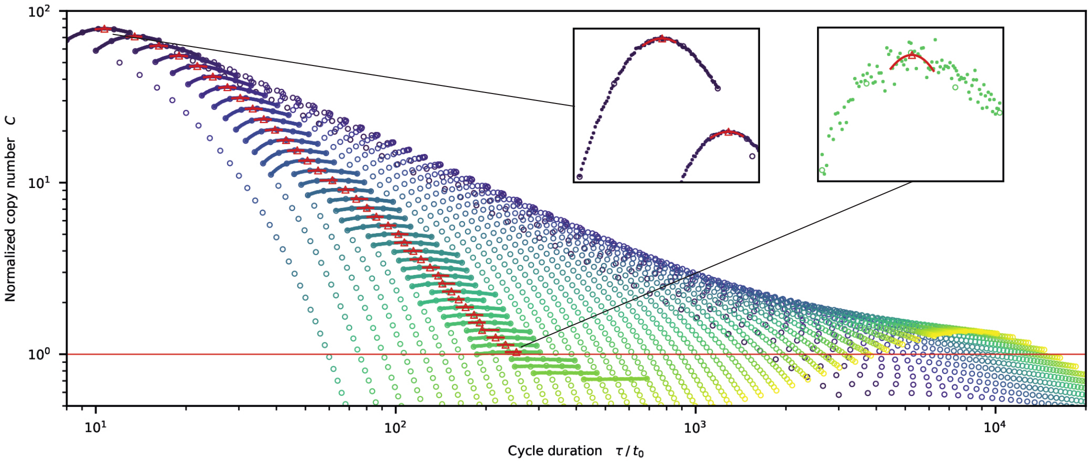


Figure A.10.: Different colors correspond to different values of the length L. Dark purple: $L = 8$, yellow: $L = 102$, L is increased by steps of two. Circles (bullets) correspond to the coarse-grained (refined) search for the maxima of the normalized copy number C. Parabolas obtained from fits to subsets of the data points for the refined search are plotted with red lines. Abscissa and ordinate of the maxima of the parabolas are identified with the τ^* and $C(\tau^*)$. A simultaneous optimization is possible for those values of L for which a maximum with $C(\tau^*) > 1$ on the left-hand side exists.

B. Growth regimes in polymers self-assembly by templated ligation

B.1. Trajectories of extension cascades

For further investigation of the assembly and growth processes in our model, we analyzed the trajectories of duplexes that undergo extension cascades resulting in a fully-hybridized duplex.

We start with a background obtained form a simulation that reached steady state. We set the complexes as background species, hence do not allow for reactions within this background. All complexes belonging to the background have a poly(A) sequence. We insert a dimer with a specific sequence (“TB”) into the system, which serves as a label for an individual tracer complex. The dimer can undergo hybridizations with all background species. Thereby the label becomes integrated into other complexes. The complex with the specific label is always kept as the only non-background species in the system (it undergoes reactions with the background). We call this specific complex the tracked complex. After each reaction we check if the complex contains a stable sub-duplex (see Fig. B.1(a)).

If one of the sub-duplexes starts to undergo an extension cascade (the dehybridization rate of its hybridization site is smaller than its extension rate), we start to track the sub-duplex within the tracked complex. We store the initial stable sub-duplex in a buffer. Whenever the stable sub-duplex is extended via templated ligation, the newly formed duplex is appended to the buffer. The stored sequential snapshots of the extensions of the stable sub-duplex is what we call a trajectory. If the duplex dehybridizes, the trajectory is deleted, and the recording restarts as soon as a new stable duplex that undergoes extension cascades is formed. When the tracked complex leaves the system via outflux, we save the trajectory to disk and the assembly process restarts with a labeled dimer. The schema is illustrated in Fig. B.2. As mentioned above, the tracked complex can undergo hybridizations with the background. There is, however, one caveat: The tracked complex is only allowed to undergo hybridizations with background complexes that do not itself contain a sub-duplex undergoing extension cascades. We further reject trajectories where two duplexes that can undergo extension cascades are formed within a complex. These restrictions guarantee the sampling of trajectories that start with the onset of a extension cascade and finish in a fully-hybridized configuration. For the rejected trajectories it would not be possible to identify a unique starting point. It can be expected that the thereby obtained strand length distribution tends to underestimate the length distribution obtained via the full simulation.

A comparison between the strand length distribution obtained via the full simulation and the sampled trajectories reveals that the latter resembles the first one, *cf.* Fig. B.1(b). Hence the sampling is consistent with the dynamics. But indeed, the length distribution obtained via sampling trajectories underestimates the concentration of strands of length $L \geq 50 \geq L_{\text{max}}$. For $L = 40$, the relative deviation is only -3%, whereas for $L = 50$ the deviation is -19%.

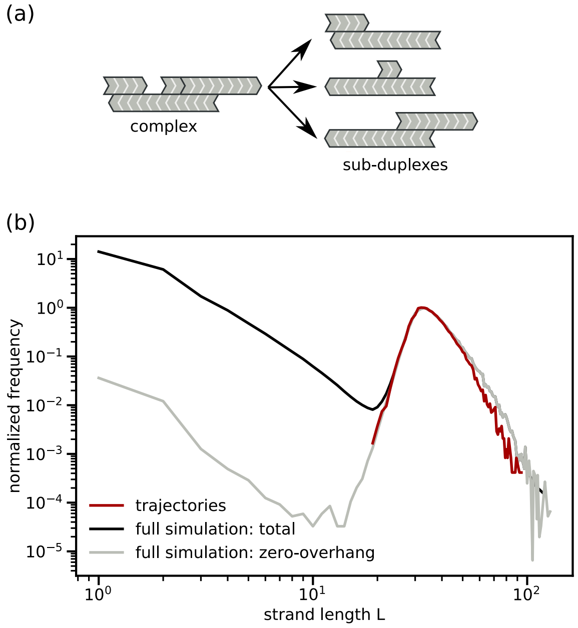


Figure B.1.: (a) The tracked complex of order n ($n-1$ hybridization-sites) can be decomposed into $n-1$ sub-duplexes. (b) The length distribution of fully-hybridized strands obtained from the trajectories (red curve) resembles the length distribution of the fully-hybridized strands obtained from the full simulation (gray curve). The length distribution for all strand lengths (black line) is plotted for orientation. The system shown here is the monomer-dimer system with a monomer fraction $f_{\text{m}} = 0.7$. The length distributions are normalized on the concentration at the maximum $L_{\text{max}} = 33$. Only at the tail the length distribution obtained via trajectory sampling underestimates the concentration of long strands. This behavior is expected since certain trajectories leading to long fully hybridized complexes are rejected.

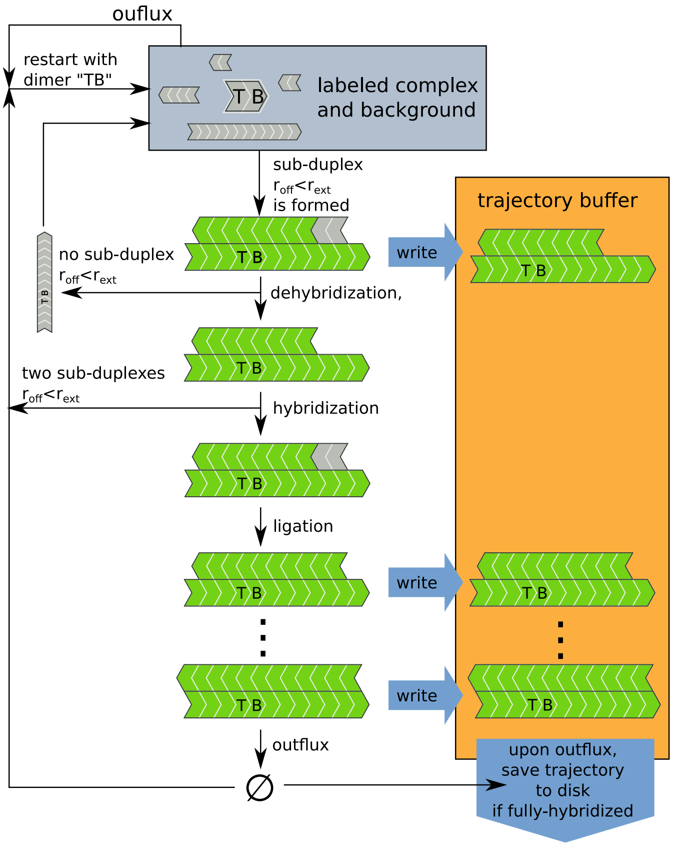

ouflux
restart with dimer "TB"
T B
labeled complex and background
sub-duplex $r_{off}<r_{ext}$ is formed
trajectory buffer
no sub-duplex $r_{off}<r_{ext}$
T B
write
dehybridization,
two sub-duplexes $r_{off}<r_{ext}$
hybridization
ligation
write
write
outflux
∅
upon outflux, save trajectory to disk if fully-hybridized

Figure B.2.: The simulation loads the complexes from a simulation that reached a steady state and utilizes them as background species. A labeled dimer (“TB”) is inserted into the reaction vessel (simulation). The complex containing the (“TB”) motif is called the tracked complex. It can undergo dehybridizations and ligations, and collisions with the background. As soon as the tracked complex contains a stable sub-duplex, the trajectory of the sub-duplex is recorded. Whenever a ligation happens, the sub-duplex structure is written to a buffer. Hybridizations and dehybridizations are not explicitly tracked, i.e. the resulting new complex structure is not written to the buffer. However, if the tracked stable duplex disassembles via dehybridization, the trajectory is rejected and the buffer is cleared. We again start saving a trajectory as soon as the complex containing the “TB” motif contains a stable sub-duplex. We also neglect trajectories that include two stable duplexes at some point, and in this case, restart with a "TB" dimer. If the complex leaves the reaction vessel via outflux and if it has reached a fully hybridized configuration, we save the trajectory to disk. The buffer gets cleared and we restart with the dimer motif.

B.2. Laboratory experiment 12 nt random A-T-DNA strands

B.2.1. Critical temperatures including salt corrections

We used *Melting* (version 5.2.0), [278], to calculate the enthalpies and entropies for various fully-hybridized duplexes based on the Santa Lucia nearest neighbor model [107]. We chose *Melting*, because it can account for salt corrections. The NEB Taq DNA ligase buffer solution contains 20 mM Tris-HCl, 25 mM potassium acetate and 10 mM magnesium acetate. The storage solution of the ligase contains 10 mM Tris-HCl and 50 mM potassium chloride. The storage solution is diluted by a factor of 25 when inserted into the buffer solution. Hence in the experiment the concentrations are: 20.4mM Tris, 27mM K, 10mM Mg. The resulting critical temperatures are listed in Table B.1.

mismatches	length	Duplex	ΔH	ΔS	T_c

0	4	`3'-TATA-5'`
		`5'-ATAT-3'`
			-17.0	-54.8	38.8

0	6	`3'-TATATA-5'`
		`5'-ATATAT-3'`
			-31.4	-95.3	56.3

0	6	`3'-AAAAAA-5'`
		`5'-TTTTTT-3'`
			-33.4	-98.4	66.3

0	8	`3'-TATATATA-5'`
		`5'-ATATATAT-3'`
			-45.8	-136.1	63.4

0	10	`3'-TATATATATA-5'`
		`5'-ATATATATAT-3'`
			-60.2	-176.9	67.2

0	12	`3'-AAAAAAAAAAAA-5'`
		`5'-TTTTTTTTTTTT-3'`
			-79	-226.2	76.1

0	12	`3'-TATATATATATA-5'`
		`5'-ATATATATATAT-3'`
			-74.6	-217.7	69.5

0	24	`3'-TATATATATATATATATATATATA-5'`
		`5'-ATATATATATATATATATATATAT-3'`
			-161.0	-462.5	75.0

1	12	`3'-TATAAATATATA-5'`
		`5'-ATATATATATAT-3'`
			-54.3	-160.9	64.3

1	24	`3'-TATATATATAAATATATATATATA-5'`
		`5'-ATATATATATATATATATATATAT-3'`
			-140.7	-405.7	73.7

Table B.1.: Standard molar enthalpy ΔH and entropy ΔS including salt corrections calculated with the program *Melting*. ΔH, ΔS and T_c are given in units of kcal/mol, cal/mol-K and $^\circ C$. The first three duplexes do not contain mismatches, while the last two include one mismatch. The critical temperature is calculated via $T_c = \Delta H / \Delta S$. T_c ranges for an overlap of $l = 6$ from $56.3^\circ C$ to $66.3^\circ C$.

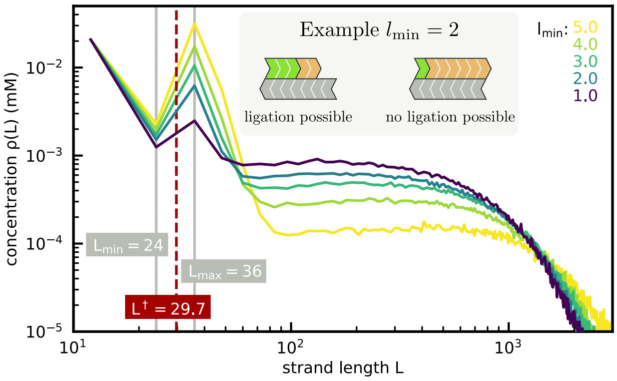


Figure B.3.: Steady state strand-length distribution for varying $l_{\min}$. The system was initialized with single-stranded 12mers. The concentration of single-stranded 12mers is kept constant. For a ligation both strands hybridized on the template need an overlap of at least $l_{\min}$.

B.2.2. Minimal overlap required for ligation

The ligase from New England Biolabs (NEB) that we used requires an overhang of at least five nucleotides to perform a ligation [265]. The usage of 12-mers in the experiment reflects this constraint: Shorter strands would require lower temperatures to form duplexes that are stable enough for the ligase to bind. However, by lowering the temperature at some point the ligase ceased to be active and hence no product can be formed. This constraint can also be rationalized by relating the ligase activity stated by NEB with the typical binding energies for tetramers obtained in nearest neighbour models: The ligase activity decreases sharply between $T = 37 - 45°C$, see [266], while for example the four nucleotide overlap duplex in Table B.1 has a $T_c = 38.8°C$. To test the robustness of our theoretical analysis with respect to the requirement of a minimal overlap, we extended our model to include a new parameter $l_{\min}$: If both strands hybridized next to each other on a template have overlaps $l_1, l_2 \geq l_{\min}$, the ligation rate is r_{lig}. If one of the strands has an overlap smaller than $l_{\min}$ a ligation is not possible (see Fig. B.3). We performed a parameter sweep of $l_{\min}$ for an open system that is initialized with single-stranded 12mers , allowing us to test the behavior of building blocks where there are multiple overlaps possible. The concentration of single-stranded 12mers is kept constant at 20 µM and the standard energy model, $\beta\Delta G_b^\circ = \gamma l$, where $\gamma = -0.5$, is used. The outflux rate was set to $r_{\text{out}} = 10^{-8}$. Fig. B.3 shows that the emergence of a non-monotonous length distribution is independent of the minimal overlap. In particular, the length-scale $L^\dagger = 29.7$ that determines the position of the maximum is unaffected by $l_{\min}$, *cf.* Eq. (15). Ceiling to the next multiple of the smallest building block length (= 12), yields $L^\dagger_\blacktriangle = 36$ which is equal to the position of the maximum,

$L_{\text{max}} = 36$. Moreover, we see that the position of the minimum $L_{\text{min}} = 24$ is unchanged. Including such a minimal overlap would complicate the calculation of the extension rate Eq. (9). Notice, that care needs to be taken in applying Eq. (9) to long building blocks, since the assumption of an approximate local equilibrium of the ligatable triplex T_i with the corresponding duplex and single strands might not hold anymore. The effect of a minimal overlap l_{min} is mostly a difference in the relation of the height of the maximum and the mass included in the tail: For $l_{\text{min}} > 1$, also duplexes consisting of two strands of length $L = 36$ with an overhang of $|o_i| < l_{\text{min}}$ are non-extendable, leading to a more pronounced maximum.

B.2.3. Initial sequence space of 12 nt A-T strand

The initial 12mer AT random DNA pool was ordered as 5′-WWWWWWWWWWWW-3′ with 5′- phosphate modification from biomers.net. The same set of DNA strands is used in Ref. [58]. There, an in-depth analysis of the initial pool revealed a small overall bias towards A-rich sequences and a lack of poly-T sequences.

B.2.4. Resulting PAGE gels and concentration quantification by Image Analysis

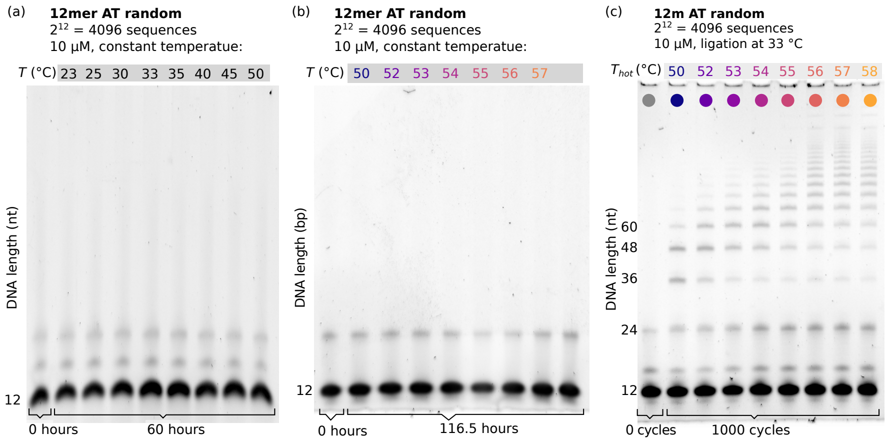

(a)
12mer AT random
2^{12} = 4096 sequences
10 µM, constant temperatue:
T (°C) 23 25 30 33 35 40 45 50
DNA length (nt)
12
0 hours
60 hours
(b)
12mer AT random
2^{12} = 4096 sequences
10 µM, constant temperatue:
T (°C) 50 52 53 54 55 56 57
DNA length (bp)
12
0 hours
116.5 hours
(c)
12m AT random
2^{12} = 4096 sequences
10 µM, ligation at 33 °C
T_{hot} (°C) 50 52 53 54 55 56 57 58
DNA length (nt)
60
48
36
24
12
0 cycles
1000 cycles

Figure B.4.: PAGE gels for experiments at constant temperature (a) and (b) and experiments exposed to 1000 temperature cycles. 12mer DNA does not any product during incubation at constant temperatures. In contrast, the experiments for temperature cycles show significant multimer production. In this case, the pattern of the PAGE gel change with temperatures.

In the experiment, a constant T_{cold} does not produce multimer products over 60 hours, see Fig. B.4(a) and (b). Temperature cycling is an easy way to "reset" dsDNA to

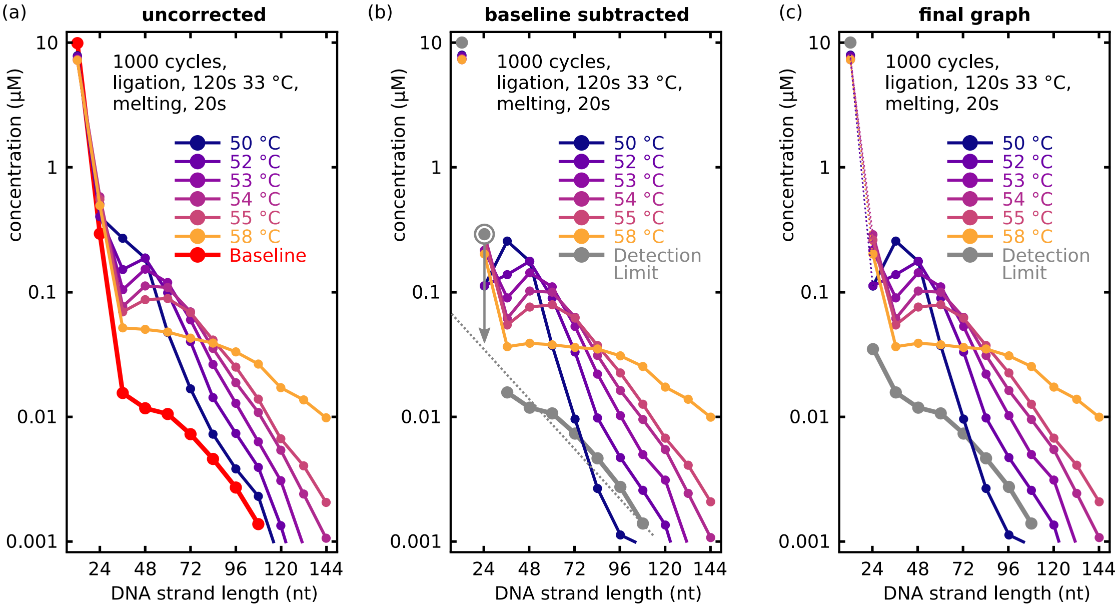


Figure B.5.: (a) Concentration estimation from the peak-areas. (b) Corrected concentration due to subtraction of the baseline signal, which is then called the detection limit. (c) Final graph, the position of the 24mer is extrapolated from the normal-log plot in (b).

their ssDNA state and to promote subsequent hybridization after the cooling. Then, novel ligation reactions are possible. Fig B.4(c) again shows the PAGE gel from Fig. 8 in the main text with the results for 1000 temperature cycles between T_{cold} and varied T_{hot}, which takes about 40 to 60 hours depending on the temperatures and the temperature ramp of the PCR thermos cycler device.

We used a custom LabView program for concentration quantification of bands on the PAGE gels.The method has limitations, as described below. A similar analysis has been carried out in [58]. First, each lane is marked with a top and a bottom cursor that span a rectangular ROI (region of interest, about 10-30 % of the lane width) on the lane. The intensity is read as mean intensity values averaged over the width of the ROI. The center region of each lane shows the lowest lateral intensity change in the band and is therefore ideal for selecting ROIs to compare different lanes. The areas in between lanes are also selected with separate ROIs. The inter-lane ROIs characterize the gel background and a possible inhomogeneous illumination. For each lane, the average background calculated from the left and the right inter-lane ROIs is calculated and subtracted to get the band intensity only. To finally quantify each band, the intensity of each lane is normalized to the reference lane. This step includes some of the limitations of this analysis:

- The total intensity per lane is homogeneous for each lane:
  - The total amount of DNA in the reference sample and in each lane is the same. We assume that all DNA are stained by SYBR gold similarly, resulting in the same total intensity per lane.
  - The increase in intensity at similar concentrations for longer bands is due to the increase in length. With the item above, an increase in length is similar to a linearly increasing probability of SYBR staining.
  - Differences in the total intensity of each lane are attributed to the pipetting error that occurs due to handling small volumes of fluid.
- There is a need for a reference sample of known length and concentration. Furthermore, the products need to be well defined as resulting from the monomers. This analysis is not suitable for pools with different illumination per length samples with similar product lengths.
- The concentration of DNA products in comparison to the monomers can only vary in the range of detection.

In the last step, the baseline is subtracted from all lanes to achieve the final concentrations. The band at a length of about 24 nt was identified as artifacts during the strand synthesis (see [58]). Each lane is then accessible as an intensity over gel-position data structure. Due to the background correction described above, there is no tilting or region-specific shift in the baseline. Each band is then either fit by a Gaussian curve, or all data points in the region of the band, from baseline to baseline, are simply summed up. The concentration is then a function of band-peak intensity and the position in the lane, encoding the length of the DNA strand.

C. Thermodynamic and kinetic sequence selection in enzyme-free polymer self-assembly

C.1. No energetic bias and no kinetic stalling: $\delta_\gamma = 0$ and $\sigma_1 = \sigma_2 = 1$ — detailed visualizations

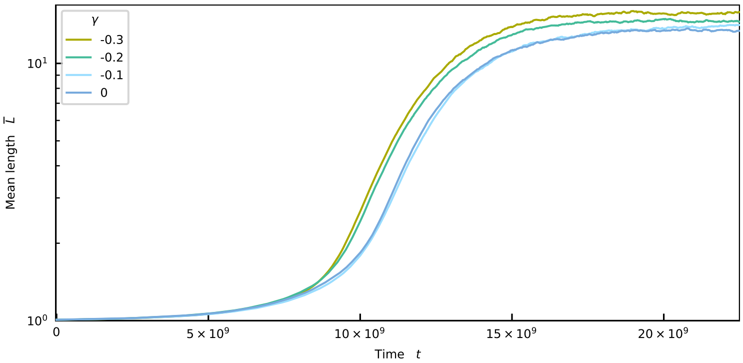


Figure C.1.: Log–linear visualization of the evolution of the mean length $\overline{L}$ for $\sigma_1 = \sigma_2 = 1$ and various values of δ_γ (see also Fig. 6.2 in Chapter 6). The initial growth of $\overline{L}$ after the short lag phase is approximately exponential.

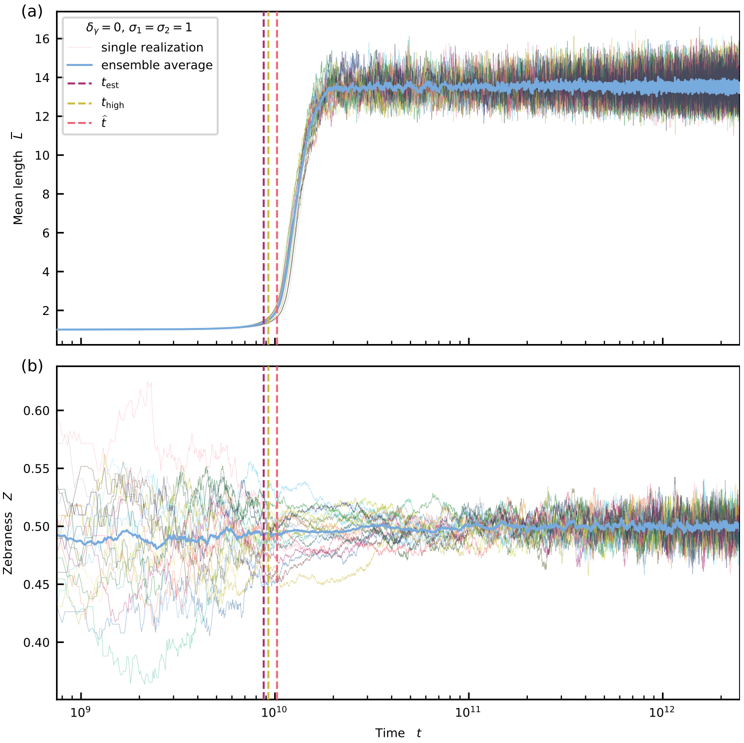


Figure C.2.: (a) and (b) Evolution of the mean length $\overline{L}$ and system-level zebraness Z for $\delta_\gamma = 0$ and $\sigma_1 = \sigma_2 = 1$ (see also Fig. 6.2 in Chapter 6). The relative deviations of t_{est} from t_{high} and $\widehat{t}$ are 5.4% and 14.6%.

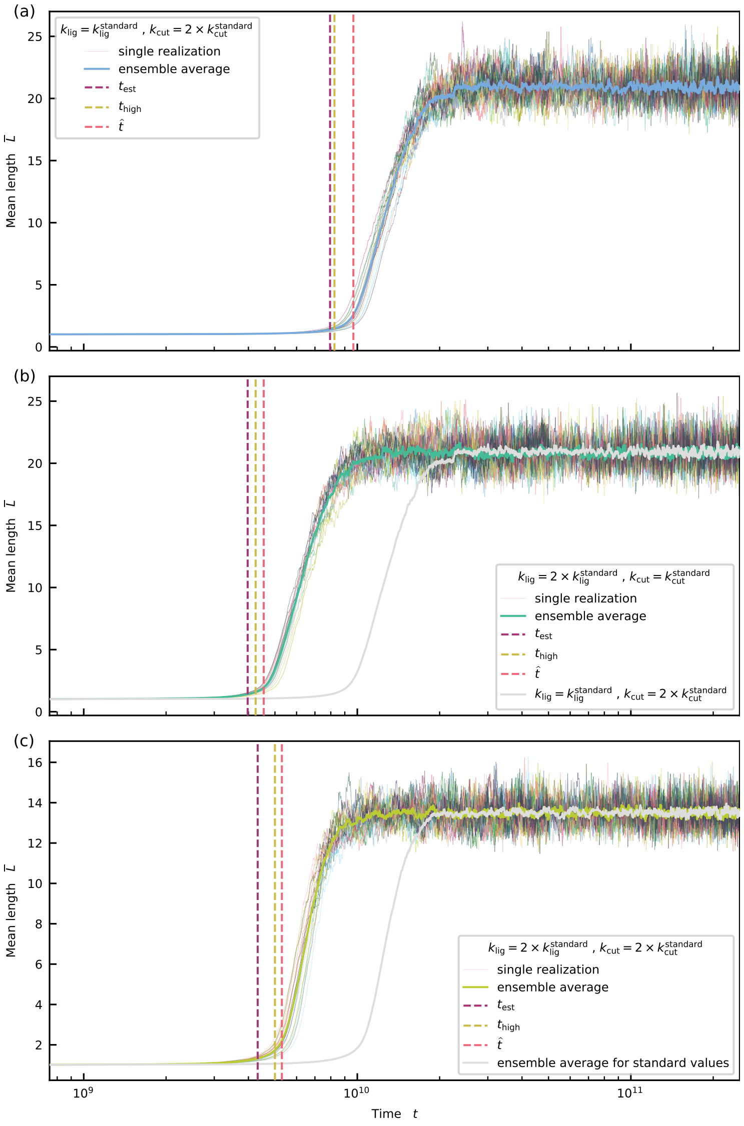


Figure C.3.: Evolution of the mean length $\overline{L}$ for $\delta_\gamma = 0$ and $\sigma_1 = \sigma_2 = 1$ for k_{lig} and k_{cut} different from the standard values $k_{\text{lig}}^{\text{standard}}$ and $k_{\text{cut}}^{\text{standard}}$ given in Tab. 6.2 (see also Fig. 6.2 in Chapter 6). (a) Reducing the rate of hydrolysis by a factor of two moves the onset slightly to the left. The relative deviations of t_{est} from t_{high} and $\widehat{t}$ are 3.6% and 17.8%. (b) Multiplying the ligation rate by a factor of two and leaving the hydrolysis rate unchanged shifts the onset of growth to the left $\widehat{t}$. The relative deviations of t_{est} from t_{high} and $\widehat{t}$ are 6.4% and 12.6%. (c) Doubling both the ligation and hydrolysis rate shifts the onset to the left and shortens the time window of the rapid growth phase. The relative deviations of t_{est} from t_{high} and $\widehat{t}$ are 13.5% and 18.4%. (a), (b) and (x) The mean length $\overline{L}$ in the steady-state is the same in the first and the second scenario. In the third scenario, the stationary value of $\overline{L}$ is the same as in the standard scenario (grey line). This observation suggests that $\overline{L}$ only depends on the ratio of k_{lig} and k_{cut}.

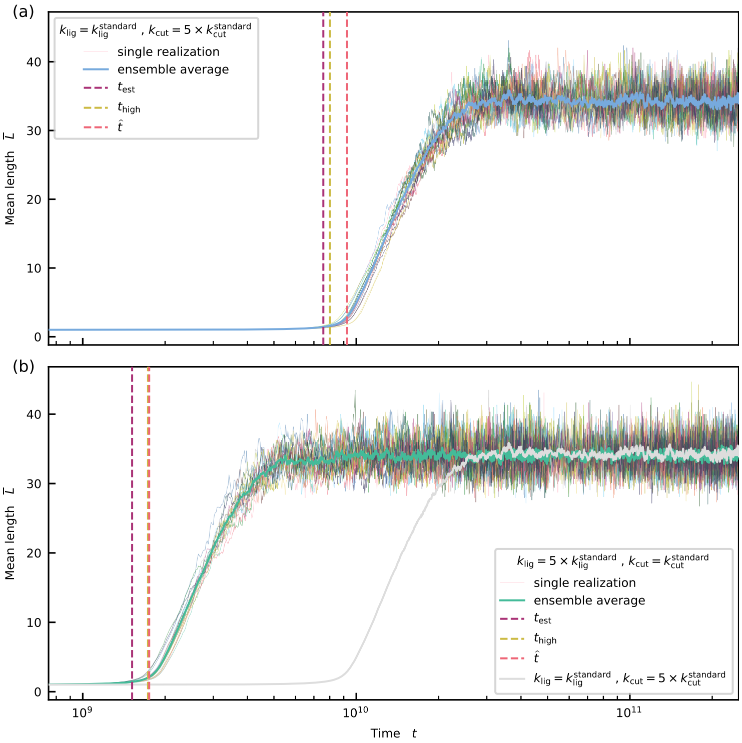


Figure C.4.: Evolution of the mean length $\overline{L}$ for $\delta_\gamma = 0$ and $\sigma_1 = \sigma_2 = 1$ for k_{lig} and k_{cut} different from the standard values $k_{\mathrm{lig}}^{\mathrm{standard}}$ and $k_{\mathrm{cut}}^{\mathrm{standard}}$ given in Tab. 6.2 (see also Fig. 6.2 in Chapter 6). (a) Reducing the rate of hydrolysis by a factor of five moves the onset slightly to the left. The relative deviations of t_{est} from t_{high} and $\widehat{t}$ are 5.2% and 18.0%. The relative deviations of t_{est} from t_{high} and $\widehat{t}$ are both 13.4%. (b) Multiplying the ligation rate by a factor of five and leaving the hydrolysis rate unchanged shifts the onset to the left $\widehat{t}$. The mean length $\overline{L}$ in the steady-state is the same in the first and the second scenario (see gray line). This observation again suggests that $\overline{L}$ only depends on the ratio of k_{lig} and k_{cut}, i.e., $\overline{L} = \overline{L}\left(k_{\mathrm{lig}}/k_{\mathrm{cut}}\right)$ (see also Fig. C.16).

C.2. Energetic bias in the absence of kinetic stalling: $\delta_\gamma < 0$ and $\sigma_1 = \sigma_2 = 1$ — detailed visualizations

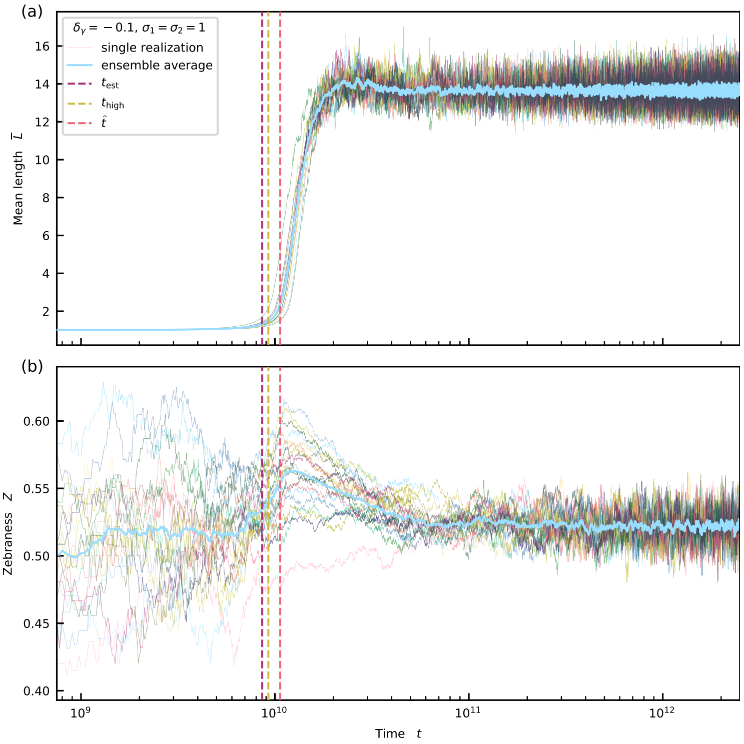


Figure C.5.: (a) and (b) Evolution of the mean length $\overline{L}$ and system-level zebraness Z for $\delta_\gamma = -0.1$ and $\sigma_1 = \sigma_2 = 1$ (see also Fig. 6.2 in Chapter 6). The relative deviations of t_{est} from t_{high} and $\widehat{t}$ are 7.0% and 19.2%.

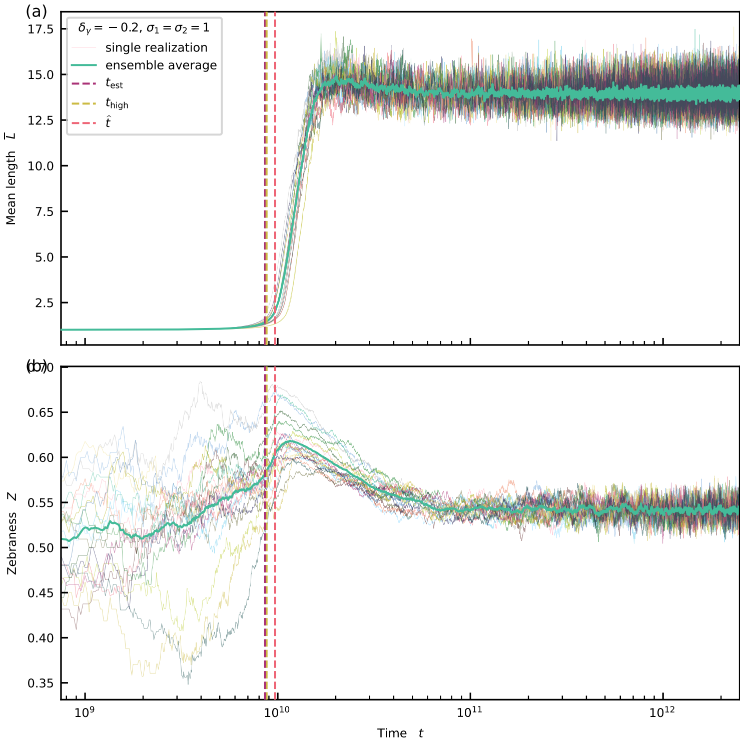


Figure C.6.: (a) and (b) Evolution of the mean length $\overline{L}$ and system-level zebraness Z for $\delta_\gamma = -0.2$ and $\sigma_1 = \sigma_2 = 1$ (see also Fig. 6.2 in Chapter 6). The relative deviations of t_{est} from t_{high} and $\widehat{t}$ are 1.7% and 11.3%.

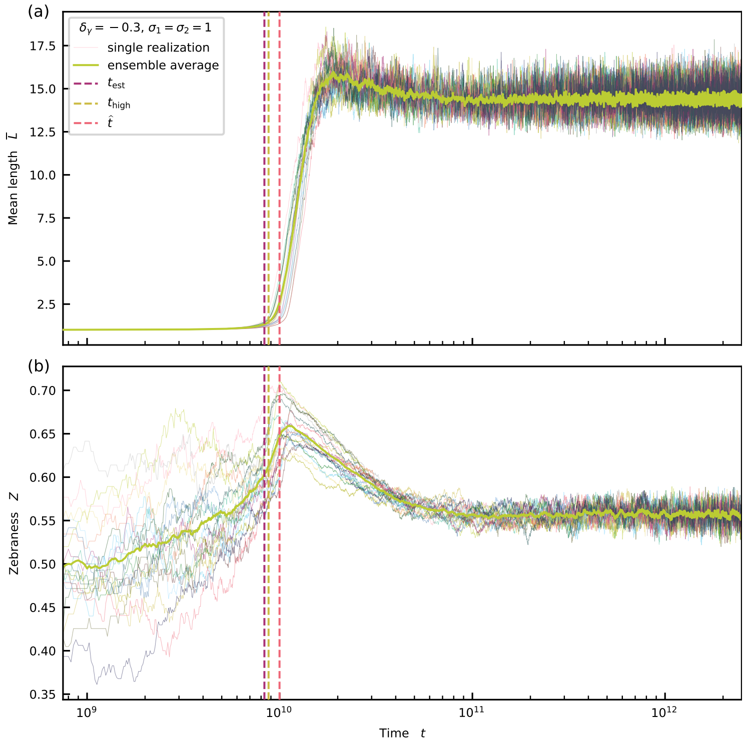


Figure C.7.: (a) and (b) Evolution of the mean length $\overline{L}$ and system-level zebraness Z for $\delta_\gamma = -0.3$ and $\sigma_1 = \sigma_2 = 1$ (see also Fig. 6.2 in Chapter 6). The relative deviations of t_{est} from t_{high} and $\widehat{t}$ are 4.9% and 16.5%.

C.3. Kinetic stalling in the absence of energetic bias: $\delta_\gamma = 0$ and $\sigma_1, \sigma_2 < 1$ — detailed visualizations

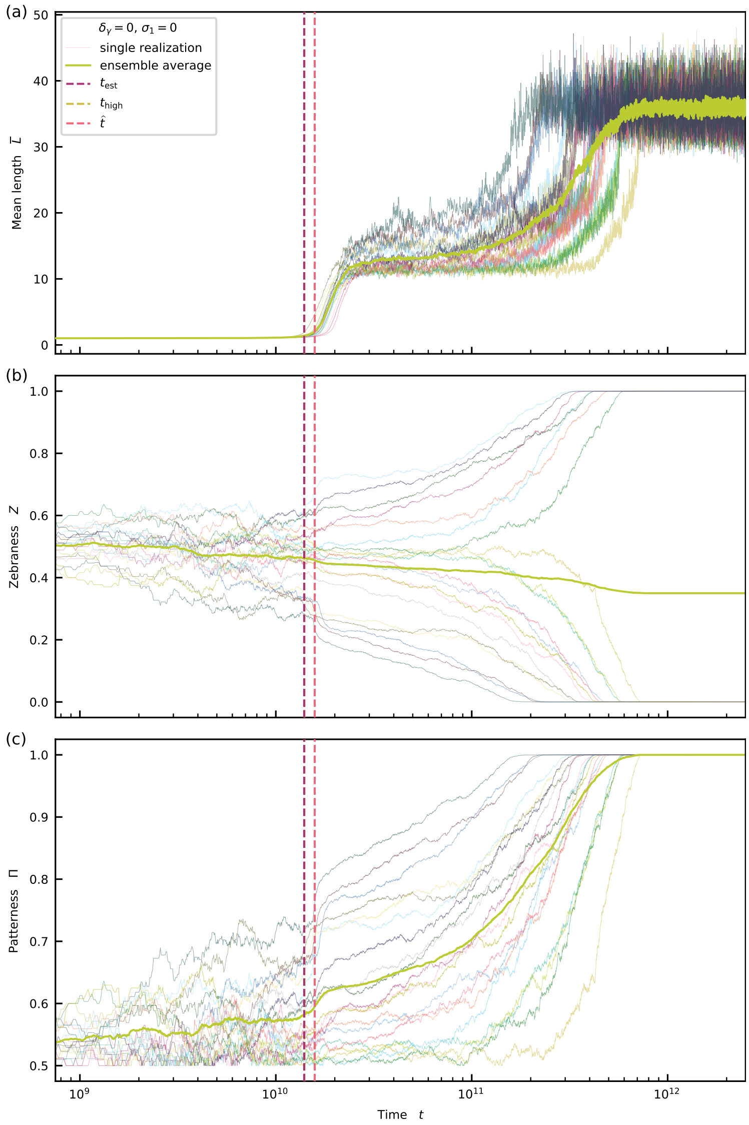


Figure C.8.: (a), (b) and (c) Evolution of the mean length $\overline{L}$, system-level zebraness Z, and patterness Π for $\delta_\gamma = 0$ and $\sigma_1 = 0$ (strong kinetic stalling). All trajectories reach a pure states $(Z = 0, 1)$. The relative deviations of t_{est} from t_{high} and $\widehat{t}$ are 0.3% and 11.7%. (See also Fig. 6.3 in Chapter 6.)

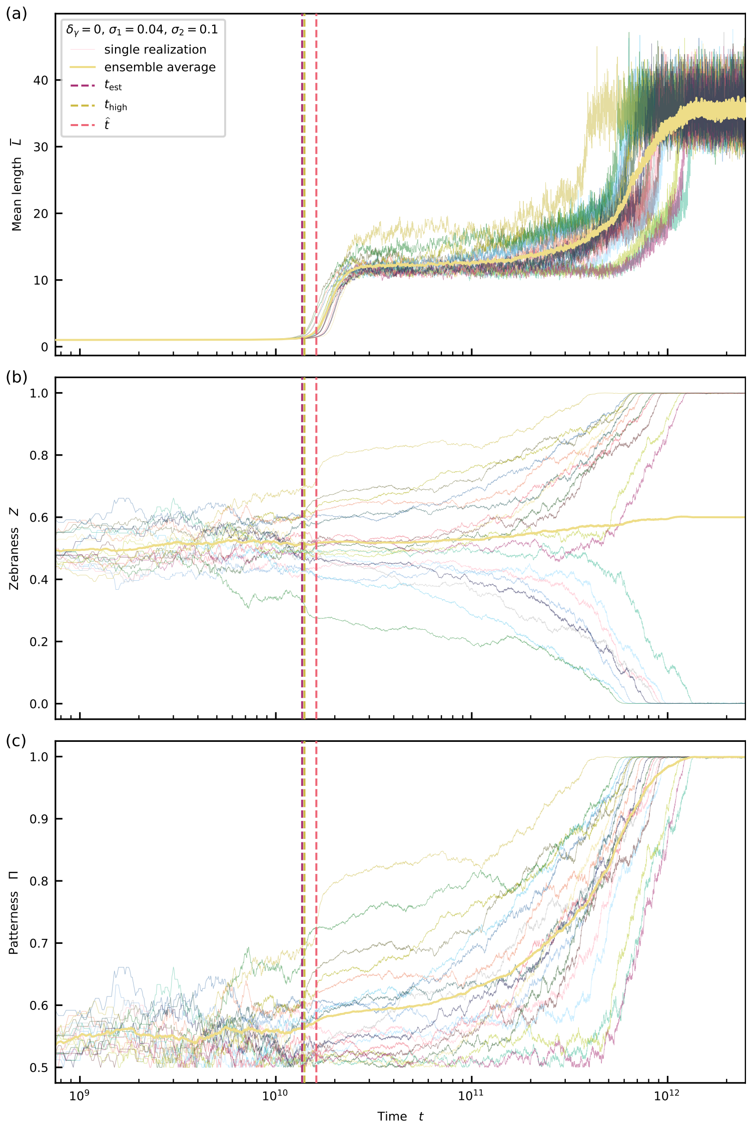


Figure C.9.: (a), (b) and (c) Evolution of the mean length $\overline{L}$ and system-level zebraness Z, and patterness Π for $\delta_\gamma = 0$ and $\sigma_1 = 0.04$ and $\sigma_1 = 0.1$ (strong kinetic stalling). All trajectories reach a pure states ($Z = 0, 1$). The relative deviations of t_{est} from t_{high} and $\widehat{t}$ are 2.6% and 15.4%. (See also Fig. 6.3 in Chapter 6.)

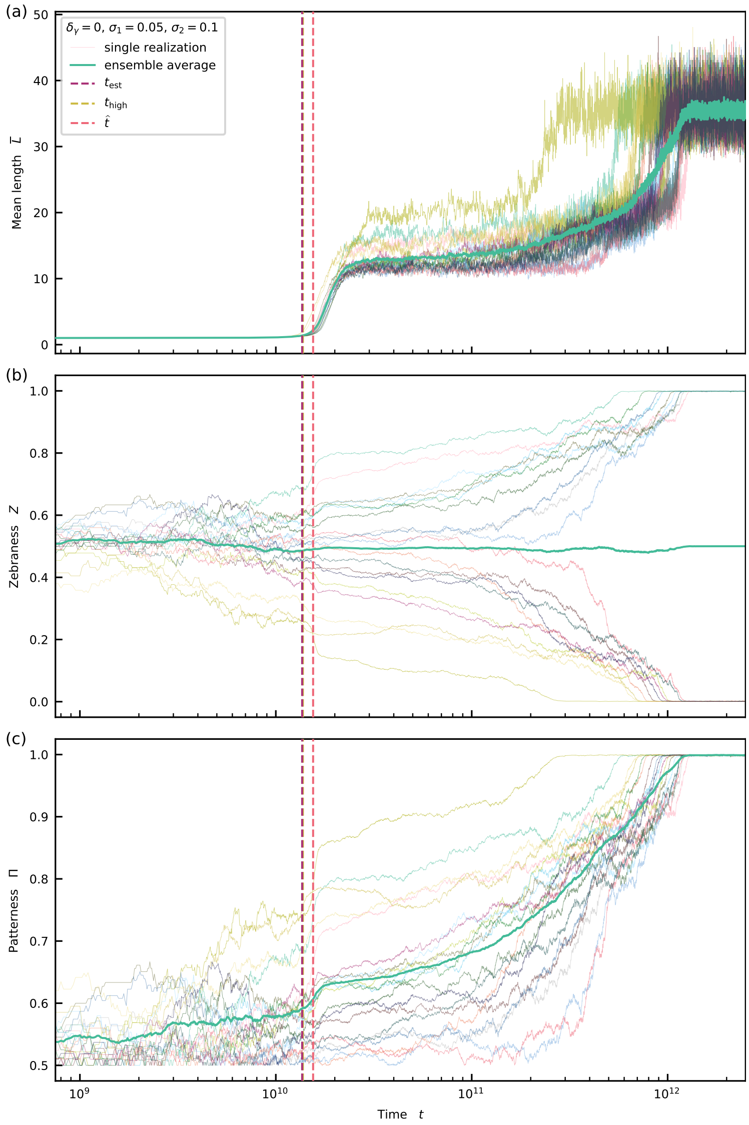


Figure C.10.: (a), (b) and (c) Evolution of the mean length $\overline{L}$ and system-level zebraness Z, and patterness Π for $\delta_\gamma = 0$ and $\sigma_1 = 0.05$ and $\sigma_1 = 0.1$ (strong kinetic stalling). All trajectories reach a pure states ($Z = 0, 1$). The relative deviations of t_{est} from t_{high} and $\widehat{t}$ are 1.6% and 12.7%. (See also Fig. 6.3 in Chapter 6.)

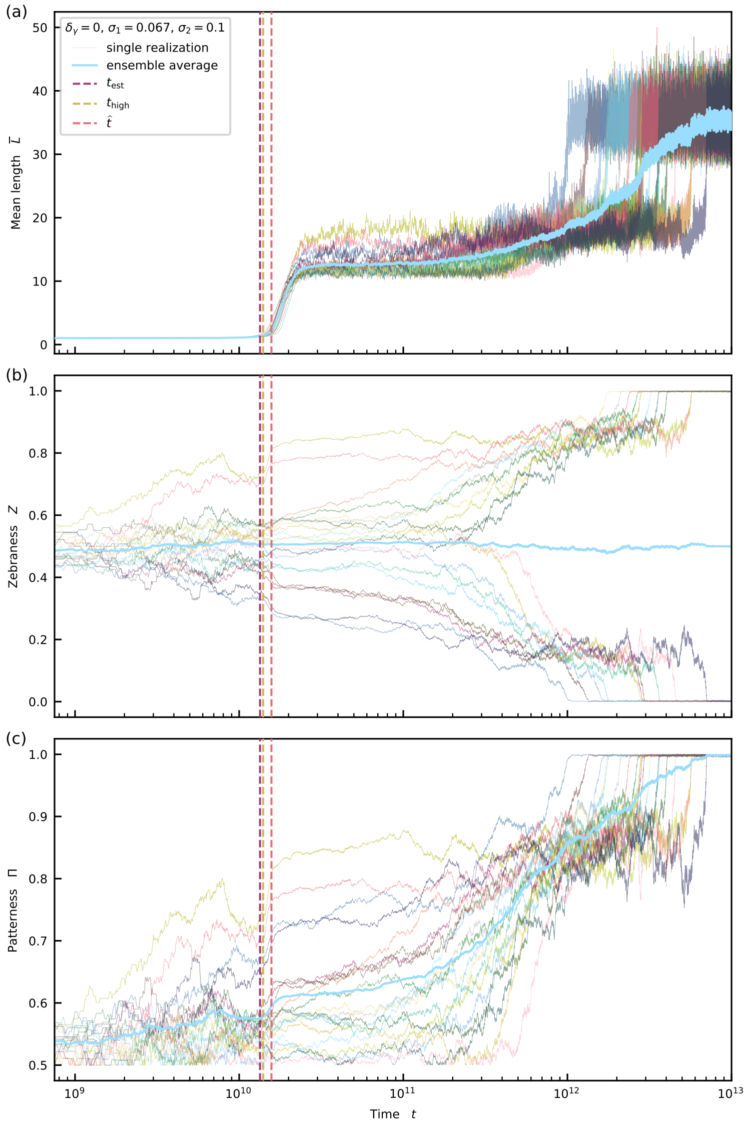


Figure C.11.: (a), (b) and (c) Evolution of the mean length $\overline{L}$ and zebraness Z, and patterness Π for $\delta_\gamma = 0$ and strong stalling, i.e., $\sigma_1 = 0.067$ and $\sigma_1 = 0.1$ (strong kinetic stalling)). All trajectories reach a pure states ($Z = 0, 1$). Note that the system needs approximately ten times longer to reach the steady-state for $\sigma_1 = 0.067$ compared to $\sigma_1 = 0, 0.04, 0.05$ (see also Fig. 6.3 in Chapter 6 and Figs. C.8–C.10. The relative deviations of t_{est} from t_{high} and $\widehat{t}$ are 4.2% and 14.8%.

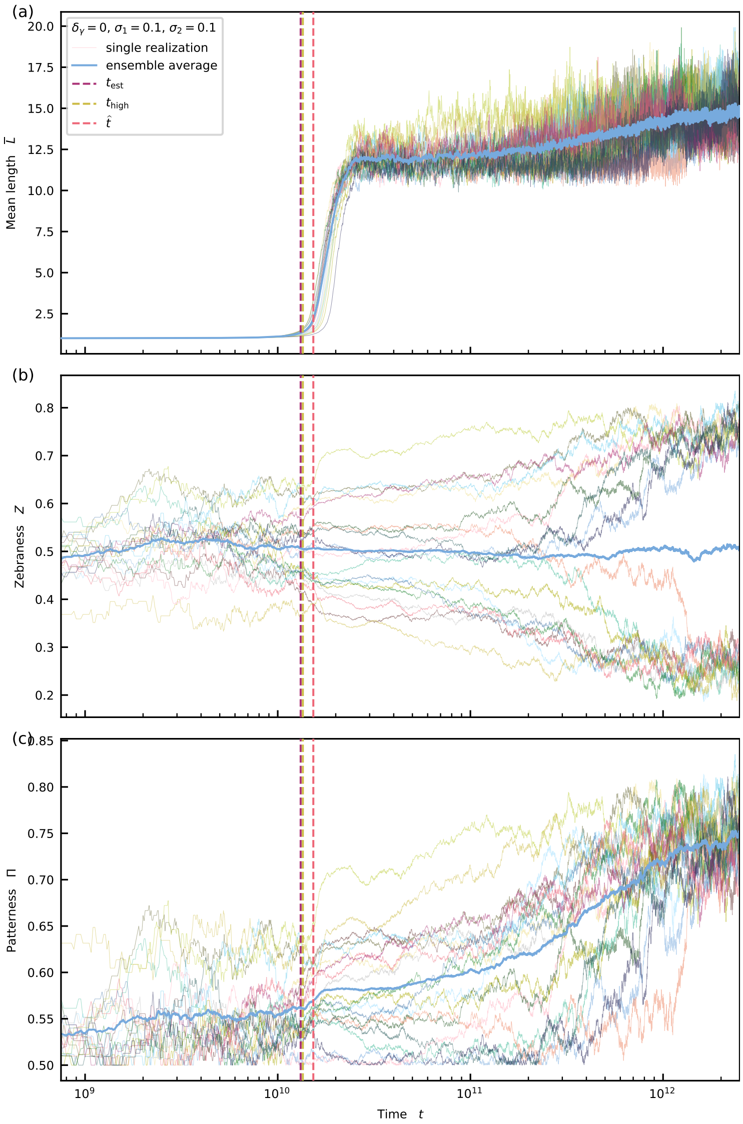


Figure C.12.: (a), (b) and (c) Evolution of the mean length $\overline{L}$ and system-level zebraness Z, and patterness Π for $\delta_\gamma = 0$ and $\sigma_1 = 0.1$ and $\sigma_1 = 0.1$ (weak kinetic stalling). In the weak kinetic stalling scenario, the trajectories do not reach a pure states ($Z = 0, 1$) for $t \to \infty$. The relative deviations of t_{est} from t_{high} and $\widehat{t}$ are 2.6% and 14.0%. (See also Fig. 6.3 in Chapter 6).

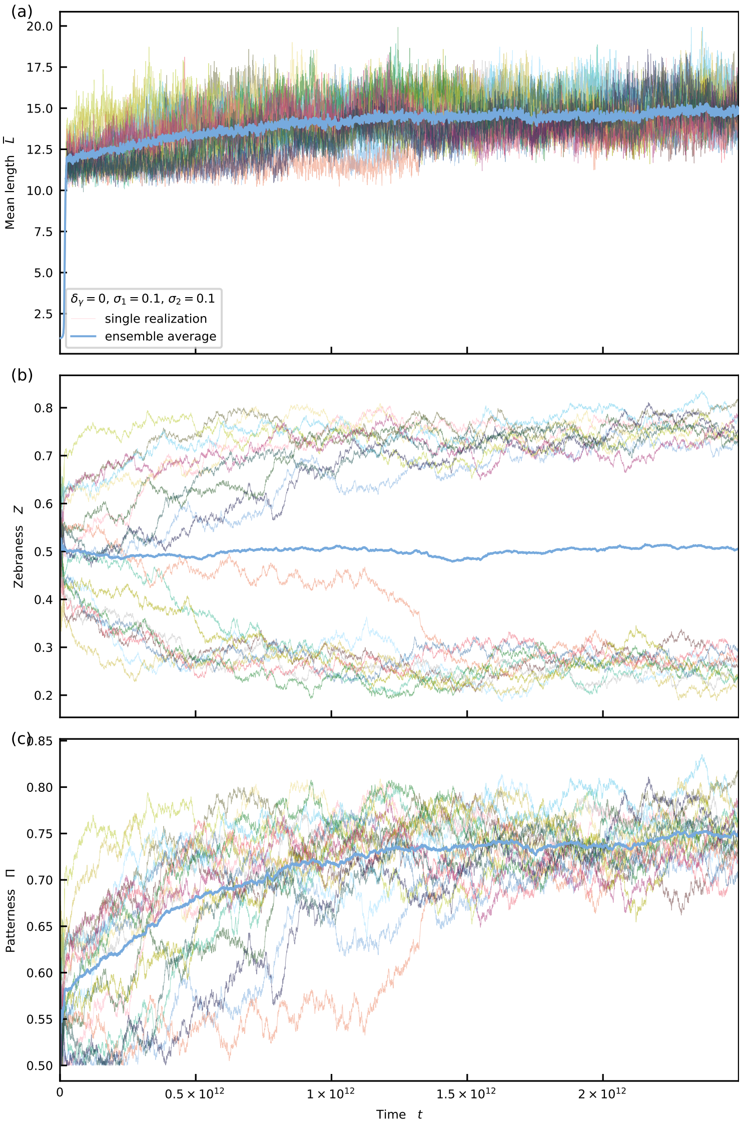


Figure C.13.: (a), (b) and (c) Visualization of the evolution of the mean length $\overline{L}$ and system-level zebraness Z, and patterness Π for $\delta_\gamma = 0$ and $\sigma_1 = 0.1$ and $\sigma_1 = 0.1$ (weak kinetic stalling) with a linear x-axis reveals that the system reaches a steady state at $t \approx 10^4$. (See also Fig. 6.3 in Chapter 6 an Fig. C.12).

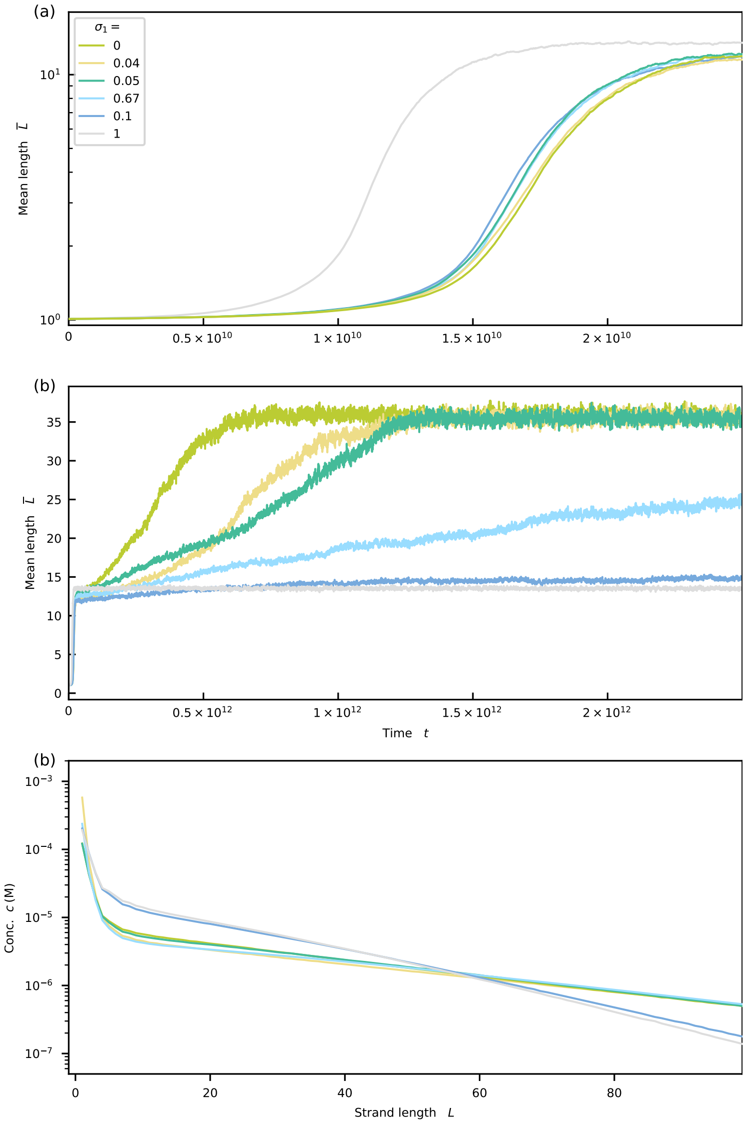


Figure C.14.: (a) and (b) The log–linear and linear–linear visualizations of the evolution of the mean length $\overline{L}$ for $\delta_\gamma = 0$ and various values of σ_1 and $\sigma_2 = 0.1$ reveal a approximately exponential first growth phase and a approximately linear second growth phase of $\overline{L}$ (see also Fig. 6.3 in Chapter 6). (c) The steady-state length distribution displays a double-exponential shape.

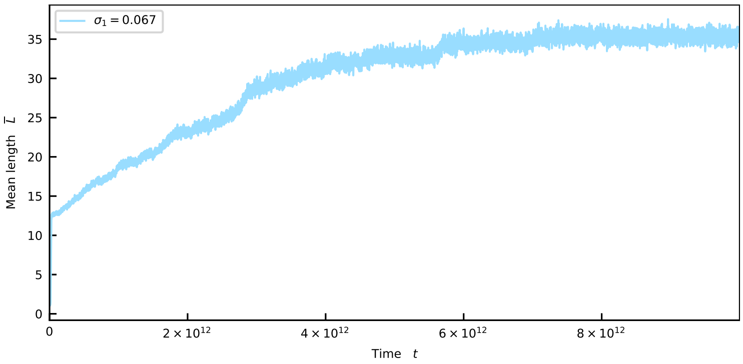


Figure C.15.: Linear–linear visualizations of the evolution of the mean length $\overline{L}$ for $\delta_\gamma = 0$ and $\sigma_1 = 0.067$ and $\sigma_2 = 0.1$ (see also Fig. C.14 and Fig. 6.3 in Chapter 6)

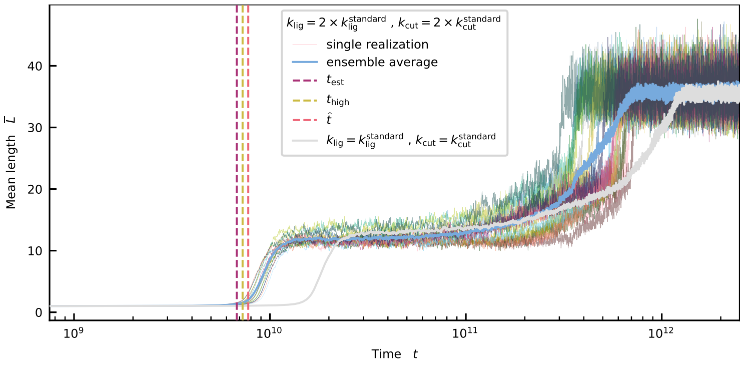


Figure C.16.: Evolution of the mean length $\overline{L}$ for $\delta_\gamma = 0.05$ and $\sigma_1 = \sigma_2 = 0.1$ for k_{lig} and k_{cut} twice as large as the standard values $k_{\mathrm{lig}}^{\mathrm{standard}}$ and $k_{\mathrm{cut}}^{\mathrm{standard}}$ given in Tab. 6.2 (see also Fig. 6.3 in Chapter 6). Doubling both rates shifts the onset to the left compared to the standard scenario (gray line). The relative deviations of t_{est} from t_{high} and $\widehat{t}$ are 6.7% and 12.7%. The mean length $\overline{L}$ in the stationary-state is the same in the standard scenario. This observation again suggests that $\overline{L}$ only depends on the ratio of k_{lig} and k_{cut}, i.e., $\overline{L} = \overline{L}\left(k_{\mathrm{lig}}/k_{\mathrm{cut}}\right)$

C.4. Hydrolysis and stalling boost sequence selection — Additional visualization

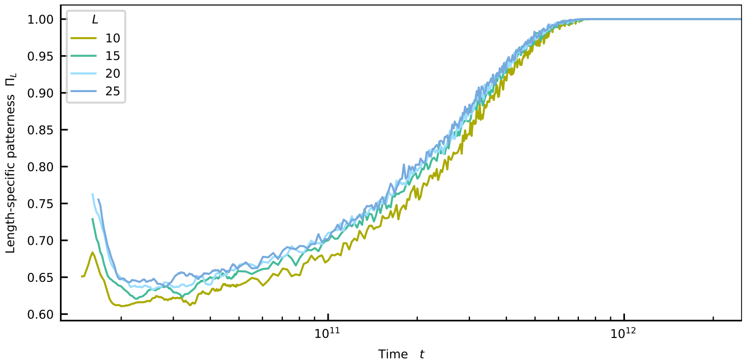


Figure C.17.: Evolution of the patterness Π_L of sequences of specific lengths $L = 10, 15, 20, 25$ for $\delta_\gamma = 0$ and $\sigma_1 = 0$. Initially, low molecule numbers lead to an increased patterness. As more strands of the given lengths form during the first rapid growth phase, their sequences become more random on average, and the Π_L decrease. During the second growth phase (see Fig. 6.3 in Chapter 6), existing strands of the given lengths break and new ones are assembled from shorter fragments. The newly assembled strands reflect the average sequence bias of the whole pool. This process becomes self-amplifying. Step by step, all strands with sequences not in line with the average sequence bias get replaced. Eventually, all strands are either fully homogeneous or zebra-like.

C.5. Energetic bias in the presence of strong kinetic stalling: $\delta_\gamma < 0$ and $\sigma_1 = 0.05, \sigma_2 = 0.1$ — Detailed visualizations

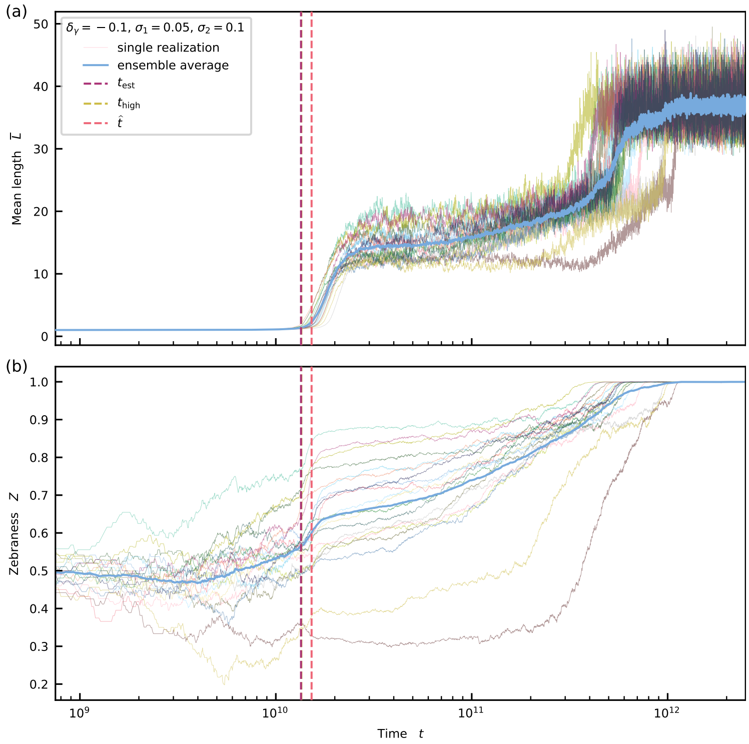


Figure C.18.: (a) and (b) Evolution of the mean length $\overline{L}$ and system-level zebraness Z, and patterness Π for $\delta_\gamma = -0.1$ and $\sigma_1 = 0.05$ and $\sigma_1 = 0.1$ (strong kinetic stalling). All trajectories reach a pure zebra states ($Z = 1$). The relative deviations of t_{est} from t_{high} and $\widehat{t}$ are 0.4% and 11.9%. (See also Fig. 6.4 in Chapter 6.)

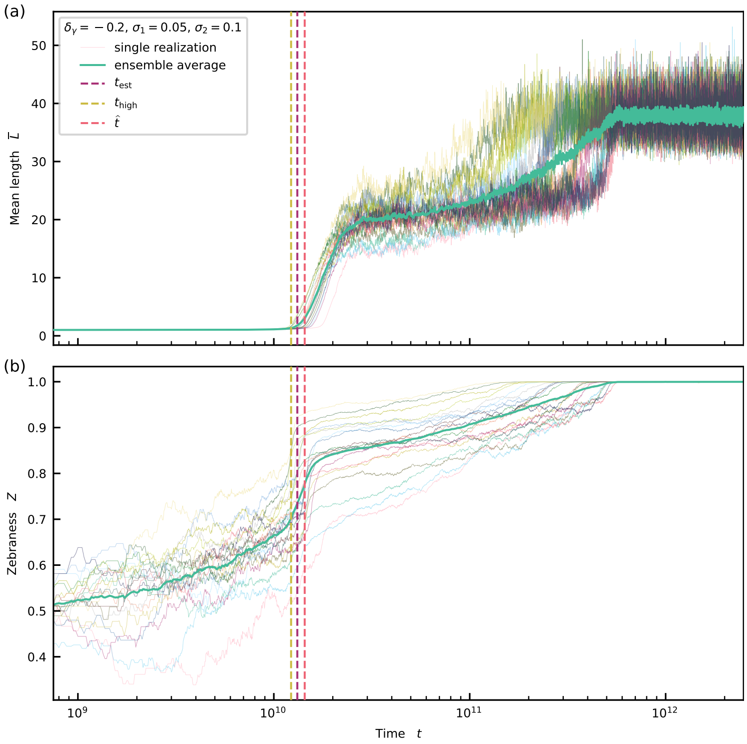


Figure C.19.: Evolution of the mean length $\overline{L}$ and system-level zebraness Z, and patterness Π for $\delta_\gamma = -0.2$ and $\sigma_1 = 0.05$ and $\sigma_1 = 0.1$ (strong kinetic stalling). All trajectories reach a pure zebra states ($Z = 1$). The relative deviations of t_{est} from t_{high} and $\widehat{t}$ are -7.6% and 8.3%. (See also Fig. 6.4 in Chapter 6.)

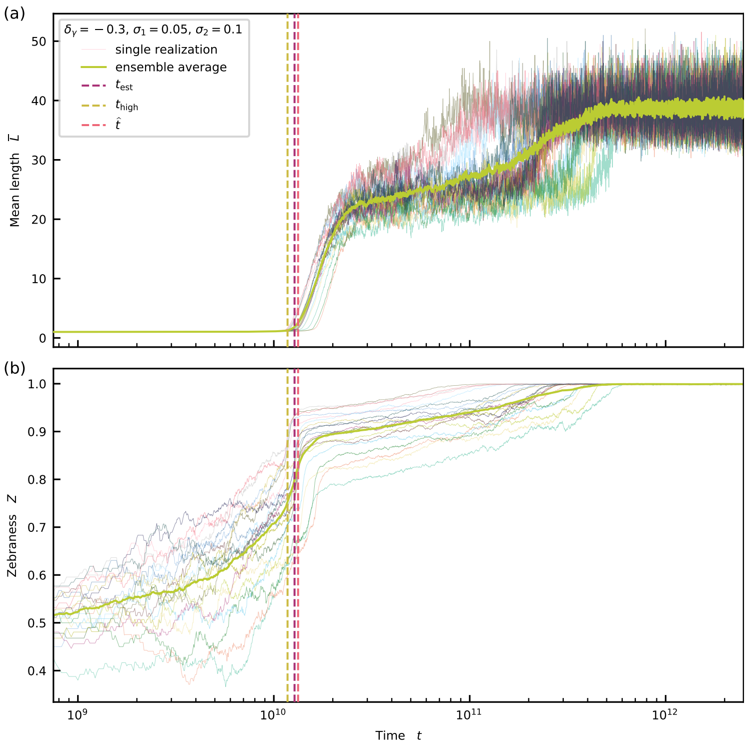


Figure C.20.: Evolution of the mean length $\overline{L}$ and system-level zebraness Z, and patterness Π for $\delta_\gamma = -0.3$ and $\sigma_1 = 0.05$ and $\sigma_1 = 0.1$ (strong kinetic stalling). All trajectories reach a pure zebra states ($Z = 1$). The relative deviations of t_{est} from t_{high} and $\widehat{t}$ are -8.6% and 4.1%. (See also Fig. 6.4 in Chapter 6.)

C.6. Energetic bias in the presence of weak kinetic stalling: $\delta_\gamma < 0$ and $\sigma_1 = \sigma_2 < 1$ — detailed visualizations

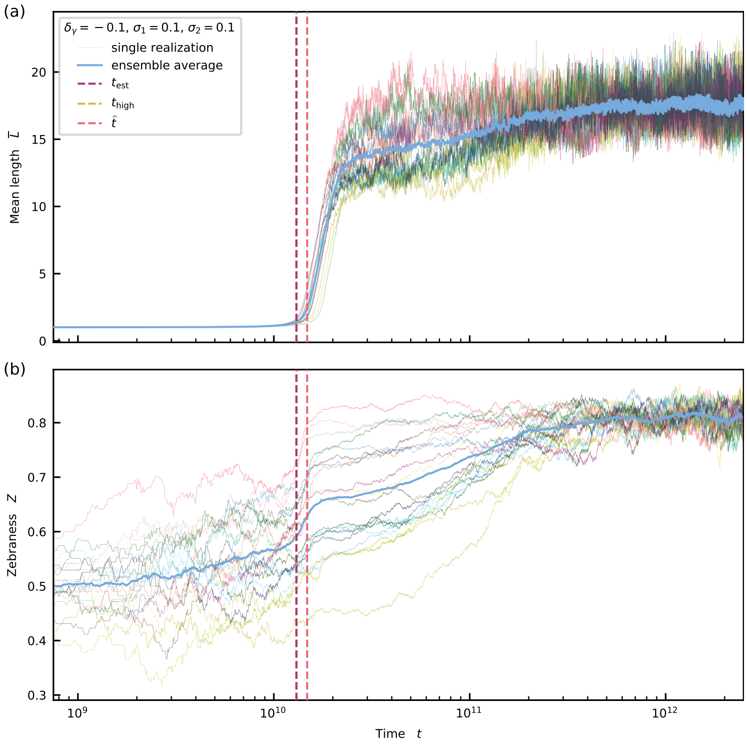


Figure C.21.: Evolution of the mean length $\overline{L}$ and system-level zebraness Z, and patterness Π for $\delta_\gamma = -0.1$ and $\sigma_1 = \sigma_1 = 0.1$ (weak kinetic stalling). All trajectories reach a non-pure zebra states ($Z < 1$). The relative deviations of t_{est} from t_{high} and $\widehat{t}$ are -0.4% and 11.8%. (See also Fig. 6.4 in Chapter 6.)

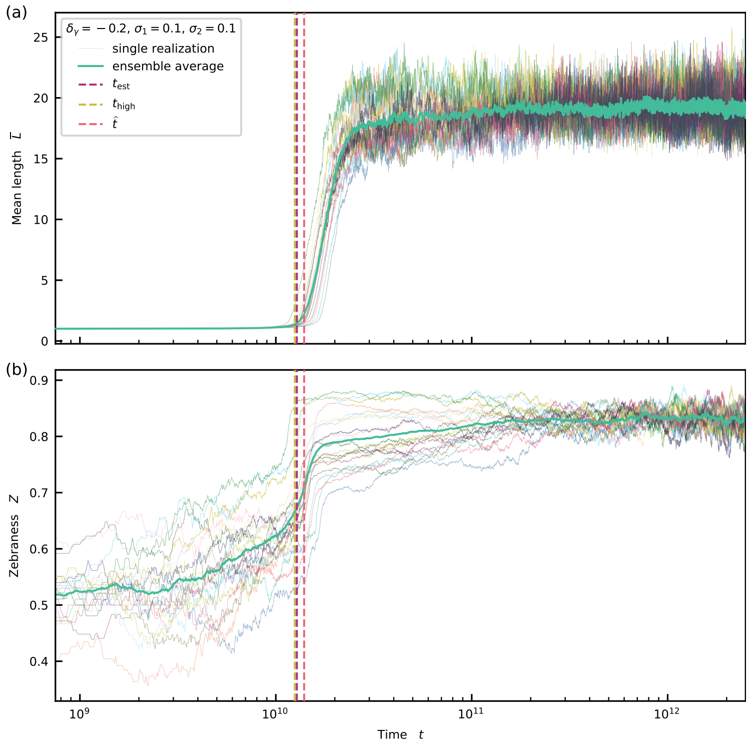


Figure C.22.: Evolution of the mean length $\overline{L}$ and system-level zebraness Z, and patterness Π for $\delta_\gamma = -0.2$ and $\sigma_1 = \sigma_1 = 0.1$ (weak kinetic stalling). All trajectories reach a non-pure zebra states ($Z < 1$). The relative deviations of t_{est} from t_{high} and $\widehat{t}$ are -2.4% and 8.2%. (See also Fig. 6.4 in Chapter 6.)

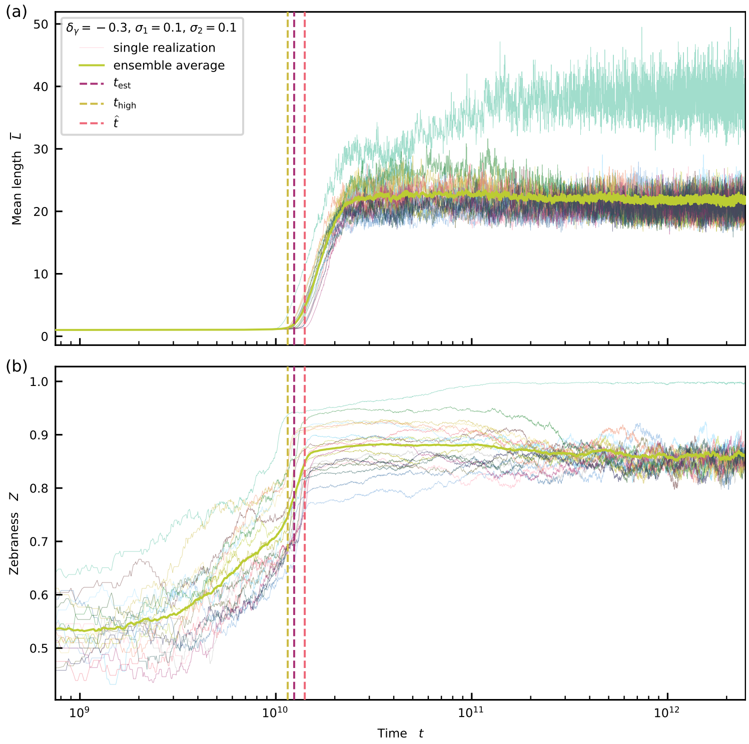


Figure C.23.: Evolution of the mean length $\overline{L}$ and system-level zebraness Z, and patterness Π for $\delta_\gamma = -0.3$ and $\sigma_1 = \sigma_1 = 0.1$ (weak kinetic stalling). While 19 out of 20 trajectories reach a non-pure zebra state, one trajectory converges to a pure zebra state. The relative deviations of t_{est} from t_{high} and $\widehat{t}$ are -7.8% and 11.7%. (See also Fig. 6.4 in Chapter 6.)

C.7. Onset of growth — Additional visualization

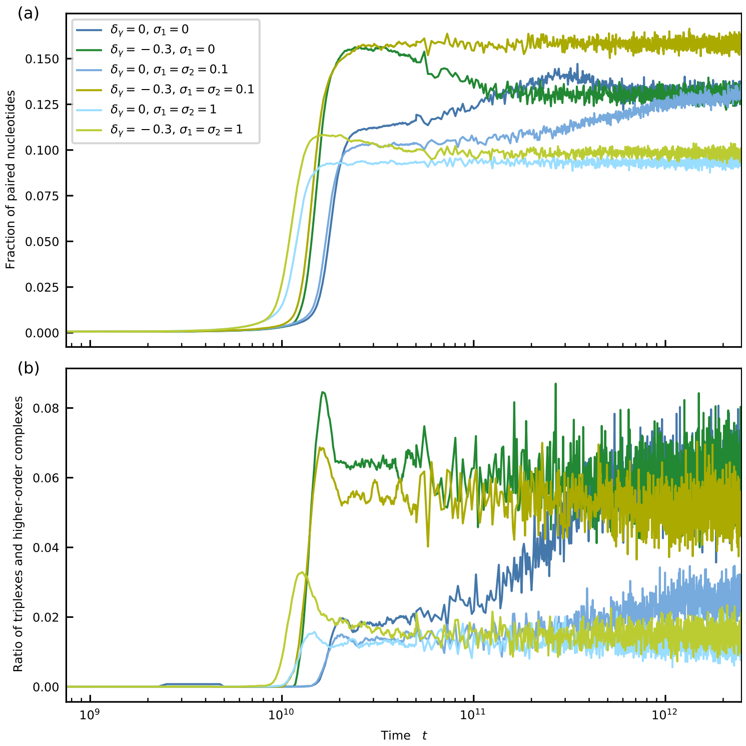


Figure C.24.: (ta) Evolution of the number of paired nucleotides for various values of δ_γ and σ_1 and $\sigma_1 = 0.1$ normalized with respect to the total number of nucleotides. Initially, the fraction of bound nucleotides is negligible. (b) Evolution of the ratio of triplexes and higher-order complexes, i.e., complexes composed of more than three strands. For early times, we can neglect higher-order complexes.

List of Figures

List of Tables

Bibliography

[1] E. Schrödinger. *What is life?* Canto classics. Cambridge: Cambridge University Press, 1948.

[2] S. I. Walker. "Origins of life: a problem for physics, a key issues review." In: *Rep. Prog. Phys.* 80.9 (Sept. 2017), p. 092601. ISSN: 0034-4885, 1361-6633. DOI: 10.1088/1361-6633/aa7804.

[3] C. P. McKay. "What Is Life—and How Do We Search for It in Other Worlds?" In: *PLoS Biol.* 2.9 (Sept. 2004), e302. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0020302.

[4] D. E. Koshland. "The Seven Pillars of Life." In: *Science* 295.5563 (Mar. 2002), pp. 2215–2216. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1068489.

[5] R. Dawkins. *The blind watchmaker: why the evidence of evolution reveals a universe without design*. New York: Norton, 1996. ISBN: 978-0-393-31570-7.

[6] N. Goldenfeld and C. Woese. "Life is Physics: Evolution as a Collective Phenomenon Far From Equilibrium." In: *Annu. Rev. Condens. Matter Phys.* 2.1 (Mar. 2011), pp. 375–399. ISSN: 1947-5454, 1947-5462. DOI: 10.1146/annurev-conmatphys-062910-140509.

[7] C. Mast, F. M. Möller, and D. Braun. "Lebendiges Nichtgleichgewicht." In: *Physik Journal* 10 (Oct. 2013), pp. 29–35.

[8] S. A. Benner. "Defining Life." In: *Astrobiology* 10.10 (Dec. 2010), pp. 1021–1030. ISSN: 1531-1074, 1557-8070. DOI: 10.1089/ast.2010.0524.

[9] F. Wachowius, J. Attwater, and P. Holliger. "Nucleic acids: function and potential for abiogenesis." In: *Q. Rev. Biophys.* 50 (2017).

[10] P. G. Higgs and N. Lehman. "The RNA World: molecular cooperation at the origins of life." In: *Nat. Rev. Genet.* 16.1 (Jan. 2015), pp. 7–17. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg3841.

[11] J. W. Szostak. "The Narrow Road to the Deep Past: In Search of the Chemistry of the Origin of Life." In: *Angew. Chem. Int. Ed.* 56.37 (Sept. 2017), pp. 11037–11043. ISSN: 14337851. DOI: 10.1002/anie.201704048.

[12] S. Ameta, Y. J. Matsubara, N. Chakraborty, S. Krishna, and S. Thutupalli. "Self-Reproduction and Darwinian Evolution in Autocatalytic Chemical Reaction Systems." In: *Life* 11.4 (Apr. 2021), p. 308. ISSN: 2075-1729. DOI: 10.3390/life11040308.

[13] L. E. Orgel. "The origin of life—a review of facts and speculations." en. In: *Trends Biochem. Sci.* 23.12 (Dec. 1998), pp. 491–495. ISSN: 09680004. DOI: 10.1016/S0968-0004(98)01300-0.

[14] R. Krishnamurthy. "Experimentally investigating the origin of DNA/RNA on early Earth." In: *Nat. Commun.* 9.1 (Dec. 2018), p. 5175. ISSN: 2041-1723. DOI: 10.1038/s41467-018-07212-y.

[15] A. Eschenmoser. "The search for the chemistry of life's origin." In: *Tetrahedron* 63.52 (Dec. 2007), pp. 12821–12844. ISSN: 00404020. DOI: 10.1016/j.tet.2007.10.012.

[16] S. A. Lanzmich. "Replication in Early Evolution." Dissertation. Ludwig-Maximilians-Universität München, 2016.

[17] P. W. Kudella. "Sequence self-selection by the network dynamics of random ligating oligomer pools." Dissertation. Ludwig-Maximilians-Universität München, 2021.

[18] M. Ohishi. "Prebiotic Complex Organic Molecules in Space." In: *Astrobiology*. Ed. by A. Yamagishi, T. Kakegawa, and T. Usui. Singapore: Springer Singapore, 2019, pp. 11–21. ISBN: 9789811336386 9789811336393. DOI: 10.1007/978-981-13-3639-3_2.

[19] Q. H. S. Chan, M. E. Zolensky, Y. Kebukawa, M. Fries, M. Ito, A. Steele, Z. Rahman, A. Nakato, A. L. D. Kilcoyne, H. Suga, Y. Takahashi, Y. Takeichi, and K. Mase. "Organic matter in extraterrestrial water-bearing salt crystals." In: *Sci. Adv.* 4.1 (Jan. 2018), eaao3521. ISSN: 2375-2548. DOI: 10.1126/sciadv.aao3521.

[20] C. Meinert, I. Myrgorodska, P. de Marcellus, T. Buhse, L. Nahon, S. V. Hoffmann, L. L. S. d'Hendecourt, and U. J. Meierhenrich. "Ribose and related sugars from ultraviolet irradiation of interstellar ice analogs." In: *Science* 352.6282 (Apr. 2016), pp. 208–212. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aad8137.

[21] T. Henning and D. Semenov. "Chemistry in Protoplanetary Disks." en. In: *Chem. Rev.* 113.12 (Dec. 2013), pp. 9016–9042. ISSN: 0009-2665, 1520-6890. DOI: 10.1021/cr400128p.

[22] A. Tupper. "Computational Modeling of RNA Replication in an RNA World." Dissertation. McMaster University, 2020.

[23] F. Crick. “Central Dogma of Molecular Biology.” In: *Nature* 227.5258 (Aug. 1970), pp. 561–563. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/227561a0.

[24] S. C. Morris. “Earth’s earliest biosphere. Its origin and evolution.” In: *Geol. J.* 20.1 (Apr. 2007), pp. 73–74. ISSN: 00721050, 10991034. DOI: 10.1002/gj.3350200107.

[25] N. H. Sleep, K. J. Zahnle, J. F. Kasting, and H. J. Morowitz. “Annihilation of ecosystems by large asteroid impacts on the early Earth.” en. In: *Nature* 342.6246 (Nov. 1989), pp. 139–142. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/342139a0.

[26] L. A. Hug, B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D. A. Relman, K. M. Finstad, R. Amundson, B. C. Thomas, and J. F. Banfield. “A new view of the tree of life.” In: *Nat.Microbiol.* 1.5 (May 2016), p. 16048. ISSN: 2058-5276. DOI: 10.1038/nmicrobiol.2016.48.

[27] C. B. Mast, S. Schink, U. Gerland, and D. Braun. “Escalation of polymerization in a thermal gradient.” In: *Proc. Natl. Acad. Sci. USA* 110.20 (May 2013), pp. 8030–8035. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1303222110.

[28] A. Salditt, L. M. Keil, D. Horning, C. Mast, G. Joyce, and D. Braun. “Thermal Habitat for RNA Amplification and Accumulation.” en. In: *Phys. Rev. Lett.* 125.4 (July 2020), p. 048104. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.125.048104.

[29] C. B. Mast and D. Braun. “Thermal Trap for DNA Replication.” In: *Phys. Rev. Lett.* 104.18 (May 2010), p. 188102. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.104.188102.

[30] B. Alberts. *Molecular biology of the cell*. Sixth edition. London; New York: Garland Science, 2015. ISBN: 978-0-8153-4432-2 978-0-8153-4464-3 978-0-8153-4524-4.

[31] R. F. Gesteland, T. Cech, and J. F. Atkins, eds. *The RNA world: the nature of modern RNA suggests a prebiotic RNA world*. 3rd ed. Cold Spring Harbor monograph series 43. Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory Press, 2006. ISBN: 978-0-87969-739-6.

[32] L. E. Orgel. “Prebiotic Chemistry and the Origin of the RNA World.” In: *Crit. Rev. Biochem. Mol. Biol.* 39.2 (Jan. 2004), pp. 99–123. ISSN: 1040-9238, 1549-7798. DOI: 10.1080/10409230490460765.

[33] G. F. Joyce. “RNA evolution and the origins of life.” In: *Nature* 338.6212 (Mar. 1989), pp. 217–224. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/338217a0.

[34] F. Crick. “The origin of the genetic code.” In: *J. Mol. Biol.* 38.3 (Dec. 1968), pp. 367–379. ISSN: 00222836. DOI: 10.1016/0022-2836(68)90392-6.

[35] L. E. Orgel. “Evolution of the genetic apparatus.” In: *J. Mol. Biol.* 38.3 (1968), pp. 381–393.

[36] W. Gilbert. “Origin of life: The RNA world.” In: *nature* 319.6055 (1986), pp. 618–618.

[37] M. P. Robertson and G. F. Joyce. “The Origins of the RNA World.” In: *Cold Spring Harb. Perspect. Biol.* 4.5 (May 2012), a003608–a003608. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a003608.

[38] K. Kruger, P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling, and T. R. Cech. “Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena.” In: *Cell* 31.1 (Nov. 1982), pp. 147–157. ISSN: 00928674. DOI: 10.1016/0092-8674(82)90414-7.

[39] W. K. Johnston. “RNA-Catalyzed RNA Polymerization: Accurate and General RNA-Templated Primer Extension.” In: *Science* 292.5520 (May 2001), pp. 1319–1325. ISSN: 00368075, 10959203. DOI: 10.1126/science.1060786.

[40] J. Attwater, A. Raguram, A. S. Morgunov, E. Gianni, and P. Holliger. “Ribozyme-catalysed RNA synthesis using triplet building blocks.” en. In: *eLife* 7 (May 2018), e35255. ISSN: 2050-084X. DOI: 10.7554/eLife.35255.

[41] A. Wochner, J. Attwater, A. Coulson, and P. Holliger. “Ribozyme-Catalyzed Transcription of an Active Ribozyme.” en. In: *Science* 332.6026 (Apr. 2011), pp. 209–212. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1200752.

[42] H. Mutschler, A. Wochner, and P. Holliger. “Freeze–thaw cycles as drivers of complex ribozyme assembly.” In: *Nat. Chem.* 7.6 (2015), pp. 502–508.

[43] T. A. Lincoln and G. F. Joyce. “Self-Sustained Replication of an RNA Enzyme.” In: *Science* 323.5918 (Feb. 2009), pp. 1229–1232. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1167856.

[44] W. K. Johnston, P. J. Unrau, M. S. Lawrence, M. E. Glasner, and D. P. Bartel. “RNA-Catalyzed RNA Polymerization: Accurate and General RNA-Templated Primer Extension.” In: *Science* 292.5520 (May 2001), pp. 1319–1325. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1060786.

[45] R. K. Saiki, S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich, and N. Arnheim. “Enzymatic Amplification of β-Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia.” en. In: *Science* 230.4732 (Dec. 1985), pp. 1350–1354. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.2999980.

[46] D. P. Horning and G. F. Joyce. "Amplification of RNA by an RNA polymerase ribozyme." In: *Proc. Natl. Acad. Sci. U.S.A.* 113.35 (Aug. 2016), pp. 9786–9791. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1610103113.

[47] N. Ban, P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz. "The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution." In: *Science* 289.5481 (Aug. 2000), pp. 905–920. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.289.5481.905.

[48] B. T. Wimberly, D. E. Brodersen, W. M. Clemons, R. J. Morgan-Warren, A. P. Carter, C. Vonrhein, T. Hartsch, and V. Ramakrishnan. "Structure of the 30S ribosomal subunit." en. In: *Nature* 407.6802 (Sept. 2000), pp. 327–339. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/35030006.

[49] M. Root-Bernstein and R. Root-Bernstein. "The ribosome as a missing link in the evolution of life." In: *J. Theor. Biol.* 367 (Feb. 2015), pp. 130–158. ISSN: 00225193. DOI: 10.1016/j.jtbi.2014.11.025.

[50] P. Nissen, J. Hansen, N. Ban, P. B. Moore, and T. A. Steitz. "The Structural Basis of Ribosome Activity in Peptide Bond Synthesis." In: *Science* 289.5481 (Aug. 2000), pp. 920–930. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.289.5481.920.

[51] H. F. Noller, V. Hoffarth, and L. Zimniak. "Unusual Resistance of Peptidyl Transferase to Protein Extraction Procedures." en. In: *Science* 256.5062 (June 1992), pp. 1416–1419. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1604315.

[52] T. A. Steitz and P. B. Moore. "RNA, the first macromolecular catalyst: the ribosome is a ribozyme." en. In: *Trends. Biochem. Sci.* 28.8 (Aug. 2003), pp. 411–418. ISSN: 09680004. DOI: 10.1016/S0968-0004(03)00169-5.

[53] K. R. Birikh, P. A. Heaton, and F. Eckstein. "The Structure, Function and Application of the Hammerhead Ribozyme." In: *Eur. J. Biochem.* 245.1 (Apr. 1997), pp. 1–16. ISSN: 0014-2956, 1432-1033. DOI: 10.1111/j.1432-1033.1997.t01-3-00001.x.

[54] C. Hsiao, S. Mohan, B. K. Kalahar, and L. D. Williams. "Peeling the Onion: Ribosomes Are Ancient Molecular Fossils." In: *Mol. Biol. Evol.* 26.11 (Nov. 2009), pp. 2415–2425. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msp163.

[55] M. W. Powner, B. Gerland, and J. D. Sutherland. "Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions." In: *Nature* 459.7244 (May 2009), pp. 239–242. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature08013.

[56] J. Xu, V. Chmela, N. Green, D. Russell, M. Janicki, R. Góra, R. Szabla, A. Bond, and J. Sutherland. "Selective prebiotic formation of RNA pyrimidine and DNA purine nucleosides." en. In: *Nature* 582.7810 (June 2020), pp. 60–66. ISSN: 0028-0836, 1476-4687. DOI: `10.1038/s41586-020-2330-9`.

[57] S. Becker, I. Thoma, A. Deutsch, T. Gehrke, P. Mayer, H. Zipse, and T. Carell. "A high-yielding, strictly regioselective prebiotic purine nucleoside formation pathway." en. In: *Science* 352.6287 (May 2016), pp. 833–836. ISSN: 0036-8075, 1095-9203. DOI: `10.1126/science.aad2808`.

[58] P. W. Kudella, A. V. Tkachenko, A. Salditt, S. Maslov, and D. Braun. "Structured sequences emerge from random pool when replicated by templated ligation." In: *Proc. Natl. Acad. Sci. USA* 118.8 (Feb. 2021), e2018830118. ISSN: 0027-8424, 1091-6490. DOI: `10.1073/pnas.2018830118`.

[59] S. Toyabe and D. Braun. "Cooperative Ligation Breaks Sequence Symmetry and Stabilizes Early Molecular Replication." In: *Phys. Rev. X* 9.1 (Mar. 2019), p. 011056. ISSN: 2160-3308. DOI: `10.1103/PhysRevX.9.011056`.

[60] J. Derr, M. L. Manapat, S. Rajamani, K. Leu, R. Xulvi-Brunet, I. Joseph, M. A. Nowak, and I. A. Chen. "Prebiotically plausible mechanisms increase compositional diversity of nucleic acid sequences." In: *Nucleic Acids Res.* 40.10 (May 2012), pp. 4711–4722. ISSN: 1362-4962, 0305-1048. DOI: `10.1093/nar/gks065`.

[61] C. Briones, M. Stich, and S. C. Manrubia. "The dawn of the RNA World: Toward functional complexity through ligation of random RNA oligomers." In: *RNA* 15.5 (May 2009), pp. 743–749. ISSN: 1355-8382, 1469-9001. DOI: `10.1261/rna.1488609`.

[62] E. Edeleva, A. Salditt, J. Stamp, P. Schwintek, J. Boekhoven, and D. Braun. "Continuous nonenzymatic cross-replication of DNA strands with *in situ* activated DNA oligonucleotides." In: *Chem. Sci.,* 10.22 (2019), pp. 5807–5814. ISSN: 2041-6520, 2041-6539. DOI: `10.1039/C9SC00770A`.

[63] F. Wachowius and P. Holliger. "Non-Enzymatic Assembly of a Minimized RNA Polymerase Ribozyme." In: *ChemSystemsChem* 1.1-2 (July 2019), pp. 12–15. ISSN: 2570-4206, 2570-4206. DOI: `10.1002/syst.201900004`.

[64] A. R. Ferre-D'Amare and W. G. Scott. "Small Self-cleaving Ribozymes." In: *Cold Spring Harbor Perspect. Biol.* 2.10 (Oct. 2010), a003574–a003574. ISSN: 1943-0264. DOI: `10.1101/cshperspect.a003574`.

[65] W. G. Scott, J. B. Murray, J. R. P. Arnold, B. L. Stoddard, and A. Klug. "Capturing the Structure of a Catalytic RNA Intermediate: The Hammerhead Ribozyme." In: *Science* 274.5295 (Dec. 1996), pp. 2065–2069. ISSN: 0036-8075, 1095-9203. DOI: `10.1126/science.274.5295.2065`.

[66] J. A. Doudna, S. Couture, and J. W. Szostak. "A Multisubunit Ribozyme That Is a Catalyst of and Template for Complementary Strand RNA synthesis." en. In: *Science* 251.5001 (Mar. 1991), pp. 1605–1608. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1707185.

[67] G. F. Joyce. "Directed Evolution of Nucleic Acid Enzymes." In: *Annu. Rev. Biochem.* 73.1 (June 2004), pp. 791–836. ISSN: 0066-4154, 1545-4509. DOI: 10.1146/annurev.biochem.73.011303.073717.

[68] M. Sosson, D. Pfeffer, and C. Richert. "Enzyme-free ligation of dimers and trimers to RNA primers." In: *Nucleic Acids Res.* 47.8 (May 2019), pp. 3836–3845. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkz160.

[69] C. Deck, M. Jauker, and C. Richert. "Efficient enzyme-free copying of all four nucleobases templated by immobilized RNA." In: *Nat. Chem.* 3.8 (Aug. 2011), pp. 603–608. ISSN: 1755-4330, 1755-4349. DOI: 10.1038/nchem.1086.

[70] M. Jauker, H. Griesser, and C. Richert. "Copying of RNA Sequences without Pre-Activation." In: *Angew. Chem. Int. Ed.* 54.48 (Nov. 2015), pp. 14559–14563. ISSN: 14337851. DOI: 10.1002/anie.201506592.

[71] N. Prywes, J. C. Blain, F. Del Frate, and J. W. Szostak. "Nonenzymatic copying of RNA templates containing all four letters is catalyzed by activated oligonucleotides." In: *eLife* 5 (June 2016), e17756. ISSN: 2050-084X. DOI: 10.7554/eLife.17756.

[72] L. Zhou, D. K. O'Flaherty, and J. W. Szostak. "Assembly of a Ribozyme Ligase from Short Oligomers by Nonenzymatic Ligation." In: *J. Am. Chem. Soc.* 142.37 (Sept. 2020), pp. 15961–15965. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/jacs.0c06722.

[73] L. Zhou, D. K. O'Flaherty, and J. W. Szostak. "Template-Directed Copying of RNA by Non-enzymatic Ligation." In: *Angew. Chem. Int. Ed.* 132.36 (Sept. 2020), pp. 15812–15817. ISSN: 0044-8249, 1521-3757. DOI: 10.1002/ange.202004934.

[74] W. S. Zielinski and L. E. Orgel. "Oligoaminudeoside phosphoramidates. Oligomeilzation of dimers of 3′-amino-3′-deoxy-nucleotides (GC and CG) in aqueous solution." In: *Nucleic Acids Res.* 15.4 (1987), pp. 1699–1715. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/15.4.1699.

[75] G. von Kiedrowski. "A Self-Replicating Hexadeoxynucleotide." In: *Angew. Chem. Int. Ed.* 25.10 (1986), pp. 932–935. ISSN: 0570-0833, 1521-3773. DOI: 10.1002/anie.198609322.

[76] D. H. Turner and D. H. Mathews. "NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure." In: *Nucleic Acids Res.* 38.suppl_1 (Jan. 2010), pp. D280–D282. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkp892.

[77] S. Rajamani, J. K. Ichida, T. Antal, D. A. Treco, K. Leu, M. A. Nowak, J. W. Szostak, and I. A. Chen. "Effect of Stalling after Mismatches on the Error Catastrophe in Nonenzymatic Nucleic Acid Replication." In: *J. Am. Chem. Soc.* 132.16 (Apr. 2010), pp. 5880–5885. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/ja100780p.

[78] K. Leu, E. Kervio, B. Obermayer, R. M. Turk-MacLeod, C. Yuan, J.-M. Luevano, E. Chen, U. Gerland, C. Richert, and I. A. Chen. "Cascade of Reduced Speed and Accuracy after Errors in Enzyme-Free Copying of Nucleic Acid Sequences." In: *J. Am. Chem. Soc.* 135.1 (Jan. 2013), pp. 354–366. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/ja3095558.

[79] L. Zhou, D. Ding, and J. W. Szostak. "The virtual circular genome model for primordial RNA replication." In: *RNA* 27.1 (Jan. 2021), pp. 1–11. ISSN: 1355-8382, 1469-9001. DOI: 10.1261/rna.077693.120.

[80] J. W. Szostak. "The eightfold path to non-enzymatic RNA replication." In: *J. Syst. Chem.* 3.1 (Dec. 2012), p. 2. ISSN: 1759-2208. DOI: 10.1186/1759-2208-3-2.

[81] M. Sosson and C. Richert. "Enzyme-free genetic copying of DNA and RNA sequences." In: *Beilstein J. Org. Chem.* 14 (Mar. 2018), pp. 603–617. ISSN: 1860-5397. DOI: 10.3762/bjoc.14.47.

[82] J. D. Watson, ed. *Molecular biology of the gene*. Seventh edition. Boston: Pearson, 2014. ISBN: 978-0-321-76243-6 978-0-321-90537-6 978-0-321-90264-1.

[83] J. March. *Advanced organic chemistry: reactions, mechanisms, and structure*. 4th ed. New York: Wiley, 1992. ISBN: 978-0-471-60180-7.

[84] B. J. Cafferty, D. M. Fialho, J. Khanam, R. Krishnamurthy, and N. V. Hud. "Spontaneous formation and base pairing of plausible prebiotic nucleotides in water." In: *Nat. Commun.* 7.1 (Sept. 2016), p. 11328. ISSN: 2041-1723. DOI: 10.1038/ncomms11328.

[85] V. Kolb, J. Dworkin, and S. Miller. "Alternative bases in the RNA world: The prebiotic synthesis of urazole and its ribosides." In: *J. Mol. Evol.* 38.6 (June 1994). ISSN: 0022-2844, 1432-1432. DOI: 10.1007/BF00175873.

[86] J. A. Piccirilli, S. A. Benner, T. Krauch, S. E. Moroney, and S. A. Benner. "Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet." In: *Nature* 343.6253 (Jan. 1990), pp. 33–37. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/343033a0.

[87] S. C. Kim, D. K. O'Flaherty, L. Zhou, V. S. Lelyveld, and J. W. Szostak. "Inosine, but none of the 8-oxo-purines, is a plausible component of a primordial version of RNA." In: *Proc. Natl. Acad. Sci. USA* 115.52 (Dec. 2018), pp. 13318–13323. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1814367115.

[88] E. Szathmary. "What is the optimum size for the genetic alphabet?" en. In: *Proc. Natl. Acad. Sci. U.S.A.* 89.7 (Apr. 1992), pp. 2614–2618. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.89.7.2614.

[89] L. Ribas de Pouplana, A. Torres, and A. Rafels-Ybern. "What Froze the Genetic Code?" In: *Life* 7.2 (Apr. 2017), p. 14. ISSN: 2075-1729. DOI: 10.3390/life7020014.

[90] S. Hoshika, N. A. Leal, M.-J. Kim, M.-S. Kim, N. B. Karalkar, H.-J. Kim, A. M. Bates, N. E. Watkins, H. A. SantaLucia, A. J. Meyer, S. DasGupta, J. A. Piccirilli, A. D. Ellington, J. SantaLucia, M. M. Georgiadis, and S. A. Benner. "Hachimoji DNA and RNA: A genetic system with eight building blocks." In: *Science* 363.6429 (Feb. 2019), pp. 884–887. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aat0971.

[91] Wikipedia. *Ribonukleinsäure — Wikipedia, die freie Enzyklopädie*. 2021.

[92] C. K. Mathews, K. E. Van Holde, D. R. Appling, and S. Anthony-Cahill, eds. *Biochemistry*. eng. Fourth Edition. Toronto: Pearson, 2013. ISBN: 978-0-13-800464-4.

[93] G. Varani and W. H. McClain. "The G·U wobble base pair: A fundamental building block of RNA structure crucial to RNA function in diverse biological systems." en. In: *EMBO reports* 1.1 (July 2000), pp. 18–23. ISSN: 1469-221X, 1469-3178. DOI: 10.1093/embo-reports/kvd001.

[94] B. Obermayer. "Mechanics and information of macromolecules," Dissertation. Ludwig-Maximilians-Universität München, 2010.

[95] E. Herrero-Galán, M. E. Fuentes-Perez, C. Carrasco, J. M. Valpuesta, J. L. Carrascosa, F. Moreno-Herrero, and J. R. Arias-Gonzalez. "Mechanical Identities of RNA and DNA Double Helices Unveiled at the Single-Molecule Level." en. In: *J. Am. Chem. Soc.* 135.1 (Jan. 2013), pp. 122–131. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/ja3054755.

[96] A. Bosco, J. Camunas-Soler, and F. Ritort. "Elastic properties and secondary structure formation of single-stranded DNA at monovalent and divalent salt conditions." en. In: *Nuc. Acids Res.* 42.3 (Feb. 2014), pp. 2064–2074. ISSN: 1362-4962, 0305-1048. DOI: 10.1093/nar/gkt1089.

[97] Q. Chi, G. Wang, and J. Jiang. “The persistence length and length per base of single-stranded DNA obtained from fluorescence correlation spectroscopy measurements using mean field theory.” en. In: *Phys. A: Stat. Mech. Appl.* 392.5 (Mar. 2013), pp. 1072–1079. ISSN: 03784371. DOI: 10.1016/j.physa.2012.09.022.

[98] E. Roth, A. Glick Azaria, O. Girshevitz, A. Bitler, and Y. Garini. “Measuring the Conformation and Persistence Length of Single-Stranded DNA Using a DNA Origami Structure.” en. In: *Nano Lett.* 18.11 (Nov. 2018), pp. 6703–6709. ISSN: 1530-6984, 1530-6992. DOI: 10.1021/acs.nanolett.8b02093.

[99] J. Abels, F. Moreno-Herrero, T. van der Heijden, C. Dekker, and N. Dekker. “Single-Molecule Measurements of the Persistence Length of Double-Stranded RNA.” en. In: *Biophys. J.* 88.4 (Apr. 2005), pp. 2737–2744. ISSN: 00063495. DOI: 10.1529/biophysj.104.052811.

[100] K. Hayashi, H. Chaya, S. Fukushima, S. Watanabe, H. Takemoto, K. Osada, N. Nishiyama, K. Miyata, and K. Kataoka. “Influence of RNA Strand Rigidity on Polyion Complex Formation with Block Catiomers.” en. In: *Macromol. Rapid Commun.* 37.6 (Mar. 2016), pp. 486–493. ISSN: 10221336. DOI: 10.1002/marc.201500661.

[101] S. I. Walker. “Homochirality.” In: *Encyclopedia of Astrobiology*. Ed. by M. Gargaud, R. Amils, J. C. Quintanilla, H. J. Cleaves, W. M. Irvine, D. L. Pinti, and M. Viso. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 759–760. ISBN: 978-3-642-11271-3 978-3-642-11274-4. DOI: 10.1007/978-3-642-11274-4_731.

[102] Y. Chen and W. Ma. “The origin of biological homochirality along with the origin of life.” In: *PLoS Comput. Biol.* 16.1 (Jan. 2020). Ed. by A. V. Morozov, e1007592. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1007592.

[103] G. Laurent, D. Lacoste, and P. Gaspard. “Emergence of homochirality in large molecular systems.” In: *Proc. Natl. Acad. Sci. U.S.A.* 118.3 (Jan. 2021), e2012741118. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2012741118.

[104] D. G. Blackmond. “The Origin of Biological Homochirality.” In: *Cold Spring Harb. Perspect. Biol.* 11.3 (Mar. 2019), a032540. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a032540.

[105] Z. Wang, W. Xu, L. Liu, and T. F. Zhu. “A synthetic molecular system capable of mirror-image genetic replication and transcription.” en. In: *Nat. Chem.* 8.7 (July 2016), pp. 698–704. ISSN: 1755-4330, 1755-4349. DOI: 10.1038/nchem.2517.

[106] J. S. McCaskill. “The equilibrium partition function and base pair binding probabilities for RNA secondary structure.” en. In: *Biopolymers* 29.6-7 (May 1990), pp. 1105–1119. ISSN: 0006-3525, 1097-0282. DOI: 10.1002/bip.360290621.

[107] J. SantaLucia and D. Hicks. "The Thermodynamics of DNA Structural Motifs." In: *Annu. Rev. Biophys. Biomol. Struct.* 33.1 (June 2004), pp. 415–440. ISSN: 1056-8700, 1545-4266. DOI: `10.1146/annurev.biophys.32.110601.141800`.

[108] T. Xia, J. SantaLucia, M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox, and D. H. Turner. "Thermodynamic Parameters for an Expanded Nearest-Neighbor Model for Formation of RNA Duplexes with Watson-Crick Base Pairs." en. In: *Biochemistry* 37.42 (Oct. 1998), pp. 14719–14735. ISSN: 0006-2960, 1520-4995. DOI: `10.1021/bi9809425`.

[109] R. Lorenz, S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. "ViennaRNA Package 2.0." en. In: *Algorithms Mol. Biol.* 6.1 (Dec. 2011), p. 26. ISSN: 1748-7188. DOI: `10.1186/1748-7188-6-26`.

[110] J. N. Zadeh, C. D. Steenberg, J. S. Bois, B. R. Wolfe, M. B. Pierce, A. R. Khan, R. M. Dirks, and N. A. Pierce. "NUPACK: Analysis and design of nucleic acid systems." en. In: *J. Comput. Chem.* 32.1 (Jan. 2011), pp. 170–173. ISSN: 01928651. DOI: `10.1002/jcc.21596`.

[111] C. Hammann, A. Luptak, J. Perreault, and M. de la Peña. "The ubiquitous hammerhead ribozyme." en. In: *RNA* 18.5 (May 2012), pp. 871–885. ISSN: 1355-8382, 1469-9001. DOI: `10.1261/rna.031401.111`.

[112] A. Jhunjhunwala, Z. Ali, S. Bhattacharya, A. Halder, A. Mitra, and P. Sharma. "On the Nature of Nucleobase Stacking in RNA: A Comprehensive Survey of Its Structural Variability and a Systematic Classification of Associated Interactions." In: *J. Chem. Inf. Model.* 61.3 (Mar. 2021), pp. 1470–1480. ISSN: 1549-9596, 1549-960X. DOI: `10.1021/acs.jcim.0c01225`.

[113] P. Yakovchuk. "Base-stacking and base-pairing contributions into thermal stability of the DNA double helix." en. In: *Nucl. Acids Res.* 34.2 (Jan. 2006), pp. 564–574. ISSN: 0305-1048, 1362-4962. DOI: `10.1093/nar/gkj454`.

[114] P. Mignon. "Influence of the π-π interaction on the hydrogen bonding capacity of stacked DNA/RNA bases." en. In: *Nucl. Acids Res.* 33.6 (Mar. 2005), pp. 1779–1789. ISSN: 1362-4962. DOI: `10.1093/nar/gki317`.

[115] E. Kervio, M. Sosson, and C. Richert. "The effect of leaving groups on binding and reactivity in enzyme-free copying of DNA and RNA." In: *Nucleic Acids Res.* 44.12 (July 2016), pp. 5504–5514. ISSN: 0305-1048, 1362-4962. DOI: `10.1093/nar/gkw476`.

[116] E. Kervio, A. Hochgesand, U. E. Steiner, and C. Richert. "Templating efficiency of naked DNA." In: *Proc. Natl. Acad. Sci. USA* 107.27 (July 2010), pp. 12074–12079. ISSN: 0027-8424, 1091-6490. DOI: `10.1073/pnas.0914872107`.

[117] E. Kervio, B. Claasen, U. E. Steiner, and C. Richert. "The strength of the template effect attracting nucleotides to naked DNA." In: *Nucleic Acids Res.* 42.11 (June 2014), pp. 7409–7420. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gku314.

[118] M.-M. Huang, N. Arnheim, and M. F. Goodman. "Extension of base mispairs by *Taq* DNA polymerase: implications for single nucleotide discrimination in PCR." en. In: *Nucleic Acids Res.* 20.17 (1992), pp. 4567–4573. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/20.17.4567.

[119] K. Leu, B. Obermayer, S. Rajamani, U. Gerland, and I. A. Chen. "The prebiotic evolutionary advantage of transferring genetic information from RNA to DNA." In: *Nucleic Acids Res.* 39.18 (Oct. 2011), pp. 8135–8147. ISSN: 1362-4962, 0305-1048. DOI: 10.1093/nar/gkr525.

[120] A. Kornberg and T. A. Baker. *DNA replication*. 2. ed., paperback ed. Sausalito, California: University Science Books, 2005. ISBN: 978-1-891389-44-3.

[121] E. Hänle and C. Richert. "Enzyme-Free Replication with Two or Four Bases." In: *Angew. Chem. Int. Ed.* 57.29 (July 2018), pp. 8911–8915. ISSN: 14337851. DOI: 10.1002/anie.201803074.

[122] Z. R. Adam. "Temperature oscillations near natural nuclear reactor cores and the potential for prebiotic oligomer synthesis." In: *Orig. Life Evol. Biosph.* 46.2-3 (Dec. 2015), pp. 171–187.

[123] L. M. R. Keil, F. M. Möller, M. Kieß, P. W. Kudella, and C. B. Mast. "Proton gradients and pH oscillations emerge from heat flow at the microscale." In: *Nat. Commun.* 8.1 (Dec. 2017), p. 1897. ISSN: 2041-1723. DOI: 10.1038/s41467-017-02065-3.

[124] A. Mariani, C. Bonfio, C. M. Johnson, and J. D. Sutherland. "pH-Driven RNA Strand Separation under Prebiotically Plausible Conditions." In: *Biochemistry* 57.45 (Nov. 2018), pp. 6382–6386. ISSN: 0006-2960, 1520-4995. DOI: 10.1021/acs.biochem.8b01080.

[125] A. Ianeselli, C. B. Mast, and D. Braun. "Periodic Melting of Oligonucleotides by Oscillating Salt Concentrations Triggered by Microscale Water Cycles Inside Heated Rock Pores." en. In: *Angew. Chem. Int. Ed.* 131.37 (Sept. 2019), pp. 13289–13294. ISSN: 0044-8249, 1521-3757. DOI: 10.1002/ange.201907909.

[126] B. Damer and D. Deamer. "The Hot Spring Hypothesis for an Origin of Life." In: *Astrobiology* 20.4 (Apr. 2020), pp. 429–452.

[127] J. Oró, B. Basile, S. Cortes, C. Shen, and T. Yamrom. "The prebiotic synthesis and catalytic role of imidazoles and other condensing agents." en. In: *Orig. Life* 14.1-4 (1984), pp. 237–242. ISSN: 0302-1688, 1573-0875. DOI: 10.1007/BF00933663.

[128] P. Nghe, W. Hordijk, S. A. Kauffman, S. I. Walker, F. J. Schmidt, H. Kemble, J. A. M. Yeates, and N. Lehman. "Prebiotic network evolution: six key parameters." In: *Mol Biosyst* 11.12 (2015), pp. 3206–3217. ISSN: 1742-206X, 1742-2051. DOI: `10.1039/C5MB00593K`.

[129] S. A. Kauffman. "Autocatalytic sets of proteins." In: *J. Theor. Biol.* 119.1 (Mar. 1986), pp. 1–24. ISSN: 00225193. DOI: `10.1016/S0022-5193(86)80047-9`.

[130] W. S. Zielinski and L. E. Orgel. "Autocatalytic synthesis of a tetranucleotide analogue." In: *Nature* 327.6120 (May 1987), pp. 346–347. ISSN: 0028-0836, 1476-4687. DOI: `10.1038/327346a0`.

[131] R. Rohatgi, D. P. Bartel, and J. W. Szostak. "Kinetic and Mechanistic Analysis of Nonenzymatic, Template-Directed Oligoribonucleotide Ligation." In: *J. Am. Chem. Soc.* 118.14 (Jan. 1996), pp. 3332–3339. ISSN: 0002-7863, 1520-5126. DOI: `10.1021/ja953712b`.

[132] P. Baaske, F. M. Weinert, S. Duhr, K. H. Lemke, M. J. Russell, and D. Braun. "Extreme accumulation of nucleotides in simulated hydrothermal pore systems." en. In: *Proc. Natl. Acad. Sci. USA* 104.22 (May 2007), pp. 9346–9351. ISSN: 0027-8424, 1091-6490. DOI: `10.1073/pnas.0609592104`.

[133] J. J. Hopfield. "Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity." In: *Proc. Natl. Acad. Sci. USA* 71.10 (Oct. 1974), pp. 4135–4139. ISSN: 0027-8424, 1091-6490. DOI: `10.1073/pnas.71.10.4135`.

[134] J. Ninio. "Kinetic amplification of enzyme discrimination." en. In: *Biochimie* 57.5 (July 1975), pp. 587–595. ISSN: 03009084. DOI: `10.1016/S0300-9084(75)80139-8`.

[135] J. J. Hopfield, T. Yamane, V. Yue, and S. M. Coutts. "Direct experimental evidence for kinetic proofreading in amino acylation of tRNAIle." en. In: *Proc. Natl. Acad. Sci. USA* 73.4 (Apr. 1976), pp. 1164–1168. ISSN: 0027-8424, 1091-6490. DOI: `10.1073/pnas.73.4.1164`.

[136] A. Murugan, D. A. Huse, and S. Leibler. "Speed, dissipation, and error in kinetic proofreading." In: *Proc. Natl. Acad. Sci. USA* 109.30 (July 2012), pp. 12034–12039. ISSN: 0027-8424, 1091-6490. DOI: `10.1073/pnas.1119911109`.

[137] S. C. Blanchard, R. L. Gonzalez, H. D. Kim, S. Chu, and J. D. Puglisi. "tRNA selection and kinetic proofreading in translation." In: *Nat. Struct. Mol. Biol.* 11.10 (Oct. 2004), pp. 1008–1014. ISSN: 1545-9993, 1545-9985. DOI: `10.1038/nsmb831`.

[138] T. W. McKeithan. "Kinetic proofreading in T-cell receptor signal transduction." In: *Proc. Natl. Acad. Sci. USA* 92.11 (May 1995), pp. 5042–5046. ISSN: 0027-8424, 1091-6490. DOI: `10.1073/pnas.92.11.5042`.

[139] R. Rao and L. Peliti. “Thermodynamics of accuracy in kinetic proofreading: dissipation and efficiency trade-offs.” In: *J. Stat. Mech. Theory Exp.* 2015.6 (June 2015), P06001. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2015/06/P06001.

[140] N. G. v. Kampen. *Stochastic processes in physics and chemistry*. Third. Amsterdam; New York: North-Holland, 1992. ISBN: 978-0-08-057138-6.

[141] B. Rauzan, E. McMichael, R. Cave, L. R. Sevcik, K. Ostrosky, E. Whitman, R. Stegemann, A. L. Sinclair, M. J. Serra, and A. A. Deckert. “Kinetics and Thermodynamics of DNA, RNA, and Hybrid Duplex Formation.” In: *Biochemistry* 52.5 (Feb. 2013), pp. 765–772. ISSN: 0006-2960, 1520-4995. DOI: 10.1021/bi3013005.

[142] T. E. Ouldridge. “The importance of thermodynamics for molecular systems, and the importance of molecular systems for thermodynamics.” In: *Nat. Comput.* 17.1 (Mar. 2018), pp. 3–29. ISSN: 1567-7818, 1572-9796. DOI: 10.1007/s11047-017-9646-x.

[143] A. Bębenek and I. Ziuzia-Graczyk. “Fidelity of DNA replication—a matter of proofreading.” en. In: *Current Genetics* 64.5 (Oct. 2018), pp. 985–996. ISSN: 0172-8083, 1432-0983. DOI: 10.1007/s00294-018-0820-1.

[144] P. Sartori and S. Pigolotti. “Kinetic versus Energetic Discrimination in Biological Copying.” en. In: *Phys. Rev. Lett.* 110.18 (May 2013), p. 188101. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.110.188101.

[145] M. Ehrenberg and C. Blomberg. “Thermodynamic constraints on kinetic proofreading in biosynthetic pathways.” In: *Biophys. J* 31.3 (Sept. 1980), pp. 333–358. ISSN: 00063495. DOI: 10.1016/S0006-3495(80)85063-6.

[146] C. Blomberg and M. Ehrenberg. “Energy considerations for kinetic proofreading in biosynthesis.” In: *J. Theor. Biol.* 88.4 (Feb. 1981), pp. 631–670. ISSN: 00225193. DOI: 10.1016/0022-5193(81)90242-3.

[147] R. R. Freter and M. A. Savageau. “Proofreading systems of multiple stages for improved accuracy of biological discrimination.” In: *J. Theor. Biol.* 85.1 (July 1980), pp. 99–123. ISSN: 00225193. DOI: 10.1016/0022-5193(80)90284-2.

[148] M. A. Savageau and D. S. Lapointe. “Optimization of kinetic proofreading: A general method for derivation of the constraint relations and an exploration of a specific case.” In: *J. Theor. Biol.* 93.1 (Nov. 1981), pp. 157–177. ISSN: 00225193. DOI: 10.1016/0022-5193(81)90062-X.

[149] C. H. Bennett. “Dissipation-error tradeoff in proofreading.” In: *Biosystems* 11.2-3 (Aug. 1979), pp. 85–91. ISSN: 03032647. DOI: 10.1016/0303-2647(79)90003-0.

[150] P. Sartori and S. Pigolotti. “Thermodynamics of Error Correction.” In: *Phys. Rev. X* 5.4 (Dec. 2015), p. 041039. ISSN: 2160-3308. DOI: 10.1103/PhysRevX.5.041039.

[151] S. Pigolotti and P. Sartori. "Protocols for Copying and Proofreading in Template-Assisted Polymerization." In: *J. Stat. Phys.* 162.5 (Mar. 2016), pp. 1167–1182. ISSN: 0022-4715, 1572-9613. DOI: 10.1007/s10955-015-1399-2.

[152] A. Murugan, D. A. Huse, and S. Leibler. "Discriminatory Proofreading Regimes in Nonequilibrium Systems." In: *Phys. Rev. X* 4.2 (Apr. 2014), p. 021016. ISSN: 2160-3308. DOI: 10.1103/PhysRevX.4.021016.

[153] V. Galstyan, K. Husain, F. Xiao, A. Murugan, and R. Phillips. "Proofreading through spatial gradients." In: *eLife* 9 (Dec. 2020), e60415. ISSN: 2050-084X. DOI: 10.7554/eLife.60415.

[154] T. Göppel, B. Obermayer, I. A. Chen, and U. Gerland. *A kinetic error filtering mechanism for enzyme-free copying of nucleic acid sequences*. preprint. Evolutionary Biology, Aug. 2021. DOI: 10.1101/2021.08.06.455386.

[155] M. Eigen. "Selforganization of matter and the evolution of biological macromolecules." In: *Die Naturwissenschaften* 58.10 (Oct. 1971), pp. 465–523. ISSN: 0028-1042, 1432-1904. DOI: 10.1007/BF00623322.

[156] R. Saiki, D. Gelfand, S Stoffel, S. Scharf, R Higuchi, G. Horn, K. Mullis, and H. Erlich. "Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase." In: *Science* 239.4839 (Jan. 1988), pp. 487–491. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.2448875.

[157] A. J. Berdis. "Mechanisms of DNA Polymerases." In: *Chem. Rev.* 109.7 (July 2009), pp. 2862–2879. ISSN: 0009-2665, 1520-6890. DOI: 10.1021/cr800530b.

[158] E. Nudler. "RNA Polymerase Active Center: The Molecular Engine of Transcription." en. In: *Annu. Rev. Biochem.* 78.1 (June 2009), pp. 335–361. ISSN: 0066-4154, 1545-4509. DOI: 10.1146/annurev.biochem.76.052705.164655.

[159] G. F. Joyce and J. W. Szostak. "Protocells and RNA Self-Replication." In: *Cold Spring Harbor Perspect. Biol.* 10.9 (Sept. 2018).

[160] J. Sulston, R. Lohrmann, L. E. Orgel, and H. T. Miles. "Nonenzymatic synthesis of oligoadenylates on a polyuridylic acid template." In: *Proc. Natl. Acad. Sci. USA* 59.3 (Mar. 1968), pp. 726–733. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.59.3.726.

[161] W. Zhang, C. P. Tam, L. Zhou, S. S. Oh, J. Wang, and J. W. Szostak. "Structural Rationale for the Enhanced Catalysis of Nonenzymatic RNA Primer Extension by a Downstream Oligonucleotide." en. In: *J. Am. Chem. Soc.* 140.8 (Feb. 2018), pp. 2829–2840. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/jacs.7b11750.

[162] M. Kreysing, L. Keil, S. Lanzmich, and D. Braun. “Heat flux across an open pore enables the continuous replication and selection of oligonucleotides towards increasing length.” In: *Nat. Chem.* 7.3 (Mar. 2015), pp. 203–208. ISSN: 1755-4330, 1755-4349. DOI: 10.1038/nchem.2155.

[163] R. Mizuuchi, A. Blokhuis, L. Vincent, P. Nghe, N. Lehman, and D. Baum. “Mineral surfaces select for longer RNA molecules.” In: *Chem. Commun.* 55.14 (2019), pp. 2090–2093.

[164] S. S. Mansy and J. W. Szostak. “Thermostability of model protocell membranes.” In: *Proc. Natl. Acad. Sci. USA* 105.36 (Sept. 2008), pp. 13351–13355.

[165] R. A. Beckman and L. A. Loeb. “Multi-stage proofreading in DNA replication.” In: *Q. Rev. Biophys.* 26.3 (1993), pp. 225–331.

[166] H. G. Hansma and D. E. Laney. “DNA Binding to Mica Correlates with Cationic Radius: Assay by Atomic Force Microscopy.” In: *Biophys J* 70.4 (Apr. 1996), pp. 1933–1939.

[167] A. Kanavarioti and D. H. White. “Kinetic analysis of the template effect in ribooligoguanylate elongation.” In: *Orig. Life Evol. Biosph.* 17.3-4 (Sept. 1987), pp. 333–349. ISSN: 0169-6149, 1573-0875. DOI: 10.1007/BF02386472.

[168] L. Li, N. Prywes, C. P. Tam, D. K. O’Flaherty, V. S. Lelyveld, E. C. Izgu, A. Pal, and J. W. Szostak. “Enhanced Nonenzymatic RNA Copying with 2-Aminoimidazole Activated Nucleotides.” In: *J. Am. Chem. Soc.* 139.5 (Feb. 2017), pp. 1810–1813. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/jacs.6b13148.

[169] D. T. Gillespie. “Exact stochastic simulation of coupled chemical reactions.” In: *J. Phys. Chem* 81.25 (Dec. 1977), pp. 2340–2361. ISSN: 0022-3654, 1541-5740. DOI: 10.1021/j100540a008.

[170] A. Kun, M. Santos, and E. Szathmáry. “Real ribozymes suggest a relaxed error threshold.” en. In: *Nat. Genet.* 37.9 (Sept. 2005), pp. 1008–1011. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng1621.

[171] S. Howorka, L. Movileanu, O. Braha, and H. Bayley. “Kinetics of duplex formation for individual DNA strands within a single protein nanopore.” In: *Proc. Natl. Acad. Sci. USA* 98.23 (Nov. 2001), pp. 12996–13001. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.231434698.

[172] I. Schoen, H. Krammer, and D. Braun. “Hybridization kinetics is different inside cells.” In: *Proc. Natl. Acad. Sci. USA* 106.51 (2009), pp. 21649–21654. ISSN: 0027-8424. DOI: 10.1073/pnas.0901313106.

[173] I. I. Cisse, H. Kim, and T. Ha. "A rule of seven in Watson-Crick base-pairing of mismatched sequences." In: *Nat. Struct. Mol. Biol.* 19.6 (June 2012), pp. 623–627. ISSN: 1545-9993, 1545-9985. DOI: 10.1038/nsmb.2294.

[174] Y. Li and R. R. Breaker. "Kinetics of RNA Degradation by Specific Base Catalysis of Transesterification Involving the 2′-Hydroxyl Group." In: *J. Am. Chem. Soc.* 121.23 (June 1999), pp. 5364–5372. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/ja990592p.

[175] G. K. Schroeder, C. Lad, P. Wyman, N. H. Williams, and R. Wolfenden. "The time required for water attack at the phosphorus atom of simple phosphodiesters and of DNA." In: *Proc. Natl. Acad. Sci. USA* 103.11 (Mar. 2006), pp. 4052–4055. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0510879103.

[176] B. Obermayer, H. Krammer, D. Braun, and U. Gerland. "Emergence of Information Transmission in a Prebiotic RNA Reactor." In: *Phys. Rev. Lett.* 107.1 (June 2011). ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.107.018101.

[177] N. Hud, B. Cafferty, R. Krishnamurthy, and L. Williams. "The Origin of RNA and "My Grandfather's Axe"." In: *Chem. Biol* 20.4 (Apr. 2013), pp. 466–474. ISSN: 10745521. DOI: 10.1016/j.chembiol.2013.03.012.

[178] N. V. Hud. "Searching for lost nucleotides of the pre-RNA World with a self-refining model of early Earth." In: *Nat. Commun.* 9.1 (Dec. 2018), p. 5171. ISSN: 2041-1723. DOI: 10.1038/s41467-018-07389-2.

[179] S. C. Kim, D. K. O'Flaherty, C. Giurgiu, L. Zhou, and J. W. Szostak. "The Emergence of RNA from the Heterogeneous Products of Prebiotic Nucleotide Synthesis." In: *J. Am. Chem. Soc.* 143.9 (Mar. 2021), pp. 3267–3279. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/jacs.0c12955.

[180] B. D. Heuberger, A. Pal, F. Del Frate, V. V. Topkar, and J. W. Szostak. "Replacing Uridine with 2-Thiouridine Enhances the Rate and Fidelity of Nonenzymatic RNA Primer Extension." In: *J. Am. Chem. Soc.* 137.7 (Feb. 2015), pp. 2769–2775. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/jacs.5b00445.

[181] M. Abramowitz and I. A. Stegun, eds. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. 9. Dover print.; [Nachdr. der Ausg. von 1972]. Dover books on mathematics. New York, NY: Dover Publ, 2013. ISBN: 978-0-486-61272-0.

[182] R. W. Butler and A. T. A. Wood. "Laplace approximations for hypergeometric functions with matrix argument." In: *Ann. Stat.* 30.4 (Aug. 2002). ISSN: 0090-5364. DOI: 10.1214/aos/1031689021.

[183] J. H. Rosenberger, T. Göppel, P. W. Kudella, D. Braun, U. Gerland, and B. Altaner. "Self-Assembly of Informational Polymers by Templated Ligation." In: *Phys. Rev. X* 11.3 (Sept. 2021), p. 031055. ISSN: 2160-3308. DOI: 10.1103/PhysRevX.11.031055.

[184] D. T. Gillespie. "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions." In: *J. Comput. Phys.* 22.4 (Dec. 1976), pp. 403–434. ISSN: 00219991. DOI: 10.1016/0021-9991(76)90041-3.

[185] M. A. Gibson and J. Bruck. "Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels." In: *J. Phys. Chem. A,* 104.9 (2000), pp. 1876–1889. DOI: 10.1021/jp993732q.

[186] P. W Atkins and J. De Paula. *Atkins' Physical chemistry*. English. New York: W.H. Freeman, 2006. ISBN: 978-0-7167-8759-4.

[187] A. V. Tkachenko and S. Maslov. "Spontaneous emergence of autocatalytic information-coding polymers." In: *J. Chem. Phys.* 143.4 (July 2015), p. 045102. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.4922545.

[188] T. E. Ouldridge, P. Šulc, F. Romano, J. P. K. Doye, and A. A. Louis. "DNA hybridization kinetics: zippering, internal displacement and sequence dependence." In: *Nucleic Acids Res.* 41.19 (Oct. 2013), pp. 8886–8895. ISSN: 1362-4962, 0305-1048. DOI: 10.1093/nar/gkt687.

[189] M. v. Smoluchowski. "Versuch einer mathematischen Theorie der Koagulationskinetik kolloider Lösungen." In: *Z. Phys. Chem.* 92U.1 (Jan. 1918). ISSN: 2196-7156, 0942-9352. DOI: 10.1515/zpch-1918-9209.

[190] "Chapter 2 Diffusion-Controlled Reactions in Solution." In: *Comprehensive Chemical Kinetics*. Ed. by C. H. Bamford, C. F. H. Tipper, and R. G. Compton. Vol. 25. Diffusion-Limited Reactions. Elsevier, Jan. 1985, pp. 3–46. DOI: 10.1016/S0069-8040(08)70252-8.

[191] S. Redner. *A Guide to First-Passage Processes*. Cambridge: Cambridge University Press, 2001. ISBN: 978-0-521-65248-3. DOI: 10.1017/CBO9780511606014.

[192] Z. Adamczyk, K. Sadlej, E. Wajnryb, M. L. Ekiel-Jeżewska, and P. Warszyński. "Hydrodynamic radii and diffusion coefficients of particle aggregates derived from the bead model." In: *J. Colloid Interface Sci.* 347.2 (July 2010), pp. 192–201. ISSN: 00219797. DOI: 10.1016/j.jcis.2010.03.066.

[193] N. C. Stellwagen, S. Magnusdottir, C. Gelfi, and P. G. Righetti. "Measuring the translational diffusion coefficients of small DNA molecules by capillary electrophoresis." In: *Biopolymers* 58.4 (), pp. 390–397.

[194] P. Reineck, C. J. Wienken, and D. Braun. "Thermophoresis of single stranded DNA." In: *Electrophoresis* 31.2 (2010), pp. 279–286. ISSN: 1522-2683. DOI: https://doi.org/10.1002/elps.200900505.

[195] A. V. Tkachenko and S. Maslov. "Onset of natural selection in populations of autocatalytic heteropolymers." In: *J. Chem. Phys.* 149.13 (Oct. 2018), p. 134901. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.5048488.

[196] E. Guseva, R. N. Zuckermann, and K. A. Dill. "Foldamer hypothesis for the growth and sequence differentiation of prebiotic polymers." In: *Proc. Natl. Acad. Sci. USA* 114.36 (Sept. 2017), E7460–E7468. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1620179114.

[197] M. L. Manapat, I. A. Chen, and M. A. Nowak. "The basic reproductive ratio of life." In: *J. Theor. Biol.* 263.3 (Apr. 2010), pp. 317–327. ISSN: 00225193. DOI: 10.1016/j.jtbi.2009.12.020.

[198] P. J. Flory. *Principles of polymer chemistry*. 19th ed. Ithaca, NY: Cornell Univ. Press, 2006. ISBN: 978-0-8014-0134-3.

[199] S. Lahiri, Y. Wang, M. Esposito, and D. Lacoste. "Kinetics and thermodynamics of reversible polymerization in closed systems." In: *New J. Phys.* 17.8 (Aug. 2015), p. 085008. ISSN: 1367-2630. DOI: 10.1088/1367-2630/17/8/085008.

[200] Y. J. Matsubara and K. Kaneko. "Optimal size for emergence of self-replicating polymer system." In: *Phys. Rev. E* 93.3 (Mar. 2016), p. 032503. ISSN: 2470-0045, 2470-0053. DOI: 10.1103/PhysRevE.93.032503.

[201] L. H. Gonçalves da Silva and D. Hochberg. "Open flow non-enzymatic template catalysis and replication." In: *Phys. Chem. Chem. Phys.* 20.21 (2018), pp. 14864–14875. ISSN: 1463-9076, 1463-9084. DOI: 10.1039/C8CP01828F.

[202] H. Fellermann, S. Tanaka, and S. Rasmussen. "Sequence selection by dynamical symmetry breaking in an autocatalytic binary polymer model." In: *Phys. Rev. E* 96.6 (Dec. 2017), p. 062407. ISSN: 2470-0045, 2470-0053. DOI: 10.1103/PhysRevE.96.062407.

[203] S. Tanaka, H. Fellermann, and S. Rasmussen. "Structure and selection in an autocatalytic binary polymer model." In: *EPL* 107.2 (July 2014), p. 28004. ISSN: 0295-5075, 1286-4854. DOI: 10.1209/0295-5075/107/28004.

[204] Y. J. Matsubara and K. Kaneko. "Kinetic Selection of Template Polymer with Complex Sequences." In: *Phys. Rev. Lett.* 121.11 (Sept. 2018), p. 118101. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.121.118101.

[205] R. Mizuuchi and N. Lehman. "Limited Sequence Diversity Within a Population Supports Prebiotic RNA Reproduction." In: *Life* 9.1 (Feb. 2019), p. 20. ISSN: 2075-1729. DOI: 10.3390/life9010020.

[206] A. Tupper, K. Shi, and P. Higgs. "The Role of Templating in the Emergence of RNA from the Prebiotic Chemical Mixture." In: *Life* 7.4 (Oct. 2017), p. 41. ISSN: 2075-1729. DOI: 10.3390/life7040041.

[207] P. W. Anderson. "Suggested model for prebiotic evolution: the use of chaos." In: *Proc. Natl. Acad. Sci. USA* 80.11 (June 1983), pp. 3386–3390. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.80.11.3386.

[208] C. Fernando, G. Von Kiedrowski, and E. Szathmáry. "A Stochastic Model of Nonenzymatic Nucleic Acid Replication: "Elongators" Sequester Replicators." In: *J. Mol. Evol.* 64.5 (May 2007), pp. 572–585. ISSN: 0022-2844, 1432-1432. DOI: 10.1007/s00239-006-0218-4.

[209] K. A. Dill and S. Bromberg. *Molecular driving forces: statistical thermodynamics in biology, chemistry, physics, and nanoscience*. 2nd ed. London ; New York: Garland Science, 2011. ISBN: 978-0-8153-4430-8.

[210] R. Rao and M. Esposito. "Nonequilibrium Thermodynamics of Chemical Reaction Networks: Wisdom from Stochastic Thermodynamics." In: *Phys. Rev. X* 6 (4 Dec. 2016), p. 041064. DOI: 10.1103/PhysRevX.6.041064.

[211] H. Subramanian and R. A. Gatenby. "Evolutionary advantage of anti-parallel strand orientation of duplex DNA." In: *Sci. Rep.* 10.1 (Dec. 2020), p. 9883. ISSN: 2045-2322. DOI: 10.1038/s41598-020-66705-3.

[212] T. Walton and J. W. Szostak. "A Kinetic Model of Nonenzymatic RNA Polymerization by Cytidine-5′-phosphoro-2-aminoimidazolide." In: *Biochemistry* 56.43 (Oct. 2017), pp. 5739–5747. ISSN: 0006-2960, 1520-4995. DOI: 10.1021/acs.biochem.7b00792.

[213] T. Walton and J. W. Szostak. "A Highly Reactive Imidazolium-Bridged Dinucleotide Intermediate in Nonenzymatic RNA Primer Extension." In: *J. Am. Chem. Soc.* 138.36 (Sept. 2016), pp. 11996–12002. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/jacs.6b07977.

[214] J. Kim and M. Mrksich. "Profiling the selectivity of DNA ligases in an array format with mass spectrometry." In: *Nucleic Acids Res.* 38.1 (Jan. 2010), e2–e2. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkp827.

[215] G. Lohman, R. J. Bauer, N. M. Nichols, L. Mazzola, J. Bybee, D. Rivizzigno, E. Cantin, and T. C. Evans. "A high-throughput assay for the comprehensive profiling of DNA ligase fidelity." In: *Nucleic Acids Res.* 44.2 (Jan. 2016), e14–e14. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkv898.

[216] L. Geyrhofer and N. Brenner. "Coexistence and cooperation in structured habitats." In: *BMC Ecol.* 20.1 (Dec. 2020), p. 14. ISSN: 1472-6785. DOI: 10.1186/s12898-020-00281-y.

[217] E. Szathmáry and J. M. Smith. "The major evolutionary transitions." In: *Nature* 374.6519 (Mar. 1995), pp. 227–232. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/374227a0.

[218] D. Sievers and G. von Kiedrowski. "Self-replication of complementary nucleotide-based oligomers." In: *Nature* 369.6477 (May 1994), pp. 221–224. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/369221a0.

[219] M. Todisco, T. P. Fraccia, G. P. Smith, A. Corno, L. Bethge, S. Klussmann, E. M. Paraboschi, R. Asselta, D. Colombo, G. Zanchetta, N. A. Clark, and T. Bellini. "Nonenzymatic Polymerization into Long Linear RNA Templated by Liquid Crystal Self-Assembly." In: *ACS Nano* 12.10 (Oct. 2018), pp. 9750–9762. ISSN: 1936-0851, 1936-086X. DOI: 10.1021/acsnano.8b05821.

[220] M. Morasch, D. Braun, and C. B. Mast. "Heat-Flow-Driven Oligonucleotide Gelation Separates Single-Base Differences." In: *Angew. Chem. Int. Ed.* 55.23 (June 2016), pp. 6676–6679. ISSN: 14337851. DOI: 10.1002/anie.201601886.

[221] D. Andrieux and P. Gaspard. "Nonequilibrium generation of information in copolymerization processes." In: *Proc. Natl. Acad. Sci. U.S.A.* 105.28 (July 2008), pp. 9516–9521. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0802049105.

[222] M. Eigen and P. Schuster. "A principle of natural self-organization: Part A: Emergence of the hypercycle." In: *Naturwissenschaften* 64.11 (Nov. 1977), pp. 541–565. ISSN: 0028-1042, 1432-1904. DOI: 10.1007/BF00450633.

[223] A. Blokhuis, D. Lacoste, and P. Nghe. "Universal motifs and the diversity of autocatalytic systems." In: *Proc. Natl. Acad. Sci. USA* 117.41 (Oct. 2020), pp. 25230–25236. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2013527117.

[224] J. R. Lorsch and J. W. Szostak. "In vitro evolution of new ribozymes with polynucleotide kinase activity." In: *Nature* 371.6492 (Sept. 1994), pp. 31–36. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/371031a0.

[225] J. A. Doudna and J. W. Szostak. "RNA-catalysed synthesis of complementary-strand RNA." In: *Nature* 339.6225 (June 1989), pp. 519–522. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/339519a0.

[226] M. O. Sinnokrot, E. F. Valeev, and C. D. Sherrill. "Estimates of the Ab Initio Limit for$\pi - \pi$Interactions: The Benzene Dimer." In: *J. Am. Chem. Soc.* 124.36 (Sept. 2002), pp. 10887–10893. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/ja025896h.

[227] F. H. C. Crick. "The complementary structure of DNA." In: *Proc. Natl. Acad. Sci. USA* 40.8 (Aug. 1954), pp. 756–758. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.40.8.756.

[228] J. C. Blain and J. W. Szostak. "Progress toward synthetic cells." In: *Annu. Rev. Biochem* 83 (2014), pp. 615–640.

[229] A. Wachtel, J. Vollmer, and B. Altaner. "Fluctuating currents in stochastic thermodynamics. I. Gauge invariance of asymptotic statistics." In: *Phys. Rev. E* 92.4 (Oct. 2015), p. 042132. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.92.042132.

[230] T. Lindahl and A. Andersson. "Rate of chain breakage at apurinic sites in double-stranded deoxyribonucleic acid." In: *Biochemistry* 11.19 (Sept. 1972), pp. 3618–3623. ISSN: 0006-2960, 1520-4995. DOI: 10.1021/bi00769a019.

[231] M. Komiyama, N. Takeda, and H. Shigekawa. "Hydrolysis of DNA and RNA by lanthanide ions: mechanistic studies leading to new applications." In: *Chem. Commun.* 16 (1999), pp. 1443–1451. ISSN: 13597345, 1364548X. DOI: 10.1039/a901621j.

[232] L. A. Basile, A. L. Raphael, and J. K. Barton. "Metal-activated hydrolytic cleavage of DNA." In: *J. Am. Chem. Soc.* 109.24 (Nov. 1987), pp. 7550–7551. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/ja00258a061.

[233] A. Luther, R. Brandsch, and G. von Kiedrowski. "Surface-promoted replication and exponential amplification of DNA analogues." en. In: *Nature* 396.6708 (Nov. 1998), pp. 245–248. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/24343.

[234] B. Damer and D. Deamer. "The Hot Spring Hypothesis for an Origin of Life." In: *Astrobiology* 20.4 (Apr. 2020), pp. 429–452.

[235] P. L. Krapivsky, S. Redner, and E. Ben-Naim. *A Kinetic View of Statistical Physics*. Cambridge: Cambridge University Press, 2010. ISBN: 978-0-511-78051-6. DOI: 10.1017/CBO9780511780516.

[236] L. Zhou, S. C. Kim, K. H. Ho, D. K. O'Flaherty, C. Giurgiu, T. H. Wright, and J. W. Szostak. "Non-enzymatic primer extension with strand displacement." In: *eLife* 8 (Nov. 2019), e51888. ISSN: 2050-084X. DOI: 10.7554/eLife.51888.

[237] H. Mutschler, A. I. Taylor, B. T. Porebski, A. Lightowlers, G. Houlihan, M. Abramov, P. Herdewijn, and P. Holliger. "Random-sequence genetic oligomer pools display an innate potential for ligation and recombination." In: *eLife* 7 (Nov. 2018), e43022. ISSN: 2050-084X. DOI: 10.7554/eLife.43022.

[238] A. S. Tupper and P. G. Higgs. "Rolling-circle and strand-displacement mechanisms for non-enzymatic RNA replication at the time of the origin of life." In: *J. Theor. Biol.* 527 (Oct. 2021), p. 110822. ISSN: 00225193. DOI: 10.1016/j.jtbi.2021.110822.

[239] A. Blokhuis and D. Lacoste. "Length and sequence relaxation of copolymers under recombination reactions." In: *J. Chem. Phys.* 147.9 (Sept. 2017), p. 094905. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.5001021.

[240] T. Göppel, V. V. Palyulin, and U. Gerland. "The efficiency of driving chemical reactions by a physical non-equilibrium is kinetically controlled." In: *Phys. Chem. Chem. Phys.* 18.30 (2016), pp. 20135–20143. ISSN: 1463-9076, 1463-9084. DOI: 10.1039/C6CP01034B.

[241] M. A. Nowak and H. Ohtsuki. "Prevolutionary dynamics and the origin of evolution." In: *Proc. Natl. Acad. Sci. USA* 105.39 (Sept. 2008), pp. 14924–14927. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0806714105.

[242] M. Manapat, H. Ohtsuki, R. Bürger, and M. A. Nowak. "Originator dynamics." en. In: *J. Theor. Biol.* 256.4 (Feb. 2009), pp. 586–595. ISSN: 00225193. DOI: 10.1016/j.jtbi.2008.10.006.

[243] G. Wachtershauser. "An all-purine precursor of nucleic acids." In: *Proc. Natl. Acad. Sci. USA* 85.4 (Feb. 1988), pp. 1134–1135. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.85.4.1134.

[244] M. Levy and S. L. Miller. "The stability of the RNA bases: Implications for the origin of life." In: *Proc. Natl. Acad. Sci. USA* 95.14 (July 1998), pp. 7933–7938. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.95.14.7933.

[245] L. Orgel. "Evolution of the genetic apparatus." In: *J. Mol. Biol.* 38.3 (Dec. 1968), pp. 381–393. ISSN: 00222836. DOI: 10.1016/0022-2836(68)90393-8.

[246] J. Rogers and G. F. Joyce. "A ribozyme that lacks cytidine." en. In: *Nature* 402.6759 (Nov. 1999), pp. 323–325. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/46335.

[247] J. S. Reader and G. F. Joyce. "A ribozyme composed of only two different nucleotides." en. In: *Nature* 420.6917 (Dec. 2002), pp. 841–844. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature01185.

[248] K. Schlosser and Y. Li. "DNAzyme-mediated catalysis with only guanosine and cytidine nucleotides." In: *Nucleic Acids Res.* 37.2 (Feb. 2009), pp. 413–420. ISSN: 1362-4962, 0305-1048. DOI: 10.1093/nar/gkn930.

[249] G. F. Joyce, A. W. Schwartz, S. L. Miller, and L. E. Orgel. "The case for an ancestral genetic system involving simple analogues of the nucleotides." In: *Proc. Natl. Acad. Sci. USA* 84.13 (July 1987), pp. 4398–4402. ISSN: 0027-8424, 1091-6490. DOI: `10.1073/pnas.84.13.4398`.

[250] K.-U. Schoning. "Chemical Etiology of Nucleic Acid Structure: The alpha - Threofuranosyl-(3'rightarrow 2') Oligonucleotide System." In: *Science* 290.5495 (Nov. 2000), pp. 1347–1351. ISSN: 00368075, 10959203. DOI: `10.1126/science.290.5495.1347`.

[251] G. F. Joyce. "The antiquity of RNA-based evolution." In: *Nature* 418.6894 (July 2002), pp. 214–221. ISSN: 0028-0836, 1476-4687. DOI: `10.1038/418214a`.

[252] L. E. Orgel. "Did template-directed nucleation precede molecular replication?" In: *Orig. Life Evol. Biosph.* 17.1 (Mar. 1986), pp. 27–34. ISSN: 0169-6149, 1573-0875. DOI: `10.1007/BF01809810`.

[253] F. Wachowius, J. Attwater, and P. Holliger. "Nucleic acids: function and potential for abiogenesis." In: *Q. Rev. Biophys.* 50 (2017), e4. ISSN: 0033-5835, 1469-8994. DOI: `10.1017/S0033583517000038`.

[254] M. M. Georgiadis, I. Singh, W. F. Kellett, S. Hoshika, S. A. Benner, and N. G. J. Richards. "Structural Basis for a Six Nucleotide Genetic Alphabet." In: *Proc. Natl. Acad. Sci. USA* 137.21 (June 2015), pp. 6947–6955. ISSN: 0002-7863, 1520-5126. DOI: `10.1021/jacs.5b03482`.

[255] P. E. Nielsen. "DNA Analogues with Nonphosphodiester Backbones." In: *Annu. Rev. Biophys. Biomol. Struct.* 24.1 (June 1995), pp. 167–183. ISSN: 1056-8700, 1545-4266. DOI: `10.1146/annurev.bb.24.060195.001123`.

[256] Y. Ura, J. M. Beierle, L. J. Leman, L. E. Orgel, and M. R. Ghadiri. "Self-Assembling Sequence-Adaptive Peptide Nucleic Acids." In: *Science* 325.5936 (July 2009), pp. 73–77. ISSN: 0036-8075, 1095-9203. DOI: `10.1126/science.1174577`.

[257] E Lescrinier, R Esnouf, J Schraml, R Busson, H. Heus, C. Hilbers, and P Herdewijn. "Solution structure of a HNA–RNA hybrid." en. In: *Chem. Biol.* 7 (Sept. 2000), pp. 719–731. ISSN: 10745521. DOI: `10.1016/S1074-5521(00)00017-X`.

[258] G. Wachtershauser. "An all-purine precursor of nucleic acids." In: *Proc. Natl. Acad. Sci. USA* 85.4 (Feb. 1988), pp. 1134–1135. ISSN: 0027-8424, 1091-6490. DOI: `10.1073/pnas.85.4.1134`.

[259] J. J. Chen, X. Cai, and J. W. Szostak. "N2′ → P3′ Phosphoramidate Glycerol Nucleic Acid as a Potential Alternative Genetic System." In: *J. Am. Chem. Soc.* 131.6 (Feb. 2009), pp. 2119–2121. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/ja809069b.

[260] D. K. O'Flaherty, L. Zhou, and J. W. Szostak. "Nonenzymatic Template-Directed Synthesis of Mixed-Sequence 3′-NP-DNA up to 25 Nucleotides Long Inside Model Protocells." In: *J. Am. Chem. Soc.* 141.26 (July 2019), pp. 10481–10488. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/jacs.9b04858.

[261] M. Winnacker and E. T. Kool. "Artificial Genetic Sets Composed of Size-Expanded Base Pairs." In: *Angew. Chem. Int. Ed.* 52.48 (Nov. 2013), pp. 12498–12508. ISSN: 14337851. DOI: 10.1002/anie.201305267.

[262] K. E. Nelson, M. Levy, and S. L. Miller. "Peptide nucleic acids rather than RNA may have been the first genetic molecule." In: *Proc. Natl. Acad. Sci. USA* 97.8 (Apr. 2000), pp. 3868–3871. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.97.8.3868.

[263] L. Orgel. "A Simpler Nucleic Acid." In: 290.5495 (2000), pp. 1306–1307. ISSN: 0036-8075. DOI: 10.1126/science.290.5495.1306.

[264] B. W. F. Colville and M. W. Powner. "Selective Prebiotic Synthesis of α-Threofuranosyl Cytidine by Photochemical Anomerization." In: *Angew. Chem. Int. Ed.* 60.19 (May 2021), pp. 10526–10530. ISSN: 1433-7851, 1521-3773. DOI: 10.1002/anie.202101376.

[265] *Do thermostable DNA ligases (such as Taq DNA Ligase, 9°N DNA Ligase, and HiFi Taq DNA Ligase) ligate sticky ends? | NEB*. Feb. 2021.

[266] *What is the activity of Taq DNA Ligase at various temperatures? | NEB*. Feb. 2021.

[267] W. Hordijk, J. Hein, and M. Steel. "Autocatalytic Sets and the Origin of Life." In: *Entropy* 12.7 (June 2010), pp. 1733–1742. ISSN: 1099-4300. DOI: 10.3390/e12071733.

[268] W. Hordijk and M. Steel. "Detecting autocatalytic, self-sustaining sets in chemical reaction systems." en. In: *J. Theor. Biol.* 227.4 (Apr. 2004), pp. 451–461. ISSN: 00225193. DOI: 10.1016/j.jtbi.2003.11.020.

[269] W. Hordijk, M. Steel, and S. Kauffman. "The Structure of Autocatalytic Sets: Evolvability, Enablement, and Emergence." en. In: *Acta Biotheor.* 60.4 (Dec. 2012), pp. 379–392. ISSN: 0001-5342, 1572-8358. DOI: 10.1007/s10441-012-9165-1.

[270] V. Vasas, C. Fernando, M. Santos, S. Kauffman, and E. Szathmáry. "Evolution before genes." en. In: *Biol. Direct* 7.1 (2012), p. 1. ISSN: 1745-6150. DOI: 10.1186/1745-6150-7-1.

[271] N. Vaidya, M. L. Manapat, I. A. Chen, R. Xulvi-Brunet, E. J. Hayden, and N. Lehman. "Spontaneous network formation among cooperative RNA replicators." en. In: *Nature* 491.7422 (Nov. 2012), pp. 72–77. ISSN: 0028-0836, 1476-4687. DOI: `10.1038/nature11549`.

[272] S. Kullback and R. A. Leibler. "On Information and Sufficiency." In: *Ann. Math. Stat.* 22.1 (Mar. 1951), pp. 79–86. ISSN: 0003-4851. DOI: `10.1214/aoms/1177729694`.

[273] H. Urata, H. Shimizu, H. Hiroaki, D. Kohda, and M. Akagi. "Thermodynamic study of hybridization properties of heterochiral nucleic acids." In: *Biochem. Biophys. Res. Commun* 309.1 (Sept. 2003), pp. 79–83. ISSN: 0006291X. DOI: `10.1016/S0006-291X(03)01531-6`.

[274] N. C. Hauser, R. Martinez, A. Jacob, S. Rupp, J. D. Hoheisel, and S. Matysiak. "Utilising the left-helical conformation of L-DNA for analysing different marker types on a single universal microarray platform." In: *Nucleic Acids Res.* 34.18 (Oct. 2006), pp. 5101–5111. ISSN: 0305-1048, 1362-4962. DOI: `10.1093/nar/gkl671`.

[275] M. Szabat, D. Gudanis, W. Kotkowiak, Z. Gdaniec, R. Kierzek, and A. Pasternak. "Thermodynamic Features of Structural Motifs Formed by β-L-RNA." In: *PLoS One* 11.2 (Feb. 2016). Ed. by H.-A. Tajmir-Riahi, e0149478. ISSN: 1932-6203. DOI: `10.1371/journal.pone.0149478`.

[276] G. F. Joyce, G. M. Visser, C. A. A. van Boeckel, J. H. van Boom, L. E. Orgel, and J. van Westrenen. "Chiral selection in poly(C)-directed synthesis of oligo(G)." In: *Nature* 310.5978 (Aug. 1984), pp. 602–604. ISSN: 0028-0836, 1476-4687. DOI: `10.1038/310602a0`.

[277] M. Bolli, R. Micura, and A. Eschenmoser. "Pyranosyl-RNA: chiroselective self-assembly of base sequences by ligative oligomerization of tetra nucleotide-2',3'-cyclophosphates (with a commentary concerning the origin of biomolecular homochirality)." In: *Chem. Biol.* 4.4 (Apr. 1997), pp. 309–320. ISSN: 10745521. DOI: `10.1016/S1074-5521(97)90074-0`.

[278] M. Dumousseau, N. Rodriguez, N. Juty, and N. L. Novère. "MELTING, a flexible platform to predict the melting temperatures of nucleic acids." In: *BMC Bioinf.* 13.1 (Dec. 2012), p. 101. ISSN: 1471-2105. DOI: `10.1186/1471-2105-13-101`.

Acknowledgments

First, I want to thank my supervisor Uli Gerland for his trust and guidance through the different research projects while giving me large freedom to follow my curiosities and design and to perform the research for the last project, presented in this thesis, also called the "zebra project" on my own.

Next, I would like to thank my co-supervisor, Dieter Braun, for the many fruitful and stimulating discussions and the good coffee in his lab. I would also like to thank Dieter for making Munich a hotspot for research on the origins of life.

I further would like to thank Joachim and Bernhard. We have been through so much together. I was often not an easy colleague and repeatedly gave free rein to my fears and insecurities about the success of our work. I know that. You had to put up with a lot. Thank you for supporting and encouraging me, and thank you for all that I was able to learn with you and from you.

I also want to say a big "thank you" to all former and current group members of the Gerland group for creating a welcoming and warm working atmosphere. Our retreats in the mountains were a lot of fun! In particular, I want to thank Ludwig, Gabin, Zara, JoHaKi, Julio, Cesar, Nanni, and Stefan Kremser for their valuable feedback and comments on my draft.

Moreover, I would like to thank the whole group of Dieter Braun. I enjoyed hanging around with you in your coffee Kitchen at the LMU, going sled riding, skiing, and hiking with you! Thank you, Patrick and Annaleand, for your feedback and comments on the draft for the "zebra project"!

I also want to thank Filiz for pulling the strings in the background! Filiz organized so many inspiring talks, events, seminars, workshops, and lectures on the theme of the origin of life. I have always found that participation in these events was very enriching.

I am also very thankful to the Quantitative Bioscience Munich (QBM) school for taking me as one of their graduate students and for supporting me financially during the first two years of my Ph.D. I also thank QBM for the great scientific environment created and the inspiring retreats and organized lecture series.

Finally, I would like to thank my family and all my friends for their great support during my Ph.D. You always believed in me, and you encouraged me when I was down and afraid.

Three people who are very important to me were unfortunately not able to experience

the successful end of my promotion. They are my dear mum, my dear grandpa and my dear friend Kathrin! I would have loved to celebrate this special event with you. I miss you very much.

Last I would like to thank Verena for always being there for me in the last five and a half years. Thank you for your patience, your love, your great support and for making me happy! I could not have done it without you!

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die bei der promotionsführenden Einrichtung

Fakultät für Physik

der TUM zur Promotionsprüfung vorgelegte Arbeit mit dem Titel:

Spatio-temporal Organization of Enzymatic Reactions

in

Fakultät für Physik, Lehrstuhl für Theoretische Physik - Theorie komplexer Biosysteme

unter der Anleitung und Betreuung durch: **Prof. Dr. Ulrich Gerland**
ohne sonstige Hilfe erstellt und bei der Abfassung nur die gemäß § 6 Ab. 6 und 7 Satz 2 angebotenen Hilfsmittel benutzt habe.

☒ Ich habe keine Organisation eingeschaltet, die gegen Entgelt Betreuerinnen und Betreuer für die Anfertigung von Dissertationen sucht, oder die mir obliegenden Pflichten hinsichtlich der Prüfungsleistungen für mich ganz oder teilweise erledigt.

☒ Ich habe die Dissertation in dieser oder ähnlicher Form in keinem anderen Prüfungsverfahren als Prüfungsleistung vorgelegt.

□ Die vollständige Dissertation wurde in ______________________________
veröffentlicht. Die promotionsführende Einrichtung ____________________
hat der Veröffentlichung zugestimmt.

☒ Ich habe den angestrebten Doktorgrad noch nicht erworben und bin nicht in einem früheren Promotionsverfahren für den angestrebten Doktorgrad endgültig gescheitert.

□ Ich habe bereits am ________ bei der Fakultät für ____________________
der Hochschule ____
unter Vorlage einer Dissertation mit dem Thema ______________________
die Zulassung zur Promotion beantragt mit dem Ergebnis: ______________

Die öffentlich zugängliche Promotionsordnung der TUM ist mir bekannt, insbesondere habe ich die Bedeutung von § 28 (Nichtigkeit der Promotion) und § 29 (Entzug des

Doktorgrades) zur Kenntnis genommen. Ich bin mir der Konsequenzen einer falschen Eidesstattlichen Erklärung bewusst.
Mit der Aufnahme meiner personenbezogenen Daten in die Alumni-Datei bei der TUM bin ich

☒ einverstanden

□ nicht einverstanden

Ort, Datum, Unterschrift