# TECHNISCHE UNIVERSITÄT MÜNCHEN

*Lehrstuhl für Proteomik und Bioanalytik*

# Cross species common gene regulatory network inference

Amin Moghaddas Gholami

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

Vorsitzender: Univ.-Prof. Dr. I. Antes
Prüfer der Dissertation:
1. Univ.-Prof. Dr. B. Küster
2. Univ.-Prof. Dr. D. Frischmann

Die Dissertation wurde am 13.12.2010 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 08.02.2011 angenommen.

# ABSTRACT

High-throughput genomic and proteomic techniques are widely used to increase our understanding of cellular processes. These technologies have generated large numbers of available data. Recent efforts are increasingly focusing on more integrated approaches to understand complex biological systems by reverse engineering gene regulatory networks. Many studies have demonstrated that large-scale networks are capable of predicting complex system behavior. Predicting complex biological systems, at system level, may help to understand how diseases like cancer develop and can lead us to better diagnosis and to detect cancer earlier.

While e.g. microarrays and mass spectrometers generate such data, there are crucial problems to be addressed before developing a predictive quantitative biology. The asymmetry of the datasets (more genes than samples) poses a problem for reverse engineering gene regulatory networks. My approach to this problem has been one of integration, bringing together a vast wealth of information from multiple datasets. Alleviating the asymmetry of the datasets considerably increases their use for systems biology. Furthermore, the ability to integrate expression experiments across species may help to identify pathways that are activated in a similar way in humans and other organisms.

Integrating data from multiple species is challenging. Automated methods are needed to extract maximum value from the mass of available data. Several meta-analysis approaches exist. Recent microarray based cross-species meta-analyses require prior affiliation of genes based on orthology information that often relies on sequence similarity. However, sequence similarity based orthology does not account for evolutionary phenomena such as sub- and neo-functionalization, thus not necessarily representing functional orthology in every case.

The computational time complexity of gene/sample affiliations is exponential in the number of genes or samples. Consequently, scoring all possible affiliations is feasible

for datasets of rather small size only. An iterative procedure is needed to approximate the global optimum in reasonable time. Prerequisite for scoring above gene affiliation solutions is to adjust different scales of the datasets. In order to gain experience by which scores (fold-changes, P-values, etc) as well as by which means of preprocessing such datasets can be best compared, I studied two single species microarray datasets. The first resembles sulfur reductase activity in *Arabidopsis Thaliana* that was recorded on the common two-channel fluorescence-tag cDNA glass platform. The second represents pooled RNAi screens on customized barcode tiling arrays.

I developed an algorithm merging microarray datasets on the basis of co-expression alone, without any requirement for orthology information. While such information can be easily incorporated to assist the process, the algorithm also performs well without being provided with any affiliations, purely driven by coherences among the data. Combining existing methods such as co-inertia analysis, back-transformation, Hungarian matching, and majority voting in an iterative non-greedy hill-climbing approach, the algorithm affiliates genes and experiments at the same time, maximizing the co-structure between the datasets.

The performance of the algorithm is demonstrated by merging datasets stemming from identical, closely related and more distantly related species. Moreover, the datasets represent different experimental contexts and had been produced on different platforms. The resulting cross-species dynamic Bayesian gene networks improve on the networks inferred from each dataset alone by yielding more significant network motifs, as well as more of the interactions already recorded in KEGG and other databases. Also, it is shown that the algorithm converges on the optimal number of nodes for network inference.

Being readily extendable to more than two datasets, it provides the opportunity to combine arbitrary numbers of e.g. microarray datasets. Furthermore, the application of the algorithm is not limited to microarray data. It could serve to integrate e.g. proteomic, transcriptomic and high-throughput methylation data recorded for the same samples.

# ZUSAMMENFASSUNG

Hochdurchsatzverfahren in Genomik und Proteomik tragen grundlegend zum besseren Verständnis zellulärer Prozesse bei. Sie erzeugen große Datenmengen. Um komplexe biologische Zusammenhänge besser zu verstehen, werden aus solchen Daten zunehmend durch sogenanntes Reverse Engineering regulatorische Netzwerke rekonstruiert. Viele Studien haben gezeigt, daß umfangreiche regulatorische Netzwerke geeignet sind, Verhalten biologischer Systeme zu prognostizieren. Solche Vorhersagen dienen letztendlich dem besseren Verständnis von Krankheitsabläufen. Sie könnten so einen Beitrag leisten zu sichereren Diagnosen oder der früheren Erkennung z. B. von Krebs.

Bis zu einer berechenbaren Biologie ist es allerdings noch ein weiter Weg. Der Verfügbarkeit geeigneter, z. B. mittels Microarrays oder Massenspektrometer erhobener Daten stehen grundlegende Probleme bei der Datenanalyse gegenüber. Die Asymmetrie der Datensätze (sehr viel mehr Gene als Experimente) steht einer zuverlässigen Schätzung regulatorischer Netze im Weg. Mein Ansatz zur Lösung dieses Problems zielt auf die Integration mehrerer Datensätze ab. Das Akkumulieren ähnlich gearteter Experimente (Beobachtungen) steigert die Signifikanz der Daten, die Robustheit der gewonnenen Netze und damit den Nutzen für systembiologische Fragestellungen. Weiterhin könnte die integrierte Analyse von Datensätzen über Artgrenzen hinweg aufdecken, welche Signalwege in Mensch und Modellorganismen gleichartig reagieren.

Eine solche Integration (Meta-Analyse) von Datensätzen erfordert komplexe automatisierte Verfahren, um größtmöglichen Nutzen aus den vorhandenen Daten zu ziehen . Mehrere solcher Methoden zur artübergreifenden Meta-Analyse von Mikroarray Datensätzen existieren bereits. Alle benötigen a priori eine Zuordnung der Gene zwischen den jeweiligen Spezies. Diese Zuordnung der orthologen Gene beruht

meist auf Sequenzhomologie. Letztere erfaßt allerdings Phänomene wie z. B. Sub- oder Neofunktionalisation nicht. Eine hierauf basierende Zuordnung repräsentiert somit nicht in jedem Fall Funktionsäquivalenz im Sinne der zu studierenden Netzwerke.

Eine Wertabschätzung aller möglichen Zuordnungen von Genen (und Proben) hat expontielle Laufzeit und wäre daher nur für sehr kleine Datensätze möglich. Ein iteratives Verfahren muß sich dem globalen Optimum in tragbarer Zeit nähern. Voraussetzung für die Wertabschätzung einer auf dem Weg vorkommenden Zuordnungslösung ist die Anpassung der unterschiedlichen Skalen der Datensätze. Welche Werte (Verhältnis, p-Wert, etc.) zum direkten Vergleich solcher Datensätze am besten geeignet sind und wie diese hierfür optimal aufbereitet werden können wurde anhand von zwei in meiner Gruppe erhobenen Einzeldatensätzen studiert. Der Schwefelmetabolismus von Arabidopsis thaliana war für den ersten Datensatz mit der verbreiteten fluoreszenz- und glasbasierten cDNA Plattform vermessen worden während der zweite Datensatz RNAi Analysen mit Pools von je fünf kuzen Haarnadelstruktur-RNS umfaßt und mithilfe sogenannter Barcode Tiling Arrays erhoben wurde.

Die von mir entwickelte Methode fusioniert Datensätze allein auf der Basis gemeinsamer Expressionsmuster, auch völlig ohne Zuhilfenahme weiterer Information. Vorabwissen über z. B. Orthologie kann zwar auf einfache Art miteinbezogen werden, der Algorithmus arbeitet aber auch bereits allein auf Basis von Koexpression erfolgreich. Er wurde durch Zusammenführen geeigneter bereits existierender Methoden als Module wie z. B. Koinertia-Analyse, Rücktransformation der Projektionskoordinaten, ungarischer Methode und Mehrheitswahl erarbeitet. Ausgehend von Datensätzen beliebiger Größen, Experiment-Reihung als auch zufälliger Anordnung der Gene in den Datentabellen wird über ein nicht-gieriges bergsteigendes Verfahren gleichzeitig sowohl die Zuordnung der Gene als auch die der Experimente hinsichtlich der Übereinstimmung (Ko-Struktur) der Datensätze optimiert.

Erfolgreiche Integration wird beispielhaft demonstriert für Datensätze aus identischen, nahe verwandten sowie aus nur entfernt verwandten Spezies. Hinsichtlich einer breiten Anwendbarkeit wurden diese Studien aus unterschiedlichen thematischen Zusammenhängen sowie beispielhaft für verschiedene Mikroarray Plattformen ausgewählt. Die resultierenden speziesübergreifenden sogenannten Dynamischen Bayes´schen Netze sind ihren aus den Einzeldatensätzen berechneten Pendants sowohl hinsichtlich des Vorkommens signifikanter Netzwerkmotive als auch beim Auffinden bereits in KEGG und anderen Datenbanken aufgeführter Interaktionen überlegen. Auch wird anhand von Beispielen gezeigt, daß das Verfahren auf einer für die Netzwerk-Inferenz optimalen Anzahl Knoten konvergiert.

Es ist weiterhin einfach auf die Zusammenführung von mehr als zwei Wertetabellen ausweitbar und eröffnet damit die Möglichkeit zur Integration beliebig vieler Datensätze. Darüber hinaus besteht keine Beschränkung auf Mikroarray Daten. In Fortführung meiner Arbeit ist selbst eine Anwendung zum integrativen Vergleich unterschiedlicher Regulationsebenen, z. B. mit aus gleichem Biomaterial gewonnenen Protein-, Transkript-, und Methylierungsdaten vorstellbar.

# ACKNOWLEDGEMENT

Rare is the dissertation that is completed without significant assistance from others and this one is certainly no exception. I owe a significant debt of gratitude to a large number of people whom I would like to recognize for their contributions to my writing, to my thinking, or to my personal development.

First, I would like to express my sincere appreciation to my supervisor and dissertation committee, Dr. Kurt Fellenberg, Professor Dr. Bernhard Küster and Professor Dr. Dmitrij Frishman.

Dr. Fellenberg has been my mentor for the past four years and it has been an absolute joy to work with him each day. His sense of humor, encouragement, creative thinking, and excitement about this area of research has served as an inspiration to me. I have benefited greatly from his wisdom and experience throughout my PhD time. He advised me on completing the writing of this dissertation as well as challenging the research that lies behind it.

Prof. Dr. Bernhard Küster has provided me opportunity to finalize this research towards my doctorate in his group. I am grateful to his invaluable ideas and support, critical reading of the thesis, the manuscript and also to create a friendly atmosphere at work. I would also like to thank Prof. Dr. Dmitrij Frishman for his insightful advice and discussions.

In addition to my committee, I would like to thank a number of other people. First, Dr. Jörg Hoheisel at the DKFZ for giving me the opportunity to initiate my research in his unique friendly research group in the beautiful Heidelberg. Yasser Riazalhosseini deserves a lot of credit for his broad and valuable biological discussions. He has always lent a willing hand or a willing ear, depending on which was more needed at the time. Additional thanks goes to the other group members at the DKFZ, Rafael Queiroz, Jorge Sozaried, Michael Böttcher, Christoph Schröder, David Jitao Zhang and

Christian Bender for helpful discussions and insights along the way and other wonderful folks in B070 who have been tremendous friends for the past four years.

I am grateful to Dr. Simone Lemeer who was extremely generous with her time in reading the thesis and providing me with helpful comments and all my current colleagues in Proteomics and Bioanalytics group at the University of Munich who helped me with this project. I am also grateful to Dr. Katharina Fellenberg for reading and correcting this thesis.

Outside Germany, I would like to acknowledge Prof. Alexander John Hartemink for providing me the source code of Banjo he developed at the Massachusetts Institute of Technology. He responded quickly and thoroughly to every email I sent, and that was a huge help whenever I was under time pressure. Special thanks also goes to Dr. Matthias E. Futschik from CBME, University of Algarve, Portugal for providing me with the source code of the 'AR' package he developed at the Institute for Theoretical Biology.

I have saved the most important acknowledgements for last. I owe everything that I am to my wife Hedieh, to my family - parents Ali and Maryam, sister Mojgan, and brother Moein – and to my new family – Parvaneh MohammadKhani, Naser and Ali Azadmehr. Their support has been unsurpassable and unflagging, their love unbounded and unconditional.

My deepest thanks goes especially to my wife Hedieh who patiently encouraged me throughout my studies, day in and day out. I especially owe her for encouragement; for unconditional support; for more happy moments than I can count; for making my life both easy and fulfilled; for believing in me when I have doubts; for making things possible that others think impossible; for motivation when I saw no light; for being there when I needed her.

Finally, I would like to express my special appreciation to my little princess Lilia for her patience and understanding while we are away from home; for keeping me optimistic and lively with her beautiful smiles.

# TABLE OF CONTENTS

**Abstract** ........ **iii**

**Zusammenfassung** ........ **v**

**Acknowledgement** ........ **ix**

**Table of Contents** ........ **xi**

**List of Figures** ........ **xv**

**List of Tables** ........ **xvii**

**Abbreviations** ........ **xviii**

**1 Introduction** ........ **1**

- 1.1 Systems biology ........ 1
- 1.2 Functional genomics ........ 3
- 1.3 Cross-species meta-analysis ........ 5
- 1.4 Machine learning in molecular biology ........ 7
- 1.5 Computational challenges ........ 9
- 1.6 Aims ........ 11
- 1.7 Dissertation overview ........ 12

**2 Single species microarray analysis** ........ **15**

- 2.1 Data collection ........ 15
- 2.2 Preprocessing ........ 16
- 2.3 Exploratory analysis ........ 19
  - 2.3.1 Cluster analysis ........ 19
  - 2.3.2 Dimension reduction ........ 25
- 2.4 Mechanistic analysis ........ 29
  - 2.4.1 Boolean networks ........ 30
  - 2.4.2 Relevance networks ........ 33
  - 2.4.3 Bayesian networks ........ 34

2.4.4 Dynamic Bayesian networks ........ 38

**3 Single species example 1: Sulfite reductase activity in *Arabidopsis Thaliana* ........ 41**

3.1 Introduction ........ 41

3.1.1 Statistical tests ........ 41

3.1.1 Importance of sulfite reductase ........ 44

3.2 Methods ........ 46

3.2.1 Analysis of transcript levels ........ 46

3.2.2 Preprocessing and statistical analysis ........ 47

3.2.3 Statistical computations ........ 47

3.2.4 FDR adjustment ........ 48

3.3 Results and discussion ........ 49

3.4 Conclusion ........ 56

**4 Single species example 2: Tiling array analysis ........ 57**

4.1 Introduction ........ 58

4.2 Methods ........ 59

4.2.1 Microarray design and hybridization ........ 59

4.2.2 Negative selection screen ........ 60

4.2.3 Data analysis ........ 60

4.2.4 Viability assay ........ 61

4.3 Results and discussion ........ 61

4.3.1 Identification of candidate genes ........ 62

4.3.2 Validation of candidate genes ........ 64

4.4 Conclusion ........ 65

**5 Cross-species microarray meta-analysis ........ 67**

5.1 Background ........ 68

5.1.1 Co-expression meta-analysis ........ 68

5.1.2 Expression meta-analysis ........ 70

5.1.3 Indirect expression meta-analysis ........ 71

5.1.4 Same array analysis ... 72
5.1.5 Combined multiple species ... 74
5.2 Methods ... 75
5.2.1 Co-inertia analysis ... 75
5.2.2 Hungarian algorithm matching ... 78
5.2.3 K-means clustering ... 80
5.2.4 Majority voting ... 80
5.2.5 Back-transformation ... 80
5.2.6 Dynamic Bayesian network ... 82
5.2.7 Evaluation ... 83
5.3 Algorithm and results ... 84
5.3.1 Algorithm ... 84
5.3.1 Single species application ... 88
5.3.2 Cross species application ... 90
5.3.3 Granularity evaluation ... 103
5.3.4 Superiority of the common network ... 108
5.4 Discussion ... 109
5.4.1 Meta analysis comparison ... 109
5.4.2 Multivariate analysis ... 110
5.4.3 Retaining information ... 111
5.4.4 Neo-functionalization ... 112
5.4.5 Comparison to KEGG ... 113
5.4.6 Granularity ... 114
5.4.7 Future work ... 115
**6 Conclusion ... 119**
**7 Bibliography ... 121**
**8 Appendices ... 133**
8.1 Banjo parameters ... 133

8.2 Supplementary Table S1 ........................................................................ 135
8.3 Publications ........................................................................................ 136
8.4 Curriculum vitae ................................................................................ 137
8.5 Erklärung ........................................................................................... 139

# LIST OF FIGURES

Figure 1. Schema of the reductionist and the integrative approach to biological research. ....3

Figure 2. Clustering with different distance measures. ....24

Figure 3. Projection of an expression matrix. ....26

Figure 4. Principal component analysis of an expression dataset. ....28

Figure 5. Different representations of Boolean networks. ....32

Figure 6. Sample Bayesian network with three nodes, two states. ....36

Figure 7. Volcano plot of fold changes versus differential expression. ....51

Figure 8. Molecular Identification of *sir* Mutants and Phenotype of *sir1-2*. ....52

Figure 9. Abundance and activity of *SiR* in the T-DNA Insertion Line *sir1-1*. ....53

Figure 10. Transcript Levels of Sulfur Metabolism-Related Genes in Leaves of *sir1-1* Plants ....54

Figure 11. Transcriptional levels by qRT-PCR. ....55

Figure 12. shRNA expression construct. ....59

Figure 13. Correspondence analysis projection. ....63

Figure 14. Cross-comparison of tiling probe performance. ....64

Figure 15. Overview of tilling array results. ....66

Figure 16. Overview for meta-analysis of microarray data. ....69

Figure 17. Affiliation of the connecting variables using Hungarian algorithm. ....78

Figure 18. Graphical representation of Majority voting. ....81

Figure 19. Overview. ....85

Figure 20. ‘*ds1*’ and ‘*ds2*’ projected by CIA. ....89

Figure 21. CIA plot of affiliated clusters. ....94

Figure 22. Common '*Sce*' and '*Spo*' regulatory network. ....97

Figure 23. Network inferred from the single ‘*Sce*’ dataset. ....99

Figure 24. Network inferred from the single ‘*Spo’* dataset. .......................................100
Figure 25. Interaction network in Human-Mouse dataset.........................................102
Figure 26. RV coefficients between ‘*Sce’* and ‘*Spo’* datasets. ....................................103
Figure 27. RV coefficients between ‘human’ and ‘mouse’ datasets. ..........................104
Figure 28. Silhouette values. .........................................................................................105
Figure 29. Verification of the termination condition...................................................106
Figure 30. ROC curves...................................................................................................107
Figure 31. Significance of motifs in the mouse/human consensus network. .............108

# LIST OF TABLES

Table 1. Significantly regulated transcripts in leaves of 7 week old *sir1-1*plants. .........50
Table 2. Affiliated clusters of *S. pombe* datasets .........................................................91
Table 3. Characterization of the affiliated gene clusters ..............................................95
Table 4. Cluster components (genes) of the affiliated gene clusters ............................96
Table 5. Transcription factor regulation references......................................................98
Table 6. Motif significance in the yeast common network. .......................................101
Table 7. Comparison of each single dataset to the common network in Yeast ..........109
Table 8. Comparison of methods developed for cross species expression analysis. ..110

# ABBREVIATIONS

| | |
|---|---|
| ANN | Artificial Neural Networks |
| ANOVA | Analysis of Variance |
| AO | Arachis Oil |
| AR(1) | First Order Autoregressive Model |
| ART | Adaptive Resonance Theory |
| BANJO | Bayesian Network Inference with Java Objects |
| BDe | Bayesian Dirichlet equivalence |
| BIC | Bayesian Information Criteria |
| BioGRID | Biological General Repository for Interaction datasets |
| BLAST | Basic Local Alignment Search Tool |
| BN | Bayesian Networks |
| CA | Correspondence Analysis |
| CCA | Canonical Correlation Analysis |
| CCA-EN | Canonical Correlation Analysis with Elastic Net |
| CIA | Co-inertia Analysis |
| DEG | Differential Expressed Genes |
| DNA | Desoxyribonucleic acid |
| DPI | Data Processing Inequality |
| ER | Estrogen Receptor |
| FDR | False Discovery rate |

| | |
|---|---|
| FN | False Negative |
| FP | False Positive |
| GNEA | Gene Network Enrichment Analysis |
| GO | Gene Ontology |
| GRN | Gene Regulatory Networks |
| GSEA | Gene Set Enrichment Analysis |
| GSVD | Generalized Singular Value Decomposition |
| intACT | Molecular Interaction Database |
| IPA | Ingenuity Pathway Analysis |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| MINT | Molecular Interactions Database |
| MRF | Markov Random Fields |
| mRNA | messenger RNA |
| nt | nucleotide |
| ODE | Ordinary Differential Equation |
| ORF | Open Reading Frame |
| PCA | Principal Component Analysis |
| PLS | partial Least Square |
| REGRN | Reverse Engineering Gene Regulatory Networks |
| RNA | Ribonucleic acid |
| RNAi | RNA interference |
| SAM | Significance Analysis of Microarrays |
| SCCA | Sparse Canonical Correlation Analysis |

| | |
|---|---|
| SCE | Saccharomyces cerevisiae |
| shRNA | short hairpin RNA |
| SOM | Self Organizing Map |
| SOTA | Self Organizing Tree Algorithm |
| SPO | Schizosaccharomyces pombe |
| TN | True Negative |
| TP | True Positive |

# 1 INTRODUCTION

The microarray technique, albeit barely older than a decade, is now both mature and widely available, accumulating an unprecedented amount of quantitative genome wide information [1]. Large scale microarray projects have revealed a comprehensive view of the transcriptome in different organisms at various stages of development, under diverse environmental conditions [2]. Efficient comparison of these data in related biological systems enables researchers to address complex biological questions [3]. While the reductionist approach to biology has proven considerably effective, recent efforts are increasingly focusing on more integrated approaches to understand complex biological systems. All of these developments point to the need for understanding the complex regulatory networks, responsible for controlling gene expression within cells.

The work presented in this dissertation addresses this need. An algorithm was developed for inferring cross-species common regulatory networks from gene expression data. The dissertation details the steps necessary for successful computational inference of cross-species genetic regulatory networks. In this introductory chapter, the necessary background is provided.

## 1.1 Systems biology

To fully understand the functioning of cellular processes, whole cells, organs, and even organisms, it is not enough to simply assign functions to individual genes, proteins, and other cellular components. We need to analyze the organization and control of the system in an integrated way by looking at the dynamic networks of genes and proteins, i.e. their interactions with each other. These interacting pathways are complex dynamic systems, and often behave in a nonlinear and adaptive way.

Nonlinearity means, for example, that doubling a stimulus does not necessarily double the response, and may even cause a qualitatively different response. Adaptive systems can modify themselves to response in a more appropriate way in the light of previous stimuli. The general goal of the theoretical systems biology is to develop computer models that predict the properties of the large, adaptive interconnected networks that are found in living cells.

Genomics, transcriptomics and proteomics have provided large datasets that can be used to describe the parts of a biological process at the gene and protein level. In systems biology descriptions of the processes under study are used to obtain a detailed description of the parts and their interactions, and then resemble them into an interconnected whole. In other words, descriptive models are applied to biological processes to identify rules about molecular or cellular associations or dependencies.

There are good reasons for the systems biology approach. First, biological systems tend to be so complex that it is difficult, without modeling, to know how they behave and to understand the actions of their control mechanisms. Second, such systems can have higher-order properties that are their main biological function, but are not apparent from properties of the separate components. Although very useful information can be obtained from the analysis of individual parts of a complex system, the ultimate aim is to understand how parts act together in real time and how the functioning of the systems is controlled (See Figure 1). This in turn will shed light on how each individual component contributes to the whole system. A system is more than the sum of its parts. It has a specific structure (the way the parts relate to each other) and dynamics (the ways in which it changes over time). A description of a fully functional system must take into account the spatial organization of elements, their interactions, and their response to external stimuli, including those processes that control and stabilize the system.

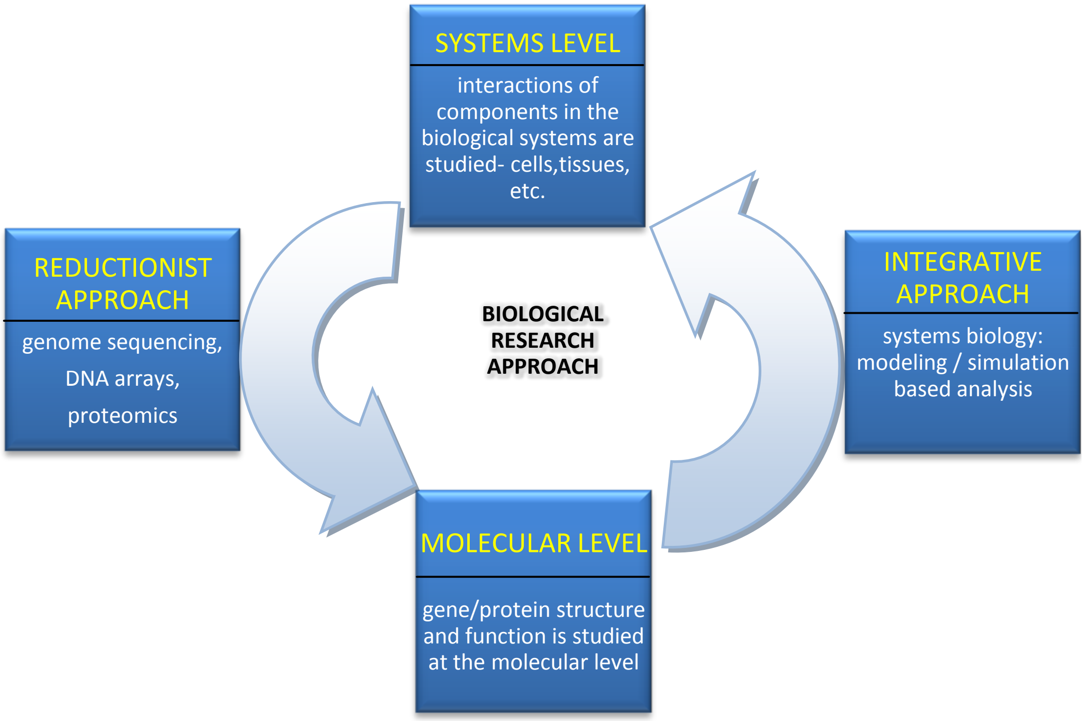

SYSTEMS LEVEL
interactions of components in the biological systems are studied- cells,tissues, etc.
REDUCTIONIST APPROACH
genome sequencing, DNA arrays, proteomics
BIOLOGICAL RESEARCH APPROACH
INTEGRATIVE APPROACH
systems biology: modeling / simulation based analysis
MOLECULAR LEVEL
gene/protein structure and function is studied at the molecular level

Figure 1. Schema of the reductionist and the integrative approach to biological research.

## 1.2 Functional genomics

Functional genomics aims to discover the biological function of particular genes and to uncover how their products work together in living organisms. What do all these genes, and by extension, what do all these proteins do? Elucidating the functions of the diverse collections of proteins within cells is the basis of functional genomics and will be a fundamental question of biology.

Several biological systems operate in similar ways in diverse species and many of the genes that play essential roles in these systems are conserved across these organisms [4]. Sequence similarity is one of the major sources of data for identifying the function of new genes, for instance, by BLAST [5]. Cross-species conservation has previously been used to delineate putative gene functions [6]. Regulatory elements have also been identified in small genomic regions by conservation analysis [7]. Comparative genomics, a term used to describe large-scale comparisons of complete genomes, has aided in the correct identification of genes and control regions in each of the sequences being compared [8] and sequence conservation analysis led to the identification of hundreds of new miRNAs [9]. Other efforts elucidate interaction data by means of network analysis, identifying core interaction modules as well as differences between regulatory programs in closely related species [10].

However, sequence conservation analyses and network comparisons can only tell part of the story. Sequence data do not change during the lifetime of a biological system. While interactions may change between conditions and over time, almost all current interaction data are static focusing on one time point and one condition [11]. Thus, using only these datasets it is often difficult to specify which genes participate in various biological processes. In addition, in some cases large changes in sequence and interactions may only have a minor effect on function, whereas in other cases, small changes in sequence between two genes may result in large changes in structure, leading to different function for the genes [12].

To address these issues researchers use microarrays to measure the dynamic, condition-specific response of complex biological systems. Examples include the cell cycle [13, 14], immune and other stress responses [15], circadian rhythm [16] and developmental processes [17]. These processes are shared between multiple, and in some cases distant species. By combining and comparing these experiments across species, essential 'core' genes can be identified. These genes are conserved both in sequence and transcription between multiple species and are thus expected to play crucial roles for the biological response of the system under study.

Although these sets of core elements play different roles within cells, one of the most challenging such roles is that of genetic regulation. Gene Regulatory Networks (GRN) control a cell at the genomic level, coordinate which genes are expressed and which remain unexpressed at any given time in the cell. Genes, RNAs and proteins interact with each other and form complex regulatory networks. RNAs are the direct products of genes. MicroRNAs (miRNAs) are small non-coding RNAs which are involved in post-transcriptional control of gene expression. MicroRNAs and small interfering RNAs (siRNA) can bind to specific other RNAs and either increase or decrease their activity, for example by preventing a messenger RNA from producing a protein. Proteins which function as transcription factors can positively or negatively influence the expression of another gene, and thus the production of other proteins. Some proteins act independently, others only become active in a complex. Gene regulatory networks describe these regulatory processes, and thus the molecular reaction of a cell to various stimuli.

Deciphering the complex structure of e.g. the transcriptional regulation of gene expression by means of computational methods is called Reverse Engineering Gene Regulatory Networks (REGRN). Advances in high-throughput biological techniques provide the basis for large scale analysis. REGRN is a quickly evolving field, with new developments and algorithms being published almost daily. It requires techniques particularly tailored to the task.

Analysis of regulatory processes within the cell will enhance our understanding of cellular dynamics. It will shed light on normal and abnormal, diseased, cellular events and may provide information on pathways that are malfunctioning in diseases such as cancer. These pathways can provide information on how the disease develops, and what processes are involved in progression. Ultimately, we can hope that this will provide us with new therapeutic approaches and targets for drug design.

## 1.3 Cross-species meta-analysis

Microarray technology measures the mRNA levels of tens of thousands of genes in tissue samples simultaneously in a high-throughput and cost effective manner. It has found widespread use in the fields of molecular genetics and functional genomics [18]. It has been applied in order to understand underlying biological mechanisms

[19], to discover novel subgroups of diseases [20, 21], to examine drug response [22, 23], to classify patients into disease groups [20], and to predict disease outcomes [24]. Despite their great promise, microarray-based studies may report findings that are not robust to data perturbations [25]. Common causes include improper analysis or validation, insufficient control of false positives, and inadequate reporting of methods [26]. The situation is provoked by the small sample size relative to large numbers of potential predictors; typically tens of thousands of probes are investigated in only tens or hundreds of biological samples.

Combining information from multiple existing studies is called ‘meta-analysis’. It can increase the reliability of results. The term meta-analysis is also widely used to describe the whole study process, not just the statistical techniques. Through meta-analysis, we can increase the statistical power to obtain more precise estimates of differential genes, and assess the overall estimate. Meta-analysis is relatively inexpensive, since it makes comprehensive use of already available data. The advantages of meta-analysis of gene expression microarray datasets have not gone unnoticed by researchers in various fields, however, most meta-analysis studies have been performed on cancer [27, 28].

Many studies combine data from multiple microarray experiments [13, 29] for one single species. Such meta-analyses have also been performed on data from multiple species. Cross-species meta-analysis can be used to utilize annotation and co-regulation information of one species to improve expression analyses of a less-studied species [14]. Further, it can serve to find common expression patterns in multiple species to reveal core gene functions [30] and to elucidate the evolution of gene expression and co-regulation [31].

Many of the successful applications of cross-species analysis to sequence and interaction data were performed using powerful computational techniques. Graph-based algorithms served to carry out whole genome alignments [32]. More recently, computational methods for cross-species comparisons of interaction networks were developed [33].

## 1.4 Machine learning in molecular biology

The term “Machine learning” refers to a set of topics dealing with the creation and evaluation of algorithms that facilitate pattern recognition, prediction and classification based on observed data. There are two main paradigms in the field of machine learning; s*upervised* and *unsupervised* learning. Both are being applied to biological questions.

In supervised learning, objects in a given collection are classified using a set of attributes, or features. In the context of gene expression, objects are often tissue samples and features are expression levels of individual genes (probes). The result of the classification process is a set of rules (classifier) that prescribe assignments of objects to classes based on values of features. In biological context, an example is to assign a tissue expression profile to disease group. The goal is to design a system that is able to accurately predict the class membership of new tissue expression profiles based on available features.

In contrast to supervised learning, in unsupervised approach no predefined class labels (categories) are available for the objects. The aim is to discover similarities/dissimilarities between objects. These are then used to define groups of objects, referred to as ‘clusters’. In supervised learning, data come with class labels, and we learn how to associate labeled data with classes, whereas in unsupervised learning, data are label free, and the learning procedure consists of both defining labels and associating objects to them.

The choice between supervised and unsupervised approach is tightly connected to the availability of prior knowledge. In supervised learning, the algorithm is tied to the specific areas of known data (training data). Therefore, careful selection of training data is vital. Furthermore training data may not represent special or unique categories that fit the classes. In contrast, in unsupervised approaches, no extensive prior knowledge is required.

Recently, different machine learning methods have been implemented to reconstruct gene regulatory networks from gene expression data. In biological context, clustering although not properly a network inference algorithm, is a method of choice to explore gene expression data. The rationale behind clustering is that coexpressed genes (i.e. genes in the same cluster) have a good probability of being functionally related [34]. However, this does not necessarily imply that there is a direct interaction between coexpressed genes, as genes separated by one or more mediators (indirect relationships) may be highly coexpressed. It is therefore important to understand what can be achieved by gene network inference algorithms, whose aim is to infer direct interactions among genes. Clustering can be used to reveal the modular structure of a network.

Inferring a gene network is defined as the process of identifying gene interactions from experimental data through computational analysis. There are two major classes of inference algorithms: those based on the ‘physical interaction' approach that aim at e.g. identifying interactions among transcription factors and their target genes and those based on ‘influence interaction' methods that try to relate the expression of a gene to the expression of the other genes in the cell. The interaction between two genes in a gene network does not necessarily imply a physical interaction, but can also refer to an indirect regulation via other proteins or metabolites that have not been measured directly. Influence interactions include physical interactions, for instance if the two interacting partners are transcription factor and its target, or two proteins in the same complex. Generally the definition of influence interactions in gene networks depends on the mathematical formalism used to model the network.

Gene networks have major practical utilities. First, to identify functional modules, that is, identifying the subset of genes that regulate each other with multiple indirect interactions, but have few interactions outside the subset. Second, to predict the behavior of the system following perturbations, where one needs to detect the genes that are directly interacting with a compound of interest. Third, to identify physical interactions by integrating the gene networks with additional information from other experimental data.

Large varieties of machine learning algorithms for gene regulatory network inference have been proposed in literature and are available as working tools. Section 2.4 presents several different approaches, discusses their strengths and weaknesses, and provides guidelines on which models are appropriate under what circumstances.

## 1.5 Computational challenges

When comparing microarray datasets across species, researchers face many of the same challenges that arise when comparing other high throughput datasets including the search issues related to the large datasets and the need to handle homology assignments between species. In addition, a good experimental design that takes into account the fact that experiments are to be compared across species is crucial for the success of such studies.

However, microarrays also raise several new challenges. Microarray data are often noisy. The agreement between experiments measuring similar processes in different labs, even within the same species is sometimes very small [35]. Another challenge results from the differences in conditions and dynamics. Unlike sequence or interaction data which are often denoted by a small number of letters (DNA) or binary edges (interactions), microarrays measure continuous values. Dynamic environments and different scales make it difficult to compare results across diverged species. For instance, while there are several similarities between human and yeast cell cycle, the duration is different (90 min for yeast vs. 24 h in human cells). Similarly, the wide range of conditions makes it difficult to use the data for direct comparisons across species. Another problem arises when comparing results from different expression analysis methods. For instance, the scoring methods in Spellman and coworkers [13] and Lelandais and coworkers [36] for cell cycle genes in budding and fission yeast are different, making direct comparison problematic. Any combination of the above may bias the analysis.

A typical feature of microarray datasets is high dimensionality of data. That is, very high number of genes simultaneously measured under few number of samples (genes >> samples). This is a problem for many machine learning techniques. For instance, in supervised learning, when the number of genes is too high, reliable estimation of the classifier’s internal parameters with a limited number of samples becomes problematic. In such situations, dimensionality reduction may be useful. One major approach is to obtain a reduced number of new features by combining the existing ones, e.g., by computing a linear combination. Principal Component Analysis (PCA) is one particular method, in which new features (principal directions) are identified and may be used instead of the original features.

In the context of network inference, high dimensionality of data particularly raises a problem when inferring the structure of dynamic graphs. Several methods have been proposed to address this problem [37-40]. In order to cope with the dimensionality problem accounted for in one GRN reverse engineering step, it seems preferable for any method to first combine the data instead of combining the resulting networks later on. Several approaches can be applied to this end [40-42]. However, all of these methods take as input the affiliation of genes between the datasets. When combining data stemming from different species, sequence homology can be used to affiliate orthologs. However, due to the ambiguity of orthology relations, mapping across species is challenging. Lineage-specific gene duplications can give rise to a different number of paralogs in one species compared to another species. One cannot tell which paralog (or in-paralog) retains the function of the ancestral gene or has been co-opted into a new function.

## 1.6 Aims

The goal of this work was to develop an integrative approach to inferring cross-species gene regulatory networks. The specific aims of my approach can be summarized as:

**Reverse engineering microarray datasets.** High-throughput genomic and proteomic techniques are widely used to increase our understanding of cellular processes. These technologies have generated large numbers of available data. Progress has been made on integrated approaches to understand complex biological systems by reverse engineering gene regulatory networks. My aim was to contribute to the methods for reconstructing regulatory networks based on these data.

**Asymmetry of datasets.** Over the past ten years, large numbers of gene expression studies have been carried out. The asymmetry of microarray datasets (genes >> samples) poses a problem for reverse engineering gene regulatory networks. My goal was to alleviate the asymmetry of the data by combining datasets.

**Cross-species comparison.** Thematically related datasets to combine are often few when searching for a single species, only. Furthermore, combining data from multiple species can lead to important findings which cannot be achieved by focusing on a single species. Combining expression experiments from multiple species may help to identify genes that are not only conserved in sequence, but also operate in a similar way in the different species. Similarly, it may help to identify pathways that are activated in the same manner in humans and other organisms. My aim was to reconstruct common gene regulatory networks by combining datasets across species.

**Gene/sample affiliations.** Recent microarray based cross-species meta-analyses require prior affiliation of genes based on orthology information that often relies on sequence similarity. The need for orthology assignments is a major drawback of the existing methods for two main reasons. First, sequence similarity based orthology does not account for evolutionary phenomena such as sub- and neo-functionalization, thus not necessarily representing functional orthology in every case. Second, when comparing cross-species datasets using one-to-one orthology information, a large amount of the probes will have to be masked, severely limiting the number of genes that can be measured in the study. My aim was to develop a systematic way for merging microarray datasets without any requirement for orthology information.

**Complexity.** The computational time complexity of gene/sample affiliations is exponential in the number of genes or samples. For instance, given an expression matrix of only 20 genes, 2.43E+18 iterations would be needed to score all possible affiliations. Therefore, an iterative hill-climbing procedure is needed to solve the problem in a reasonable time. However, hill-climbing approaches may obtain the local optimal solution and it is not guaranteed that it will obtain the global optimum. In order to overcome this problem, my aim was to use a non-greedy approach to affiliate genes and samples between datasets.

**Adjusting different scales.** Prerequisite for scoring above gene affiliation solutions (to find the optimum affiliation) is to adjust different scales of the datasets. Therefore, my aim was to gain experience by which scores (signal intensities, fold-changes, P-values, etc) as well as by which means of preprocessing (normalization, filtering, scale adjustments) such datasets can be best compared.

## 1.7 Dissertation overview

Chapter 2 presents perquisites for the analysis of single species microarray experiments followed by related biological concepts, techniques, computational models and methods to extract meaningful biological results. In the context of unsupervised analysis, chapter 2 reviews existing alternatives for distance measures in clustering approaches and concludes that the measures based on “variance standardization” perform better when it comes to variance analysis. This section also shows how dimensions of expression datasets can be reduced to minimize the loss of information regarding the relationships between genes and samples.

Chapter 3 discusses various methods for differential gene expression detection of single microarray datasets. This chapter presents the application of empirical Bayesian methods to a set of microarray experiments to detect response of sulfur metabolism at transcriptional level in *Arabidopsis thaliana.* The results show the performance of linear models by experimentally validating a set of genes that are identified as significantly differentially expressed genes. The results support that the downregulation of *SiR* causes severe adaptive reactions of primary and secondary metabolism and is essential for growth and development in *Arabidopsis thaliana*.

Chapter 4 describes high-level analysis of tiling arrays to decode pooled RNAi screens. This chapter demonstrates how barcode tiling arrays can be used to predict anti-proliferative effects of individual shRNAs from pooled negative selection screens.

Chapter 5 begins with a brief literature review on four relevant approaches for meta-analysis of microarray data. This chapter particularly emphasizes the application of Co-inertia analysis (CIA) to cross-platform comparisons of gene expression data. Incorporating CIA with existing methods represents an iterative procedure to match genes and samples, at the same time. This chapter also demonstrates different ways to assess the procedure by means of reverse engineering approaches, followed by the application of the algorithm on several microarray datasets. In this chapter the performance of the algorithm on two independent but closely related experiments is demonstrated. These datasets served as verification datasets since, considerable prior knowledge and direct evidence on the reliable genes and samples in both datasets were known. In order to show that the algorithm is not limited to a certain level of similarity, its application to distantly related datasets is demonstrated.

Chapter 6 closes the dissertation with a summary of the lessons learned. The chapter also includes comprehensive itemization of ways in which this work could be extended in the future.

# 2 SINGLE SPECIES MICROARRAY ANALYSIS

Many techniques have become available that produce vast amounts of quantitative biological data. As microarrays have become more commonplace, the challenges associated with collecting, managing, and analyzing the data from each experiment have increased substantially. Falling prices for commercial platforms, robust laboratory protocols, and more complex experimental designs all lead to the generation of large amounts of data. A few years ago, microarray studies typically included 10 hybridizations; now, studies tend to have hundreds or more such assays. Interpretation of the large datasets produced by microarray experiments can be time consuming. Moreover, different methods can yield different conclusions. These experiments aim at extracting biological or functional meaning, either by identifying critical genes that might be responsible for a biological effect or by finding patterns that point to an underlying biological process. This chapter discusses issues associated with analyzing such data to extract meaningful biological results.

## 2.1 Data collection

DNA microarrays use gene-specific probes that represent thousands of individual genes. The probes are arrayed on an inert substrate, each confined to a separate surface area to discriminate their hybridization signals. RNA is extracted from tissues of interest, labeled with a detectable marker, typically a fluorescent dye, and allowed to hybridize to the array. Messenger RNA (mRNA) molecules hybridize to their complementary probes on the array. Images are rendered with the use of laser scanning. The relative fluorescence intensity of each gene-specific probe is used as a measure of the expression level of that gene. A more intense signal is caused by a higher degree of hybridization which in turn, is caused by higher expression levels.

There are two basic approaches to generate microarray data. In single-channel arrays, such as the GeneChip (Affymetrix), only one sample is hybridized to one array. After hybridization sample is removed by washing. The level of expression of each gene is summarized into one value. For two-channel (two-color) arrays, two samples of RNA, each labeled with a different dye, are simultaneously hybridized to the array. The sample of interest (for example, a sample of cancer tissue), is labeled with one dye, and a reference sample (normal tissue) is labeled with a different dye. Two samples are mixed in an approximate ratio of 1:1 on the basis of dye incorporation. This compares paired samples and reports expression as the ratio of RNA in a query sample to that in a control sample.

Expression data are typically represented as “expression matrix” in which each row represents a particular gene and each column represents a specific biological sample. Each row is a “gene expression vector” where the individual entries are its expression levels in different samples. Each column is a “sample expression vector” that records the expression of all genes in that sample. Any data which can be placed into this “genes by samples” matrix format can be analyzed using the same techniques.

## 2.2 Preprocessing

The hypothesis underlying microarray analysis is that the measured intensities for each gene are proportional to its transcriptional level in the cell. Relevant expression patterns are typically identified by comparing measured expression levels across different samples. But before the levels can be compared appropriately, a number of transformations must be carried out on the data to eliminate systematic errors, low-quality measurements and to adjust the measured intensities to facilitate comparisons. Some of these transformations can be highlighted as background correction, ratio transformation and normalization.

**Background correction:** Image analysis software returns foreground and background intensities for each spot. The foreground is an overall measure of the intensity of the spot while the background is a measure of the ambient signal. Background fluorescence can arise from many sources. For instance from non-specific binding of labeled sample to the array surface, processing effects such as deposits left after the wash stage or optical noise from the scanner. Removal of ambient, non-specific signal from the total intensity is known as 'background correction'.

Most image analysis programs return 'local' background intensities, obtained from the mean or median of the pixel intensity values surrounding each spot. Local background is an estimate of the local non-specific signal, so subtracting it from the foreground intensity gives an estimator of the true signal. This approach produces negative intensities whenever the background intensity is larger than the foreground intensity. Data for such spots are leading to missing log-ratios, sometimes for a substantial proportion of probes on an array.

**Ratio transformation:** Most microarray experiments investigate relationships between related biological samples based on patterns of expression. The simplest approach looks for genes that are differentially expressed. Let us assume an array of $N$ distinct elements, and compare a query and a reference sample as $R$ and $G$, respectively, then the ratio ($T$) for the $i$th gene (where $I$ is an index of all the arrayed genes from 1 to $N$) can be written as

$$T_i = \frac{\mathrm{R}_i}{\mathrm{G}_i} \qquad \text{EQ 1}$$

The measures $R_i$ and $G_i$ can be made on either a single array or on two separate arrays.

Although ratios provide a natural measure of expression changes, they have the disadvantage of treating up- and down regulated genes differently. Genes up regulated by a factor of two have an expression ratio of two, whereas those down regulated by the same factor have an expression ratio of (–0.5). The most widely used transformation of the ratio is the logarithm, which has the advantage of treating up- and down regulated genes in a similar fashion.

**Normalization:** Normalization methods are applied to expression data to aid comparison between individual hybridizations to compensate differences in labeling, hybridization, and detection efficiencies between fluorescent dyes. There are different normalization methods. The approaches used depend on platform and the assumptions made regarding the biases in the data [43-47].

Normalization adjusts the fluorescence intensities on each array and can change the relative intensity difference observed between samples – the fold change. Normalization methods are mainly based on knowledge of the particular experimental methodology and potential sources of systematic error (such as using different quantities of different samples). Their application often involves taking averages, yet can achieve good results. Many of the techniques correct the mean but do not pay much attention to the variance. It has often been observed in general experimental work that the variance tends to increase as the signal increases. Several normalization methods have been proposed that correct this. Systematic errors can in the best case be completely removed, whereas it is only possible to approximate the form of random noise and not remove it entirely.

While normalization method is always necessary to compensate for systematic errors [43, 46, 48] that are introduced during the experimental process, over-normalizing the data can deform the final outcome of the analysis. Similarly, the way in which the data are filtered can generate different results [49].

**Filtering:** There are different filtering approaches applied to the data using a variety of methods to eliminate a) genes that have minimal variance across the collection of samples, b) those that fail to provide data in a majority of the experiments and c) whose signal lack reproducibility. The value of these filtering methods is that they reduce the noise within the dataset by eliminating those genes that are not likely to contribute to any high level analysis. The manner in which the data are filtered can produce very different results. Therefore, appropriate means of dealing with “high dimensional” datasets should be considered.

## 2.3 Exploratory analysis

### 2.3.1 Cluster analysis

Cluster analysis is usually the first step in any genomics experiment as it takes an unbiased approach to look for new groups in the data. For instance, one might examine a group of cancer patients to see if their expression profiles allow them to be placed into distinct groups without using any prior knowledge of their disease progression. Clustering is an unsupervised method that does not use any predefined class labels for the samples to explore expression patterns. They group samples based on some measure of similarity. In other words, unsupervised learning is used to unveil natural groupings in the data. After finding new groups based on expression profiles, the challenge then becomes to find a link to clinical or biological factors that can explain the difference.

There are many approaches that have been applied to unsupervised analysis, including self-organizing maps (SOM) [50-52], self-organizing trees (SOTA) [53], relevance networks [54], force-directed layouts [55], principal component analysis [56], and others. Each of these algorithms uses some attribute of the data and a set of rules for determining relationships between group of genes (or samples) that are similar in expression patterns. All of these algorithms are able to separate data into some clusters, but the evaluation of the results requires expert input and analysis.

Two of the most widely used approaches are hierarchical clustering [34, 57, 58] and k-means clustering [59]. Hierarchical clustering creates a hierarchical, tree-like structure of the data. This can be constructed using either a bottom-up or a top-down approach. In a bottom-up approach, each data point is initially taken as a cluster. Subsequently, the clusters are iteratively merged based on their similarity. In contrast, the top-down approach starts with a unique cluster containing all data points. This initial unique cluster is iteratively divided into smaller clusters until each cluster contains a single data point.

A potential problem with many hierarchical clustering methods is, as the size of each cluster grows, the expression profile of the cluster representative may no longer represent any of the genes within the cluster. This poses a problem when calculating distances between clusters. Consequently, as clustering progresses, the actual expression patterns of the genes become less relevant. Furthermore, if a bad choice is made at an early stage it cannot be corrected. An alternative, which can avoid these artifacts, is to use iterative refinement approaches, such as *k*-means clustering, to partition either genes or samples into groups having similar expression patterns.

#### 2.3.1.1 K-means clustering

In k-means clustering, objects are partitioned into a fixed number (*k*) of clusters such that objects within each cluster are more similar to one another than those assigned to different clusters. K-means clustering can be computationally intensive:

1. All initial objects are randomly assigned to one of *k* clusters (where *k* is specified as input).
2. An average expression vector (centroid) is calculated for each cluster to compute the distances between clusters.
3. Using an iterative method, objects are moved between clusters and intra and inter-cluster distances are measured with each move. Objects are allowed to remain in the new cluster only if they are closer to it than to their previous cluster.
4. Following each move, the expression vectors for each cluster are recalculated.
5. The shuffling proceeds until moving any more objects would make the clusters more variable, increasing intra-cluster distances and decreasing inter-cluster dissimilarity.

It is possible to adapt the k-means clustering method to vary the number of clusters automatically. Once the set of *k* clusters has been identified, data can be examined to identify any data points that are relatively distant from the centroid (outlier). In such cases, an extra cluster centroid can be added at that data point. An alternative would be to ignore outliers, but in the case of expression data analysis such outliers may be of great interest and are best left in the analysis. A different initial location of the cluster centroids can result in a different final partition, so one can use several different starting points, generating several partitions.

One run of the k-means clustering produces a single partition of the data into *k* clusters. Although there are ways of changing the value of *K* during a run, in general if there is advance knowledge regarding the number of clusters that should be represented in the data, k-means clustering is superior to hierarchical methods [59, 60].

**Distance measure.** All clustering techniques identify clusters according to a distance between each pair of data points (genes or samples) and therefore need a definition of this distance called distance measure. The distance must be defined as a single number, and therefore each gene or sample in the experiment requires a set of quantitative parameters. There are several alternative distance measures. Some of the widely used distance measures are Euclidean, Pearson correlation coefficient and Chi-square.

Having *N* different genes measured for each of the *M* samples so that *A*th sample has a value $X_{i,A}$ for the *i*th gene. The Euclidean distance between samples A and B is defined as

$$d_{AB} = \sqrt{\sum_{i=1}^{N} (X_{i,A} - X_{i,B})^2} \qquad \text{EQ 2}$$

The Euclidean distance measure is commonly used because it is easy to evaluate. A key feature of the Euclidean distance is that all parameters are treated in an identical way without any modification. This is not necessarily appropriate in the case of expression measurements. To give an example, why should doubling the expression level of a kinase and of a cytochrome contribute equivalently to the distance? In Euclidean measure, quantitative changes in expression ratios are treated equally for all genes. Genes whose transcription is correlated will not necessarily produce such equivalent responses, and it may be useful to give greater emphasis to the observation that the two genes have correlated expression changes. In other words, two such genes' expression levels may not increase to the same degree, but their correlation is still an important and useful observation.

The Pearson correlation coefficient is a commonly used method to measure a correlation between two series of numbers. The definition of correlation coefficient between two samples A and B is given by

$$r_{AB} = \frac{1}{(N-1)} \sum_{i=1}^{N} \left(\frac{X_{i,A} - \bar{X}_A}{S_A}\right)\left(\frac{X_{i,B} - \bar{X}_B}{S_B}\right) \qquad \text{EQ 3}$$

Where $\overline{X_A}$ is the average of the values $X_A$, and $S_A$ is the standard deviation of these values; similarly for $\overline{X_B}$. The values range from -1 for a completely negative correlation between two sets, through 0 for no correlation to +1 for perfect correlation. The Pearson correlation coefficient measures distance in terms of the shape of the patterns, and not in absolute values. Therefore it identifies two genes or proteins as similar if their expression pattern across samples is similar.

Figure 2 shows the effect of using different distance measures to interpret an experiment comparing gene expression patterns of four genes measured on four consecutive days. In this case all the gene expression levels are used to define all pair wise gene expression distances. In the dendogram (B) genes labeled with 'up' and 'down' are clustered together since their absolute values across all samples are similar. In contrast, dendogram (C) clusters genes with similar trends; 'up' and 'extreme_up' as opposed to 'change' and 'down'.

When discussing the Euclidean distance measure a problem was mentioned in that the different genes are treated equally even though some may show much greater absolute variations than others. The contribution of these genes to Euclidean calculation is huge. In other words, the larger expression values have larger intersample differences, so they will dominate in the calculation of Euclidean distances. One way to overcome this problem is to balance out the contributions using weighted Euclidean distance. This is called ‘Chi-square’ distance and the definition between two samples A an B is given by

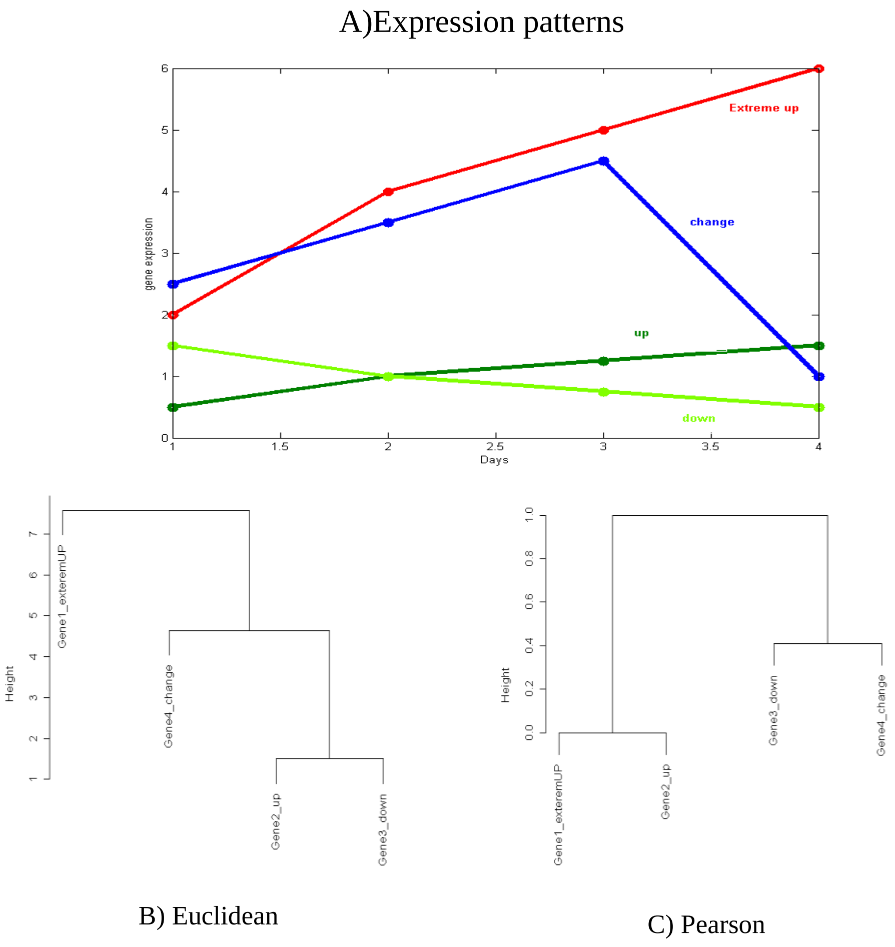
$$C_{AB} = \sqrt{\frac{\sum_{i=1}^{N}(X_{i,A} - X_{i,B})^2}{X_{.N}}} \qquad \text{EQ 4}$$

where $X_{.N}$ is the average row profile.

In chi-square distance, the division of each squared values by the row profiles (relative frequency distributions of the genes for each sample) is denoted as “variance standardization" and compensates for the larger variance in high frequencies and the smaller variance in low frequencies. Without such standardization, the differences between larger proportions would tend to be large and thus dominate the distance calculation, while the differences between the smaller proportions would tend to be filled up. The advantage of using Chi-square distance is that, it satisfies the principle of distributional equivalence. If two profiles are identical, they can be combined into one without affecting the result for the other profiles. In the case of large microarray datasets it may often be expedient to combine variables (e.g. genes) having almost identical profiles, thus making it easier to interpret the results. This stability in distances is a unique property of the chi-square measure.

#### 2.3.1.2 Quality assessment

Many heuristic approaches compare the quality of the clustering results for different numbers of clusters [61-65]. Moreover, many solutions to systematically evaluate the quality of the clusters have been reported [66-68]. The estimation of the number of clusters in a dataset is a major problem in unsupervised learning. The applications of several validation techniques such as Silhouette values [69], Dunn’s based index [70] and Davies-Bouldin index [71] have been previously studied [63, 66, 67]. It has been shown that the Silhouette method [69] is suitable for estimating the best partition.


A)Expression patterns

B) Euclidean

C) Pearson

Figure 2. Clustering with different distance measures.

Four samples are analyzed from a time series experiment in which microarray were used to compare gene expression patterns in four consecutive days. The genes are labeled from 'down' to 'extreme high' and marked with colors. Panel A) Shows the expression patterns of all genes measured in four days. In (B) Euclidean distance is used, whereas in (C) Pearson correlation coefficient.

It has been successfully used in combination with other validation techniques (Dunn's and Davies-Bouldin indices) for predicting different optimal clustering partitions [66]. The silhouette value is a measure of how similar an object (e.g. gene or sample) is to other objects in its own cluster compared to objects in other clusters. It is defined as

$$S(i) = \frac{\min\{b(i,k) - a(i)\}}{\max\{a(i), \min(b(i,k))\}} \qquad \text{EQ 5}$$

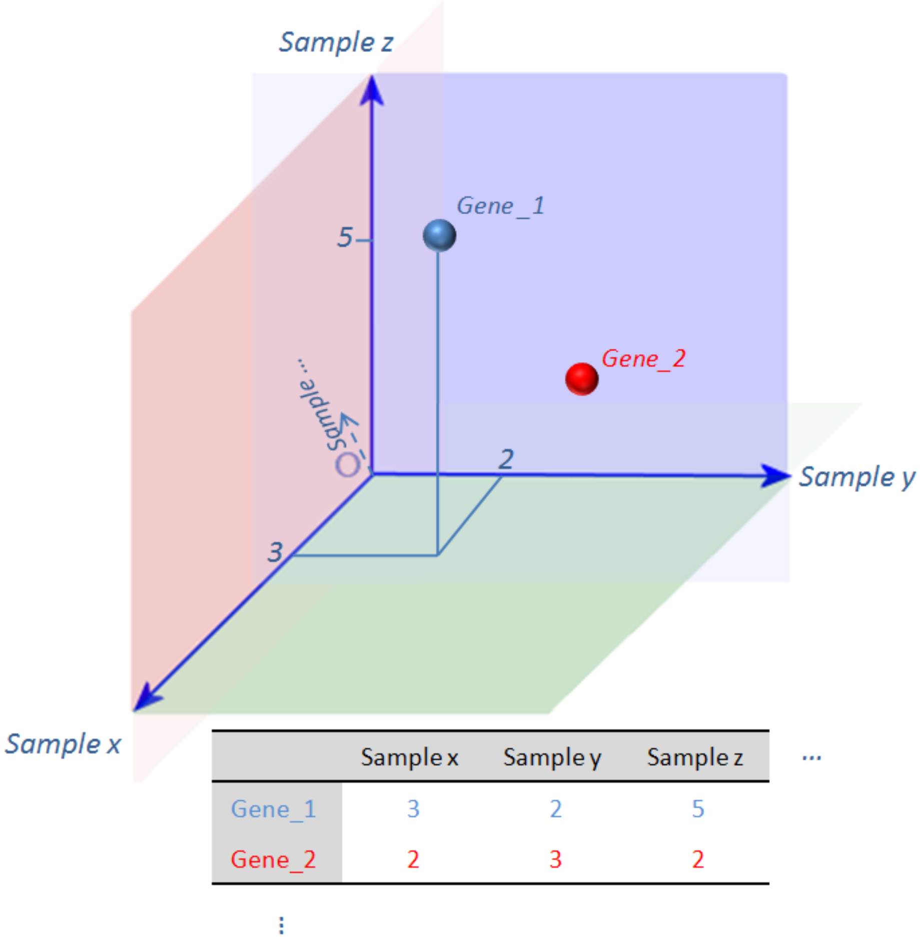
where *a(i)* is the average distance from the *i*th object to the other objects in its cluster, and *b(i)* is the average distance from the *i*th object to objects in another cluster *K*. Silhouette values ranges from -1 to 1. Average silhouette width of all objects can be used to score a clustering of *n* clusters. A score with *S(i)* close to 1 indicate a good clustering result, whereas -1 shows the clustering to be unsuccessful. Comparing Silhouette scores of repeatedly performed clustering over various *n* can reveal the optimal number of clusters that can be discriminated on the datasets under study.

## 2.3.2 Dimension reduction

Several measurement techniques such as DNA microarrays and mass spectrometers can measure levels of thousands of mRNAs or proteins in hundreds of samples. Such high-dimensionality makes visualization of features (genes, proteins or samples) difficult and limits simple exploration of the data. Dimension reduction techniques can be used to reduce the dimensionality, making it possible to project features into low dimensional subspaces. For example, given an original dataset, one can represent genes as numerical vectors with the number of elements of each vector being the number of samples. Therefore those vectors could be plotted as points in sample dimensional space, if only the number of dimensions were small enough to visualize. Dimensionality reduction techniques can be used to project these points into a two or three dimensional subspace so that they can be plotted.

Figure 3 shows an example of the above mentioned genes as vectors in sample space. Vice versa, the columns of the data table (samples) can be represented in gene space. Such a projection plot is an explorative way to visualize the underlying structure of a data set and can be used to visually assess classes or groups of objects. It also allows visual judgment of the number of clusters.

There are two main categories of approaches for dimensionality reduction. The first one is to obtain a reduced number of new features by combining the existing ones, e.g., by computing a linear combination. Principal component analysis (PCA) is one particular method. The second type of dimensionality reduction involves feature selection that seeks subsets of the original features that are adequately predictive.



| | Sample x | Sample y | Sample z |
|---|---|---|---|
| Gene_1 | 3 | 2 | 5 |
| Gene_2 | 2 | 3 | 2 |

Figure 3. Projection of an expression matrix.
The three columns of an expression matrix of two genes three samples are represented in 3-dimensional gene space (for simplicity). Typical microarrays dataset consist of a few hundreds of samples and several thousands of genes.

### 2.3.2.1 Principal Component Analysis

Principal component analysis (PCA) is a mathematical algorithm that reduces the dimensionality of the data while retaining most of the variation in the data set. It accomplishes this reduction by identifying directions, called principal components, along which the variation in the data is maximal. By using a few components, each sample can be represented by relatively few numbers instead of by values for thousands of variables. Samples can then be plotted, making it possible to visually assess similarities and differences between samples and determine whether samples can be grouped.

In order to explain PCA with simple geometrical interpretations, let us assume an example of microarrays measured the expression levels of 27,648 genes in 105 breast tumor samples [72]. The gene expression data set is available through the Gene Expression Omnibus database (accession no. GSE5325). This dataset is used to demonstrate how PCA can represent samples with a smaller number of genes, visualize samples and genes, and detect dominant patterns of gene expression. In order to simplify plotting the breast cancer samples according to their expression profiles, let us take only two genes '*GATA3'* and '*XBP1'* (Figure 4a). Breast cancer samples are classified as being either positive or negative for the estrogen receptor. These genes have been selected since their expression is known to correlate with estrogen receptor status [72].

PCA identifies new features, the principal components, which are linear combinations of the original features. The two principal components for the two-dimensional gene expression profiles are shown in (Figure 4b). The first principal component is the direction along which the samples show the largest variation. The second principal component is the direction uncorrelated to the first component along which the samples show the largest variation. If data are standardized such that each gene is centered to zero average expression level, the principal components are normalized eigenvectors (See EQ 6) of the covariance matrix of the genes and ordered according to how much of the variation present in the data they contain. Each component can then be interpreted as the direction, uncorrelated to previous components, which maximizes the variance of the samples when projected onto the component. Here, genes were centered in all examples before PCA was applied to the data (Figure 4b).

The dimensionality of two-dimensional expression profiles can be reduced to a single dimension by projecting each sample onto the first principal component (Figure 4c). This one-dimensional representation of the data retains the separation of the samples according to estrogen receptor status. The projection of the data onto a principal component can be viewed as a gene-like pattern of expression across samples, and the normalized pattern is sometimes called an eigengene. So for each sample-like component, PCA reveals a corresponding gene-like pattern containing the same variation in the data as the component. Moreover, provided that data are standardized so that samples have zero average expression, the eigengenes are eigenvectors to the covariance matrix of the samples.

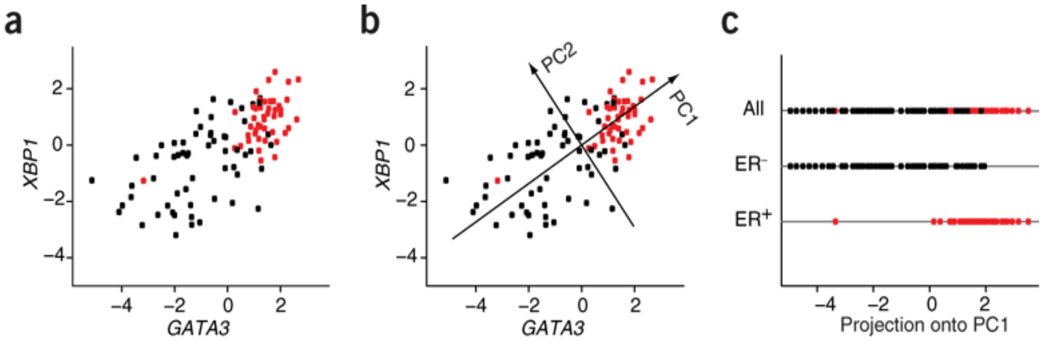


Figure 4. Principal component analysis of an expression dataset.

**a)** Each dot shows a sample plotted against its expression levels for two genes. (Samples are colored according to estrogen receptor (ER) status: ER+, red; ER-, black). **(b)** PCA identifies the two principal components (PC1 and PC2) along which the data have the largest divergence. **(c)** Samples plotted in one dimension using their projections onto the first principal component (PC1) for ER+, ER-, and all samples separately.

The calculation of PCA is the diagonalization of a data matrix, $X$. If an experiment consists of $N$ genes or proteins and $M$ different samples, $X$ will be an $N \times M$ matrix. A row of this matrix corresponds to a gene and a column represents a sample. A matrix element $X_{i,A}$ is the expression of gene *i* under condition *A*. In the calculation, matrix *X* will be re-expressed as a product of three new matrices:

$$X = U \varepsilon \ V^{\mathrm{T}} \qquad \text{EQ 6}$$

The outcome of the calculation is two sets of *M* principal components, one set for the genes (referred as eigengenes) and one set for samples (eigensamples). The expression level of every gene in an eigensample is given in matrix *U*, and expression level of every eigengene in each sample is given in matrix $V^T$. The first is a matrix, *U* that is also *N*×*M*; the second, *ε*, is a square matrix of dimensions *M* × *M* (assuming that M<N); and the third, $V^T$ is of dimensions *M*×*M*. The *A*th eigengene is only expressed in the *A*th eigensample, with the eigenexpression level $\varepsilon_A$. Each eigenvector defines a principal component. The expression data can then be plotted for each gene/protein *i* along the axis defined by *p*th principal component.

### 2.3.2.2 Correspondence Analysis

Correspondence Analysis (CA) additionally captures the correspondence between columns and rows (samples and genes). It is conceptually similar to PCA, but scales the data so that rows and columns are treated equivalently, thus visualizing genes and samples at the same time (Figure 20). Whilst, as with PCA, similarity among genes as well as similarity among samples is depicted as proximity, a gene that is particularly up-regulated under a certain condition will be located in the direction of this condition. The farther away from the centroid in this direction (towards the outer margin of the plot) it is displayed, the stronger the association [73, 74].

## 2.4 Mechanistic analysis

Biology has undergone a transition from focusing on single components of cells, such as single gene, RNAs and proteins, to the analysis of relationships and interactions between these parts. The traditional approach of molecular biology breaks up a system into its various parts, analyzes each part in turn, to gain knowledge about the system. In contrast, systems biology aims at understanding and modeling the entire system quantitatively, proposing that the system is more than the sum of its parts and can only be understood as a whole.

Gene regulatory networks explain which genes are to which extent transcribed to RNA, which in turn functions as a template for protein synthesis. High throughput experimental techniques to measure RNA and protein concentrations enable new approaches to the analysis of such networks. The analysis of these data requires techniques particularly designed for the task. Starting with models which allow for qualitative statements only, in recent years there are methods to describe the dynamic response of a system in more detail.

These models are often represented as graphs, with nodes corresponding to genes, and edges indicating interactions between genes. Expression measurements may consist of time-series gene expression data (i.e. gene expression changing dynamically with time) or measurements taken at steady-state in different conditions (i.e. gene expression levels in homeostasis). Some inference algorithms can work on both kinds of data, whereas others have been specifically designed to analyze one or the other. Depending on the inference algorithm, the resulting gene network can be either an undirected graph, that is, the direction of the interaction is not specified, or a directed graph specifying the direction of the interaction. A directed graph can also be labeled with a sign and strength for each interaction, where each edge has a positive, zero or negative value indicating activation, no interaction and repression, respectively.

The following sections provide an overview over the field. In particular, several different approaches to gene regulatory network inference are presented, discussing their strengths and weaknesses, and providing guidelines on which models are appropriate under what circumstances.

### 2.4.1 Boolean networks

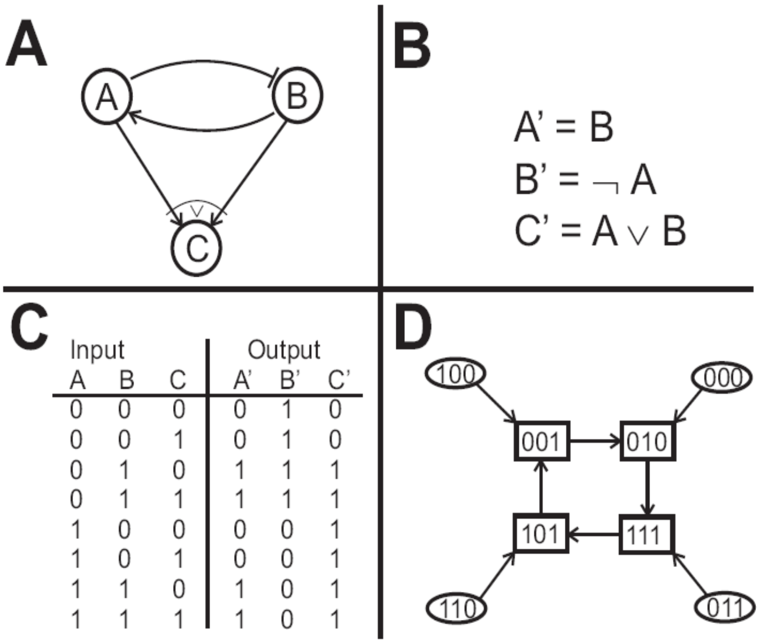
Boolean networks are probably the simplest models conceivable for regulatory networks. They offer a binary, discrete-time description of a system. They can be seen as a generalization of Boolean cellular automata [75], and have been introduced as models of genetic regulatory networks by Kauffman [76] in 1969. Boolean networks assume that each gene is in one of two states, either active or inactive.

Interactions between genes are modeled through Boolean logic functions, and updates are carried out simultaneously for all genes in discrete time steps. The updates are deterministic, and Boolean networks provide only a qualitative description of a system. A Boolean network can be graphically represented in several ways, emphasizing different aspects of the network. An example is shown in Figure 5 for a small sample network consisting of three nodes A, B and C. In Figure 5A, pointed arrows indicate activation. For example, gene A will be activated if gene B is active. Flat arrows indicate an inhibition. For instance, gene B will be deactivated if gene A is active. Gene C is activated if either gene A or gene B is active, as denoted by the "or" symbol "⋁" in the figure. In Figure 5B, the same relationships are expressed by Boolean logical rules; the second line specifying B' (i.e. the state of B in the next time point as a negation of A). Figure 5C shows a tabular representation of all possible input states and the resulting next states of the network. Figure 5D visualizes the state space in a graphical form, showing how the eight possible states of the network are interconnected. For example, if the network is in state (A = 1, B = 0, C = 0), then the next state of the network will be (A = 0, B = 0, C = 1).

Recent studies show that many biologically relevant phenomena can be explained using the Boolean formalism [77]. Focusing on generic principles, Boolean networks can capture switch-like behavior, oscillations [78-80], and providing a qualitative description of a system [81].

Recent modeling results combine data from living cells with experimental techniques to validate genetic models, showing that such simple models can indeed predict the overall dynamics of a biological genetic circuit [82]. It has been shown that for understanding the general dynamics of a regulatory network, usually detailed dynamic parameters are not needed. It is the wiring that is most important [83]. For instance, Albert and Othmer [84] have predicted the segment polarity network in *Drosophila melanogaster* solely on the basis of discrete binary models. Similarly, Li and coworkers [85] have constructed the genetic regulatory network controlling the yeast cell cycle using a binary model.

A drawback of Boolean networks is that they are deterministic in nature. However, true biological networks are known to have stochastic (i.e. non-deterministic) components. For instance, proteins produced from an activated promoter in short bursts seem to occur at random time intervals [86]. Furthermore, we are often dealing with noisy inputs and experimental errors which may lead to data inconsistency.



| Input | | | Output | | |
|---|---|---|---|---|---|
| A | B | C | A' | B' | C' |
| 0 | 0 | 0 | 0 | 1 | 0 |
| 0 | 0 | 1 | 0 | 1 | 0 |
| 0 | 1 | 0 | 1 | 1 | 1 |
| 0 | 1 | 1 | 1 | 1 | 1 |
| 1 | 0 | 0 | 0 | 0 | 1 |
| 1 | 0 | 1 | 0 | 0 | 1 |
| 1 | 1 | 0 | 1 | 0 | 1 |
| 1 | 1 | 1 | 1 | 0 | 1 |

Figure 5. Different representations of Boolean networks.

(A) Graph representation, (B) Logical Boolean rules, (C) State transition table and (D) State transition graph.

Boolean networks are attractive due to their simplicity. However, the underlying assumptions appear to be very strict. In particular, modeling genes as being in one of two states either *‘on’* or *‘off’*, certainly is an oversimplification of a true biological network. Similarly, true networks are time continuous, whereas Boolean networks assume time discrete.

## 2.4.2 Relevance networks

While the assumption of Boolean networks is that genes can only be in one of two states, expressed or not expressed, relevance networks [87] look at similarity or dissimilarity between pairs of genes on a continuous scale. In the network inference context, relevance networks are often known as “correlation networks”. In this approach two major steps are involved:

1. All pairs of genes are compared using some measure of similarity or dissimilarity. For example, pair wise correlation coefficients [88-90], or mutual information [91].
2. The complete set of pair wise comparisons is filtered to determine the relevant connections, corresponding to either positive or negative associations between genes.

The resulting network can then be represented in a graphical form. This section only presents one representative example, the ARACNe (Algorithm for the Reconstruction of Accurate Cellular NEtworks) by Basso *et al*. [92, 93] as one of the most successful algorithms representing the relevance network approach. It identifies statistically significant gene-gene co-regulation by mutual information. Mutual information is estimated using Gaussian kernel estimators for discrete and continuous random variables [92]. Relevant edges in the network are determined by statistical test. Monte Carlo randomization of the data is used for the computation of p-values, and edges are filtered based on a p-value threshold. ARACNe simplifies the network by eliminating indirect relationships, in which two genes are co-regulated by one or more intermediary genes. This is done using the data processing inequality (DPI), which essentially states that if three random variables X, Y and Z depend on one another in a linear fashion (X → Y → Z), then the mutual information

$M(X,Z) \leq \min[M(X,Y), M(Y,Z)]$. This is used to remove indirect edges X → Z from the network. ARACNe has been performed on microarray gene expression data from human B cells, reconstructing a network with approximately 129,000 interactions from 336 expression profiles [92].

Similar to Boolean networks, relevance networks are relatively simple models of gene regulatory networks. In contrast to Boolean networks, however, they are continuous models, i.e., genes can have expression values on a quantitative scale.

One drawback of Relevance networks is that they do not consider time, and thus disregard any dynamic aspects of gene expression. Hence, it is not clear how to carry out simulations with an inferred network. Algorithms such as ARACNe are based on pair wise similarity only, and it may thus miss interactions between multiple genes. Furthermore, the choice of threshold for the inclusion of edges is arbitrary in that varying threshold parameters may change the network considerably. Depending on the similarity/dissimilarity measure used, relevance network approaches are less sensitive to noise.

## 2.4.3 Bayesian networks

While previously described network models assume functional dependence between different nodes, conditional models consider statistical correlation between genes. Conditional models explain the correlation between two genes by other genes in the network. These models are particularly simple in the Gaussian setting, since in this case networks can be learned from data using classical statistical tests [94]. The most popular conditional model is the Bayesian network, which is widely used to model and infer gene regulatory networks [95].

Bayesian networks are probabilistic models. They model the conditional dependence structure between genes in the network. Edges in a network correspond to probabilistic dependence relations between nodes (genes), described by conditional probability distributions. Distributions used can be discrete or continuous, and

Bayesian networks can be used to compute likely successor states for a given system in a known state.

In a Bayesian network the probability relationship among a set of random variables $X_i$, where $I = 1, \ldots, n$ are encoded in the structure of a directed acyclic graph $G$, whose nodes (genes) are the random variables $X_i$. The relationships between the variables are described by a joint probability distribution $P(X_1, \ldots, X_n)$ that is consistent with the independence assertions embedded in the graph $G$ and has the form:

$$p(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid parents(X_i)) \qquad \text{EQ 7}$$

Given a simple Bayesian network (Figure 6) with three nodes (A→C, B→A and B→C) the joint probability distribution is computed by:

$$p(A, B, C) = p(B)p(A|B)p(C|A, B) \qquad \text{EQ 8}$$

Figure 6 shows a simple Bayesian example network with three nodes A, B and C, each assumed to be in one of two states, either *on* or *off*. The conditional probabilities *p(A|B)*, *p(C|A,B)* and the unconditional probability *p(B)* in this binary case are easily tabulated, as shown in the figure. Note that the probability distributions of the nodes in Bayesian networks can be of any type, and need not necessarily be restricted to discrete or even binary values as in this example. The joint distribution of a set of variables $X_1, X_2, \ldots, X_n$ is the product of the local distributions described in EQ 7. For example, the joint probability that all nodes are *on* is p(A = on, B = on, C = on) = p(B = on)p(A = on | B = on)p(C = on | A = on, B = on) = 0.2 × 0.2 × 0.0 = 0.0. It is important to mention at this point that the joint probability distribution can only be resolved this way if the network does not contain any 'directed' cycles.

In order to reversely engineer a Bayesian network model of a gene network, the directed acyclic graph *G* (i.e. the regulators of each transcript) must be found that best describes the gene expression data *D*, where *D* is assumed to be a steady-state data set. This is performed by choosing a scoring function that evaluates each graph *G* with respect to the gene expression data *D*, and then searching for the graph *G* that maximizes the score.

The score can be defined using the Bayes rule: $P(G \mid D) = \frac{P(D|G)\,P(G)}{P(D)}$, where *P*(*G*) is 'prior probability', that does not take into account any information about *D*, and can either contain some apriority knowledge on network structure, if available, or can be a constant non-informative prior. *P*(*D*|*G*) is the conditional probability of *D*, given *G*. That is a function to be chosen by the algorithm that evaluates the probability that the data *D* has been generated by the graph *G*. The most popular scores for the likelihood are Bayesian Information Criteria (BIC) or Bayesian Dirichlet equivalence

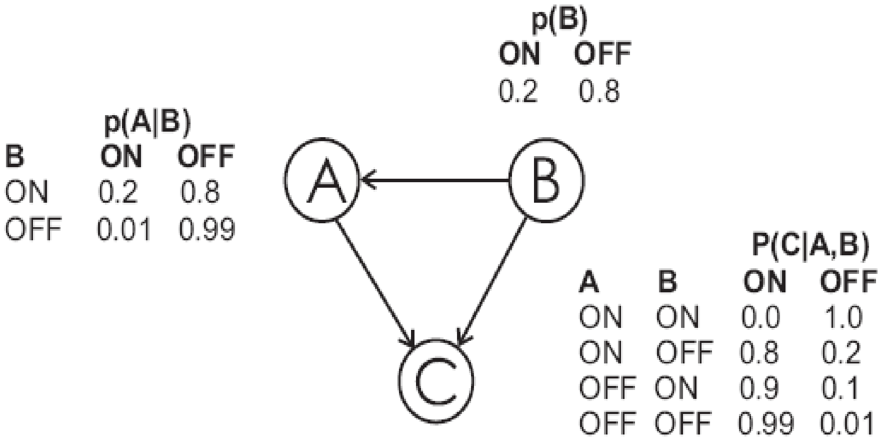


Figure 6. Sample Bayesian network with three nodes, two states.

There are two possible states (ON or OFF). Given next to each node are the conditional distributions for the node, conditioned on its parents, as indicated by the arcs. For example, the probability that A is off given that B is on, p(A = off|B = on) is 0.8.

(BDe).

BIC performs a search through the space of possible networks and scores each structure. The aim is to identify the network with the maximum score. A variety of search strategies can be used, the simplest being a greedy hill-climb. The search begins with an empty network. At each stage of the search, networks in the current neighborhood are found by applying operators such as 'add edge', 'remove edge' and 'reverse edge' to the current network. This may overfit the data because any edge which improves the fit will be added, and so the measure tends to make the graph too dense. Therefore a penalty can be introduced to prevent overfitting. BIC function is a combination of the model log-likelihood and a penalty term that favors less complex models. The BIC is shown in EQ 9.

$$BIC \ = \ -2\ log\ p(G|D) + \ klog(n) \qquad \text{EQ 9}$$

In the above equation, $n$ is the number of observations (sample size) and $k$ is the number of parameters. $log(G|D)$ is the log-likelihood while the term $klog(n)$ is the penalty term.

BDe is a multinomial distribution that describes the conditional probability of each node in the network. Because there are so many possible graphs, this usually needs to be calculated using a Monte Carlo search method.

$$\log \mathrm{p}(G|D) = \log \int p(D|G, \theta) p(\theta|G)\, d\theta \qquad \text{EQ 10}$$

In EQ 10, θ denotes the parameter for the conditional probability distribution for graph *G*. Both scores incorporate a penalty for complexity to guard against over-fitting of data. BDe substantially outperforms the BIC when the training data is limited. This is because the BIC over penalizes complexity relative to the BDe. Training data for GRN inference is typically very limited.

In Bayesian networks, several high-scoring networks are found. One can use model averaging or bootstrapping to select the most probable regulatory interactions and to obtain confidence estimates for the interactions. Alternatively, one can augment an incomplete data set with prior information to help select the most likely model structure.

Bayesian networks can deal with noisy data and stochastic aspects of gene expression in a natural way [86, 96], and they are extendable to deal with missing data [97]. Furthermore, they provide an intuitive and simple visualization of the conditional dependence structure in given data, and are much easier for humans to understand than conditional distributions. At the same time, depending on the probability distributions used (continuous or discrete), they can model quantitative aspects of gene regulatory networks. Bayesian networks have been used to e.g., infer regulatory interactions of the yeast cell cycle [13, 98].

Since learning Bayesian networks is NP-hard, heuristic search methods have to be used, which do not guarantee that the globally optimal solution is found [96]. Probably their main disadvantage is that they disregard dynamic aspects, and require the network structure to be acyclic, since otherwise the joint distribution cannot be decomposed as in equation EQ 7. Moreover, feedback loops are known to play key roles in causing certain kinds of dynamic behavior such as oscillations or multi-stationary [13, 79, 80, 99, 100], which cannot be captured by the Bayesian network model.

### 2.4.4 Dynamic Bayesian networks

Efforts have been made to overcome these limitations of Bayesian networks. Bayesian networks can be extended to capture the dynamic aspects of regulatory networks by assuming that the system evolves over time. Thus, gene expression can be revealed as a time series, considering different vectors $X(1), ...,X(T)$ at $T$ consecutive time points. One can assume that a variable $X_i(t)$ of a particular gene $i$ at time $t$ can have parents only at time $t-1$. Thus the cycles in the Bayesian network unroll, the resulting network becomes acyclic and the joint probability in EQ 8 becomes tractable again. The resulting networks are called Dynamic Bayesian Networks [39, 101].

Dynamic Bayesian Networks can handle noisy data to capture the architecture of regulatory networks from microarray data [102, 103]. These models have been combined with hidden variables to capture non-transcriptional effects [38]. Similarly, aiming at the integration of information from additional data sources into the Bayesian network learning process, Bernard and Hartemink [104] include transcription factor binding location data through the prior distribution, while evidence from gene expression data is considered through the likelihood.

# 3 SINGLE SPECIES EXAMPLE 1: SULFITE REDUCTASE ACTIVITY IN *ARABIDOPSIS THALIANA*

A common goal in DNA microarray experiments is to identify those genes that are significantly expressed at different levels in different samples. This involves comparing the expression levels of genes for different phenotypic groups (treated, disease tissue versus normal) in order to discover the genes that best distinguish the groups. For example, the data may represent samples treated with two different drugs to investigate different responses to disease and normal tissues. The question to be asked is: “Which genes best distinguish the various classes in the data?” The goal is to identify those genes that are most informative for distinguishing the samples based on classes.

This section presents a procedure for statistical inference using linear models along with an appropriate interpretation of a set of microarray experiments on the sulfur metabolism of *Arabidopsis Thaliana*. Experimental and computational results presented in this section have been published in The Plant Cell [105].

## 3.1 Introduction

### 3.1.1 Statistical tests

The simplest technique to identify differentially expressed genes is to look at the ratios of expression in different samples. A gene with a ratio exceeding a given threshold can be considered as differentially expressed. A threshold is often set to two-fold change. The larger the threshold, the more confident the assignment, but also the more differentially expressed genes will be missed. On the other hand, if a low threshold is applied more differentially expressed genes will be identified, but with greater probability of false-positives. There is a wide variety of statistical tests can be applied to this question, including t-tests (for two classes) and analysis of variance (ANOVA; for three or more classes) that assign *p*-values to genes based on their ability to distinguish between groups. The benefit of using such tests is that they provide a quantitative measure of the significance of the difference.

A major problem of identifying differentially expressed genes (DEG) across specified conditions in microarray experiments is that the classical distribution assumptions do not usually hold and in most expression experiments the small numbers of measurements prevent the assumption being properly tested. There are tests that do not rely on any distribution assumptions. The standard nonparametric statistical test that is equivalent of the t-test for parametric data is Mann-Whitney or Wilcoxon test and empirical Bayes approach. Another method is the Significance Analysis of Microarrays (SAM) [106], which uses the gene expression measures to estimate the significance by means of permutation test. A permutation test randomly swaps measurements between different groups obtaining the distribution of the test statistic from the data observed. Each time, a value for the test statistic (*t*) is calculated. Total number of times, by chance, a value for *t* occurs that is equal to or more than that measured for the real data, allows estimating of the probability that the dataset shows a significant separation between the classes.

Linear statistical models, among other techniques, are often used to overcome distribution assumptions by capturing the variation specific to all genes. A linear statistical model is an algebraic equation containing the variables whose parameters are to be estimated from the mean gene expression values. Once the linear model has been fitted to the data, residuals (the part of the intensity values that are not explained by the model) can be used globally for the variation of intensity specific for each gene. Once the individual gene model has been obtained, two types of statistical hypothesis can be performed. The first hypothesis is, “Is at least one of the treatment levels different within this gene data set?” The second relates to testing for changes in expression levels of each gene in response to the specific individual treatments. The test statistics used to address the first question is the F-statistics that involves an analysis of variance (ANOVA), and t-test is often used for the second question. ANOVA involves estimating the mean and standard deviation for different groupings of data including all of the gene values and different subsets (dyes, treatments slides, etc). Mean square values (mean square = sum of squares of deviations of each value from mean divided by degrees of freedom) are then calculated for each group. If the mean square value for one of the groups is greater than the residual error for the gene then the treatment is having a greater overall effect on the intensity. In this case the residuals are the deviations of the observations from the sample mean, not the population means. The F-test estimates the probability of such a variation within treatment effects, observed by chance (given the degree of freedom associated with each mean square value).

One concern with the above statistical approaches is the problem of multiple testing. In most biological experiments, one measures a small number of observations across a relatively large number of samples. However, in most microarray experiments, there are thousands of gene expression levels across a relatively small number of samples, and this can lead to the misidentification of genes being differentially expressed even when they are not—the problem of false positives.

This chapter presents the application of empirical Bayesian methods to a set of microarray experiments to detect response of sulfur metabolism at transcriptional level in *Arabidopsis thaliana.* The results show the performance of linear models by experimentally validating a set of genes that are identified as significantly differentially expressed genes.

## 3.1.1 Importance of sulfite reductase

Plants take up the essential macronutrient sulfur from the soil in the form of sulfate. The uptake of sulfate and its subsequent assimilatory reduction into organic sulfur compounds proceed through a highly coordinated mechanism. First, uptake of sulfate is catalyzed by specific proton cotransporters in root epidermal cells. They belong to the group of high affinity sulfate transporters (SULTR group 1) and are inducible by external sulfate deprivation [107]. Internal allocation of sulfate is catalyzed by members of the low affinity SULTR groups 2 and 3 [108]. Next, assimilatory reduction of sulfate is initiated by ATP-dependent activation of sulfate to adenosine 5'-phosphosulfate (APS) catalyzed by ATP sulfurylase (ATPS). Further activation with ATP is catalyzed by APS kinase and yields 3'-phosphoadenosyl-5'- phosphosulfate (PAPS). APS kinase is present in plastids and the cytosol to provide PAPS for sulfation reactions by sulfotransferases [109].

APS reductase (APR) in plastids from *Arabidopsis thaliana* and other plants strongly prefers APS instead of PAPS as a substrate, its expression responds to sulfate and nitrate availability, and a number of stress factors result in regulation of its activity [110]. In addition, flux analysis using $^{35}$S-labeled sulfate hinted that APR, after sulfate uptake, exerts the strongest control over flux through the sulfate reduction pathway in *Arabidopsis* [111] and is responsible for genetically determined variation in sulfate content in *Arabidopsis ecotypes* [112].

In contrast with APR, the second enzyme of the free reduction pathway, sulfite reductase (SiR), has received little attention. Plant SiR is a plastid-localized soluble enzyme of two 65-kD subunits, contains a single siroheme and (4Fe-4S) cluster as prosthetic groups, and has a high affinity ($K_m^{sulfite}$ ~ 10µM) for sulfite [113]. Ferredoxin acts as the physiological donor of the six electrons required for sulfite reduction, whereas bacterial SiR uses NADPH [114]. The structure, sequence, and ligands of SiR in bacteria, archea, and eukaryotes are similar to those of nitrite reductase, which catalyzes an equivalent reduction step in nitrate assimilation (i.e., a six-electron reduction of nitrite to ammonia) [115]. In *Arabidopsis*, SiR shows 19% identity with nitrite reductase (NiR) at the amino acid level. Phylogenetic analysis showed that both SiR and NiR arose from an ancient gene duplication in eubacteria, before the primary endosymbiosis that gave rise to plastids [116]. SiR is able to reduce nitrite as well as sulfite, and substrate preferences can be converted by a single amino acid mutation [113].

SiR is encoded by the only single-copy gene in primary sulfur metabolism in *Arabidopsis*, whereas the rice (*Oryza sativa*) and poplar (*Populus spp*) genomes each contain two copies [117]. It is expressed in nearly all tissue types and shows the least transcriptional responses among sulfur-related genes in *Arabidopsis* in classical sulfate starvation experiments or under other stress conditions, according to a survey in microarray databases [118]. Expression changes were observed after treatment with $SO_2$ [119], but these were not translated into significant changes of SiR enzyme activity under similar conditions [120]. Activity of SiR is generally believed to be maintained in excess to scavenge potentially toxic sulfite [117], based on flux control and APR over-expression experiments in *Arabidopsis* and maize [121]. However, the bulk of sulfite is normally channeled into the assimilatory reduction pathway for sulfur amino acid and protein biosynthesis.

Here, we investigated two *Arabidopsis* lines with T-DNA insertions in the promoter region of SiR. Mutant line *sir1-2* is early seedling lethal and unequivocally demonstrates that the free sulfate reduction pathway is essential for survival and cannot be compensated for by any other enzymatic process. In mature leaves, mutant *sir1-1* has 28% of SiR activity and 3.6% of flux in the assimilatory reduction pathway *in vivo* compared with the wild type. *sir1-1* has a strongly retarded growth phenotype, showing that in contrast with general assumptions, SiR can easily become limiting for growth. *sir1-1* mutant plants are sensitive to cadmium due to lack of GSH for phytochelatin synthesis. Carbon, nitrogen, and sulfur composition are severely altered with a shift toward carbon-bound reduced nitrogen, indicating that lowered sulfite reduction leads to comprehensive reprogramming of primary metabolism.

## 3.2 Methods

### 3.2.1 Analysis of transcript levels

Microarray and qRT-PCR experiments were performed by the group of Professor Dr. Rüdiger Hell, Heidelberg Institute for Plant Sciences, University of Heidelberg. The microarray datasets have been deposited in the Gene Expression Omnibus under the series number GSE20670.

For qRT-PCR and microarray analyses, RNA was isolated from 200 mg leaf material of 7-week-old soil-grown homozygous *sir1-1* and Col-0 plants with the RNeasy kit (Qiagen) according to manufacturer's protocol. For microarray analyses of transcript levels of 1920 selected genes, the RNA was converted to cDNA, hybridized with a custom-made microarray, and evaluated as described by [122]. The transcript levels of sulfur metabolism related genes in leaves of *sir1-1* and wild-type plants (*Col-0*) grown on soil under short-day conditions for 7 weeks were compared using a targeted microarray approach. Total mRNA was extracted from three individuals of each plant line, labeled independently two times with Cy3 and Cy5, and hybridized with the microarray twice (n = 12). Data were normalized and examined for statistical significance as described in section 3.2.2.

Abundances of selected transcripts were independently confirmed from the same RNA preparation by qRT-PCR after cDNA conversion with the SuperScript VILO cDNA synthesis kit (Invitrogen). The qRT-PCR reaction was set up by mixture of 10 ng of freshly synthesized cDNA with 1.6 pmol of each specific primer in onefold EXPRESS Two-Step SYBR GreenER Universal mixture (Invitrogen). The reaction was performed in the LightCycler 480 (Roche Diagnostics) according to the EXPRESS Two-Step SYBR GreenER protocol and evaluated with Light-Cycler software 4.0 (Roche Diagnostics) using elongation factor 1a (EF1a) as reference for normalization. Each analysis consisted of three biological replicas. Each replica was tested three times, and this test was repeated once (i.e., n = 6 per replica).

## 3.2.2 Preprocessing and statistical analysis

The microarray data sets were analyzed with M-CHiPS software [74]. Signal intensities were normalized by log-linear regression as described [74]. All hybridizations showed correlation coefficients higher than 0.8 between the two channels.

The normalized data were used to compute p-values. A contrast matrix which defines the comparisons of interest between samples in the experiment was set up (as described in section 3.2.3) to test differential expression across all samples. Regression coefficients were estimated using a least squares linear model fitting procedure and tested for differential expression with moderated Student's t-statistic via the empirical Bayesian statistics described in the limma package [123]. P-values computed for the F-statistic were adjusted for multiple testing to control the FDR at 5% [124]. The adjusted p-values can serve to accept or reject the null hypothesis based on a significance level. Genes showing intensity levels of more than 1000 in at least one of the conditions and also exhibit p-values smaller than 0.05, were selected as differentially expressed genes.

## 3.2.3 Statistical computations

In order to detect which treatment is responsible for the difference in expression level between each sample a complete pair wise comparison was performed as follows: Let $y_{gj}$ be the expression values for genes $g = 1,...,G$ and arrays $j = 1,...,J$, pre-processed, background-corrected and normalized, then systematic effect for each gene by a linear model $E(y_g) = X\beta_g$, where $y_g = (y_{g1}, ...,y_{gJ})^T$ is the vector of expression values for gene $g$, $X$ is a known design matrix and $\beta_g = (\beta_{g1},...,\beta_{gK})^T$ is a gene-specific vector of regression coefficients. The design matrix depends on the experimental design and choice of parameterization, and the regression coefficients represent comparisons of interest between measurements in the experiment. These coefficients were estimated with the least squares linear model fitting procedure of the Bioconductor package limma [123] and tested for differential expression (testing any particular $\beta_{gk}$ equal to 0) with moderated Student's t-statistics [123]. I accepted or rejected (equal means for all groups) the null hypothesis on the basis of P-values computed for the omnibus F-statistic via limma as described above, at a specified significance level.

### 3.2.4 FDR adjustment

The adjustment performed here is based on a method developed by Benjamini and Hochberg [124] called FDR. The FDR is based on certain statistical considerations. In statistical inference, there is often concern for minimizing the risk associated with choosing one hypothesis over the other. The risk generally minimized is the type I error, known as ‘false positive’. That is, the probability of falsely rejecting the null hypothesis when the null hypothesis is actually true. The 0.05 level for p-values has historical significance and corresponds to the probability of making a type I error that is 5% of all experiments of the same size and type performed by the experimenter. The null hypothesis will at least once be falsely rejected when in reality the null hypothesis is true. Each time the hypothesis test is applied, there is a risk of making the type I error, and the probability of that accumulates the more hypotheses that are tested. This means that type I error rate for the overall experiment utilizing all genes could be much higher than 5%, even though each test was controlling the type I error at the 5% level. These *P*-values, with appropriate multiple testing adjustment to control the False Discovery Rate (FDR) at 5% [124], allow us to identify differentially expressed genes.

## 3.3 Results and discussion

The impact of reduced *SiR* activity on the transcription of sulfur metabolism-related genes in leaves of 7-week-old soil-grown plants was investigated with microarrays carrying 1920 genes related to primary metabolism and stress responses as described by Haas *et al.* [122].

Based on three biological repetitions of the wild type and *sir1-1* with four technical replicates, each including dye swaps for each set, 67 genes were found to be significantly up- or downregulated in the leaves of hydroponically grown *sir1-1* plants compared with *Col-0* according to p-values of <0.05. Most regulated genes were related to redox homeostasis (20), while genes of sulfur metabolism (11), pathogen resistance (11), glucosinolate synthesis (10), hormone synthesis and signaling (5), GSH transfer activity (4), sulfur-induced nonsulfur genes (3), and amino acid synthesis (3) were also found to be significantly changed in abundance (Table 1). Figure 7 gives an overview of the top 67 differentially expressed genes. It shows the fold changes versus a measure of statistical significance of the changes.

Two independent *Arabidopsis* T-DNA insertion lines, further annotated as *sir1-1* and *sir1-2*, were identified in the GABI-Kat collection center. In both lines, the T-DNA was inserted in the promoter region of *SiR* (Figure 8A). Flanking sequences of the T-DNA were PCR amplified and sequenced to characterize the insertion sites [105]. Quantitative real-time PCR (qRT-PCR) detected successful transcription initiation corresponding to 14% of mature *SiR* transcript in the early seedling stage that may account for embryo development and germination of *sir1-2* (Figure 8C). For genetic complementation, the heterozygous *sir1-2* plants were transformed with a construct that expressed the *SiR* cDNA along with its plastid transit peptide under control of the constitutive 35S cauliflower mosaic virus promoter. The phenotype of *sir1-2* was completely restored (Figure 8B), demonstrating that the loss of function of *SiR* was the cause of the early seedling lethality observed in the homozygous *sir1-2* plants.

Table 1. Significantly regulated transcripts in leaves of 7 week old *sir1-1*plants.

| MIPS | NCBI ID | Description | MF* col-0 | MF sir | change | Change | p-value |
|---|---|---|---|---|---|---|---|
| AT5G04590 | GeneID:830336 | SiRSulfur-metabolism | 16134 | 5008 | -3.22 | 31.1 | 0 |
| At5g24770 | GeneID:832546 | VEGETATIVE STORAGE PROTEIN 2Sulfur-induced | 2675 | 30425 | 11.37 | 1137.0 | 0.001 |
| At5g24780 | GeneID:832547 | VEGETATIVE STORAGE PROTEIN 1Sulfur-induced | 2104 | 18883 | 8.98 | 898.0 | 0.001 |
| AT1G19670 | GeneID:838554 | Chlorophyllasepathogens-related | 6112 | 13274 | 2.17 | 217 | 0.002 |
| At5g43780 | GeneID:834400 | APS4Sulfur-metabolism | 3554 | 1932 | -1.84 | 54.3 | 0.003 |
| At3g19710 | GeneID:821508 | branched-chain amino transferaseGlucosinolat-synthesis | 5594 | 9639 | 1.72 | 172.0 | 0.004 |
| At5g26000 | GeneID:832669 | myrosinaseGlucosinolat-synthesis | 5726 | 9084 | 1.59 | 159.0 | 0.004 |
| At5g06290 | GeneID:830517 | 2-Cys Prx BREDOX | 42866 | 31255 | -1.37 | 73.0 | 0.004 |
| At3g57260 | GeneID:824893 | PR2REDOX | 18484 | 10710 | -1.73 | 57.8 | 0.004 |
| At2g44290 | GeneID:819037 | Lipid transfer proteinREDOX | 6112 | 3441 | -1.78 | 56.2 | 0.004 |
| At4g03060 | GeneID:828102 | AOP2Glucosinolat-synthesis | 1948 | 2476 | 1.27 | 127.0 | 0.006 |
| AT1G75040 | GeneID:843842 | PR5pathogens-related | 6586 | 4134 | -1.59 | 62.9 | 0.006 |
| AT3G44300 | GeneID:823555 | nitrilaseAuxin-Biosynthese | 6014 | 9007 | 1.5 | 150.0 | 0.006 |
| AT3G20770 | GeneID:821625 | ETHYLENE-INSENSITIVE3pathogens-related | 9919 | 6722 | -1.48 | 67.6 | 0.007 |
| At4g02520 | GeneID:827931 | ATGST F2GSH-Transfer | 33328 | 15340 | -2.17 | 46.1 | 0.009 |
| AT4G16860 | GeneID:827395 | RPP4pathogens-related | 7455 | 4951 | -1.51 | 66.2 | 0.009 |
| AT5G15230 | GeneID:831375 | gibberellin-regulated (GASA4)Hormon induced/related | 2778 | 3778 | 1.36 | 136.0 | 0.01 |
| AT2G06050 | GeneID:815160 | OPR3pathogens-related | 2692 | 3296 | 1.22 | 122 | 0.011 |
| At2g05380 | GeneID:815086 | glycine-rich proteinREDOX | 19826 | 16057 | -1.23 | 81.3 | 0.012 |
| AT4G16950 | GeneID:827403 | RPP5pathogens-related | 16675 | 10827 | -1.54 | 64.9 | 0.012 |
| AT4G12470 | GeneID:826859 | lipid transfer protein family proteinSulfur-induced | 3877 | 3121 | -1.24 | 80.6 | 0.014 |
| At1g24100 | GeneID:839022 | glucosyl transferaseGlucosinolat-synthesis | 3025 | 3591 | 1.19 | 119.0 | 0.016 |
| At4g31870 | GeneID:829316 | AtGpx7REDOX | 9410 | 7139 | -1.32 | 75.8 | 0.017 |
| At5g10180 | GeneID:830882 | Sultr2;1 Sulfur-metabolism | 3738 | 2728 | -1.37 | 73.0 | 0.018 |
| At2g29580 | GeneID:817507 | zinc finger family protein REDOX | 4009 | 3118 | -1.29 | 77.5 | 0.018 |
| At2g25080 | GeneID:817046 | putative glutathione peroxidase AtGpxGSH-Transfer | 13950 | 9001 | -1.55 | 64.5 | 0.021 |
| At4g13770 | GeneID:827011 | cytochrom P450Glucosinolat-synthesis | 7864 | 12020 | 1.53 | 153.0 | 0.022 |
| At1g16400 | GeneID:838210 | cytochrom P450. CYP79F2Glucosinolat-synthesis | 2083 | 2957 | 1.42 | 142.0 | 0.022 |
| At4g03520 | GeneID:825653 | ThioredoxinREDOX | 19556 | 13243 | -1.48 | 67.6 | 0.022 |
| AT3G25250 | GeneID:822119 | OXI1REDOX | 2742 | 1716 | -1.6 | 62.5 | 0.022 |
| At2g14610 | GeneID:815949 | PR1 REDOX | 2742 | 1716 | -1.6 | 62.5 | 0.022 |
| At5g18170 | GeneID:831935 | glutamate dehydrogenaseamino acid synthesis | 3356 | 2820 | -1.19 | 84.0 | 0.022 |
| AT3G44310 | GeneID:823556 | NIT1Auxin-Biosynthese | 7552 | 9348 | 1.24 | 124.0 | 0.022 |
| AT4G24620 | GeneID:828564 | Phosphoglucose isomerase Asc-biosynthesis | 5388 | 4293 | -1.26 | 79.4 | 0.022 |
| At2g02930 | GeneID:814822 | putative glutathione S-transferaseGSH-Transfer | 3718 | 2873 | -1.29 | 77.5 | 0.024 |
| AT1G59870 | GeneID:842281 | ATP binding cassette transporterpathogens-related | 21777 | 15576 | -1.4 | 71.4 | 0.024 |
| At2g43570 | GeneID:818959 | chitinaseREDOX | 3112 | 2681 | -1.16 | 86.2 | 0.025 |
| AT3G44480 | GeneID:823573 | RPP10pathogens-related | 6297 | 4170 | -1.51 | 66.2 | 0.025 |
| At3g11630 | GeneID:820335 | 2-Cys Prx AREDOX | 33458 | 23286 | -1.44 | 69.4 | 0.027 |
| At5g25980 | GeneID:832667 | myrosinaseGlucosinolat-synthesis | 35276 | 56689 | 1.61 | 161.0 | 0.03 |
| At1g54040 | GeneID:841842 | epithiospecifier protein.Glucosinolat-synthesis | 2408 | 2751 | 1.14 | 114.0 | 0.03 |
| AT3G22740 | GeneID:821845 | AtHMT-3Sulfur-metabolism | 1580 | 1937 | 1.23 | 123.0 | 0.03 |
| AT1G13420 | GeneID:837902 | Sulfotransferase family proteinSulfur-metabolism | 12743 | 8940 | -1.43 | 69.9 | 0.03 |
| At2g32880 | GeneID:817849 | MATH domain-containing proteinREDOX | 2032 | 2324 | 1.14 | 114.0 | 0.03 |
| At5g63030 | GeneID:836423 | GlutaredoxinREDOX | 3430 | 2900 | -1.18 | 84.7 | 0.03 |
| AT2G41680 | GeneID:818766 | dihydrolipoyl dehydrogenase REDOX | 12743 | 8940 | -1.43 | 69.9 | 0.03 |
| At1g72260 | GeneID:843558 | THI2.1pathogens-related | 1486 | 1767 | 1.19 | 119 | 0.031 |
| At5g44070 | GeneID:834430 | Phytochelatin synthaseSulfur-metabolism | 4958 | 4081 | -1.21 | 82.6 | 0.032 |
| At2g47880 | GeneID:819400 | GlutaredoxinREDOX | 2618 | 4220 | 1.61 | 161.0 | 0.032 |
| AT4G01850 | GeneID:826987 | SAM2 Sulfur-metabolism | 5572 | 6779 | 1.22 | 122.0 | 0.033 |
| At5g05730 | GeneID:830457 | ASA1amino acid synthesis. | 4182 | 3420 | -1.22 | 82 | 0.033 |
| At4g39540 | GeneID:830108 | shikimate kinase - like proteinamino acid synthesis | 3686 | 2922 | -1.26 | 79.4 | 0.033 |
| AT4G14560 | GeneID:827103 | auxin induced gene (IAA1)Hormon induced/related | 3725 | 2836 | -1.31 | 76.3 | 0.033 |
| At3g56060 | GeneID:824772 | mandelonitrile lyase-like proteinGlucosinolat-synthesis | 3126 | 2400 | -1.3 | 54.3 | 0.034 |
| At1g62180 | GeneID:842514 | APR2Sulfur-metabolism | 6278 | 5390 | -1.16 | 86.2 | 0.034 |
| At2g03980 | GeneID:814924 | hydrolase family proteinREDOX | 2887 | 3421 | 1.18 | 118.0 | 0.034 |
| At2g29450 | GeneID:817494 | ATGST U5GSH-Transfer | 4869 | 7458 | 1.53 | 153.0 | 0.035 |
| AT2G25450 | GeneID:817083 | similar to ACC oxidase Sulfur-induced | 6902 | 11797 | 1.71 | 171.0 | 0.036 |
| At1g03680 | GeneID:839436 | ThioredoxinREDOX | 21720 | 18502 | -1.17 | 85.5 | 0.038 |
| At1g66100 | GeneID:842924 | THI1.1pathogens-related | 3790 | 5138 | 1.36 | 136 | 0.038 |
| AT5G13160 | GeneID:831155 | AVRPPHB SUSCEPTIBLE 1pathogens-related | 2697 | 2342 | -1.15 | 87 | 0.038 |
| At3g54660 | GeneID:824631 | Glutathione Reductase IREDOX | 7168 | 6124 | -1.17 | 85.5 | 0.039 |
| At2g22330 | GeneID:816765 | cytochrome P450 CYP79B3Glucosinolat-synthesis | 1627 | 1829 | 1.12 | 112.0 | 0.04 |
| ATCG01270 | GeneID:668617 | chloroplast encoded hypothetical proteSulfur-induced | 15336 | 9567 | -1.66 | 2.5 | 0.042 |
| At5g65720 | GeneID:836701 | cysteine desulfhydrylaseSulfur-metabolism | 3258 | 2731 | -1.19 | 84.0 | 0.043 |
| At4g23150 | GeneID:828414 | pad2 regulatededREDOX | 1753 | 1587 | -1.1 | 90.9 | 0.047 |
| AT3G56300 | GeneID:824797 | Cysteinyl-tRNA synthetaseSulfur-metabolism | 37265 | 29949 | -1.24 | 80.6 | 0.048 |

The table represents significantly up-(white) and down-regulated (red) transcript levels in leaves of *sir1-1*in comparison to the wild type (*Col-0*). MIPS code (unique identifier of genes in Arabidopsis thaliana), NCBI GeneID (unique identifier of cDNAs), description (common name or function of gene product),Category (classification in metabolic network), MF ( Median fitted raw data sets), Change (difference in transcript level in x-fold of wild type), Change in % of wild type (difference in transcript level in percent of wild type), p-value (statistical significance).

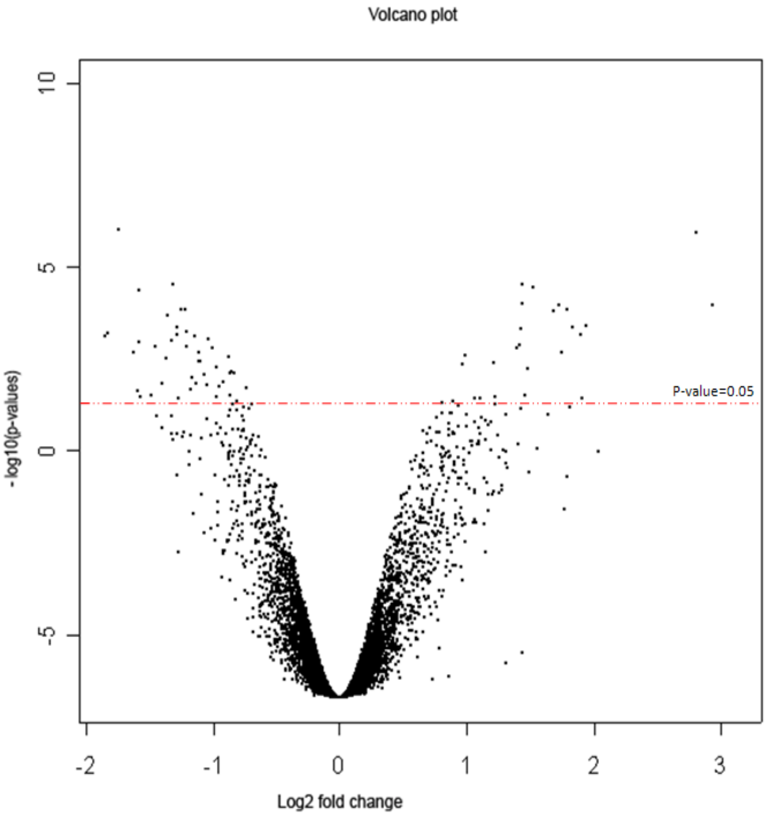


Figure 7. Volcano plot of fold changes versus differential expression.

The log2 fold changes in leaves of *sir1-1* in comparison to the wild type (*Col-0*) were plotted against their significance. The red dashed line represents a p-value of 0.05 which was used as cut off.

Transcript levels of *SiR* in mature leaves of 7-week-old soil-grown homozygous *sir1-1* plants were decreased to 17% compared with the wild type (Figure 9). Accordingly, the amount of *SiR* protein was significantly reduced, and *SiR* activity also was lowered to 28% of the wild-type level (Figure 9A). This provides an explanation for the slower vegetative growth in comparison to the wild type that became more pronounced with time (Figure 9B).

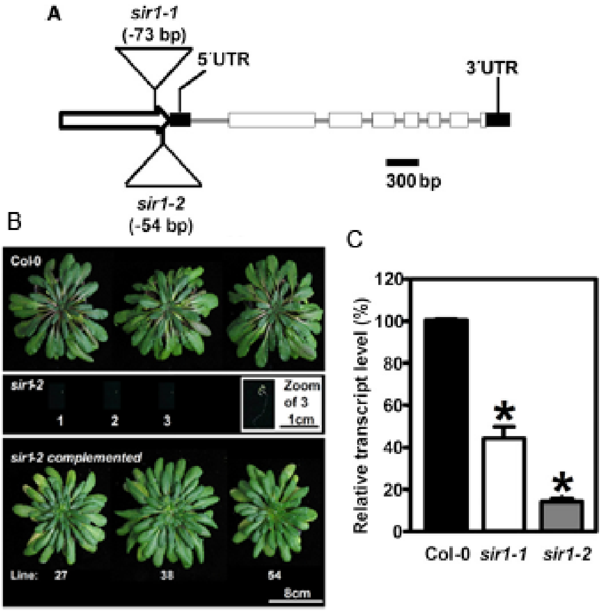


Figure 8. Molecular Identification of *sir* Mutants and Phenotype of *sir1-2*.

(A) Structure of the *SiR* locus with the T-DNA insertion sites in *sir1-1* and *sir1-2*. The putative promoter is marked by a white arrow, exons are indicated as white boxes, and untranslated regions by black boxes. (B) Growth phenotype and genetic complementation of *sir1-2*. For genetic complementation, the cDNA of SiR was fused with the 35S promoter and introduced in *sir1-1* plants by Agrobacterium tumefaciensmediated transformation. (C) Transcript levels of *SiR* in the wild type, *sir1-1*, and *sir1-2* determined by qRT-PCR at the developmental stage of five to six leaves of the wild type. The homozygous *sir1-2* plants arrested at the two cotyledon stage. Bars represent the mean of pooled individuals (n = 6), while error bars show standard errors of technical replicates (n = 3). Asterisks indicate statistically significant (P < 0.05) differences from wild-type values. Figure from [105] by permission of American Society of Plant Biologists.

The microarray analysis confirmed independently the downregulation of *SiR* transcript in mature leaves of *sir1-1*. Besides *SiR*, three genes of the primary sulfur assimilation pathway were significantly downregulated: *ATPS4, APS REDUCTASE2 (APR2), and SULFATE TRANSPORTER 2.1* (*SULTR2.1*; Figure 10). *SULTR2.1* is known to be specific for the vasculature and downregulated in leaves upon sulfur deficiency [107]. ATPS4 catalyzes the activation of sulfate in plastids, which leads to formation of

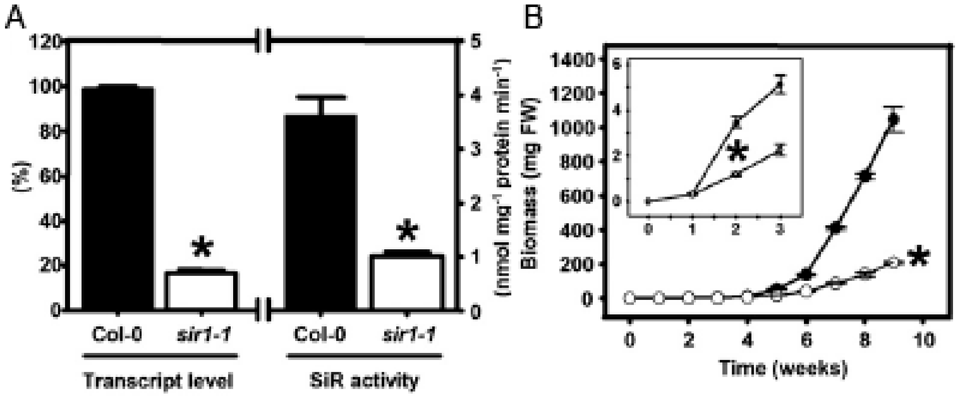


Figure 9. Abundance and activity of *SiR* in the T-DNA Insertion Line *sir1-1*.

(A) Determination of *SiR* activity (n = 5, mean 6 ± SE) and relative *SiR* transcript levels by qRT-PCR (n = 3, mean 6 ± SE) in leaves of 7-week-old plants. Amplification of Ef1a from the same cDNA preparations was used as a control for qRT-PCR. (B) Growth curve of wild-type (black circles) and *sir1-1* plants (white circles). Growth retardation of *sir1-1* was statistically significant from week 2 on, as detailed in the inset. All plants were grown on soil in a growth chamber under short-day conditions (n = 5). Means 6 ± SE are shown. Asterisks indicate statistically significant ($P < 0.05$) differences from wild-type values. Figure from [105] by permission of American Society of Plant Biologists.

*APS*. APR2 is the key APR isoform in leaves for reduction of activated sulfate in *APS* to sulfite that is further reduced by *SiR*. Downregulation of *APR2* was independently confirmed by real-time PCR and revealed 58% mRNA content in *sir1-1* compared with wild-type plants (Figure 11A). Most likely *ATPS4* and *APR2* are downregulated to avoid extensive accumulation of toxic sulfite, which cannot be incorporated into Cys as a result of reduced *SiR* activity in *sir1-1*. The stable total SAT and OAS-TL enzymatic activities and abundances of analyzed SAT and OAS-TL isoforms shown in leaves of *sir1-1* (Figure 11D) were further supported by unchanged SAT and OAS-TL transcript levels in *sir1-1* (Figure 10).

In accordance with unchanged SO protein contents, SO was not upregulated at the transcriptional level despite clearly increased enzymatic activity, leaving the possibility of posttranslational activation of SO in peroxisomes of *sir1-1*. Multiple genes involved in pathogen defense, such as *PATHOGENESIS-RELATED1 (PR1), PR2, PR5, RECOGNITION OF PERONOSPORA PARASITICA1 (RPP1), RPP4, RPP5* as well as *ETHYLENE INSENSITIVE3 (EIN3)*, and *PENETRATION3 (PEN3)* were significantly downregulated along with the sulfur assimilation pathway, indicating a coregulation of both processes, as predicted by the sulfur-enhanced defense hypothesis [125].

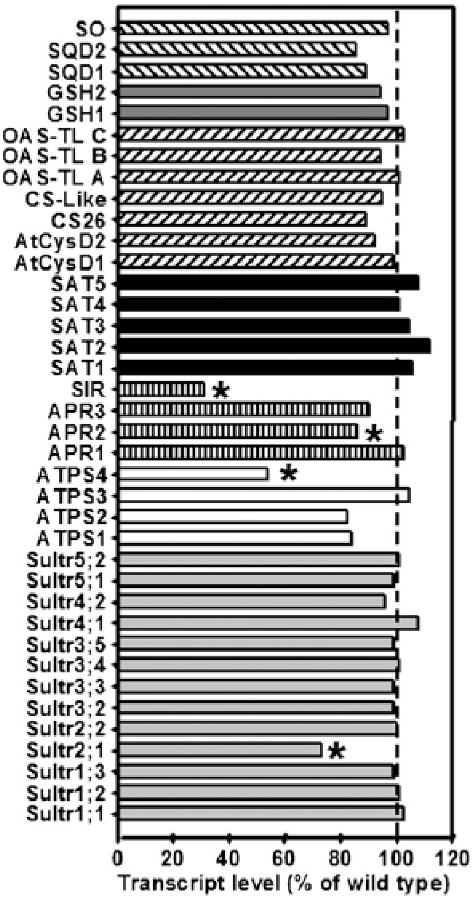


Figure 10. Transcript Levels of Sulfur Metabolism-Related Genes in Leaves of *sir1-1* Plants

From bottom to top: The transcript levels of genes encoding sulfate transporters (light-gray bars), ATPS (white bars), sulfate-reducing enzymes (striped bars), SATs (black bars), OAS-TLs (inclined dashed bars), proteins participating in GSH synthesis (dark gray bars), and sulfolipid biosynthesis enzymes (declined dashed bars) in sir1-1 plants are shown as percentage of wild-type levels. Asterisks indicate statistically significant ($P < 0.05$) differences from wild-type expression levels of the same gene.

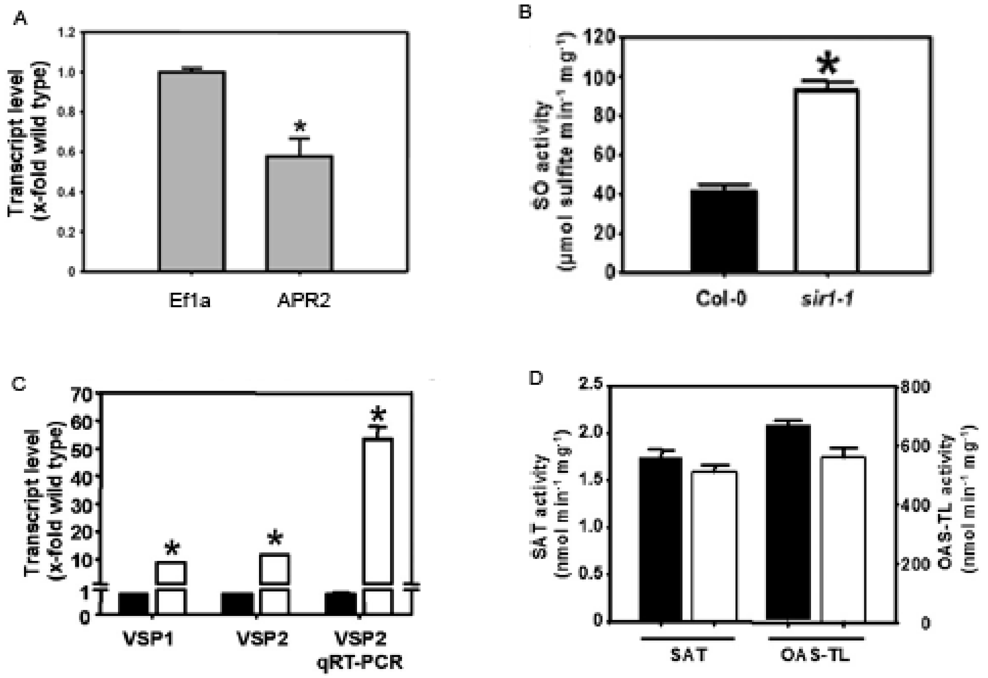


Figure 11. Transcriptional levels by qRT-PCR.

(A)Transcript levels of APR2 in *sir1-1* determined by qRT-PCR. Total mRNA was extracted from leaves of *sir1-1* and wild type plants grown on soil under short day conditions for 7 weeks (n = 3) and analyzed for transcript levels of APR2 by qRT-PCR as described in the Methods. Ef1a served as reference. Means ± SE are shown. (B) Abundance and Activity of SO in Leaves of *sir1-1* plants. Specific activity of SO in protein extracts of 7-week-old wild-type (black) and *sir1-1* (white) plants that were grown on soil under short-day conditions. (C) Transcript levels of VSP1 and VSP2 in leaves of wild-type (black) and *sir1-1* (white) plants were determined by microarray hybridization (VSP1 and VSP2, n = 12) and qRT-PCR (VSP2 qRT-PCR, n = 3). Material was harvested from wild-type and *sir1-1* plants that were grown for 7 to 8 weeks on soil under short-day conditions. (D) Abundance and activity of SAT and OAS-TL in Leaves of *sir1-1* plants from the same extracts as in (B). The specific activities of SO, SAT, and OAS-TL were determined for each extract in triplicates with varying amounts of proteins to prove time and protein linearity of measurement (n = 5 to 7). Means 6 ± SE are shown. The asterisk indicates a statistically significant ($P < 0.05$) difference from the wild-type value. Figure from [105] by permission of American Society of Plant Biologists.

The genes encoding two vegetative storage proteins (VSP1 and VSP2) that are supposed to serve as transient reservoirs for surplus of amino acids in vegetative tissues were upregulated in expression in *sir1-1* leaves. VSP2 upregulation was also confirmed by real-time PCR (Figure 11C). In agreement with the lower contents of total glucosinolates in the mutant, genes encoding thioglucoside glucohydrolase (TGG1 and TGG2; myrosinases), enzymes that are potentially involved in breakdown of glucosinolates upon sulfur and the usage of respective breakdown products (*NITRILASE1* [*NIT1*] and *NIT2*), were significantly upregulated (Table 1). The significant upregulation of the gene for the chlorophyll degrading enzyme CHLOROPHYLLASE1 in *sir1-1* (Table 1) could be a hint toward the pale phenotype.

## 3.4 Conclusion

The data presented here demonstrate the performance of linear models for analyzing designed experiments and the assessment of differential expression. Linear models of empirical eBayes approach applies equally well to both single channel and two color microarray experiments and it is more stable when the number of arrays is small [123]. Identification and analysis of differentially gene expression revealed the importance of the sulfite reductase activity for growth and development in *Arabidopsis thaliana*. The results support that optimal activity of *SiR* is essential for normal growth, and its downregulation causes severe adaptive reactions of primary and secondary metabolism [105].

# 4 SINGLE SPECIES EXAMPLE 2: TILING ARRAY ANALYSIS

While conventional microarrays can provide insights into cellular regulatory networks, other valuable sources of data are increasingly becoming available to aid the learning process. Tiling arrays are a subtype of microarrays. Similar to conventional microarrays, labeled target molecules are hybridized to unlabeled probes immobilized on a solid surface. However, tiling arrays differ in the nature of the probes. Short fragments are designed to cover the entire genome or contiguous regions of the genome. Depending on the probe lengths and spacing different degrees of resolution can be achieved. Tiling arrays are also used for gene expression. DNA microarrays designed to look at gene expression use a few probes for each known or predicted gene. In contrast, tiling arrays can provide an unbiased view at gene expression because previously unidentified genes can still be incorporated. On top of individual gene expression analysis, other uses of tiling arrays are in transcriptome mapping, ChIP-chip and array CGH among others. Tiling arrays are quickly becoming one of the most powerful tools in genome-wide investigations.

Tiling arrays have much wider applications, and researchers might use them for different experiments and informatically select a subset of the probes for analysis. RNAi screens via pooled short hairpin RNAs (shRNAs) have recently become a powerful tool for the identification of essential genes in mammalian cells. In the past years, several pooled large-scale shRNA screens have identified a variety of genes involved in cancer cell proliferation. All of those studies employed microarray analysis.

This section describes high-level analysis of tilling arrays to decode pooled RNAi screens. Experimental and computational results presented in this section have been published [126].

## 4.1 Introduction

In the past two decades, extensive efforts have been undertaken to characterize genes involved in breast cancer development. Gene expression signatures associated with breast cancer and chemotherapy response have been identified [127]. One way to identify such essential genes is the inhibition of their expression via RNA interference (RNAi) followed by the analysis of the resulting 'loss-of-function' phenotype. RNAi screens are commonly used to analyze gene function in a variety of model organisms, the most popular ones being *C. elegans* and *Drosophila*. More recently, shRNA libraries targeting the human and mouse genome have become available. These libraries allow RNAi mediated 'loss-of-function' screens in mammalian cell lines. Pooled RNAi screens have been performed by several groups and revealed a number of cancer cell essential genes [128]. The decoding of such pooled RNAi screens by means of microarray analysis has been described previously [129]. While some groups employed probe sequences complementary to each shRNAs' specific 21 nt half-hairpin stem sequence others used unique barcode sequences to analyze pooled shRNA screens [126]. These 60 nt barcode sequences were cloned adjacent to each shRNA template, allowing the determination of the abundance of individual shRNA templates from a complex pool. Up until now analysis of pooled RNAi screens via barcode sequences was performed by probes complementary to the full length barcode. Here, in order to analyze pooled shRNA screens the concept of barcode tiling is introduced.

To assess the performance of the barcode tiling approach for the detection of essential genes in the breast carcinoma cell line MDA-MB-231, a negative selection screening system was established. A-apoptotic genes which were previously shown to be expressed in either breast carcinoma tissues or normal human breast were targeted. From a pooled RNAi screen, 28 different shRNA sequences were identified which were depleted from a pool of lentiviral infected cells over a period of four weeks. Potentially inhibitory as well as non-inhibitory shRNAs were selected for individual analysis of their effects on the proliferation of the cell line MDA-MB-231. Validation assays revealed the genes *BIRC5*, *BRCA1*, *HSPA8* and *NUP62* to be essential for the viability of the cell line.

# 4.2 Methods

## 4.2.1 Microarray design and hybridization

The microarray experiments, viability assays and validation of the candidate genes were performed in the group of Dr. Joerg Hoheisel, German Cancer Research Center. The Geniom One microarray [130] is divided into eight individually accessible subarrays allowing the analysis of eight samples in parallel. Half hairpin probes were synthesized in quadruplicates as 21 nt sequence as well as 25 nt sequences containing additional 4 nt from the common mir-30 sequence at their 3' end. As for the barcode sequences, probes the length of 25 nt were synthesized complementary to each 60 nt barcode. Every barcode was covered by six probes in seven nucleotide jumps (Figure 12). Three replicates of each probe were synthesized in each subarray, resulting in 18 probes representing one barcode. In total 5490 probes were synthesized to detect barcodes associated with 305 different shRNA expressing constructs. Additionally eleven half hairpin and 66 tiling probes that did not match any barcode sequence

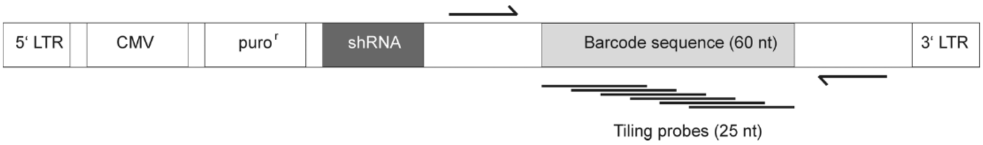


Figure 12. shRNA expression construct.

Each shRNA template is associated with a unique 60 nt barcode sequence. For analysis of pooled RNAi screens, six overlapping tiling probe sequences (25 nt) complementary to each barcode were synthesized on a Geniom One microarray.

were synthesized in triplicates as negative controls.

### 4.2.2 Negative selection screen

A negative screening system was established to detect essential genes in the breast carcinoma cell line MDA-MB-231. For that purpose, lentiviruses carrying each of the 305 different shRNA expression constructs, targeting 121 individual antiapoptotic genes were pooled. This lentiviral mix was used to infect MDA-MB-231 breast carcinoma cells at a low multiplicity of infection (MOI) of 0.3. After five days, total high molecular weight (HMW) DNA was extracted and served as a reference pool ($t_{zero}$). Another cell fraction was cultured for an additional four weeks and then subjected to HMW DNA extraction, representing the test pool ($t_{end}$). The barcode sequences from $t_{zero}$ and $t_{end}$ of the pooled screen were labeled and hybridized to two individual barcode tiling arrays. All probe signal intensities from the test pool ($t_{end}$) were normalized to the reference pool from time point zero ($t_{zero}$) by calculating the ($t_{end}/t_{zero}$) ratio.

### 4.2.3 Data analysis

Median background signal intensities were determined from half hairpin or barcode tiling probe sequences complementary to shRNA expression constructs that were absent in the analyzed pools. Signal intensities from each probe after local background subtraction were normalized to the median signal intensity of each subarray, and the mean ratio from all tiling probes representing one barcode was determined. The analysis of the negative selection screen was performed in three independent replicates.

Candidates with significant signals were identified using linear models in the limma package [123]. Coefficients, moderated t-statistics and corresponding p-values for testing all possible contrasts were calculated using Empirical Bayesian methods. Appropriate design matrices were constructed for the linear model fitting. Complete pair-wise comparisons between time points were performed by means of a contrast matrix. The p-values for the coefficients of interest were adjusted for multiple testing by means of Benjamini and Hochberg's algorithm [124], which controls the expected false discovery rate (FDR) below the specified value.

### 4.2.4 Viability assay

MDA-MB-231 cells were seeded in 96 well microplates at 300 cells per well. After 24 h, 15 µl of lentiviral supernatant (approx. 1000 units) in culture medium containing 8 µg/ml polybrene was added to the cells to achieve a MOI > 1. Twenty four hours later the viral medium was aspirated and replaced by culture medium containing 0.5 µg/ml puromycin or culture medium without puromycin, respectively. 72 h post-infection puromycin selected and non-selected cells were assayed by resazurine assay in triplicate measurements ($t_{zero}$). The fluorescence intensity ratio from puromycin selected cells divided by the intensity from unselected cells was used as quality control for efficacy of lentiviral infection. Another triplicate was allowed to proliferate in fresh puromycin culturing medium for another five days before resazurine measurement ($t_{end}$). The fluorescence intensity ratio [$t_{end}/t_{zero}$] served as a relative measure for the anti-proliferative effect of tested shRNA constructs. All values were normalized to a non-silencing control (NSC) as well as an empty-pGIPZ vector control. For the viability assays at sixteen days post infection total cells were transferred from a well of a 96 well plate to that of a six well plate at six days post infection.

## 4.3 Results and discussion

Associations between tiling probes and barcode sequences were analyzed by means of Correspondence Analysis. Correspondence analysis aims to separate dissimilar objects, in this case tiling probe sequences as well as barcode sequences, from one another [74]. Thus, similar objects are clustered together resulting in small distances, whereas dissimilar objects are located further apart. A projection of this analysis is shown in Figure 13 where signal intensities from all 305 barcodes were used to determine the association between each of the six different tiling probes representing every barcode, marked as colored squares.

As expected, contiguous tiling probes, sharing the highest similarity with one another, are located closer to each other than tiling probes sharing no sequence similarity. All barcodes, represented as black dots, are located in the projection plot according to their association with each of their six tiling probes. Strongest signal intensity from one particular tiling probe as compared to the remaining five, means strongest association of the barcode with this tiling probe. In case of a positive association of a barcode with a particular tiling probe, both objects are located in the same direction from the centroid. The larger the distance from the centroid, the stronger the association between the barcode and the given tiling probe. For negative associations, the two objects are located on opposite sides of the centroid.

An example of strong association is given by the barcode sequences from constructs BIRC5-A and HSPA8-B, highlighted in the projection. Both barcodes show a positive association with tiling probe two and, at the same time, a negative association with tiling probes four, five and six. In other words, signal intensities detected from tiling probe two were much stronger for both barcodes than signal intensities detected from tiling probes four, five and six. Interestingly, no general preference for any of the tiling probes was detected, as represented by the equal distribution of all vector profiles in the projection.

## 4.3.1 Identification of candidate genes

The depletion of a certain barcode over the time of the screen is expected to result in a decreased ($t_{end}/t_{zero}$) ratio and thus indicate that the associated shRNA targeted a gene which was essential for the proliferation of the cell line MDA-MB-231. Therefore, log2 signal intensity ratios ($t_{end}/t_{zero}$) were calculated from all signals for each tiling probe sequence individually. In total, three independent replicate microarray experiments were carried out, resulting in a maximum of nine signal intensity ratios for each tiling probe.

Tiling probes represented by less than four out of the possible nine replicate signal ratios were discarded. A summary of all determined log2 ratios and microarray data is accessible through ArrayExpress [131]. Expression constructs represented by at least two barcode tiling probes were considered for further analysis. Altogether, out of 305 shRNA expression constructs included in the pool, 278 (91%) could be analyzed.

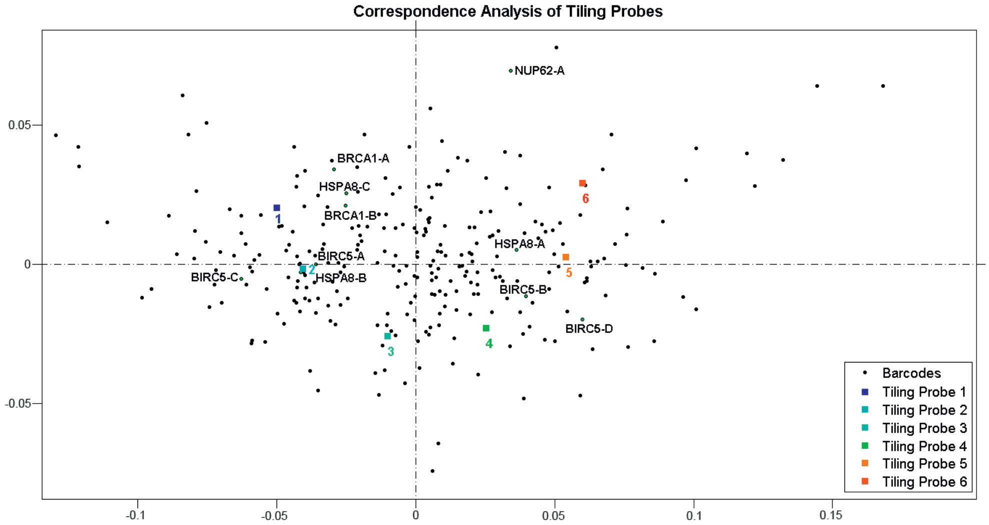


Figure 13. Correspondence analysis projection.

Colored squares represent the six tiling probes complementary to each barcode sequence. Black dots represent signal intensity profiles at time point zero from each of the barcode sequences included in the pooled screen.

A heat map of all log2 ratios is shown in Figure 14. Lines represent the 278 shRNAs sorted by the mean value of their corresponding log2 ratios from tiling probes retained after filtering.

A ranking of the mean log2 ratios, representing the abundance of each shRNA in the pool after four weeks of screening, is shown in Figure 15 (top). Those log2 ratios were then plotted against their significance. The volcano plot in Figure 15 (bottom) gives an overview of the results from the pooled screen. It shows the distribution of log2 ratios determined for each shRNA, relative to their calculated p-values. 28 candidate constructs showed negative log2 ratios together with a p-value < 0.05, indicating their depletion from the pool.

Individual validation of a subset of eleven shRNA expression constructs (Figure 15) with potential inhibitory, as well as non-inhibitory effects on the cell line proliferation provides further evidence for the accuracy of the barcode tiling approach.

## 4.3.2 Validation of candidate genes

Validation assays were performed in the group of Dr. Joerg Hoheisel, German Cancer Research Center. To verify the potential anti-proliferative effects of candidates identified through the analysis of the pooled RNAi screen, eleven shRNA expression constructs were selected for closer analysis in an arrayed 96-well format. First of all, two shRNA expression constructs, termed *BRCA1-A* and *BRCA1-B*, both encoding identical shRNA sequences targeting the expression of *BRCA1*, but associated with

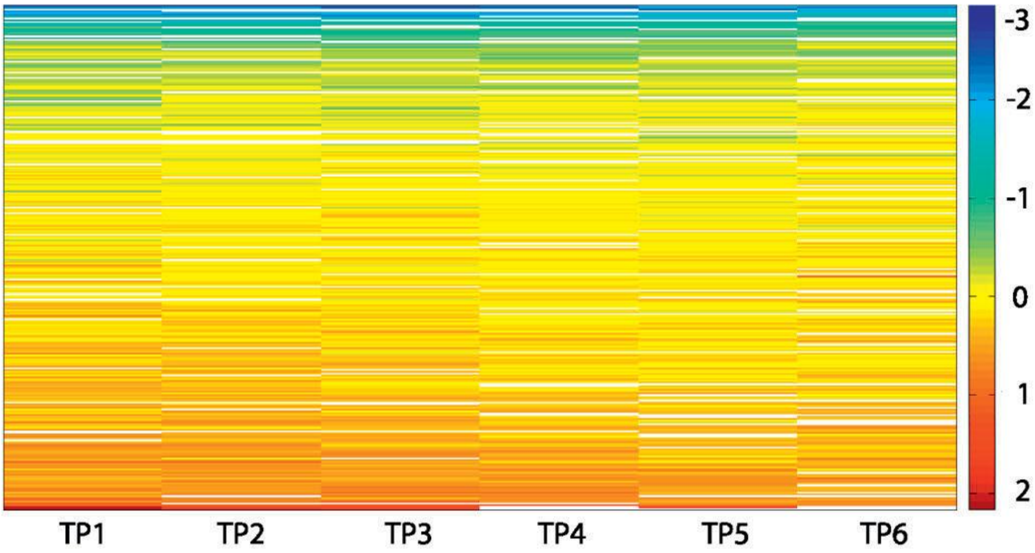


Figure 14. Cross-comparison of tiling probe performance.

A heat map was generated of log2 ratios obtained from each tiling probe (TP) that passed the filter criteria. Columns represent the six different tiling probes and lines the 278 barcode sequences retained after filtering sorted by their TP mean values. White cells represent TPs which did not pass filter criteria.

two different 60 nt barcode sequences were selected for validation. The log2 ratios from both constructs indicated a significant anti-proliferative effect [(*BRCA1-A* (-1.706, p = 1.3e-4)/*BRCA1-B* (-1.145, p = 1e-5)]. The constructs were transducted individually into the host cell line to reduce target mRNA abundance, inhibit cell viability and induce apoptosis. For *BRCA1-A* as well as for *BRCA1-B* we detected close to equal reduction of *BRCA1* expression [126].

In much the same way as for *BRCA1*, further constructs targeting expression of the genes *BIRC5* (*BIRC5-A-D*), *NUP62* (*NUP62-A-B*) and *HSPA8* (*HSPA8-A-C*) were analyzed. For each of the three genes we identified at least one construct with a significant log2 ratio below -0.5 and one construct showing a ratio greater than -0.5. Expression levels were reduced below 0.4-fold that of the non-silencing control (NSC) by at least one construct targeting each of the three mentioned genes. Cells with efficient reduction of *BIRC5* and *NUP62* expression were strongly impaired in their viability when assayed eight days post-infection (*BIRC5-A-C/NUP62-A-B*). In the case of *HSPA8*, a reduction of mRNA expression to 0.1-fold that of the NSC caused only a mild reduction in cell viability (*HSPA8-A*).

## 4.4 Conclusion

The work presented here demonstrates analysis of pooled RNAi screens by means of barcode tiling arrays. We demonstrate how pooled RNAi screens can be quantitatively and reproducibly analyzed by this method. Barcode tiling arrays have been used to predict anti-proliferative effects of individual shRNAs from pooled negative selection screens. The presented experimental approach, coupled with commercially available lentiviral vector shRNA libraries, has the potential to greatly facilitate the discovery of putative targets for cancer therapy.

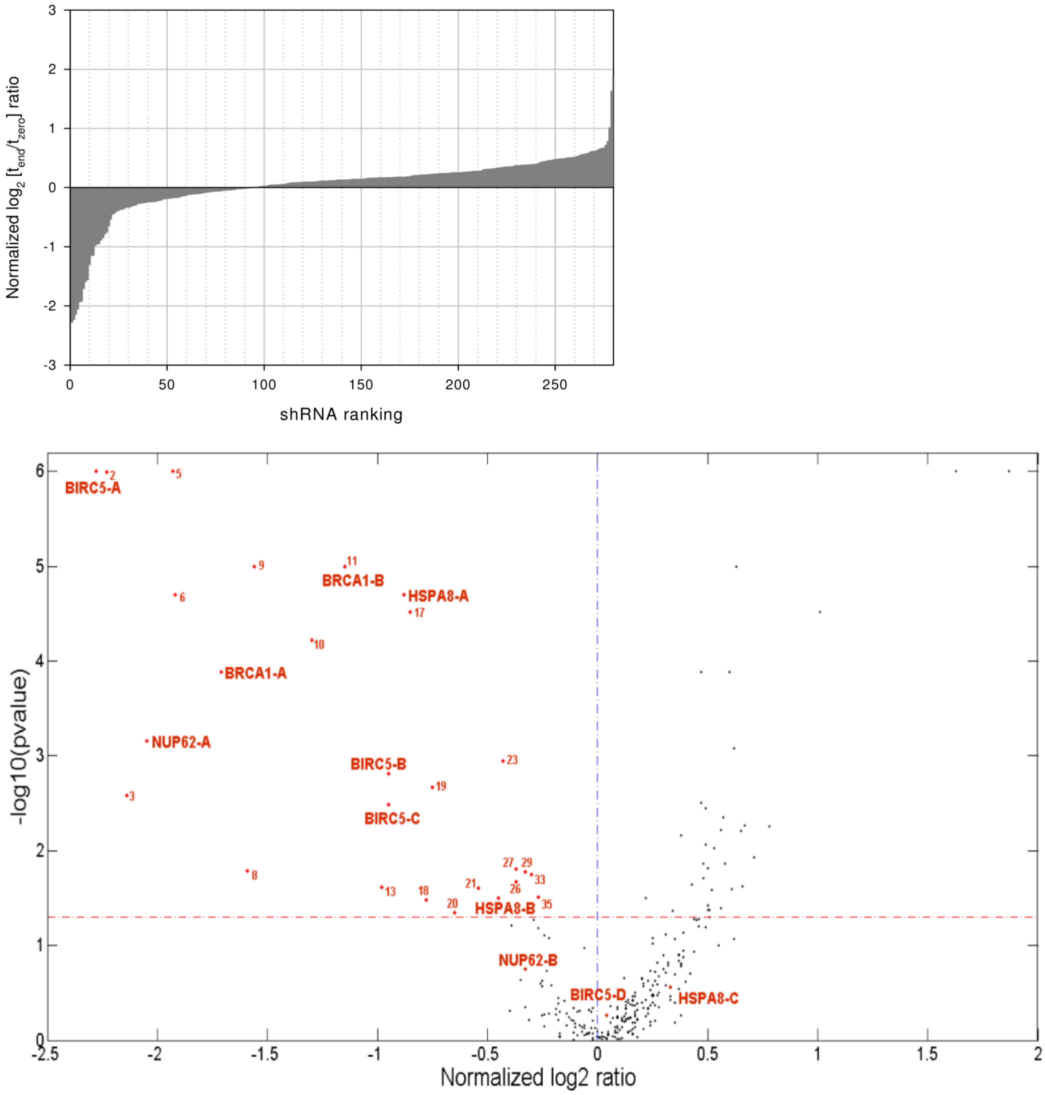


Figure 15. Overview of tilling array results.

**Top** - The log2 ($t_{end}/t_{zero}$) signal intensity ratios were calculated from all probes that passed the described filter criteria and averaged for each shRNA expression construct. Negative log2 ratios indicate the depletion of cells expressing a particular shRNA from the pool of cells, following the four weeks of the screen. **Bottom** - The log2 ratios determined for each shRNA expression construct were plotted against their significance. The red dashed line represents a p-value of 0.05 which was used as cut off. Highlighted in red are significant candidate shRNAs (numbers) and validated candidate constructs. The indicated numbers correspond to the numbers given in the Appendix 3.

# 5 CROSS-SPECIES MICROARRAY META-ANALYSIS

Combining expression data from multiple species often leads to important findings which cannot be achieved by focusing on a single species. The first chapter of this section discusses meta-analysis methods and their applications to cross-species analysis of microarray data. The following chapter summarizes studies that are carried out on individual species and then combined in a post-processing approach. A short review will be given to applications that use the same microarray to study different species. Finally, methods will be presented that use a separate microarray for each species but unlike the first set of methods the analysis of data from all species is carried out concurrently. Each of these methods has its advantages and disadvantages.

One of the major contributions of this thesis is an algorithm to estimate the common regulatory network. This chapter represents an iterative procedure combining existing methods to integrate information from different species. The efficiency of the algorithm is demonstrated by analyses of pairs of example datasets. The common regulatory network was obtained by reverse engineering the combined set. Reverse engineering a common network provides the opportunity to determine statistical significance of network motifs. Comparing reversely engineered gene regulatory networks from each individual and combined dataset with known interactions supplied by KEGG and other repositories provided an additional means to evaluate the performance of the algorithm. The algorithm and results presented in this chapter has been published [132].

# 5.1 Background

Cross-species meta-analysis can be divided into two types: co-expression meta-analysis and expression meta-analysis. These two approaches address different questions. Co-expression meta-analysis asks whether genes co-expressed in one species are also co-expressed in another species. Expression meta-analysis directly analyzes the similarity between expression profiles of homologous genes in different species. While expression analysis identifies homologous genes that respond in the same way to specific stimuli in multiple species, co-expression analysis can result in genes with very different response patterns in each of the species. Co-expression analysis allows the use of different conditions for the different species whereas expression analysis requires the use of the same conditions in all species. The next section discusses both of these methods in turn.

## 5.1.1 Co-expression meta-analysis

The first method that has been applied to study cross-species microarray experiments is co-expression meta-analysis (Figure 16). Rather than comparing expression data directly between species, evidence for gene co-expression is derived separately in each individual species, and then combined to infer gene modules. The advantage of co-expression meta-analysis is that the experiments can be combined even under different experimental conditions for different species. Stuart and coworkers [30] were among the first authors to introduce the concept of metagenes. A metagene is a set of strictly orthologous genes among multiple species. Two metagenes are defined to be co-expressed if their constituent genes are significantly co-expressed in each species. In Bergmann *et al.* [31] a ‘signature algorithm’ is presented in a way that maps an annotated gene module in one species to its set of homologous genes in other species, and then co-expressed genes are extended to a gene module that is co-expressed. Both methods identified significant co-expression sets of genes but also revealed modules of divergent co-expression between species.

Choi and coworkers [88] used a direct approach of co-expression network inference to search for differences in expression between cancer and normal tissues by comparing co-expression networks extracted from multiple studies containing expression data from the respective tissues. Co-expression networks have also been used to refine

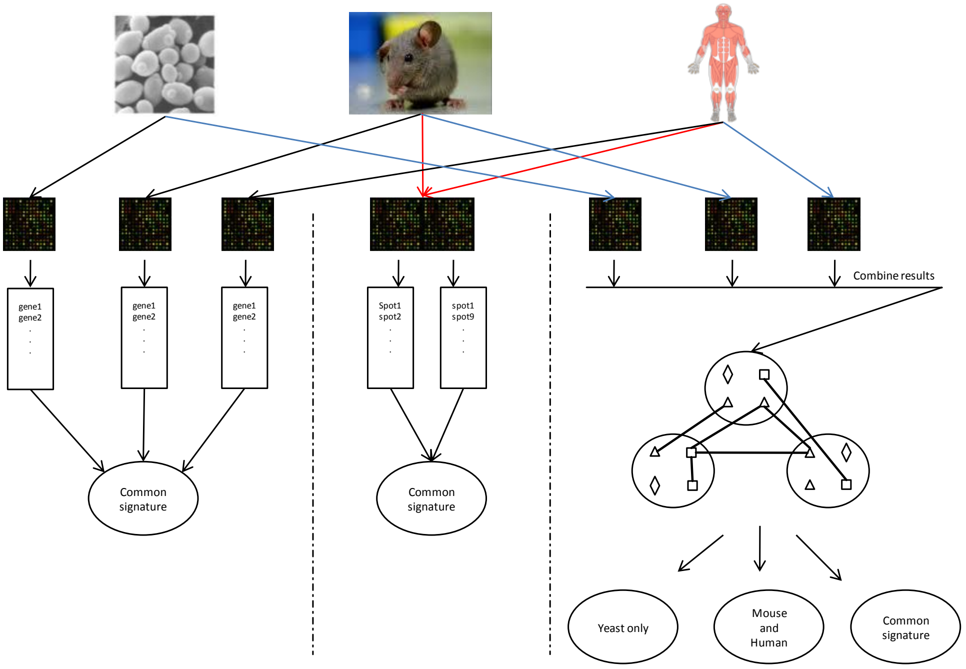


Figure 16. Overview for meta-analysis of microarray data.

**Left:** Expression meta-analysis, Samples from each species are hybridized to different arrays and each array is independently analyzed. Lists of differentially expressed genes are later compared to identify the overlap. **Middle:** using the same array for all species. Samples from all species are hybridized to the same array and all arrays are analyzed using the same method. The list of differentially expressed probes can then be compared. Note that this method can only be used to compare closely related species. **Right:** Samples from each species are hybridized to separate arrays but are analyzed together so that extra information can be used to improve the assignment of genes.

gene annotation by investigating the condition dependencies of a particular interaction in the co-expression network [89].

### 5.1.2 Expression meta-analysis

Expression meta-analysis compares the expression of orthologous genes under similar conditions. It should be noted that directly comparing the data tables is impossible due to their different scales. In contrast to what is frequently assumed, not even the ratios are comparable. Therefore, efforts have been made to compare multiple datasets on the basis of statistical significance [133]. However, these p-values cannot be taken at face value, either. Comparing lists of 'significant' genes (when using the same P-value cutoff for all studies) often results in high disagreement between studies. Because the genes identified as significant in one study might not be significant in other studies. This is especially true if the different studies use different methods to determine a P-value, which is often the case.

Most expression meta-analyses focus on comparing lists of differentially expressed genes (DEGs) reported from published papers [134, 135]. For instance, DEGs from 22 studies in different organisms were investigated by Han and Hickey [135]. They found no agreement in DEGs for different species, and very little agreement even within species.

Expression meta-analysis can so far be categorized into three approaches: effect size models, rank-based and P-value based. These methods can be adapted to cross-species data, if a "one-to-one orthology relationship" is known. Effect size methods combine the expression data for a gene from each microarray study, and then estimate the significance of the combined expression data. Error variables can be included to identify global systematic differences between different experiments. Choi and coworkers [136] developed such a method, termed t-based, and applied it to various cancer datasets. They demonstrated that consistent expression changes can be identified for some genes, which neither of the individual microarray studies alone could identify as significant. Rank-based methods sort the genes in each study according to the significance of differential expression and then aggregate the rankings of each gene across studies, which allow overcoming differences in P-value comparisons. This allows comparing the P-values. Permutation tests can then be used to identify genes whose combined rankings are significant. Finally, P-value aggregation methods combine the P-values for each gene's expression obtained from different microarray experiments, to obtain an aggregate P-value. P-values for a gene can be aggregated by various methods, e.g. by taking the minimum or the product of the P-values.

Recent studies have shown that utilizing differentially expressed genes and strict cutoff methods may lead to underestimating expression conservation. Analysis of a comparison of gene expression in more than 50 mouse and human tissues [137] identified only a small intersection between the expression of orthologous genes. However, later analysis of these data focusing on expression correlation found a considerable intersection between orthologs [138].

### 5.1.3 Indirect expression meta-analysis

Direct cross-species comparison is more difficult in distant species. Instead, Subramanian and coworkers [139] proposed gene set enrichment analysis (GSEA). In GSEA, a set of genes in a microarray are first sorted according to a differential expression score. Then, for a predefined subset of genes of interest (e.g. a GO category or pathway data), an enrichment score is calculated which essentially measures how high the scores of the top-ranked genes in the subset are among the top-ranked genes. In GSEA the sets of genes can be defined in different ways, e.g. by GO categories, results of previous microarray studies, or pathway data. This allowed comparing a list of DEGs to predefined sets of genes, to identify significantly enriched annotations. Enriched annotations can then be compared across species. Identifying enriched pathways from lists of DEGs has been shown to improve reproducibility and comparability of microarray studies of prostate cancer [140] and may also help facilitate comparisons of microarray studies across species.

Liu and coworkers [141] extended GSEA to gene network enrichment analysis (GNEA). In GNEA, a highly scoring sub-network is found within a given protein–protein interaction network. Then predefined gene sets (e.g. GO categories) are tested for enrichment within these sub-networks. Applying GNEA to type 2 diabetes datasets from human and mouse, Liu and coworkers implicated gene sets involved in insulin signaling and nuclear receptors as differentially expressed in diabetes.

### 5.1.4 Same array analysis

A problem with the above methods arises from the different probe sets used on each array which may have different hybridization properties. These probe-dependent expressions bias the estimation of gene expression peaks [142] which adds to the disagreement between species. For more efficient comparison of gene expression across species, it is beneficial to control for platform-related variations. One way to circumvent this issue is to use the same microarray to study different species.

There are two approaches for using a single array type when studying multiple species: using an array constructed for a single species [143] and constructing a customized array containing probes for every species studied [144] . The following two sections detail the technical and computational issues involved.

#### 5.1.4.1 Single species array

Microarrays constructed for one species can be used to measure gene expression in another species because orthologous genes are likely to share high sequence similarity. Thus, probes designed for a gene in one species are able to hybridize with its ortholog.

There are several advantages, especially if genomic data are available only for closely related species. For instance, Nuzhdin and coworkers used microarrays designed for

*Drosophila melanogaster* to study gene expression in *D. simulans* [145]. Another application is in preclinical cancer drug screening, where animal models with transplanted tumors are often used to study the progression of tumors, identify therapeutic targets and test treatment response. Single species microarrays can be used to validate animal model by comparing gene expression in both the animal model and the primary tumor in human tissues [146].

There are several issues to consider when using single species arrays to study multiple species. The first is to decide which species the microarrays should be based on. Because the probes are designed for one species and used to detect genes in another species, in some cases the sequence of a probe does not match that of the target gene, which may result in weaker hybridization. This sequence mismatch effect complicates the estimation of gene expression levels [147]. To alleviate this effect, it is desirable to select a species that is as close as possible to the target species [148]. Sequence mismatch effect varies for each species due to different evolutionary distances between them. Gilad and coworkers compare the sequence mismatch effect on four species (human, chimpanzee, orangutan and rhesus macaque), and show that sequence mismatch effect becomes more severe as sequence divergence increases. This variations make it more difficult to directly compare gene expressions [149].

A possible way to avoid this problem has been reported by masking probes with a sequence mismatch between orthologous genes [150]. The drawback is that when comparing more than two species, a large amount of the probes will have to be masked, limiting the number of genes that can be measured in the study [144].

#### 5.1.4.2 Multi species array

In multi-species arrays, samples from two species are competitively hybridized to the probes. The expression level of a gene is then estimated by averaging the log-ratios of probes from every species. Oshlack and coworkers show that multi-species microarrays can alleviate the problem of sequence mismatch effects, when compared with single species arrays [144]. They have shown that using the average log-ratio will eliminate the bias caused by cross-species hybridizations. These arrays have also been used to study expression of evolutionarily conserved genes [151].

Multi-species arrays are only suitable for genes with an ortholog in every species under study. Consequently, the genes become limited in distant species. Liao and Zhang determined that only less than half ortholog pairs (10,670 pairs) between human and mouse are suitable for a multi-species array study, which covers less than 40% of the known mouse genes.

### 5.1.5 Combined multiple species

Recent methods combine ideas from both single and multi species approaches. Similar to the first set of methods they use a species-specific array for each species. Similar to the second set of methods the gene expression levels are estimated by averaging the log-ratio of probes from every species. These methods have initially been applied to study the cell cycle. Early applications of microarrays to study the cell cycle focused on the identification of cycling genes in different species including budding yeast [13], human [152], plants [153] and bacteria [154]. These studies differ in the approach to determine cyclic genes but they all used similar expression meta-analysis techniques.

Alter and coworkers [155] were among the first authors to analyze cell cycle expression data from multiple species. Generalized Singular Value Decomposition (GSVD) has been used to compare human and budding yeast cell cycle experiments. The authors were able to identify more accurate cyclic expression profiles for some of the genes based on the information from the other species. Recently, Lu and coworkers [156] used Markov random fields (MRF), which is an undirected graphical model, to concurrently analyze cell cycle data from human and budding yeast. In these models both sequence and expression data are used to construct a graph in which gene expression is represented by nodes and edges represent homology. The set of edges can be derived from a curated database or from sequence analysis methods including BLAST.

Combining multiple species microarray datasets into simultaneous analysis is challenging. It is crucial to capture the associations between variables from different high-throughput multidimensional datasets. Another challenge is the need to define a one-to-one orthology relationship to carry out such an analysis**.** The binary assignment (ortholog or not) in databases cannot account for more complex similarity measures which are often represented using a more continuous value (e.g. Blast *e*-value).

My algorithm solves this problem by not requiring any orthology relations as perquisite. The next chapter discusses the multivariate analysis methods that represent the 'building blocks' of the algorithm.

# 5.2 Methods

## 5.2.1 Co-inertia analysis

Co-inertia Analysis (CIA) is a multivariate approach that can identify co-relationships within multiple datasets by finding successive principal axes of maximum co-variance. It was first introduced applying to ecological data [157]. Co-inertia is a score as a measuring co-structure between two data matrices. When the matrices are centered, co-inertia is a sum of square covariances. A formal definition is given in the next section.

Culhane and co-workers demonstrated the efficiency of CIA on cross-platform comparisons of gene expression data [73]. CIA is often used in combination with Principal Component Analysis (PCA) or Correspondence Analysis (CA), the latter being capable of visualizing genes and hybridizations at the same time [74]. Whilst, as with PCA, similarity among genes as well as similarity among hybridizations is depicted as proximity, a gene that is particularly up-regulated under a certain condition will be located in the direction of this condition. The farther away from the centroid in this direction (towards the outer margin of the plot) it is displayed, the stronger the association [73, 74]. If used together with CIA, genes and hybridizations are shown simultaneously for both datasets, projecting their common variance or co-inertia (Figure 21). Here, proximity among objects and directions can be interpreted as aforementioned, now highlighting common trends and patterns. Overall similarity of the datasets is captured by the RV-coefficient (RV) which is a commonly used matrix correlation [158]. In CIA, the RV is calculated as the co-inertia (sum of eigenvalues of a co-inertia analysis) divided by the square root of the product of the square inertias (sum of the eigenvalues) from the individual Correspondence Analyses [73]. Much like a correlation coefficient, the stronger the joint trends between two datasets agree, the closer to 1 the RV score becomes. A zero RV score indicates no similarity. Prerequisite for CIA is that either the genes or the hybridizations are affiliated between the two datasets. Therefore, either the columns or the rows of the tables must match (and have equal weights). In the following text, ‘connecting variables’ are used to refer to the variables (genes or samples) needed to be affiliated beforehand and ‘projected distance’ refers to the distances between objects in a CIA output (projection). Hungarian algorithm is used to affiliate connecting variables in CIA as detailed in next section.

### 5.2.1.1 Co-inertia calculation

The mathematical basis of CIA, following the notation of Dolédec and coworkers [73, 157, 159] is summarized below.

Let X and Y be the original data tables, with *n* rows, and respectively *p* and *q* columns. The two statistical triplets produced by the ordination methods performed on the datasets are denoted $(X, D_n, D_p)$ and $(Y, D_n, D_q)$, with $D_n$ and $D_p$ being diagonal matrices containing row and column weights for $X$, and $D_n$ and $D_q$ diagonal matrices containing row and column weights for *Y*. After diagonalization let *u* and *v* be a pair of eigenvectors for ($X, D_n, D_p$) and ($Y, D_n, D_q$), respectively. The projection of the multidimensional space associated with *X* onto vector *u* generates *n* coordinates in a column matrix:

$$\alpha = XD_p u \qquad \text{EQ 11}$$

The projection of the multidimensional space associated with table *Y* on to vector *v* generates *n* coordinates in a column matrix:

$$\psi = YD_q \, v \qquad \text{EQ 12}$$

Co-inertia associated with the pair of vectors *u* and *v* can be written as

$$\mathrm{H(u,v)} = \alpha^t \mathrm{D} \psi \qquad \text{EQ 13}$$

If the initial data tables are centered, then the co-inertia is the covariance between the two new scores:

$$Cov(\alpha,\psi) = Corr(\alpha,\psi)\sqrt{\eta_1(u)\eta_2(v)} \qquad \text{EQ 14}$$

with $\eta_1(u)$ denoting the projected inertia on to vector *u* (i.e. the variance of the new scores on *u*), $\eta_2(v)$ the projected inertia on to vector *v* (i.e. the variance of the new scores on *v*), and $Corr(\alpha,\psi)$ the correlation between the two coordinate systems. A CIA axis associated with a pair of eigenvectors *u* and *v* will maximize $Cov(\alpha,\psi)$.

### 5.2.1.2 Co-inertia affiliations

As aforementioned, measuring the associations between samples by CIA requires affiliation of each gene from one dataset to one gene of the other dataset as a prerequisite. Or, projecting common variance between the genes of both datasets requires prior matching between the samples of the two datasets. Either the columns or the rows of the tables (connecting variables) must be matchable and have to be weighted similarly. As a basis for a successful co-inertia analysis, this matching needs to be both complete and reliable. Hungarian algorithm can be used to affiliate connecting variables in CIA.

## 5.2.2 Hungarian algorithm matching

Two sets of objects (here genes or samples of the two datasets to be combined) can be matched by the Hungarian algorithm, also called Kuhn–Munkres algorithm [160]. It takes as input a penalty weight matrix of all possible pairwise projected distances and computes the pairs summing up to minimal penalty (Figure 17). The original publication refers to a quadratic penalty matrix. However, the Hungarian algorithm

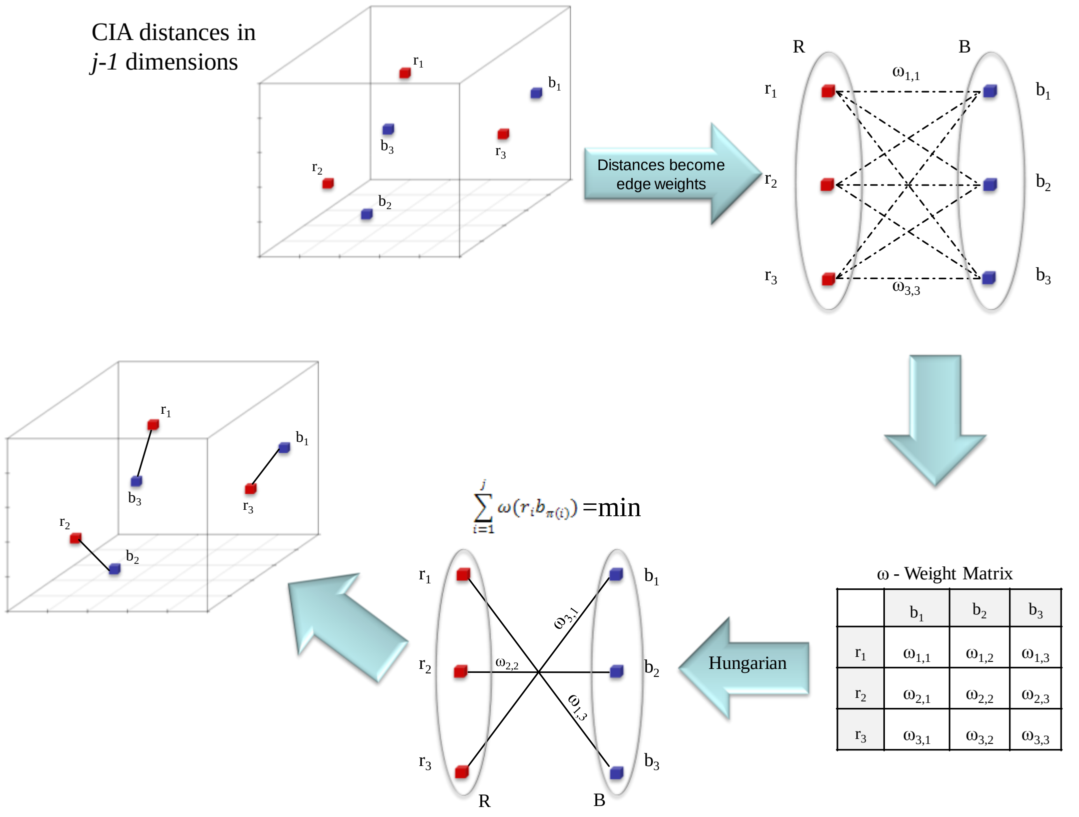

CIA distances in
*j-1* dimensions
$r_1$
$b_1$
$b_3$
$r_3$
$r_2$
$b_2$
Distances become edge weights
R
B
$\omega_{1,1}$
$\omega_{3,3}$
$\sum_{i=1}^{j} \omega(r_i b_{\pi(i)}) = \min$
$\omega_{3,1}$
$\omega_{2,2}$
$\omega_{1,3}$
Hungarian
ω - Weight Matrix
$\omega_{1,1}$ $\omega_{1,2}$ $\omega_{1,3}$
$\omega_{2,1}$ $\omega_{2,2}$ $\omega_{2,3}$
$\omega_{3,1}$ $\omega_{3,2}$ $\omega_{3,3}$

Figure 17. Affiliation of the connecting variables using Hungarian algorithm.

Samples from datasets R and B are represented as red and blue squares, respectively. Only samples are projected into 3-dimensional space for simplicity (top left). In the bipartite graph (top right) edges correspond to all pair-wise projected distances (weights) from every element of R to all elements of B. Each edge corresponds to one element in the weight matrix ω recording these distances. The Hungarian algorithm computes a matching of minimal distances (lower bipartite graph and lower 3d plot). Figure from [132] by permission of Oxford University Press.

can also be applied to sets of different cardinalities by adding virtual objects of highest penalty to the smaller set until its cardinality matches the larger one [161]. Here, virtual genes (or samples) have been added to the penalty matrix showing the maximum of all occurring pairwise projected distances to all other genes (or samples).

**Distance matrix.** Given two microarray datasets, let us assume $r$ samples $\{r_1, r_2, \ldots, r_j\}$ of dataset R and $b$ samples $\{b_1, b_2, \ldots, b_j\}$ of dataset B. Projection of CIA sample distances on $j-1$ dimensions can be plotted (Figure 17, top left). $\omega$ could be an $(i \times j)$ distance matrix derived from CIA recording the distance $\omega(r_i b_j)$ between each pair of elements of R and B.

$$\omega = \begin{bmatrix} r_1 b_1 & \cdots & r_1 b_j \\ \vdots & \ddots & \vdots \\ r_i b_1 & \cdots & r_i b_j \end{bmatrix} \tag{EQ 15}$$

**Weighted bipartite graph.** The above distance matrix can be seen as a bipartite graph where each vertex belongs to dataset R or B, and each edge corresponds to one element of the distance matrix ($\omega$) representing all inter-set distances of the CIA coordinates.

Here, the affiliation problem could be stated as given an $\omega$, find $j$ independent elements of permutation π of $\{1, \ldots, j\}$ such that the sum of edge weights (EQ 16) is minimal for the selected edges.

$$\sum_{i=1}^{j} \omega(r_i b_{\pi(i)}) \tag{EQ 16}$$

Given a weighted bipartite graph where edge $r \to b$ has weight $\omega(rb)$, the optimal assignment minimizes the overall weights. Figure 17 shows a graphical representation of the above.

## 5.2.3 K-means clustering

Data can be subdivided into pre-defined numbers of homogenous gene or sample (array) clusters by the k-means algorithm. K-means can be performed on the $\chi^2$ (*Chi-square*) distance, the same distance measure that governs CIA (see section 2.3.1.1.

## 5.2.4 Majority voting

The optimal assignment produced by Hungarian algorithm allows connecting each element of one dataset to one element of the other dataset minimizing the overall weights. While Hungarian algorithm can be used to address the affiliation problem, there is still need to compute the cluster affiliations of these datasets from those of the individual cluster members. Given two datasets with *k* number of clusters each, any two clusters (across datasets) with the highest number of connections between their components become paired. Here, the pairing problem can again be solved by Hungarian algorithm, negating the number of pairings to yield a penalty matrix as input (Figure 18).

## 5.2.5 Back-transformation

CIA projection reduces the dimensionality of the original data tables to a few principal axes of maximum co-variance. While an, e.g. two-dimensional projection is ideal for visual inspection, the corresponding data table of only two rows (or columns) would be too small for any reverse engineering GRN method. However, CIA results can be back-transformed to yield tables of the original format whose content is solely based on the selected eigenvectors [162]. Mathematical basis of the back-transformation method following the notation of [162] is described below. Given a data table of *X* with *n* rows, *p* columns and *nf* kept axes, the approximated data table can be obtained from the following equations:

$$K\,\Lambda_{[r]}^{1/2}A^t = KK^tDX \tag{EQ 17}$$

And with the left multiplication of $K^tD$ we will have:

$$K^tDK\Lambda_{[r]}^{1/2}A^t = K^tDX \tag{EQ 18}$$

$$K\Lambda_{[r]}^{1/2}A^t = X \tag{EQ 19}$$

where D is a vector of row weights with length *n* . Λ is a diagonal matrix of eigenvalues with length *r. r* is called the rank of the diagram where the nonzero eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_r > 0$ are stored in the diagonal matrix $\Lambda_{[r]}$. K is a data matrix of *n* rows and *nf* columns and A is a matrix with the principal axes of *p* rows and *nf* columns. The details for reconstitution of these data are described by [162]. The derivation of the duality diagram concept is also described by [157, 163].

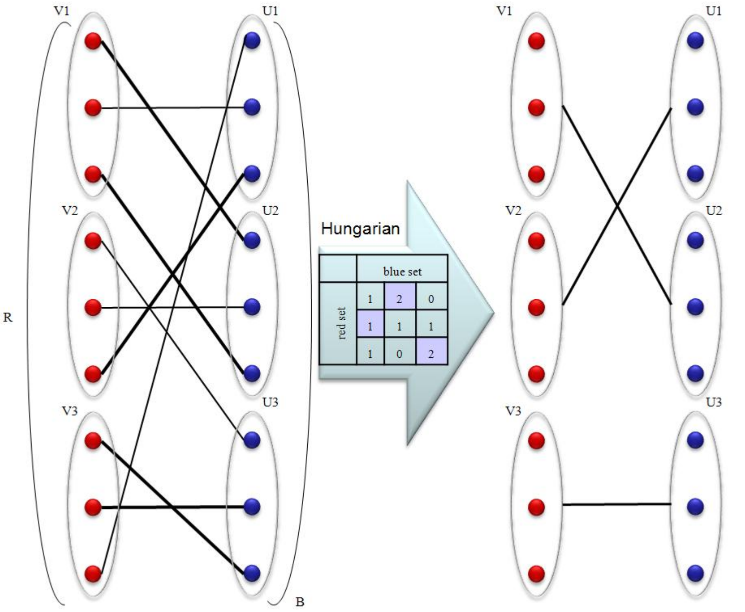


Figure 18. Graphical representation of Majority voting.

Two datasets (R and B) are subdivided into same number of clusters. Cluster components from R and B are represented as red and blue dots, respectively. In the bipartite graph (left) edges correspond to a connection between elements of R and B. The Hungarian algorithm computes an inter-cluster matching of maximal number of connections.

### 5.2.6 Dynamic Bayesian network

Banjo (Bayesian Network Inference with Java Objects) was used for the inference of Dynamic Bayesian Networks. It is freely available, easy to use and has only a few prior parameters to adjust. It focuses on score-based structure inference. For each network structure explored, the parameters of the conditional probability density distribution are inferred and an overall network's score is computed using the Bayesian Dirichlet scoring metric (BDe). In Banjo, heuristic approaches, such as greedy with random restart or simulated annealing, are used to search for the highest scoring graph among a set of networks. The output network will be either the top graph (highest score) or consensus network. The consensus network is computed based on the N top-scoring networks by assigning exponentially weighted probabilities to the individual edges in each of the high-scoring networks, based on the ranking of each network in the set. The probability of edges being present in the consensus network is computed using the weighted average approximation among N highest scoring models. The background for the concept of the consensus graphs is described by [164].

Banjo was run using default parameters (Appendix section 8.1). To identify robust interactions among a set of top-scoring networks, the consensus network was used. The output was rendered with *dot*, a graph layout visualization tool by AT&T[1]. Since it is possible to run java from within MATLAB, BANJO release 2.0 was run in MATLAB version 6.0.4 (R12). DBN algorithm performance was compared when each model dataset was discretized into three, four and five bins. Regulatory networks close to a "true" network compiled from KEGG and other databases were obtained when three categories were selected.

---

[1] http://www.graphviz.org/

## 5.2.7 Evaluation

To assess sensitivity, specificity and accuracy of the approach, the resulting gene networks were compared to a gold-standard of known interactions. A true positive (TP) was counted as interaction that is present both in an observed and an expected network, a false positive (FP) for any edge that was predicted in the learnt network but does not exist in the expected network, a false negative (FN) as an edge that is present in the expected network but not in the learnt network, and a true negative (TN) when an interaction does not exist in either learnt or expected networks. To construct an expected network, we merged all pathways involved in our gene lists into a new graph containing all nodes and edges. Therefore, the expected network represents comprehensive regulatory paths and physical interactions, accounting for the fact that many KEGG [165] pathways embed other pathways. Ingenuity Pathway Analysis (IPA)[2] was used to account for experimental findings reported in a variety of data resources, such as BioGRID, IntAct, MINT, KEGG and others as detailed in Appendix 8.2. In the expected network, all edges were supported by at least one published reference or from canonical information stored in the protein interaction databases.

### 5.2.7.1 Significance analysis of network motifs

To uncover the structural design principles of the reversely engineered GRN, the comprised network motifs were assessed. Network motifs are patterns occurring significantly more frequently than at random [166] in complex biological networks. A large number of comprised motifs indicate authenticity and robustness. Motif detection was carried out using a so-called *rand-esu* algorithm [167], generating the random networks from the reversely engineered consensus network by a series of edge switching operations as the default randomization model. Ten million random

---

[2] http://www.ingenuity.com/products/pathways_analysis.html

networks were searched to obtain a comparison to the consensus network. The higher the number of randomized networks, the more accurate the results. Significance analysis of the motifs was carried out by comparing the occurrence of a motif in the consensus network to the occurrence of the same motif in the randomized network. *Z*-scores were calculated as the occurrence of a motif in the consensus network minus its random frequency divided by the standard deviation in random networks. The higher the *Z*-score, the more significant is a motif. *P*-values correspond to the number of random networks in which the motif occurred more often than in the original network, divided by the total number of random networks.

## 5.3 Algorithm and results

### 5.3.1 Algorithm

An overview is given in Figure 19. Starting on a pair of preprocessed datasets A and B of no particular numbers of genes and also differing in the numbers of samples (arrays), methods described in the previous section (CIA, Hungarian matching and k-means clustering) are iteratively applied in the following manner (Algorithm 1).

**Initialization:** As an initial step, A and B are (separately) divided into *n* gene clusters, each. For the results presented, I initialized with *n=3*.

Each cluster is represented by its centroid (weighted average) as if it were only one gene representing a typical transcription profile for this cluster. Each cluster centroid of A is paired with one cluster centroid of B. There are *n!* possible ways to combine A and B, each of which is subjected to CIA to determine the one of highest co-inertia. This affiliation, albeit of low granularity (only three connections), is used as a starting point for iteration.

**Iteration:** The remaining procedure consists of two consecutive parts that are iterated with increasing *n* (until *n* reaches the number of samples). Both parts are identical in that they take as input an existing CIA, using its projected distances as weight matrix for Hungarian matching and let the resulting matches vote for cluster affiliations that are in turn basis for the next CIA. However, the two parts differ in that the first part starts on an affiliation of sample clusters in order to improve affiliation of gene clusters and vice versa. This is implemented by calling ‘doMatching’ subroutine.

The first part uses the previously performed CIA, collecting the projected distances between the samples of A and B into a sample(A) × sample(B) matrix which is then subjected to the Hungarian algorithm as a penalty matrix. The resulting matching preferentially pairs samples of low distance (resembling co-ordination). Subsequently, samples of A and B are separately clustered into *n* sample clusters. Each sample

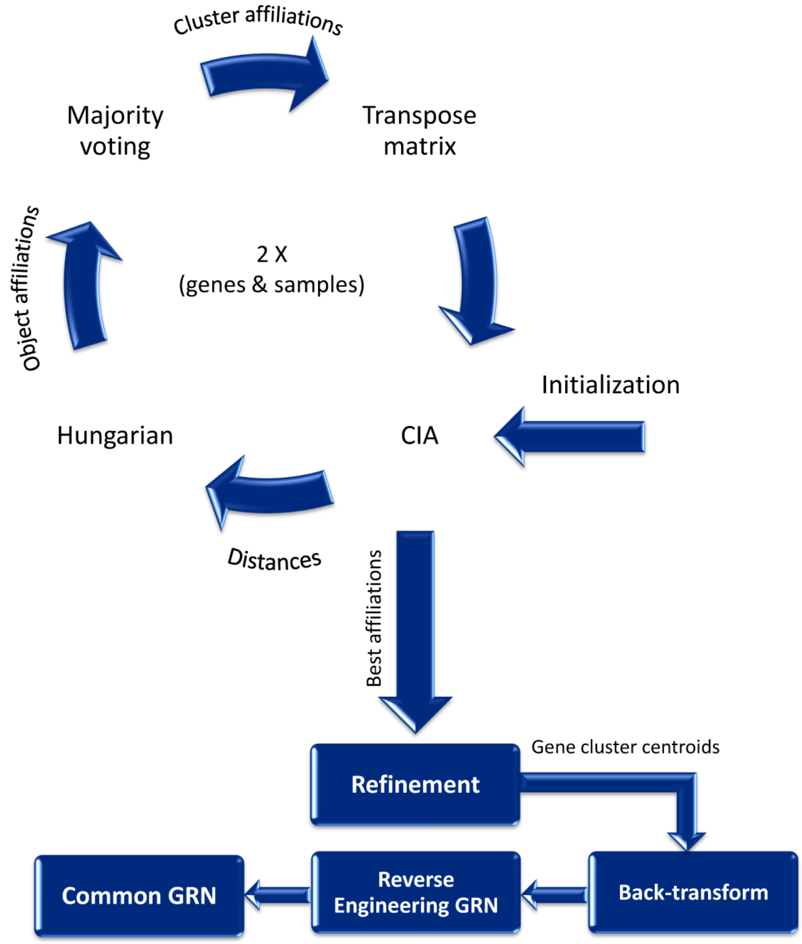


Figure 19. Overview.
Figure from [132] by permission of Oxford University Press.

Algorithm 1. Pseudocode

```
Input: Preprocessed tables A and B of microarray data, differing in numbers of
genes (rows) and samples (columns)
Initialization:
        Cluster each table (A and B) into same small no. of gene clusters n (e.g. 3)
        Represent each cluster by its centroid
        Affiliate each cluster centroid of A to a cluster centroid of B yielding a
        pairing
        for each possible pairing do
            Use these gene cluster pairs as connecting variables for a CIA of
            samples
            to identify the pairing with highest co-inertia
        end for
Iteration:
        while number of clusters <= no. of samples do
            while RV increases or remains constant do
                    Call doMatching ( samples, connecting variables )
                    Call doMatching ( genes, connecting variables )
            end while
            increase number of clusters n by 1
        end while
  doMatching:   Compute weight matrix (objects × objects) containing
                penalties for high distances
                Use Hungarian algorithm to compute optimal matching
                between objects
                Cluster each data set into n clusters
                Affiliate cluster centroids by majority voting of object matches
                Use these pairs as connecting variables for the next CIA
                Return : RV, connecting variables
  Refinement:   Decider: define the number of clusters to be matched based
                on silhouette values
                Rearrangement: Recall doMatching for m gene clusters out of
                decider module, obtaining m paired gene cluster centroids

→ back-transformation → reverse engineering GRN
→ verification of common model
```

cluster is represented by its cluster centroid (typical sample) and each cluster centroid of A is paired to a cluster centroid of B. The pairs are determined by majority voting of above matches, i.e. any two clusters with the highest number of connections between the comprised samples (arrays) become paired. The paired sample cluster centroids serve as connecting variables for a CIA projecting the genes.

The second part uses these projected distances between the genes of A and B, collecting them into a gene(A) × gene(B) matrix which is then subjected to the Hungarian algorithm as penalty matrix. All operations of the second part resemble those in the first part, but the roles of genes and samples are switched. In practice, the second part can be performed after transposing both A and B. Please note that the two parts are consecutively iterated until the co-inertia stops increasing (inner loop) before increasing *n*. The algorithm terminates as *n* approaches the number of samples (arrays) of the smaller data set.

**Refinement:** The motivation for this step is to allow a larger n (exceeding the number of samples) for the genes. ‘Refinement’ consists of two modules, ‘Decider’ and ‘Rearrangement’. The ‘Decider’ determines whether the number of clusters proposed by the iteration part is accepted as the optimum or if there is room for improvement by further increasing *n* for the genes. The choice of the decider is tightly connected to the Silhouette values [69] of gene clusters. If a larger *n* (maybe even larger than the number of samples) improves clustering, ‘Decider’ will proceed to determine the optimum number of clusters. Subsequently, ‘Rearrangement’ generates *m* pairs of gene cluster centroids by calling the ‘doMatching’ subroutine.

**Reverse Engineering gene regulatory networks:** For the resulting gene cluster centroids, the CIA coordinates are back-transformed into a data table. Its format and scale resemble that of a conventional microarray data table but it comprises only the variance that is common to both input data tables (of gene cluster centroids). It is subjected to DBN inference resulting in a graph each node of which represents a (cross-species) pair of gene clusters, while its edges stand for inter-dependencies detected for both species.

Back-transformation, DBN, as well as motif analysis were not used as parts of the algorithm. Apart from that revealing the underlying common gene regulatory network can be rewarding in and of itself, the resulting networks served as a means to validate my algorithm.

## 5.3.1 Single species application

To demonstrate applicability of the algorithm, its performance on two independent but closely related experiments on the same species is presented. These datasets provide more prior knowledge and direct evidence for correct matching of genes and samples of both datasets. The matching algorithm performed on two *S. pombe* cell cycle experiments described by [35] and [14]. Both samples and genes were permuted for one of the two *S. pombe* cell cycle experiments. The algorithm was performed after estimating the periodic genes in the cell cycle.

### 5.3.1.1 Preprocessing

Two different laboratories used DNA microarrays to study periodic gene expression of the fission yeast cell cycle. The normalized datasets named “elutriation 1” [14] and “elutriation A” [35], each relating to individual experiments were used. In the following text these data sets will be referred to as *‘ds1’* and *‘ds2’,* respectively. For perfect synchronization, the first cell cycle period (10 time points) was selected from both datasets. The total number of genes found to oscillate with a FDR of 0.1 in the first study was 1060, whereas in the second study 360 genes were identified. Of those, 337 were found in both. These genes, as well as the samples, were permuted (anonymized) and used as an input to the algorithm.

### 5.3.1.2 Running the algorithm

The algorithm terminated after 12 iterations, with maximal two inner loops, resulting in an RV coefficient of 0.9356. In theory, the optimal pairing could be determined by evaluating all possible pairings as in the initialization step. However, $n \times n!$ evaluation steps are not feasible for larger $n$. This would require $3.6\times10^7$ instead of the 35 CIA executed until convergence.

The result was visualized by CIA (Figure 20). The first two (x and y) axes of '*ds1*' explain 72% and 25% of the total inertia, respectively. The first two axes of '*ds2*' represent 71% and 22% of the total variance within '*ds2*'. Thus more than 90% of the variance of the CIA was accounted for by the first two co-inertia axes and thus presents a good summary of the co-structure between the two datasets.

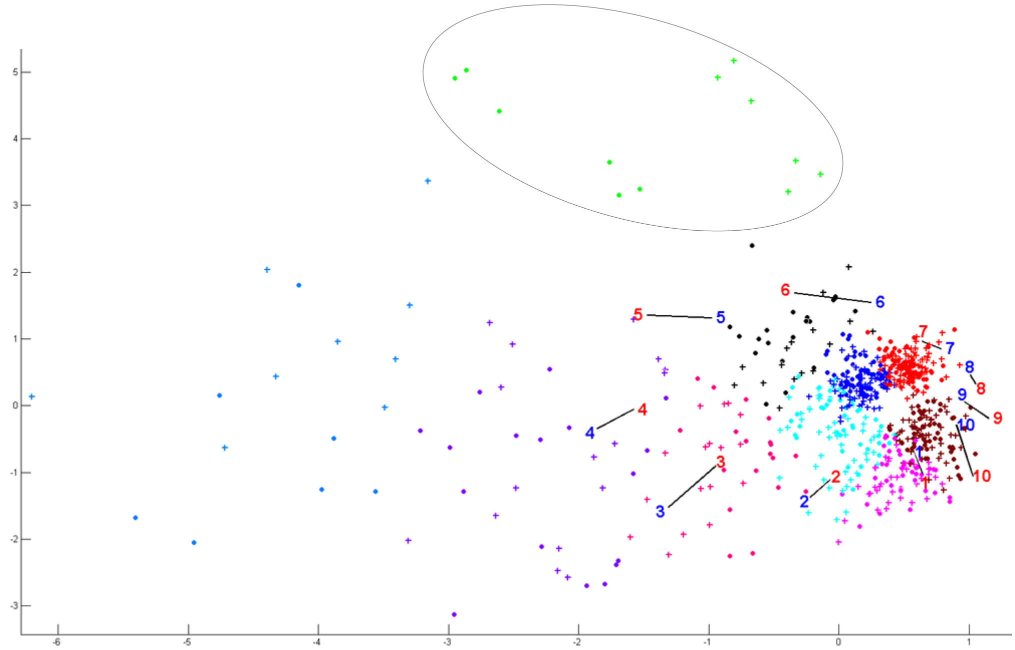


Figure 20. '*ds1*' and '*ds2*' projected by CIA.

Affiliated samples of both datasets are connected by lines. Each red or blue number represents samples (time points), and each dot or '+' depicts a gene of '*ds1*' or '*ds2*', respectively. Affiliated gene clusters are highlighted in same colors. Histones are encircled in grey. Figure from [132] by permission of Oxford University Press.

### 5.3.1.3 Refinement

A maximum of 10 well separated gene clusters were obtained in the refinement step with overall silhouette values of 0.42 and 0.44 for '*ds1*' and '*ds2*', respectively. Increasing the number of gene clusters from 10 to 17, the overall silhouette values would almost remain the same (differing by less than 0.01 for each dataset). However, the best RV was obtained for n=10. This suggests an optimum similarity when datasets are clustered into 10 gene clusters only.

#### 5.3.1.4 Co-inertia on cluster affiliations

In the shown example the algorithm terminated with an RV coefficient of 0.9356 and with 10 affiliated clusters. The algorithm was able to reconstruct correct affiliations of all samples (Figure 20) as well as for 87% of all genes (Table 2).

### 5.3.2 Cross species application

In order to further demonstrate the performance of the algorithm, it was applied to two yeast cell cycle studies [13, 14] comprising nearly identical experimental conditions (*Saccharomyces cerevisiae, Schizosaccharomyces pombe*) and two estrogen-regulated gene expression studies of *Homo sapiens* and *Mus musculus* [168, 169]. These datasets were selected to support general applicability of the algorithm to different levels of similarity, underlying structure and experimental platform (both two-channel cDNA and Affymetrix chips). Each dataset was pre-processed separately.

#### 5.3.2.1 Preprocessing

Spellman and coworkers recorded mRNA levels for 6,178 open reading frames (ORFs) of *Saccharomyces cerevisiae* over two cell-cycle periods in a yeast culture synchronized initially in the cell-cycle stage M/G1 at 7 minute intervals for 119 minutes. Rustici and coworkers monitored mRNAs whose levels oscillate during the cell-cycle for 6,978 ORFs of *Schizosaccharomyces pombe* as a function of time in cells synchronized through centrifugal elutriation for 285 minutes and temperature-sensitive cell-cycle mutants for 270 minutes at 15 minute intervals. Both datasets were recorded on glass-slides using two-channel fluorescent labeling.

Table 2. Affiliated clusters of *S. pombe* datasets

| *'ds1'* gene clusters | *'ds2'* gene clusters | *'ds1'* %† | *'ds2'* % |
|---|---|---|---|
| ***hht1,h4.1,h3.2,h4.3,ams3,htb1*** | ***hht1,h4.1,h3.2,h4.3,ams3,htb1*** | 100% | 100% |
| **SPAC11E3.10,arb1SPAC1565.02c,mug99,ssl3,fin1,pkd2,SPAC24C9.05c,SPAPB21F2.01,SPAC26A3.11,spk1,SPAC343.13,ppc89,SPAC513.07,cfr1,ndk1,spn2,SPAC823.13c,SPAC926.06,cdc4,SPBC1271.03c,top1,arg7,SPBC31F10.16,cdc13,mto2,apc2,cyp4,emp24,bis1,SPCC553.07c,SPCC553.12c,sly1,psy1**,SPAC343.20,SPAC1002.17c,SPAC23A1.02c,pom1,SPBC18E5.07,SPBC21B10.07,pob1 | **SPAC11E3.10,arb1SPAC1565.02c,mug99,ssl3,fin1,pkd2,SPAC24C9.05c,SPAPB21F2.01,SPAC26A3.11,spk1,SPAC343.13,ppc89,SPAC513.07,cfr1,ndk1,spn2,SPAC823.13c,SPAC926.06,cdc4,SPBC1271.03c,top1,arg7,SPBC31F10.16,cdc13,mto2,apc2,cyp4,emp24,bis1,SPCC553.07c,SPCC553.12c,sly1,psy1**,pho2,SPBC1773.02c,SPBC1773.03c,mob1,SPBC215.11c,SPCC1840.04,coq10,SPCC1259.12c,SPCC1223.09,aph1,SPCC1020.07 | 83% | 76% |
| **adg2,eng1,rpc17,cig2,adg1,cfh4,adg3** | **adg2,eng1,rpc17,cig2,adg1,cfh4,adg3**,SPAC644.05c,ams2 | 100% | 78% |
| **cdc42,pub1,SPAC1486.06,SPAC1639.01c,SPAC17G8.11c,SPAC18G6.09c,SPAC19B12.06c,nup40,pas1,rer1,SPAC23D3.07,mug66,tbp,pam2,cam1,ptc4,SPAC4G9.15,SPAC57A10.09c,acp2,SPAC644.16,pro1,sod2,SPBC119.10,SPBC119.17,SPBC13E7.08c,myo1,erg6,ksp1,trp4,rsc9,ape1,SPBC21B10.09,suc22,psm1,php5,dsd1,lac1,mts2,SPBC428.19c,SPBC543.02c,SPBC902.04,SPCC1020.08,alp8,dga1,SPCC1450.02,SPCC1620.08,taf50,caf1,cwf13,spp27,ins1,SPCC306.08c,cap,SPCC364.07,SPCC4F11.03c,atg1,prp28,bpb1**,SPAC6B12.07c,sfp1,SPBC660.15,his2,pho2,suc1,cwl1,SPBC1271.09 | **cdc42,pub1,SPAC1486.06,SPAC1639.01c,SPAC17G8.11c,SPAC18G6.09c,SPAC19B12.06c,nup40,pas1,rer1,SPAC23D3.07,mug66,tbp,pam2,cam1,ptc4,SPAC4G9.15,SPAC57A10.09c,acp2,SPAC644.16,pro1,sod2,SPBC119.10,SPBC119.17,SPBC13E7.08c,myo1,erg6,ksp1,trp4,rsc9,ape1,SPBC21B10.09,suc22,psm1,php5,dsd1,lac1,mts2,SPBC428.19c,SPBC543.02c,SPBC902.04,SPCC1020.08,alp8,dga1,SPCC1450.02,SPCC1620.08,taf50,caf1,cwf13,spp27,ins1,SPCC306.08c,cap,SPCC364.07,SPCC4F11.03c,atg1,prp28,bpb1**,SPAC607.07c,SPAC23A1.02c,pom1,SPAC977.09c,SPBC18E5.07,SPBC21B10.07 | 88% | 89% |
| **SPAC17H9.18c,cdc22,bet1,SPAP27G11.01,spo12,myo3,SPAP14E8.02,mid2,exg1,cdc18,chs2,rid1,SPBC27.04,SPCC1322.10,rad21**,etd1,SPAC644.05c,ams2 | **SPAC17H9.18c,cdc22,bet1,SPAP27G11.01,spo12,myo3,SPAP14E8.02,mid2,exg1,cdc18,chs2,rid1,SPBC27.04,SPCC1322.10,rad21**,lad1 | 83% | 94% |
| **SPAP4C9.01c,SPAC1F7.10,alg1,SPAC26H5.02c,uch1,SPAC29B12.13,SPAC29E6.05c,SPAC2E1P3.01,SPAC2F3.05c,SPAC513.06c,SPAC6B12.05c,SPAC7D4.05,SPAC7D4.13c,SPAC8C9.16c,SPAC922.07c,SPAPYUG7.06,rex2,nta1,SPBC25B2.08,SPBC409.17c,tas3,apc15,mog1,SPCC285.04,SPCC2H8.05c,SPCC320.14,tip41**,res1,csn3,fta1,SPCC1840.04,coq10,SPCC1259.12c,SPCC1223.09,aph1,SPBC1773.02c,SPBC1773.03c,SPBC215.11c | **SPAP4C9.01c,SPAC1F7.10,alg1,SPAC26H5.02c,uch1,SPAC29B12.13,SPAC29E6.05c,SPAC2E1P3.01,SPAC2F3.05c,SPAC513.06c,SPAC6B12.05c,SPAC7D4.05,SPAC7D4.13c,SPAC8C9.16c,SPAC922.07c,SPAPYUG7.06,rex2,nta1,SPBC25B2.08,SPBC409.17c,tas3,apc15,mog1,SPCC285.04,SPCC2H8.05c,SPCC320.14,tip41**,SPCC584.03c,thp1,fft1,SPAC17A5.09c,SPAC1002.17c,SPAC4H3.14c,SPBC19G7.07c,mug117 | 71% | 77% |
| **mde6,SPAC1705.03c,hri1,pol1,slp1,nrm1,fkh2,klp5,SPBC31F10.17c,rum1,SPBC32F12.10,msh6,SPCC338.08,SPCC63.13,SPCC757.12**,SPAP27G11.01,SPAC2E1P5.03,SPBC83.18c,mob1 | **mde6,SPAC1705.03c,hri1,pol1,slp1,nrm1,fkh2,klp5,SPBC31F10.17c,rum1,SPBC32F12.10,msh6,SPCC338.08,SPCC63.13,SPCC757.12**,etd1,SPAC343.20,lad1 | 79% | 83% |
| **zpr1,spb1,SPAC16C9.03,ssr1,uap56,sup45,rrn3,SPAC19A8.07c,SPAC1B3.13,SPAC1F7.02c,SPAC20G8.09c,SPAC212.10,lys3,SPAC23C4.05c,sxa1,SPAC26H5.07c,hal4,rbp28,prh1,SPAC6F12.16c,SPAC6F6.03c,ppa1,SPAC890.05,SPAC926.08c,SPAPB17E12.14c,pdr1,mae1,SPAPB8E5.07c,SPBC11G11.03,fkbp39,SPBC13G1.09,SPBC14F5.06,SPBC16H5.08c,SPBC17D1.05,tif213,edc3,ppp1,utp10,grn1,SPBC365.14c,SPBC3D6.12,nuc1,SPBC4F6.13c,SPBC4F6.14,int6,uvi15,nog1,SPBP8B7.20c,SPBPB10D8.04c,SPCC1183.07,SPCC11E10.07c,SPCC1442.04c,SPCC1672.07,SPCC18.05c,tif6,cgs2,SPCC320.08,SPCC330.09,cfh2,SPCC550.11,SPCC550.15c,SPCC63.06,SPCC663.10,rnc1,SPCC830.08c,SPCP1E11.08,SPCP1E11.11**,SPAC607.07c,sds22 | **zpr1,spb1,SPAC16C9.03,ssr1,uap56,sup45,rrn3,SPAC19A8.07c,SPAC1B3.13,SPAC1F7.02c,SPAC20G8.09c,SPAC212.10,lys3,SPAC23C4.05c,sxa1,SPAC26H5.07c,hal4,rbp28,prh1,SPAC6F12.16c,SPAC6F6.03c,ppa1,SPAC890.05,SPAC926.08c,SPAPB17E12.14c,pdr1,mae1,SPAPB8E5.07c,SPBC11G11.03,fkbp39,SPBC13G1.09,SPBC14F5.06,SPBC16H5.08c,SPBC17D1.05,tif213,edc3,ppp1,utp10,grn1,SPBC365.14c,SPBC3D6.12,nuc1,SPBC4F6.13c,SPBC4F6.14,int6,uvi15,nog1,SPBP8B7.20c,SPBPB10D8.04c,SPCC1183.07,SPCC11E10.07c,SPCC1442.04c,SPCC1672.07,SPCC18.05c,tif6,cgs2,SPCC320.08,SPCC330.09,cfh2,SPCC550.11,SPCC550.15c,SPCC63.06,SPCC663.10,rnc1,SPCC830.08c,SPCP1E11.08,SPCP1E11.11**,SPAPB1A10.01c,Tf2-11,SPAC6B12.07c,Tf2-3,Tf2-4,sfp1,SPAC4F10.03c,SPBC660.15,his2,cwl1,suc1 | 97% | 86% |
| **SPAC13G6.10c,msa1,alm1,SPAC14C4.12c,grx2,erg3,SPAC16E8.02,SPAC1786.02,SPAC17A2.08c,cki3,SPAC19A8.02,SPAC20H4.02,etr1,SPAC29B12.05c,SPAC328.05,erg8,met11,SPAC3G9.05,cdr2,pan6,SPAC5H10.09c,gmh2,SPAC6F12.08c,sen1,SPAP7G5.03,rpb9,zas1,SPBC1347.09,SPBC13A2.03,SPBC1711.03,SPBC2G2.13c,SPBC409.08,mis15,SPCC1672.04c,SPCC1682.09c,SPCC18.15,nup61,gyp3,swc5,ksg1,myo2,ubc12,alr1**,SPCC1020.07,fft1,SPAC17A5.09c,SPAPB1A10.01c,SPAC4H3.14c,SPAC4F10.03c,SPBC19G7.07c,mug117,SPCC584.03c,thp1 | **SPAC13G6.10c,msa1,alm1,SPAC14C4.12c,grx2,erg3,SPAC16E8.02,SPAC1786.02,SPAC17A2.08c,cki3,SPAC19A8.02,SPAC20H4.02,etr1,SPAC29B12.05c,SPAC328.05,erg8,met11,SPAC3G9.05,cdr2,pan6,SPAC5H10.09c,gmh2,SPAC6F12.08c,sen1,SPAP7G5.03,rpb9,zas1,SPBC1347.09,SPBC13A2.03,SPBC1711.03,SPBC2G2.13c,SPBC409.08,mis15,SPCC1672.04c,SPCC1682.09c,SPCC18.15,nup61,gyp3,swc5,ksg1,myo2,ubc12,alr1**,fta1,csn3,sds22,res1 | 81% | 92% |
| **SPAC11E3.13c,SPAC8C9.05,pht1,SPBC1306.01c,SPBC17G9.06c,csx2,SPBC19C7.04c,SPBC28F2.11,SPBPB2B2.19c,SPBPJ4664.02,sap1,SPCC1795.10c,SPCC18.02,SPCC338.12**,Tf2-3<br>Tf2-4,SPAC977.09c,Tf2-11 | **SPAC11E3.13c,SPAC8C9.05,pht1,SPBC1306.01c,SPBC17G9.06c,csx2,SPBC19C7.04c,SPBC28F2.11,SPBPB2B2.19c,SPBPJ4664.02,sap1,SPCC1795.10c,SPCC18.02,SPCC338.12**,SPAC2E1P5.03,SPBC83.18c,SPBC1271.09,pob1 | 74% | 78% |

†Percentage of correctly affiliated genes in each cluster

Histones are shown in Italic; matched clusters are represented as rows; orthologs are sorted and represented as bold in each row.

Generally, synchronization substantially decreased after two periods. In order to maximize similarity, 10 time points of highest synchronization and quality from either dataset were selected. I will refer to these data as *'Sce'* and *'Spo'* respectively.

The pair of normalized yeast microarray gene expression datasets (*Saccharomyces cerevisiae and Schizosaccharomyces pombe*) was obtained from the publically accessible ArrayExpress data repository [131, 170] as well as from the authors' web resource. After log2-transformation, genes for which more than 50% of the data were missing were discarded. The remaining missing values were imputed by k-nearest neighbor algorithm [171] and Spline interpolation [172] which are commonly used for time-series data. Periodically expressed genes of significance with false discovery rate (FDR) of 0.05 were extracted by AR(1)-based background model [173].

The second two datasets (human/mouse) were obtained from Gene Expression Omnibus [174]. The studies investigated the effect of Estradiol on human [168] and murine cells [169]. Stossi *et al.*, examined U2OS osteosarcoma cells after treatment with either estrogen receptor (ER) alpha or beta for various periods of time up to 48 hours (10 time points in total).  They generated U2OS human osteosarcoma cells stably expressing ESR1 or ESR2, at levels comparable to those in osteoblasts. The characterization of the response to estradiol (E2) over time was measured using Affymetrix GeneChip microarrays. Moggs and coworkers recorded the uterus response of immature mice subcutaneously injected with 17β-estradiol (E2) or arachis oil (AO) at various time points up to 72 hours following treatment.

Datasets were normalized by variance stabilization [175]. Differentially expressed genes were extracted using the eBayes method of the limma package [123]. For multiple testing adjustments, the FDR was calculated using the algorithm of Benjamini and Hochberg [124].

#### 5.3.2.2 Cell cycle data – cerevisiea vs. pombe

The algorithm succeeded in producing the correct matching of time points after 20 iterations. Challenging the ability of the algorithm to reconstruct the correct order of time points without any knowledge about affiliation of neither time points nor gene orthologs, the sequence of the time points and the genes were randomly permutated. Typically after 16 to 35 iterations the algorithm converged to the very same result.

In the shown example the algorithm terminated with an RV coefficient of 0.8983. While the algorithm’s outer loops improved the matching score with increasing granularity, the inner loops optimized overall co-structure for a given *n* (Figure 26). The algorithm gradually increased the matching score in minimum two consecutive inner loops and identified the best similarity score by finding the correct affiliations of the connecting variables in seven outer loops. The result was verified in terms of optimal co-inertia and granularity as detailed in section 5.3.3. The result was visualized by CIA (Figure 21).

Here, the two pairs of projection coordinates are highly correlated and the overall similarity in the structure of the dataset was very high resulting in a RV coefficient of 0.8983. Clearly, the algorithm was able to detect and highlight the similarity between histones in these datasets, projecting them all in a cluster of histones differentiated from other functionally related genes (Figure 21(a), encircled in black).

In order to characterize the affiliated gene clusters, GO term enrichment analysis [176] was performed. Table 3 shows that the other affiliated clusters comprise common functionalities and orthologs. It lists significant common terms along with the percentage of the involved genes in each cluster.

Top common functions are represented by significant associations of p-value<0.05. In the same manner, Table 3 summarizes the percentages of correctly affiliated orthologs. A complete table listing all genes is given in Table 4. Based on the two back-transformed data tables, each cluster is represented by the gene-wise sum of all comprised genes across both tables. Subjecting this combined table to DBN algorithm, these cluster representatives became the nodes of a common gene network.

A graphical representation of the resulting common network is shown in Figure 22. To illustrate it by example, I follow its edges from the smallest to the largest cluster, moving through the cell-cycle from S phase towards mitosis (see Table 3). The histones (cluster g8) play an important role in transcriptional regulation. I observe an edge from g8 to g9 comprising the cyclin *CLN2* comparing favorably with work of Santisteban and coworkers [177].

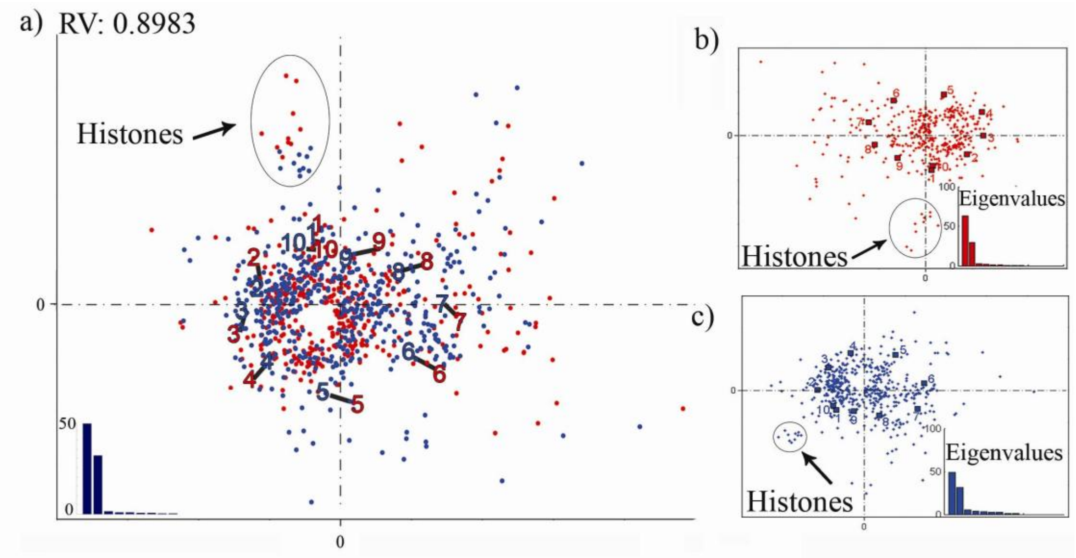


Figure 21. CIA plot of affiliated clusters.

**a)** '*Sce*' and '*Spo*' projected by CIA. The affiliated samples of both datasets are connected by lines, the lengths of which indicate the divergence between the two datasets. Each end of a line marks the position of a sample (time point) in the projection. Each blue or red dot represents a gene of '*Sce*' or '*Spo*', respectively, its position determined by its relative expression across all samples. The genes that are projected in the same direction from the centroid are those which are highly expressed in that sample. **b)** '*Spo*' dataset projected by CA. **c)** '*Sce*' dataset projected by CA. Eigenvalues are shown in the bottom corner for each dataset, normalized to 100%. The first two (x and y) axes of '*Sce*' explain 49% and 33% of the total inertia within this dataset. The first two axes of '*Spo*' represent 64% and 30% of the total variance within '*Spo*'. Thus more than 80% of the variance of the CIA was accounted for by the first two co-inertia axes and thus presents a good summary of the co-structure between the two datasets. Figure from [132] by permission of Oxford University Press.

Table 3. Characterization of the affiliated gene clusters

| Node‡ | Genes | | Orth. Counts* | | Top common over-represented biological functions | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | ‘Spo’ | ‘Sce’ | ‘Spo’ | ‘Sce’ | Category | Spo† | Sce† | Spo % | Sce % | Spo pvalue | Sce pvalue |
| g1(7) | 64 | 63 | 88% | 91% | organelle organization and biogenesis | 23 | 22 | 35.94 | 34.92 | 2.17E-02 | 1.59E-02 |
| | | | | | non-membrane-bound organelle | 17 | 21 | 26.56 | 33.33 | 6.31E-03 | 5.30E-03 |
| | | | | | intracellular non-membrane-bound | 17 | 21 | 26.56 | 33.33 | 6.31E-03 | 5.30E-03 |
| | | | | | chromosome organization and biogenesis | 15 | 15 | 23.44 | 23.81 | 4.29E-04 | 8.99E-05 |
| | | | | | cell cycle | 12 | 16 | 18.75 | 25.40 | 8.98E-03 | 8.98E-03 |
| | | | | | cell cycle process | 11 | 14 | 17.19 | 22.22 | 3.87E-02 | 1.66E-02 |
| | | | | | meiotic cell cycle | 8 | 15 | 12.50 | 23.81 | 2.46E-03 | 8.99E-05 |
| | | | | | cell cycle phase | 10 | 12 | 15.63 | 19.05 | 3.48E-02 | 1.57E-02 |
| | | | | | chromosomal part | 8 | 14 | 12.50 | 22.22 | 1.23E-02 | 4.29E-04 |
| | | | | | DNA binding | 15 | 16 | 23.44 | 25.40 | 1.81E-03 | 3.59E-02 |
| | | | | | M phase | 9 | 12 | 14.06 | 19.05 | 1.42E-02 | 8.80E-03 |
| | | | | | DNA packaging | 8 | 11 | 12.50 | 17.46 | 2.08E-02 | 4.23E-04 |
| g2 (9,M/G1) | 54 | 43 | 85% | 97% | cellular component organization | 23 | 23 | 42.59 | 53.49 | 5.24E-03 | 2.72E-03 |
| | | | | | biopolymer metabolic process | 22 | 23 | 40.74 | 53.49 | 9.53E-03 | 7.77E-03 |
| | | | | | ribonucleotide binding | 10 | 14 | 18.52 | 32.56 | 6.02E-03 | 9.03E-03 |
| | | | | | cell cycle | 11 | 5 | 20.37 | 11.63 | 7.77E-03 | 5.45E-03 |
| | | | | | DNA binding | 5 | 10 | 9.26% | 23.26 | 3.77E-02 | 3.77E-02 |
| | | | | | mitotic cell cycle | 7 | 6 | 12.96 | 13.95 | 3.69E-02 | 3.69E-02 |
| | | | | | M phase of mitotic cell cycle | 6 | 8 | 11.11 | 18.60 | 5.45E-03 | 4.21E-02 |
| g3 (8) | 60 | 84 | 68% | 84% | protein binding | 39 | 19 | 65.00 | 22.62 | 2.47E-02 | 1.92E-03 |
| | | | | | biological regulation | 20 | 24 | 33.33 | 28.57 | 8.09E-03 | 1.84E-03 |
| | | | | | G1/S transition of mitotic cell cycle | 9 | 7 | 15.00 | 8.33% | 3.37E-02 | 5.87E-03 |
| | | | | | developmental process | 8 | 13 | 13.33 | 15.48 | 3.98E-02 | 1.26E-02 |
| | | | | | cell cycle | 15 | 5 | 25.00 | 5.95% | 6.16E-03 | 3.34E-02 |
| | | | | | cell division | 10 | 10 | 16.67 | 11.90 | 8.30E-03 | 1.05E-02 |
| | | | | | site of polarized growth | 11 | 7 | 18.33 | 8.33% | 4.62E-05 | 4.25E-05 |
| | | | | | regulation of progression through cell cycle | 7 | 8 | 11.67 | 9.52% | 4.31E-02 | 1.96E-02 |
| | | | | | cytoskeleton | 7 | 8 | 11.67 | 9.52% | 1.96E-02 | 7.65E-03 |
| | | | | | cellular bud | 10 | 5 | 16.67 | 5.95% | 5.73E-03 | 2.26E-04 |
| | | | | | negative regulation of biological process | 9 | 6 | 15.00 | 7.14% | 5.73E-03 | 3.37E-02 |
| | | | | | regulation of cell cycle | 7 | 8 | 11.67 | 9.52% | 4.31E-02 | 1.96E-02 |
| g4 (10) | 36 | 54 | 94% | 90% | regulation of biological process | 14 | 9 | 38.89 | 16.67 | 8.65E-03 | 7.95E-03 |
| | | | | | regulation of cellular process | 14 | 9 | 38.89 | 16.67 | 7.43E-04 | 7.12E-03 |
| | | | | | biological regulation | 16 | 5 | 44.44 | 9.26 | 1.89E-03 | 1.89E-03 |
| | | | | | cell cycle process | 9 | 7 | 25.00 | 12.96 | 9.05E-04 | 7.31E-03 |
| | | | | | mitotic cell cycle | 8 | 6 | 22.22 | 11.11 | 1.85E-02 | 1.62E-02 |
| | | | | | regulation of S phase | 2 | 3 | 5.56% | 5.56% | 7.26E-04 | 2.53E-02 |
| g5 (6, M) | 49 | 32 | 91% | 86% | cell division | 6 | 9 | 12.24 | 28.13 | 3.11E-02 | 1.53E-02 |
| | | | | | mitotic cell cycle | 5 | 7 | 10.20 | 21.88 | 9.78E-04 | 4.13E-02 |
| | | | | | M phase | 8 | 3 | 16.33 | 9.38% | 3.11E-02 | 7.39E-03 |
| | | | | | cytoskeletal part | 4 | 7 | 8.16% | 21.88 | 9.8E-03 | 8.68E-03 |
| | | | | | cell cycle control | 3 | 6 | 6.12% | 18.75 | 1.27E-02 | 4.88E-03 |
| g6 (5) | 31 | 41 | 64% | 88% | transmembrane protein | 14 | 10 | 45.16 | 24.39 | 5.7E-03 | 2.62E-02 |
| | | | | | cell division | 6 | 6 | 19.35 | 14.63 | 9.3E-03 | 4.12E-02 |
| | | | | | cytoskeleton organization | 5 | 5 | 16.13 | 12.20 | 6.E-04 | 3.49E-02 |
| g7 (11) | 13 | 10 | 89% | 95% | mitotic cell cycle | 5 | 5 | 38.46 | 50.00 | 7.8E-03 | 1.82E-03 |
| | | | | | cell cycle | 4 | 4 | 30.77 | 40.00 | 1.9E-02 | 4.55E-03 |
| | | | | | cell division | 4 | 4 | 30.77 | 40.00 | 3.2E-02 | 2.42E-02 |
| g8 (1, S) | 10 | 10 | 100% | 100% | **Histones** | | | | | | |
| g9 (2) | 13 | 21 | 75% | 83% | cellular component organization | 9 | 13 | 69.23 | 61.90 | 8.48E-04 | 6.75E-03 |
| | | | | | cell cycle | 5 | 6 | 38.46 | 28.57 | 1.02E-02 | 2.90E-03 |
| g10 (12, G1) | 22 | 27 | 90% | 89% | protein binding | 17 | 8 | 77.27 | 29.63 | 3.95E-02 | 5.39E-03 |
| | | | | | cell cycle | 6 | 9 | 27.27 | 33.33 | 8.58E-03 | 6.94E-03 |
| | | | | | DNA damage | 3 | 5 | 13.64 | 18.52 | 4.95E-02 | 3.40E-02 |
| | | | | | DNA-dependent DNA replication | 3 | 4 | 13.64 | 14.81 | 5.90E-03 | 1.68E-02 |
| g11 (3, G2) | 17 | 13 | 81% | 94% | cell cycle | 3 | 3 | 17.65 | 23.08 | 6.71E-04 | 2.60E-02 |
| | | | | | cell wall organization and biogenesis | 3 | 3 | 17.65 | 23.08 | 6.89E-03 | 2.58E-02 |
| g12 (4) | 24 | 46 | 87% | 100% | cellular component organization | 12 | 26 | 50.00 | 56.52 | 7.08E-05 | 2.74E-02 |
| | | | | | biological regulation | 8 | 19 | 33.33 | 41.30 | 9.06E-04 | 4.17E-03 |
| | | | | | phosphoprotein | 3 | 17 | 12.50 | 36.96 | 9.43E-03 | 5.88E-03 |
| | | | | | developmental process | 4 | 10 | 16.67 | 21.74 | 8.38E-02 | 1.38E-02 |

‡ The sequence of the nodes in the cell cycle is provided in brackets along with their cell cycle affiliations.

* This column shows the number of correctly affiliated orthologues as a percentage of all orthologues “available” for this gene cluster.

†Number of genes known to be involved in the same functional category (GO-term) in each individual gene cluster.

Table 4. Cluster components (genes) of the affiliated gene clusters

| node† | 'Sce' gene clusters | 'Spo' gene clusters |
|---|---|---|
| g1 (7) | **YOR292C,YJL123C,YFL034W,YCL047C,VIP1,TRA1,THG1,TCB3,SRM1,SPE1,SEC11,RPL30,PUS7,PRO1,PGM2,PDS5,OSH3,MOT1,HOG1,GPI1,ERJ5,CSE4,CDC20,(BCH1,BUD7)**,ISR1,HSP150,MCM5,YRF11,YPR203W,MCM3,CWP2,SED1,CTI6,SYP1,SOK1,YPL014W,TOS2,SMF3,GPI13,SPF1,CCR4,ACS2,HOL1,COS8,BIO2,FIN1,CIN8,SCY1,ERG3,YNL176C,AXL2,TOF2,CSM2,YGR035C,WSC2,PSA1,SEC53,FLC3,YEH1,SVS1,YNK1 | **SPAC3G6.05,SPCC1322.09,SPAC6F6.13c,SPAC694.03,asp1,SPBP16F5.03c,SPCC63.07,SPAPYUK71.03c,dcd1,spe1,sec11,rpl30,SPBC1A4.09,SPAC17H9.13c,SPBC32F12.10,pds5,SPAP27G11.01,mot1,phh1,gpi1,SPAC2E1P5.03,cnp1,slp1,SPBC31F10.16**,SPBC19G7.04,atl1,mde6,SPAP27G11.08c,tim22,ask1,SPAC144.08,SPCC736.02,SPCC63.10c,mac1,SPCC1682.13,cki3,SPCC338.08,SPAC9.11,SPBP22H7.03,nup132,SPBC19C2.10,tas3,SPAC630.12,mrpl28,meu29,mrc1,cfh3,SPBC83.18c,vps8,phf2,set3,SPBC31F10.02,SPCC63.13,SPBC21C3.04c,bet5,SPBC27.04,SPBC428.06c |
| g2 (9, M/G1) | **ALG11,GAS5,HTZ1,KAP114,LEO1,MRM1,RAD54,RBG1,RPT5,SEY1,SSK2,YLL023C**,GRX6,IRC19,SPP381,SNT309,GLO3,PIR3,YRF13,HXT7,GPA1,YIL177C,YRF12,PIG1,COQ4,SWF1,PCM1,TEL2,OCH1,ECM25,ERP3,GWT1,CHS6,PMT4,YIR043C,YOX1,ADK2,REB1,YKL069W,ERP2,WHI5,YHP1,TAO3 | **alg11,SPAC11E3.13c,mal1,kap114,SPBC13E7.08c,SPBC1347.13c,rad54,SPAC9.07c,pam2,SPAC222.14c,SPAPJ730.01,SPBC1539.04**,chs2,klp5,SPCC320.03,SPBC16E9.07,SPBC15D4.01c,SPCC550.11,SPBC582.04c,csk1SPBC2F12.12c,urb2,SPBC1105.07c,SPAPB18E9.06c,nnf1,SPCC1223.04c,spc25,klp8,fkh2,SPAC19B12.08,cfh1,SPAC521.02,SPCP31B10.09,cdt1,imp2,arp2,mid2,mis17,SPBC8E4.04,SPAC3F10.08c,rid1,adg2,pht1 |
| g3 (8) | **BEM1,DID4,DUT1,(EXG1,SPR1),HOF1,MRN1,MSY1,(MYO3,MOY5),RSP5,SSO1,STU2,VPH1,YPL206C**,FAR8,SPH1,TGL2,HEK2,POL12,GCR1,ALG14,PET112,RAD53,ANP1,HST4,YHR126C,ADE12,GLK1,UBP11,AAD10,SPT3,RED1,JSN1,YRF1_7,GYP7,PFK1,FAA1,EXO1,RMA1,MDS3,SWE1,SUB1,YDL027C,YPT11,CSH1,CDC9,PXL1,CDC45,EMP70,MSH2,ELG1,FRE6,VID22,SPO16,EXG2,SLD2,UBP3,POL30,ASF1,OST2,BNI5,PDR16,GIC1,PHS1,RRN9,GPI16,SFB3,OPY2,FHL1,SHO1,CKS1,SVL3,ATF1, FIR1 | **ral3,did4,SPAC644.05c,exg1,cdc15,msa2,SPCC576.06c,myo1,pub1,psy1,alp14,vph1,SPAC4D7.02c**,etd1,ntf1,pof3,SPBC24C6.10c,cdt2,SPBC405.02c,SPAC4H3.06,SPAPJ698.04c,set8,SPAC19G12.05,SPAC17H9.18c,ppk3,fin1,alp1,orc5,SPCC794.08,rpl7,nse5,SPAC30D11.01c,psc3,SPAC3G9.05,SPCC320.12,cfh4,SPAC24H6.08,ppc89,par2,SPBC1306.01c,set2,SPAC9.10,SPBC29A3.03c,SPAC2E1P5.02c,SPAC1F12.05,SPAC227.05,fep1,lad1,cfh2,SPAC1639.01c,pmp31,ulp1,SPAC977.01,SPBC14F5.10c,cdm1 |
| g4 (10) | **AIR2,BMH1,DDP1,ENT1,GAS1,HEM4,MDR1,MOB1,MSE1,RTT106,SNQ2SSK22,STE11,TFB4,TRP4**,SPO12,KIP3,YLR462W,FAT1,HXT10,EMI2,GEF1,YER071C,MSB1,DDI2,PRI2,SHE3,YMR31,CLB5,SNO3,PHO3,MSA1,YEL077C,YMR258C,HAA1,PEX7,YJL218W,YLR464W,AVT2,YGL036W,IST2,BNI4,RTT109,CTF4,HXT2,RCO1,GCN5,PDR5,YMR118C,SNT1,CLB6,EPT1,LPP1,RAD27 | **SPBP35G2.08c,rad25,aps1,ent1,SPAC19B12.02c,ups,SPBC215.01,mob1,SPAPB1A10.11c,SPAC6G9.03c,pdr1,wak1,byr2,tfb4,trp4**,apc15,mog1,SPAC13G6.10c,SPAC821.03c,hrr1,meu34,SPCC613.07,SPAC1705.03c,SPAPB17E12.10c,SPBC1861.07,SPAC5D6.02c,SPCC16C4.02c,SPAC11E3.10,SPBC17G9.06c,SPBC15D4.02,SPAC1565.02c,efc25,hmt1,SPBC3H7.13,SPBP4H10.16c,SPBPB2B2.19c |
| g5 (6, M) | **ACF2,ATG8,CLB2,ERG6,HSL1,LSM2,MCD1,MUC1,NOB1,RER1,RFA1,RNR1,RRP45,RVS161,SEN1,TOP2,UBP6**,YCK1,CAF120,HCM1,STB1,LSP1,ERS1,YRF15,ALD6,STE2,RGA1,CMK1,YPR157W,YCR102C,YDL118W,YBR071W | **(eng1,eng2),atg8,cdc13,erg6,cdr1,lsm2,rad21,SPBPJ4664.02,SPAC1486.09,rer1,rad11,ssb1,dc22,SPCC757.08,hob3,(sen1,SPBC29A10.10c),ptr11,ubp6**,SPAC3C7.07c,hcn1,mcp1,pus2,chp2,SPBC947.14c,SPCC576.12c,SPAP8A3.11c,SPCC794.03,SPAC15A10.09c,SPBC29A10.08,SPBC14C8.13,bet1,SPAC630.04c,rho4,cdc25,SPBC36.06c,SPAC26H5.11,SPAC14C4.05c,SPAC24B11.07c,SPBC16H5.12c,vip1,SPCC594.04c,SPAP14E8.02,cam2,csx2,ucp10 |
| g6 (5) | **CDC5,CLB3,ENT3,MYO1,POL31,RKM1,SAC6,SLY41,TFB3,VID27,VRG4,YDL124W**,FMP45,KEL2,VHT1,YFL067W,RLF2,DPH1,CUE4,GET1,KCC4,REV7,SKG6,SUT1,GYP6,YHL026C,PMA2,IRC4,YPR202W,RAD2,NUP170,APA2,NRG2,UIP5,WTM2,STV1,ALE1,GGA2,TPK1,TRK2,MSB4 | **plo1,cig1,SPCC794.11c,myo3,cdc1,SPBC1709.13c,fim1,SPBC83.11,mcr1,SPBC1685.14c,SPAC144.18,SPAC19G12.09**,ace2,SPAC14C4.12c,SPBC4F6.12,mrpl4,SPAC4G9.19,rpc31,aph1,mrp51,SPAC27D7.11c,agn1,SPBC1198.07c,SPBC31F10.10c,SPAC688.07c,dga1,SPCC1795.10c,spTrap240,af1,SPAC637.13c |
| g7 (11) | **ADY2,CDC7,COG3,GPI8,SMC4**,GTT3,GAS3,GAS4,GAS2,GLE1 | **SPAC5D6.09c,hsk1,SPBC1539.05,gpi8,cut3**,sfc2,clr8,myo52,mbx1,SPCC1020.12c,rum1,SPCC4F11.03c,SPAC20H4.05c |
| *g8 (1, S)* | ***HHT1,HHT2,(HHF1,HHF2),(HTA2,HTA1),(HTB2,HTB1),HDA1,RPD3*** | ***(hht1,hht2),(hht1,hht3),(hhf1,hhf2),(hta1,hta2),htb1,clr3,clr6*** |
| g9 (2) | **DBF2,INP53,OAC1,(ODC1,OCD2),RAI1,SEC4**,COP1,ECO1,ARV1,RSF2,LYS1,DPB2,HSD1,CDC11,CLN2,PMT5,PMT1,RFA3,PIN2,SAS3,FKH2 | **sid2,SPBC2G2.02,oac1,SPAC328.09,din1,ypt2**,sec72,sad1,SPCC970.08,SPCC553.07c,prp28,cam1,spd1 |
| g10 (12, G1) | **EMP24,KAP122,LCB2,MLC1,MSH6,MSW1,POL1,RUP2,SKI3,YMR259C**,HSL7,HAP2,YDL089W,QDR3,TOS4,UBX7,MCM2,MMS4,GGA1,PHO8,COS4,SVF1,HIF1,YFL042C,HOS3,AIM34,YLR455W | **emp24,kap111,lcb2,cdc4,msh6,msw1,pol1,SPAPJ696.03c,SPCC1919.05,SPCC1494.07**,cdc18,pnk1,SPBC25B2.07c,SPAC8F11.06,dfp1,SPCC1235.09,pkd2,SPBC660.06,SPCC188.10c,rpc17,SPAC2C4.17c,cut2 |
| g11 (3, G2) | **MSS51,RPC11,(SIM1,SUN4),SLA2,SMC3,YML096W,YPT52**,KAR5,DSN1,DAM1,FKH1,FIP1,MSA2 | **SPAC25B8.04c,rpc11,psu1,end4,psm3,SPBC4F6.11c,ypt5**,SPBC4C3.04c,SPAC10F6.07c,SPCC757.12,SPCC1259.08,ams2,pmc2,SPAC27D7.09c, pdf1 |
| g12 (4) | **CCC2,CDH1,CHS5,IPL1,MUS81,NUP2,PTM1,RHO1,RPP1,YMR244W,YOR291W**,GFA1,MSB2,CLN3,CYS3,ESC8,ALY2,HXT4,ROD1,MAL33,TCM62,TPS2,PAN2,HST2,PUF4,CAC2,ORC1,PEX11,HXT5,TUB2,SLN1,RDS2,SPT21,CTF18,INP2,EUG1,SFG1,MCD4,PRY2,NPP2,GLG2,FCP1,PBI2,SGS1,CRH1,RHO3 | **SPBC29A3.01,srw1,cfr1,aim1,mus81,nup61,SPAC26H5.07c,(rho1,rho5),SPAC3A12.04c,adg3,SPCC1672.11c**,spo12,SPCC126.01c,SPAC3H8.03,SPAC10F6.14c,ssp1,klp6 SPAC11H11.02c,zds1,cwf25 |

† The sequence of the nodes in the cell cycle is provided in brackets along with their cell cycle affiliations.

Following the cell-cycle from S to G2, the cohesin complex is required to hold together the sister chromatids. This process is mediated by the acetyltransferase *ECO1* of cluster g9 (S-phase) directly interacting with cohesion complex subunit *SMC3* (G2-phase) of cluster g11 [178-180]. The edge linking g9 and g11 suggests an according tight transcriptional regulation of acetylase *ECO1* preceding *SMC3.*

Following the cell-cycle from G2 to M, the common transcription network shows node g11 (G2) to regulate both g12 and g3, whereas g12 itself also regulates g3, forming a

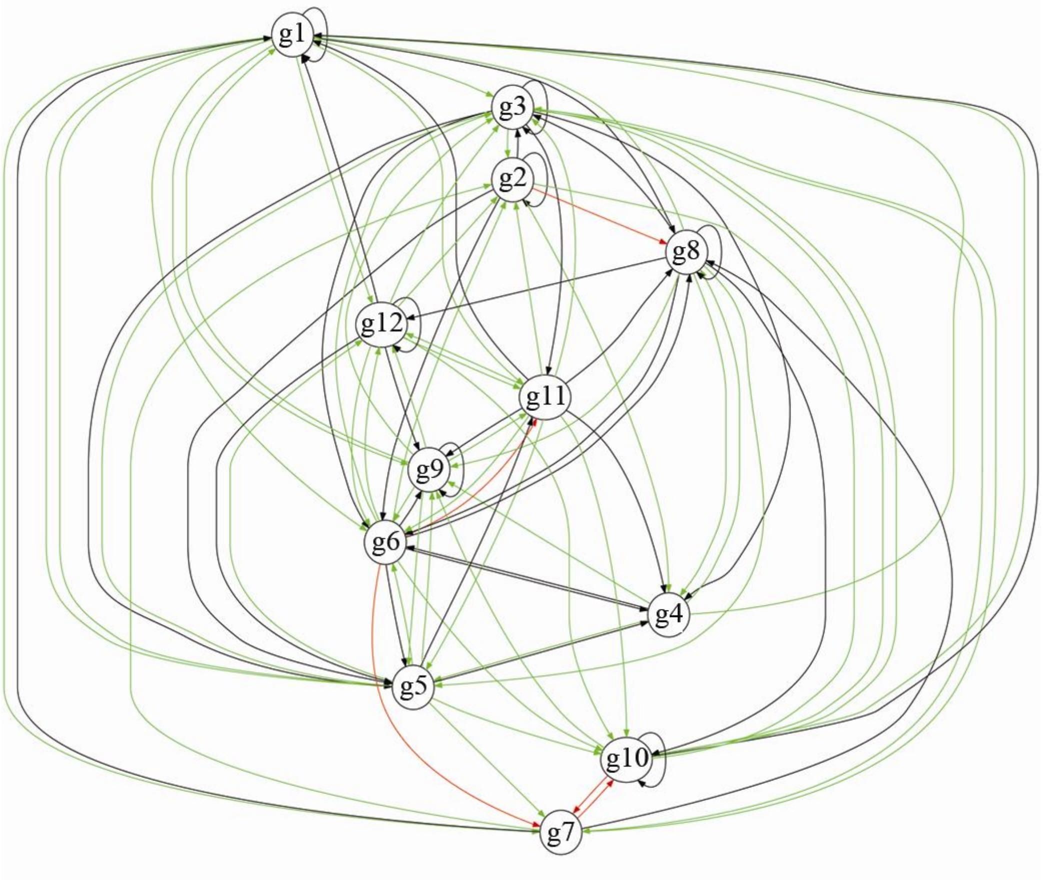


Figure 22. Common '*Sce*' and '*Spo*' regulatory network.

Affiliated gene clusters are represented as nodes, their interactions as edges. These interactions are color-coded according to their occurrence in KEGG or one of the other pathway databases listed in the methods section. True positive (TP) edges are shown in green, missing edges (FN) are shown in black, incorrect or previously unknown interactions (FP) are shown in red. Any green or black edge is supported by at least one publication. Figure from [132] by permission of Oxford University Press.

network motif referred to as feedforward loop [181]. It is often found in the context of signal amplification. The increase in cellular activity during G2/M transition is also reflected by increased glycolysis (*GLK1, PFK1*) and by g3 being the largest cluster. While still transcribing genes important for G2/M transition (*SWE1*) and DNA repair (*RAD53, EXO1, MSH2, POL3*), the cell already prepares for budding (*BEM1, SPH1, FAA1, BNI5, GIC1*) and cytokinesis (*HOF1, MYO3, STU2, YPT11*).

The observed edges can be explained by transcription factor activity. For the direct edge from g11 to g3, transcription factors *SIM1* and *FKH1* of g11 have been shown to regulate 2 and 8 genes in g3, respectively (genes and literature are provided in (Table 5).

Table 5. Transcription factor regulation references

| **Transcription Factor** | **Regulated by** | | **Targets** | |
|---|---|---|---|---|
| | **gene** | **In node** | **gene** | **In node** |
| SIM1 | † | g11 | ATF1, SPR1 [182] | g3 |
| FKH1 | † | g11 | FIR1, FLH1 [183]<br>SVL3, JSN1 [183, 184]<br>OPY2, RSP5 [183, 185]<br>ALG14 [184]<br>HOF1 [186] | g3 |
| SFP1 | KAR5, SMC3 [170, 187]<br>RPC11 [187, 188] | g11 | IPL1, GFA1, ESC8, ALY2, HXT4, ROD1, PUF4, SPT21, CTF18, MCD4, NPP2, FCP1, RHO3 [187]<br>RPP1 [187, 189]<br>YMR244w, ORC1 [188]<br>SFG1 [183] | g12 |
| MAL33 | † | g12 | ALG14, OPY2, RAD53 [183] | g3 |
| RDS2 | † | g12 | PDR16 [190] | g3 |

† The transcription factor itself is cluster member

For the path from g11 to g3 via g12, 17 genes of g12 are targets of the transcription factor *SFP1*. While *SFP1* itself was filtered out for showing unreliably small signals, it is regulated by *KAR5, RPC11* and *SMC3* of cluster g11 (Table 5). From g12, the comprised transcription factors *RDS2* and *MAL33* are known to regulate *PDR16* and *ALG14*, *OPY2*, and *RAD53* of g3, respectively (Table 5).

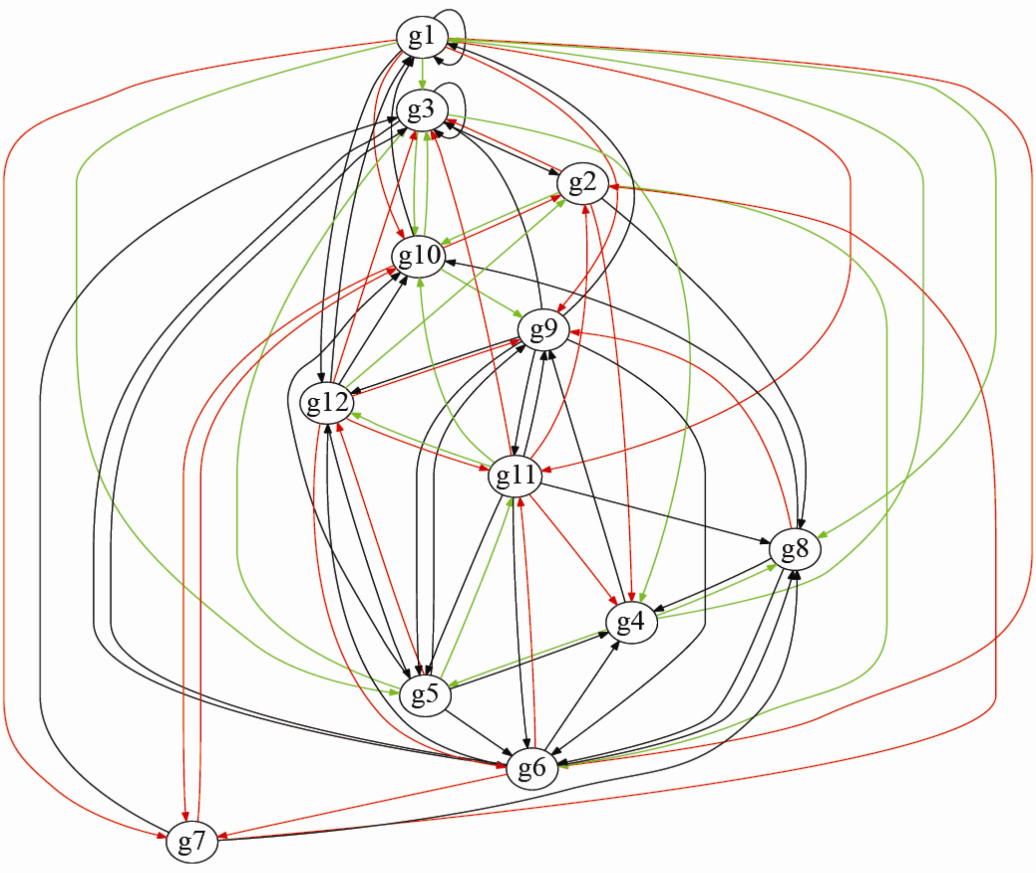


Figure 23. Network inferred from the single '*Sce*' dataset.

The layout follows Figure 22. Figure from [132] by permission of Oxford University Press.

While the edge from g11 to g12 can also be obtained from the '*Sce*' dataset alone (Figure 23), the edge from g6 to g12 is not present in either single network (Figure 23 and Figure 24) but is only detected by combining the datasets (Figure 22). The same is true for the above described edge between g9 and g11. The superiority of the common network is quantified in section 5.3.4.

Out of 144 possible directed interactions, 53 true positives, 5 false-positives, 36 false-negatives and 50 true-negatives were detected. Assuming that any interaction listed in any database for these genes would be detectable from these small datasets, sensitivity is 60%. Thus, most (more than half) interactions in pathway databases are present in these data, common to both datasets, and successfully detected here, with 72% accuracy and a specificity of 91%.

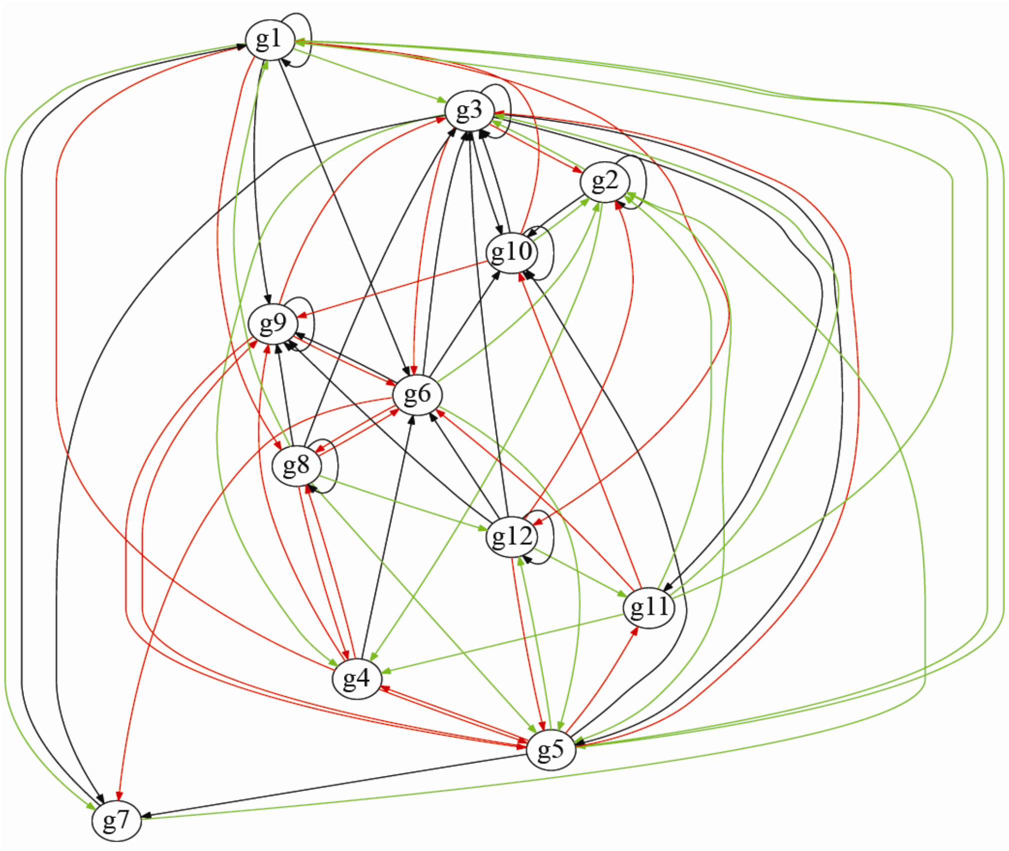
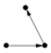
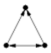
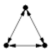
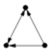
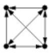
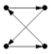
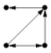
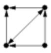
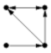
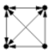
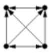
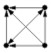
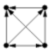
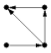
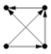
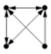
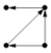
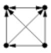
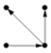
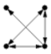
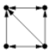


Figure 24. Network inferred from the single '*Spo*' dataset.

Figure from [132] by permission of Oxford University Press.

Furthermore, the coherence of the interactions found were assessed, i.e. their tendency to form sound regulatory modules, by network motif analysis (Table 6). Size-3 and size-4 sub, graph frequencies were determined by generating ten million directed random graphs with same sample probabilities and in which cases the probability that a given edge exists was preserved. For this, all 13 non-isomorphic directed size-3 sub graphs as well as 199 non-isomorphic directed size-4 sub graphs were calculated. All 21 network motifs listed in Table 6 exhibit p-values smaller than 0.05 as well as Z-scores greater than two.

Table 6. Motif significance in the yeast common network.

| Motif | Frequency | Mean-Freq | Standard-Dev | Z-Score | p-Value |
|---|---|---|---|---|---|
|  | 40% | 13.529% | 0.020248 | 13.074 | 0 |
|  | 17.714% | 6.7986% | 0.010763 | 10.142 | 0 |
|  | 1.7143% | 0.058427% | 0.0017218 | 9.617 | 0 |
|  | 16% | 6.1522% | 0.010424 | 9.4476 | 0 |
|  | 2.0785% | 0.0054736% | 0.00034742 | 59.67 | 0 |
|  | 7.8522% | 0.1405% | 0.0020045 | 38.472 | 0 |
|  | 10.393% | 0.26417% | 0.0027108 | 37.363 | 0 |
|  | 11.778% | 0.41705% | 0.0037092 | 30.63 | 0 |
|  | 10.162% | 0.52822% | 0.0037828 | 25.466 | 0 |
|  | 1.8476% | 0.017144% | 0.00075102 | 24.373 | 0 |
|  | 4.6189% | 0.34613% | 0.0029107 | 14.68 | 0 |
|  | 5.3118% | 0.87444% | 0.0042215 | 10.511 | 0 |
|  | 5.7737% | 0.91135% | 0.0047198 | 10.302 | 0 |
|  | 9.4688% | 1.573% | 0.007716 | 10.233 | 0 |
|  | 4.6189% | 0.89778% | 0.0052242 | 7.1229 | 0 |
|  | 0.69284% | 0.025932% | 0.0011094 | 6.0116 | 0.003921 |
|  | 3.2333% | 0.65543% | 0.0045951 | 5.61 | 5.8e-005 |
|  | 3.0023% | 0.73487% | 0.0043461 | 5.2172 | 2.2e-005 |
|  | 1.8476% | 0.52102% | 0.0039299 | 3.3755 | 0.003337 |
|  | 0.23095% | 0.0156% | 0.00071087 | 3.0294 | 0.014266 |
|  | 1.8476% | 0.84962% | 0.0044075 | 2.2642 | 0.016043 |

### 5.3.2.3 Esteradiol effect on human – mice

After combining highly similar datasets, two datasets of lesser, i.e. more natural, relatedness were examined. The effect of Estradiol on human [168] and murine cells [169] was studied measuring various time points up to 48 hours and 72 hours on HG-U95A and MG-U74A Affymetrix Gene Chips, respectively.

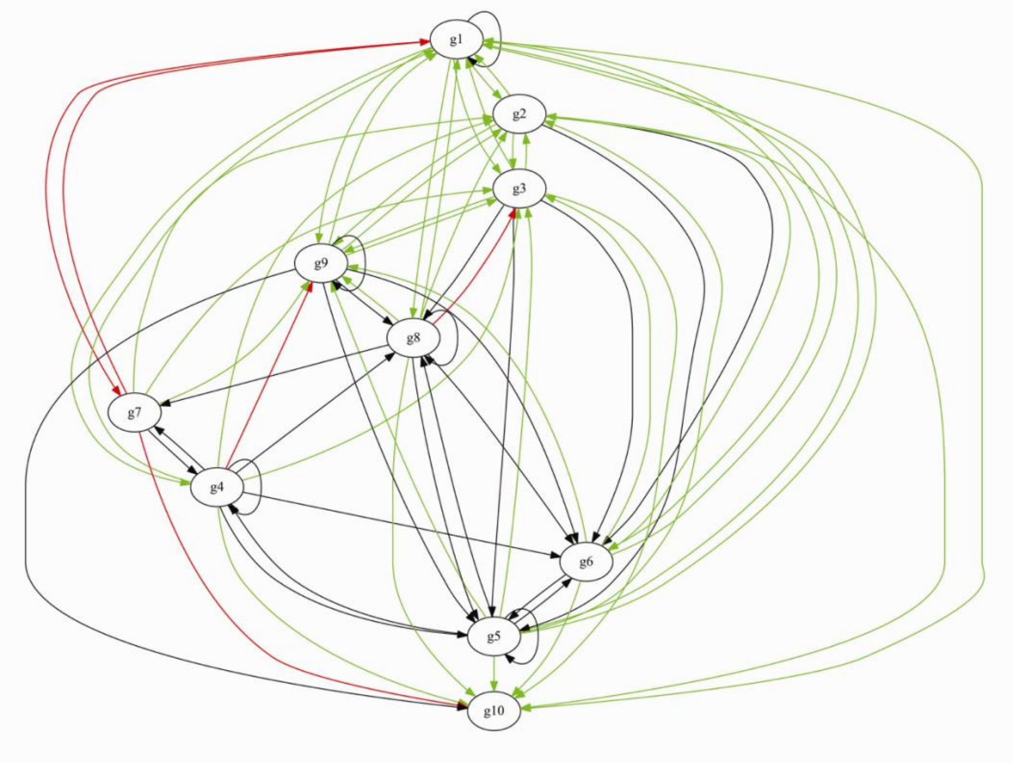


Figure 25. Interaction network in Human-Mouse dataset

The algorithm converged after 18 iterations in a maximum local swap of 6 (inner loops) between genes and samples as connecting variables. A dramatic increase of RV coefficient (from 0.37 to 0.78) was observed when the algorithm switched from 6 to 7 connecting variables (both genes and samples) (Figure 27). Proceeding from n=7 to 10, the algorithm consolidates this co-structure, further improving RV by 0.1025. Termination at n=10 clusters was verified as before (Section 5.3.3). The result shows an RV coefficient of 0.8194 (Figure 27). More than 87% of the co-inertia was accounted for by the first two principal axes for *Homo sapiens* and 74% by the first two axes of the *Mus musculus* dataset (52% and 22% by the first and second axis respectively). These eigenvectors were selected for back-transformation. To determine the maximal number of well-distinguished clusters, the back-transformed data were assessed with respect to silhouette values. 10 clusters were confirmed as optimal, showing an overall mean Silhouette value of 0.2645 for *Mus musculus* and 0.2115 for *Homo sapiens* (Figure 28b). Clusters were matched as before, i.e. by majority voting of a Hungarian match based on distances from the CIA.

## 5.3.3 Granularity evaluation

### 5.3.3.1 RV coefficients improvement

Figure 27 (next page) shows the gradual improvement in matching scores by optimizing the overall co-structure of the datasets. While the affiliation of samples was considerably enhanced by increasing to 7 clusters, it took a second iteration with

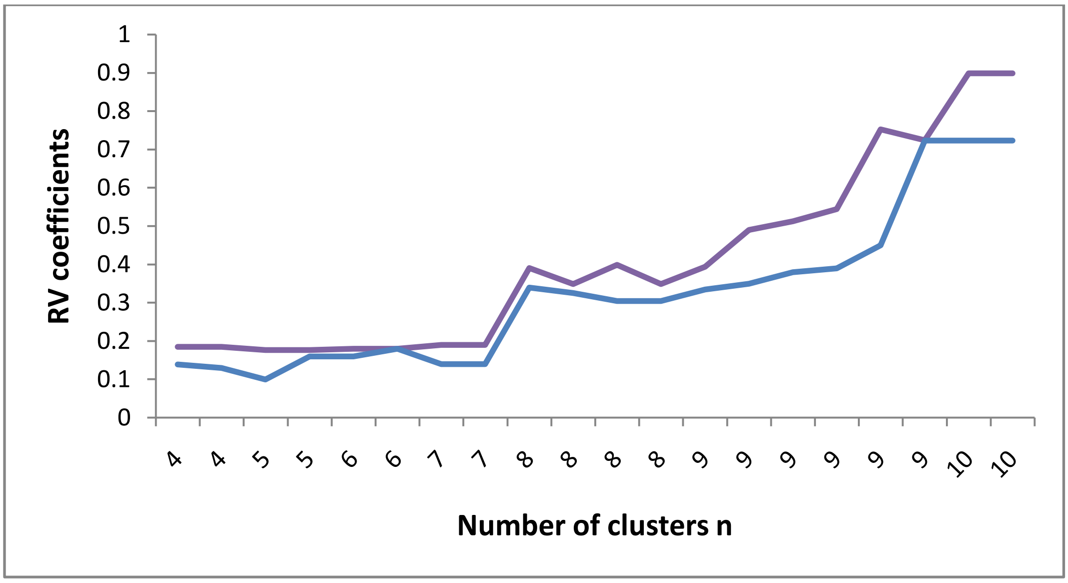


Figure 26. RV coefficients between ‘*Sce*’ and ‘*Spo*’ datasets.

The co-structure (RV coefficient) is plotted versus the number of connecting clusters n in two colors. The purple line shows the RV when the samples are the connecting variables. The light blue line refers to the genes as the connecting variables.

7 clusters for the genes to get affiliated accordingly well. In my experience “leaps” in RV increase tend to be initiated by sample affiliations, which could probably be due to their generally much smaller numbers.

Figure 26 shows an exception in that gene and sample affiliations may also be considerably improved at the same time. This happened upon increasing to 8 clusters without any further delay. However, the largest one-step increase in RV of the same example occurred in the transition from the last but three to the last but two iterations. Here, a large increase for the genes is made possible by a slight decrease in terms of the samples, speaking in favor of my non-greedy approach. It only occurred after six rounds maintaining the same number of nine clusters. This emphasizes the importance of performing several optimizing steps before further increasing the number of clusters.

Figure 27 shows the algorithm proceeding from low to high RV correlation coefficients

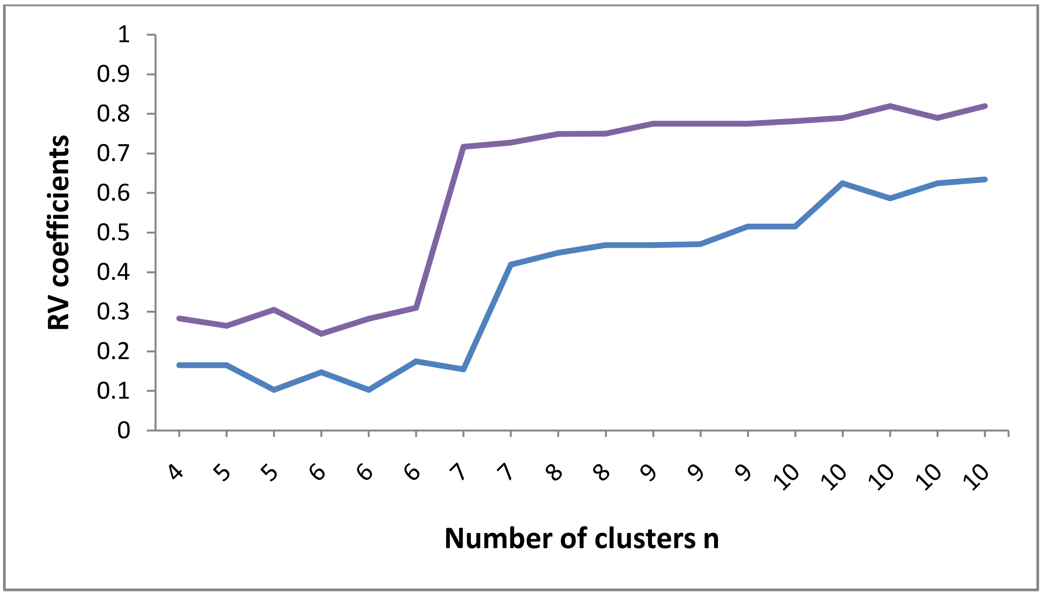


Figure 27. RV coefficients between ‘human’ and ‘mouse’ datasets.

The co-structure between the two sets considerably increases beyond six clusters. The best matching is achieved after 17 iterations (10 clusters). The layout follows Figure 26.

while merging the human and murine datasets. Alternating between samples (purple) and genes (blue) as connecting variables, it performs several optimizing steps before further increasing the number of clusters.

From this step, the algorithm proceeds to ten clusters, performing two more rounds and then stops. How to know that termination is not premature (triggered by a local optimum)? In order to demonstrate that the terminal matching is indeed optimal, the stop flag in the outer loop of the algorithm was removed. While the number of samples limits the number of sample clusters, the number of gene clusters can be further increased (up to individual genes, Figure 29).

The optimal number of gene clusters for the combination depends on how many clusters can be discriminated in each dataset (Figure 28). Figure 28a shows silhouette values for the yeast datasets. As the number of clusters increases silhouette values decrease from 0.3935 (for 2 clusters) down to -0.2154 (140 clusters) in "*Sce"* and from 0.4987 to -0.0731 in "*Spo*". The overall optima of 0.3504 and 0.3921 were obtained for maximally 12 well separated gene clusters.

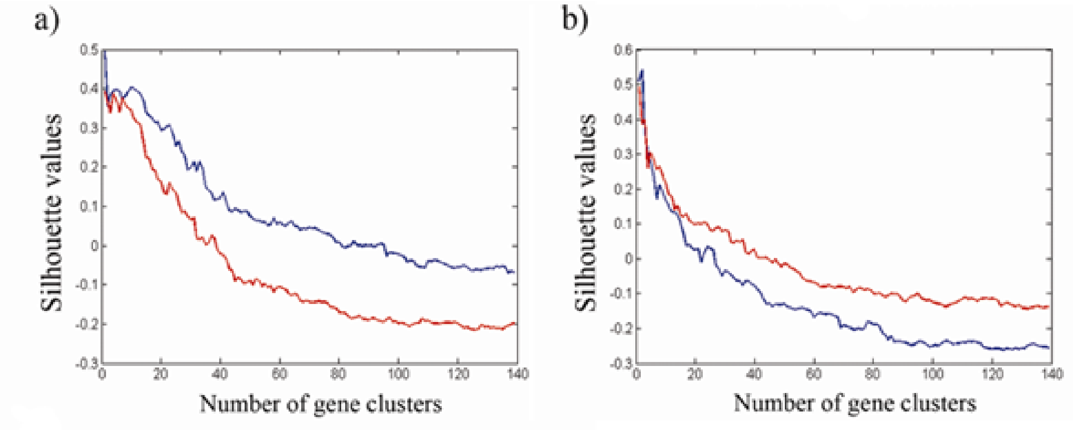


Figure 28. Silhouette values.

Panel a) plots the quality of the clustering depending on the number of clusters for '*Sce'* (red line) and '*Spo'* (blue). '*Homo sapiens'* (red) and '*Mus musculus'* (blue) are shown in panel b.

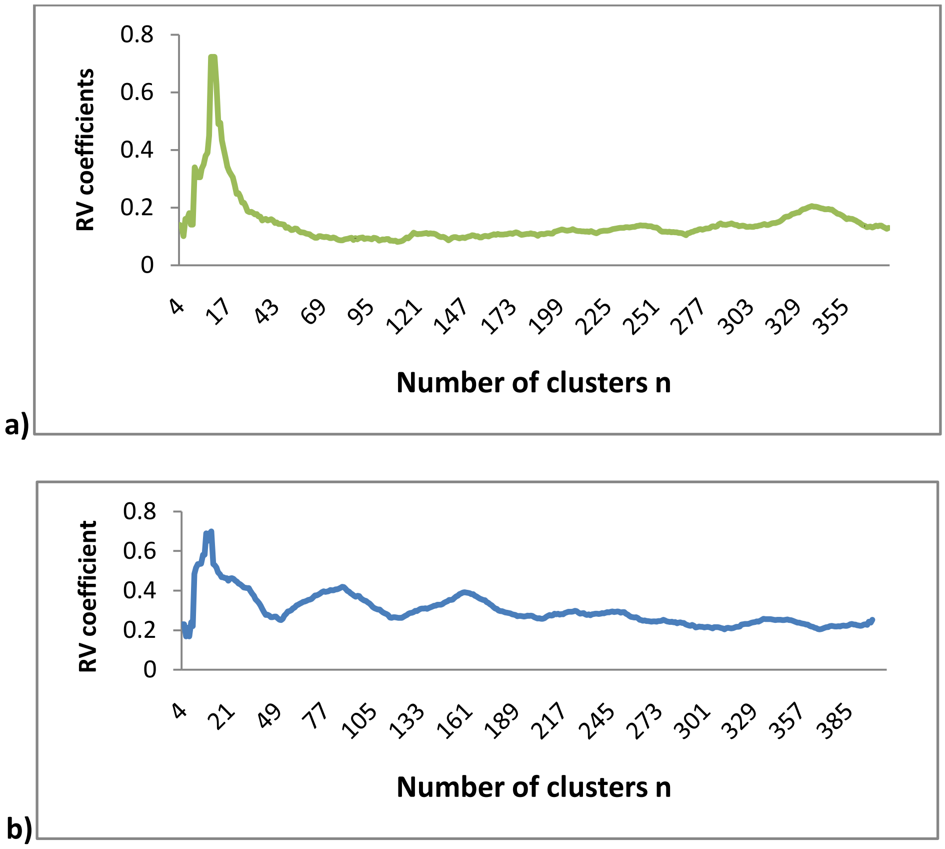


Figure 29. Verification of the termination condition.

Gene cluster numbers have been increased beyond termination for ‘*Sce*’ and ‘*Spo*’ datasets (panel a) and ‘*Homo sapiens*’ and ‘*Mus musculus*’ datasets (panel b).

### 5.3.3.2 Receiver operating characteristics

In order to study different granularities for both the *Saccharomyces/pombe* and the *human/mouse* merges, ROC curves were applied (Figure 30). The method is described in detail by Swets and Pickett (1982). Here, each point depicts a common network compared to known interactions (expected network) derived from various data sources (Appendix Table S1). The pathway databases were queried using the Ingenuity Pathway Analysis in order not to miss any existing interaction as described in the method section.

The cluster number *n* is plotted next to each data point. Both curves show similar increases in false positive rates when progressing to larger *n*.

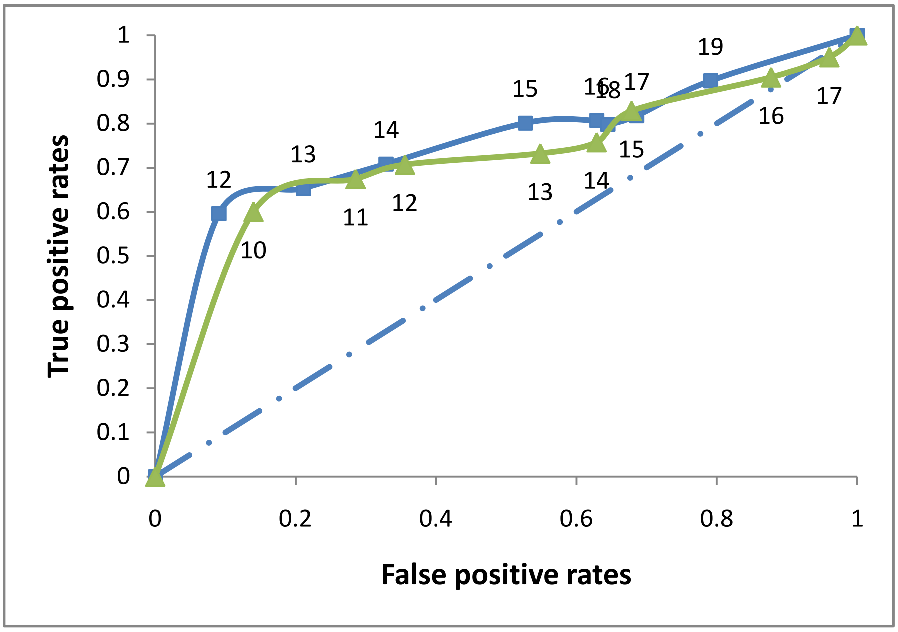


Figure 30. ROC curves.

The sensitivity is plotted against the specificity for '*Sce*' /'*Spo*' (blue) and mouse/human (green) common networks. Numbers annotate the number of clusters *n* (granularity) used for the common network. The diagonal dashed line is the expected ROC curve of a random predictor.

#### 5.3.3.3 Significance analysis of network motifs

Significance of network motifs for a specific network was studied by Z-Score and p-value (as described in section 5.2.7.1). Significance profiles on the basis of the Z-scores can be used to compare different networks. Furthermore, the frequency of motifs can directly be used for network evaluation.

In the analysis, motifs of sizes three and four are assessed to compare different directed networks. While the number of non-isomorphic motifs grows exponentially with the size of the motifs, in practice, only a fraction of all possible motifs is implemented by real biological networks. Up to the present time, known network motifs are small and usually comprise three to five vertices only.

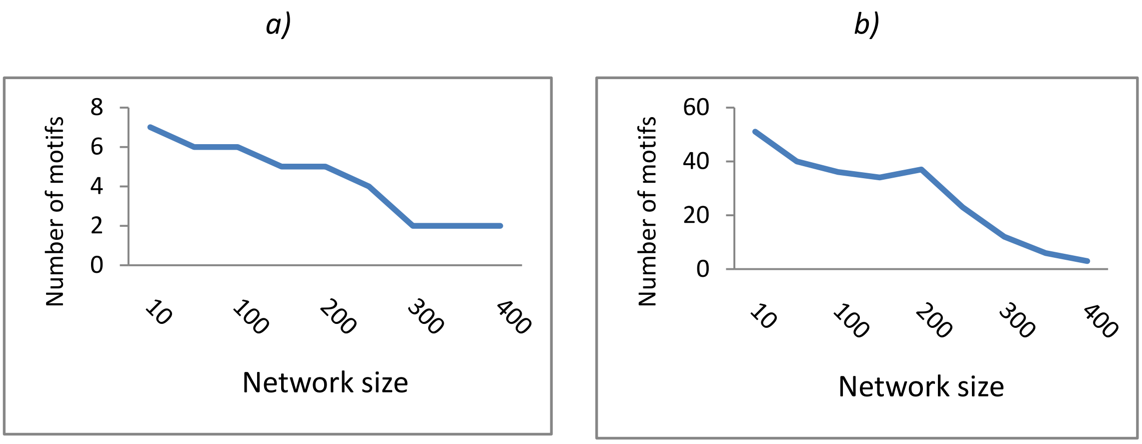


Figure 31. Significance of motifs in the mouse/human consensus network.

The number of motifs (size-3 and size-4 in panels 'a' and 'b', respectively) is plotted versus the network size. The number of discovered motifs decreases with the number of clusters n used for combining the datasets (network size).

For the calculation of the statistical significance of network motifs a commonly used method was used that compares the number of the observed network motifs to the concentration of the motifs in an ensemble randomized networks. Therefore, calculation of the motif statistics requires the consideration of several hundreds to thousands of randomized networks.

The size of the observed network determines the number of motifs. In general it is expected for the number of significant motifs to increase with the size of the network. Interestingly, highest numbers of significant motifs in the smaller networks was obtained, corroborating the authenticity of the common network.

### 5.3.4 Superiority of the common network

In order to assess the advantage of combining datasets using the algorithm, the common network was compared to the networks obtained from each single dataset. The networks inferred from "*Sce*" and "*Spo*" datasets are shown in Figure 23 and Figure 24 respectively.

The specificity and sensitivity of the networks compared to the common network is summarized in Table 7. The common network improves upon the single (“*Sce*” and “*Spo*”) networks in terms of absolute numbers of true positive and false positive edges, as well as in sensitivity, specificity, accuracy and the number of network motifs (Table 7).

Table 7. Comparison of each single dataset to the common network in Yeast

| | “*Sce*” network | “*Spo*” network | Common network |
|---|---|---|---|
| True positive edges | 17 | 21 | 53 |
| False positive edges | 22 | 25 | 5 |
| False negative edges | 31 | 26 | 36 |
| True negative edges | 74 | 72 | 50 |
| Sensitivity | 35% | 44% | 60% |
| Specificity | 77% | 74% | 91% |
| Accuracy | 63% | 64% | 72% |
| Number of network motifs | 13 | 18 | 21 |

## 5.4 Discussion

This section discusses peculiarities, pitfalls and computational challenges of meta-analysis in general, with a focus on combining information across organisms to model genetic regulatory networks in particular. Building on this discussion, last section suggests directions for future research to extend the work presented here.

### 5.4.1 Meta analysis comparison

The exponential growth in microarray datasets over the last decade opens the door for large-scale, cross-species comparisons. Analysis of sequence and interaction data have led to many important findings, and the algorithms and computational tools developed for these comparisons are routinely used.

Analysis of cross-species microarray data is challenging. Direct comparison of these experiments would require them to be carried out in a very similar manner in all species and temporal differences between the species would need to be accounted for prior to the actual comparisons. Still, many such studies were carried out in closely related species. These studies identified common and unique expression patterns in specific tissue types. They have uncovered conserved functional categories and interaction networks that are commonly activated in the different species. Table 8 summarizes the methods that have been suggested for analyzing such experiments.

Table 8. Comparison of methods developed for cross species expression analysis.

| | Co-expression meta-analysis | Expression meta-analysis | Indirect | Single species array | Multi species array | Combined analysis |
|---|---|---|---|---|---|---|
| Single platform (all experiments) | No | No | No | Yes | Yes | Yes |
| Customized array | No | No | No | No | Yes | No |
| Similar experimental conditions | No | Yes | Yes | Yes | Yes | Yes |
| Direct comparison of expression profiles | No | No | No | Yes | Yes | Yes |
| Applicable on distant species | Yes | Yes | Yes | No | No | Yes |
| Require orthology information | Yes | Yes | Yes | No | Yes | No |
| Separate p-value cutoff for each species | No | Yes | Yes | No | No | No |

## 5.4.2 Multivariate analysis

Integrating datasets into simultaneous analysis is a major challenge in systems biology. It is crucial to capture the associations between variables from different high-throughput multidimensional datasets. Different techniques exist to investigate the associations between large-scale datasets. Canonical Correlation Analysis (CCA; [191]), Partial Least Square (PLS; [192]) and Co-inertia Analysis (CIA; [157]) transform high-dimensional data into few, usually two or three dimensions for visualization.

PLS is a correlation based method. It explains relationships between two datasets by simultaneously decomposing the data matrices into low-dimensional vectors. CCA is a special case of PLS, identifying linear combinations of variables from each set such that they have maximum correlation. However CCA and PLS often suffer from the asymmetry of microarray datasets where the number of variables exceeds the number of samples.

Penalized CCA adapted with Elastic Net (CCA-EN; [193, 194]) and Sparse CCA (SCCA; [195]) are derivatives of the classical CCA. They incorporate variable selection to address above limitations of earlier approaches. However, if invoked repeatedly from within an unsupervised iterative algorithm, this becomes computationally infeasible for large numbers of variables.

In contrast, CIA can cope with both asymmetry and large numbers of variables without becoming computationally infeasible. CIA is a multivariate coupling approach measuring the adequacy between datasets. It was first introduced applying ecological data [157], and amino acid properties [196]. Culhane and co-workers demonstrated the efficiency of CIA on cross-platform comparisons of gene expression data, applying it to both cDNA and Affymetrix microarrays [73]. An extension of CIA that links more than two tables has been reported [197]. Fagan and co-workers combined information from multiple layers (genes, samples and GO terms) by CIA [198].

Throughout this work, I used CIA because of its visual interpretability, its speed and because its applicability to asymmetric microarray data had been demonstrated in many studies [73, 159, 198, 199].

## 5.4.3 Retaining information

Co-expression has been widely used to reveal, amongst others, functional relationships [200, 201] or to identify common regulatory motifs [202, 203]. Much like conserved sequence motifs, important regulatory patterns can be observed across species borders. In order to account for different scales such datasets may have, co-expression can be determined on the basis of intermediate results such as vote counting [41, 204], probabilities [205] or ranks. However, in order not to lose any information beforehand, I perform information reduction in the very process of combination. Co-inertia analysis [157] is particularly well–suited for this task, reducing dimensions based on the common variance (co-inertia) of two datasets. It can deal with datasets whose variables (genes) far exceed the number of observations (samples) and its use for microarray data has been demonstrated before [73].

### 5.4.4 Neo-functionalization

Most microarray based cross-species analyses rely on the mapping of orthologous genes between different organisms. Bergmann and coworkers [31], for instance, developed to this end a two-step approach in which they first, starting from a group of coexpressed genes in one organism, identified the corresponding homologs in a second organism. In a second step only homologs that also appeared coexpressed in the second reference organism are retained as functional homologs.

Studies which use homogeneous experiments, i.e. datasets that for both organisms contain similar conditions, rely on differences of gene expression to compare the changes in the transcriptional response between organisms. The correlation between the log ratios of all genes is used as a global indication of how much the conditions are comparable between the different organisms. Rifkin and coworkers [143] for example studied “evolutionary variation” of gene expression in *Drosophila* at the onset of metamorphosis by comparing to what extent orthologous genes exhibiting developmental changes during metamorphosis in one species were no longer differentially expressed during the same process in other members of the species. Lelandais and coworkers [36], compared the sporulation network between budding and fission yeasts using for both organisms similarly designed time series experiments. The authors proposed a method that superimposes the two species-specific coexpression networks by taking into account the structure of each individual network and the orthologous relations between the species.

All of these methods take as input the affiliation of genes (orthology information) between the datasets. However, sequence similarity based orthology does not account for evolutionary phenomena such as sub- and neo-functionalization, thus not necessarily representing functional orthology in every case [206]. In the course of this work, an algorithm has been developed that instead of identifying orthologs beforehand, affiliates genes on the basis of the expression data (section 5.2.2).

### 5.4.5 Comparison to KEGG

In an approach solely based on co-expression, genes that show identical expression behavior are indistinguishable, thus becoming one single entity. This entity can be viewed as a node in a GRN. Comparing such networks with known interactions supplied by KEGG and other repositories can provide an additional means to evaluate the performance of the algorithm. To this end, out of many algorithms proposed for network inference, I picked DBN as one of the successful algorithms to date for time-series [102, 207]. Non-time series data can be handled, for example by information-theoretic approaches [92] or algorithms based on ordinary differential equations (ODE) following transcriptional perturbations [208].

For DBN inference, as for other GRN inference methods, the number of observations is critical. The very high number of genes simultaneously measured for only a few samples ($g >> s$) raises a dimensionality problem. Moreover, a large majority of time series gene expression data contain no or very few repeated measurements of the expression level of the same gene at a given time. Even under the assumption of homogeneity, which enables to use the pairs of successive time point gene expression as repeated measurements, we still have to deal with the dimensionality problem ($g >> s$), when inferring the structure of dynamic acyclic graphs.

Several inference methods have been proposed to overcome the dimensionality problem. To name a few, Ong and coworkers [37], reduce the dimension of the problem by considering prior knowledge; Zou and Conzen [39] limit potential regulators of the genes with either earlier or simultaneous expression changes and estimate the transcription time lag; and Opgen-Rhein and Strimmer [40] proposed a model selection procedure based on an analytic shrinkage approach. All of these approaches either use prior knowledge for the network inference or estimate undirected gene networks from microarray gene expression.

In this thesis, I proposed an algorithm to obtain optimal numbers of clusters (nodes) sufficient for the common network inference. In general, due to a lack of samples, only few genes can make it as nodes for stable network inference.

### 5.4.6 Granularity

In order to obtain number and composition of nodes optimal for inferring a common network, the algorithm increases the granularity step by step (outer loop). For each *n*, the inner loop pairs the *n* clusters of each dataset, seeking for an inter-datasets affiliation of optimal co-inertia. It does so via a combination of CIA, Hungarian matching and majority voting, alternating between ‘connecting variable’ affiliations. While each of these steps “learns” from the previous one, the approach is non-greedy in that each decision on e.g. affiliating two genes (or clusters thereof) may be reversed in the next iteration.

Gene cluster affiliations were evaluated directly (within the same species), by counting gene orthologs, and by inferring common GRN. Although the third pair of datasets shows less similarity, the common GRN does not appear less accurate than that for the first and second. As the terminal granularity (final number of distinguishable clusters) is crucial for network inference, I carefully evaluated the termination point for the algorithm. The largest (optimal) numbers for RV, for Silhouette values, for true positive interactions and for the yield of network motifs all coincide for n=12 in the second example as well as for n=10 in the third. In the second example, the iteration part of the algorithm could not come up to the final *n* because the number of clusters cannot exceed the number of samples in the smaller dataset (here both 10). Instead, n=12 was determined by the refinement part while for the third dataset the decider module determined that further refinement was not beneficial. Generally, in my hands a yielding granularity never exceeded the number of samples (arrays) by far if at all. However, after back-transformation and RE, the inferred network comprises 21 significant network motifs and the delineated edges show 72% accuracy, 91% specificity and, remarkably, 60% sensitivity in comparison to known interactions. Thus, the chosen granularity, although it is small in comparison to the number of genes, resulted in a robust and most informative network.

Furthermore, this common network shows increased specificity, sensitivity, and accuracy, as well as more significant network motifs if compared to the networks inferred from the single datasets. This demonstrates that it is possible to successfully combine datasets solely on the basis of co-expression, without applying any further information. To my knowledge, the algorithm represents a novelty in this respect.

## 5.4.7 Future work

### 5.4.7.1 Information fusion

Information fusion of diverse data sources gives the opportunity to reveal insights not readily apparent when sources are examined individually. Towards information fusion, one would use publicly available RNA-seq data from GEO/SRA to advance the algorithm for identifying expression networks that are conserved among a variety of species as well as species specific ones. RNA-seq data are particularly useful for such a comparative profiling approach. Recent advances in RNA-seq have opened the way for comprehensive analysis of any transcriptome [209]. In principle, RNA-seq allows analysis of all expressed transcripts, with annotating the structures of all transcribed genes, quantifying expression of each transcript and measuring the extent of alternative splicing. To this end, one would first concentrate on the expression data alone and then incorporate phylogenetic foot printing information of regulatory elements in order to strengthen the prediction with additional evidence. Once a core set of conserved expression modules have identified, one can use them to predict putative functions for their poorly characterized members or novel regulatory pathways. This should greatly reduce the noise and false-positive predictions compared to single organism gene expression profiling approaches. For instance, in case of treatments of different organisms with the same drug, one could use this method to identify drug target sites much more reliably.

#### 5.4.7.2 Further applications

The algorithm does not only identify a correspondence between the genes, but also between the conditions. It would therefore provide the opportunity to study of more poorly characterized cells and relate to another study of better-characterized cells. For instance, it can be used relating different cancers to different developmental stages or stem cell features to elucidate underlying mechanisms in cancer progression.

External knowledge can be made available to the method via the penalty matrices. These can be weighted according to known similarities between genes and/or between samples. Here, however, all external knowledge is used for evaluation purposes.

Requiring no beforehand affiliation, the algorithm can be used for automated large-scale combination of microarray datasets. Back-transformation results in an artificial data table containing only the variance common to the two initial tables while retaining the scale of the first table. Thus, it can be handled like any real data table, e.g. for subsequent GRN inference or for combining it with yet another real data table or a combination of such. Thus, the method can be extended to linking more than two datasets, either hierarchically merging back-transformed data tables, or by using multiple co-inertia analysis.

With increasing numbers of datasets and integrating external knowledge to be summarized in one model, the common variance will decrease. Generally speaking, I would expect a tendency for such a model to be small, widely applicable, robust, and relatively free of noise and systematic errors when multiple experimental platforms are mixed.

When comparing many different datasets, I would expect many genes/samples show changes that are not to be matched across all datasets. The more experiments the more objects lacking their variant to be matched. If these objects remain in the iteration step of the algorithm, they may add considerable amount of noise to the model. Therefore a strategy for Garbage Collection would be needed to increase the overall RV-coefficient of the common model.

Extensive cross-species models could be useful in a pharmacological context in order to predict if a model organism closely resembles a human regulatory mechanism to interfere with. Furthermore, the application of this algorithm is not limited to microarray data. It could serve to integrate proteomic, transcriptomic, and high-throughput methylation data recorded for the same samples.

# 6 CONCLUSION

In this work an algorithm has been developed that contributes to the meta-analysis of high throughput expression data. It advocates the use of data-driven algorithms to combine multiple expression data in cross species studies. It is an iterative procedure using existing methods to estimate the common regulatory network from data of different species. It affiliates arrays and genes at the same time without any requirements of gene/sample affiliations or any other prior knowledge. It provides the opportunity to systematically combine arbitrary number of microarray datasets.

Using co-inertia analysis, the algorithm can deal with both asymmetric microarray datasets and large number of variables without becoming computationally intensive. In particular, it can cope with datasets whose genes far exceed the number of samples (arrays).

In terms of cluster granularity, the algorithm also provides a way to obtain the optimal number of nodes for the common network inference. This is an essential step prior to any reverse engineering approach. In general, due to the sample size limit, only few genes can make it as nodes for stable network inference.

Successful application of the algorithm is demonstrated on two independent but closely related experiments on *Schizosaccharomyces pombe*. These datasets provide more prior knowledge and most reliable gene affiliations since both datasets stem from the same species. Extending applicability of the algorithm, its performance on two yeast cell cycle studies comprising nearly identical experimental conditions but on two different species (*Saccharomyces cerevisiae, Schizosaccharomyces pombe*) is presented. In order to support its application to different levels of similarity, underlying structure and experimental platform, it was further applied to Estrogen-regulated gene expression studies of *Homo sapiens* and *Mus musculus.*

The performance of the algorithm was demonstrated by reversely engineering the combined dataset. In particular, reverse engineering the common network provided the opportunity to compare the occurrence of a motif in the reversely engineered network to the occurrence of the same motif in the randomized network. The resulting network constructed on the combined datasets yielded more significant network motifs than for the single dataset.

Moreover, comparing reversely engineered gene regulatory networks of each individual and combined dataset with respect to known interactions supplied by KEGG and other repositories provided an additional means to evaluate the performance of the algorithm. The resulting cross-species networks improve on the networks inferred from each dataset alone by yielding more of the interactions already recorded in KEGG and other databases. This success advocates a purely data-driven approach to combining multiple datasets across species. Being readily extendable to more than two datasets, the algorithm provides the opportunity to infer extensive gene regulatory networks.

# 7 BIBLIOGRAPHY

1. Quackenbush, J., *Microarray analysis and tumor classification.* N Engl J Med, 2006. **354**(23): p. 2463-72.
2. Hughes, T.R., et al., *Functional discovery via a compendium of expression profiles.* Cell, 2000. **102**(1): p. 109-26.
3. Ramasamy, A., et al., *Key Issues in Conducting a Meta-Analysis of Gene Expression Microarray Datasets.* PLoS Med, 2008. **5**(9): p. e184.
4. Wilkins, A., *The Evolution of Developmental Pathways*. 2001, Sunderland, MA, USA: Sinauer Associates.
5. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
6. McGuire, A.M., J.D. Hughes, and G.M. Church, *Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes.* Genome Res, 2000. **10**(6): p. 744-57.
7. Pennacchio, L.A. and E.M. Rubin, *Genomic strategies to identify mammalian regulatory sequences.* Nat Rev Genet, 2001. **2**(2): p. 100-9.
8. Kellis, M., et al., *Sequencing and comparison of yeast species to identify genes and regulatory elements.* Nature, 2003. **423**(6937): p. 241-54.
9. Stark, A., et al., *Identification of Drosophila MicroRNA targets.* PLoS Biol, 2003. **1**(3): p. E60.
10. Loots, G.G., et al., *Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons.* Science, 2000. **288**(5463): p. 136-40.
11. Krogan, N.J., et al., *Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.* Nature, 2006. **440**(7084): p. 637-43.
12. Alexander, P.A., et al., *The design and characterization of two proteins with 88% sequence identity but different structure and function.* Proc Natl Acad Sci U S A, 2007. **104**(29): p. 11963-8.
13. Spellman, P.T., et al., *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.* Mol Biol Cell, 1998. **9**(12): p. 3273--3297.
14. Rustici, G., et al., *Periodic gene expression program of the fission yeast cell cycle.* Nat Genet, 2004. **36**(8): p. 809--817.
15. Nau, G.J., et al., *Human macrophage activation programs induced by bacterial pathogens.* Proc Natl Acad Sci U S A, 2002. **99**(3): p. 1503-8.
16. Correa, A., et al., *Multiple oscillators regulate circadian gene expression in Neurospora.* Proc Natl Acad Sci U S A, 2003. **100**(23): p. 13597-602.
17. Arbeitman, M.N., et al., *Gene expression during the life cycle of Drosophila melanogaster.* Science, 2002. **297**(5590): p. 2270-5.
18. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.

19. DeRisi, J., et al., *Use of a cDNA microarray to analyse gene expression patterns in human cancer.* Nat Genet, 1996. **14**(4): p. 457-60.
20. Golub, T.R., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**(5439): p. 531-7.
21. Perou, C.M., et al., *Distinctive gene expression patterns in human mammary epithelial cells and breast cancers.* Proc Natl Acad Sci U S A, 1999. **96**(16): p. 9212-7.
22. Staunton, J.E., et al., *Chemosensitivity prediction by transcriptional profiling.* Proc Natl Acad Sci U S A, 2001. **98**(19): p. 10787-92.
23. Dan, S., et al., *An integrated database of chemosensitivity to 55 anticancer drugs and gene expression profiles of 39 human cancer cell lines.* Cancer Res, 2002. **62**(4): p. 1139-47.
24. Chang, H.Y., et al., *Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival.* Proc Natl Acad Sci U S A, 2005. **102**(10): p. 3738-43.
25. Michiels, S., S. Koscielny, and C. Hill, *Prediction of cancer outcome with microarrays: a multiple random validation strategy.* Lancet, 2005. **365**(9458): p. 488-92.
26. Dupuy, A. and R.M. Simon, *Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting.* J Natl Cancer Inst, 2007. **99**(2): p. 147-57.
27. Ma, S. and J. Huang, *Regularized gene selection in cancer microarray meta-analysis.* BMC Bioinformatics, 2009. **10**: p. 1.
28. Marot, G. and C.-D. Mayer, *Sequential analysis for microarray data based on sensitivity and meta-analysis.* Stat Appl Genet Mol Biol, 2009. **8**: p. Article3.
29. Ramaswamy, S., et al., *Multiclass cancer diagnosis using tumor gene expression signatures.* Proc Natl Acad Sci U S A, 2001. **98**(26): p. 15149-54.
30. Stuart, J.M., et al., *A gene-coexpression network for global discovery of conserved genetic modules.* Science, 2003. **302**(5643): p. 249--255.
31. Bergmann, S., J. Ihmels, and N. Barkai, *Similarities and differences in genome-wide expression data of six organisms.* PLoS Biol, 2004. **2**(1): p. E9.
32. Zhu, W. and C.R. Buell, *Improvement of whole-genome annotation of cereals through comparative analyses.* Genome Res, 2007. **17**(3): p. 299-310.
33. Sharan, R., et al., *Conserved patterns of protein interaction in multiple species.* Proc Natl Acad Sci U S A, 2005. **102**(6): p. 1974-9.
34. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863--14868.
35. Oliva, A., et al., *The cell cycle-regulated genes of Schizosaccharomyces pombe.* PLoS Biol, 2005. **3**(7): p. e225.
36. Lelandais, G., et al., *Comparing gene expression networks in a multi-dimensional space to extract similarities and differences between organisms.* Bioinformatics, 2006. **22**(11): p. 1359-66.
37. Ong, I.M., J.D. Glasner, and D. Page, *Modelling regulatory pathways in E. coli from time series expression profiles.* Bioinformatics, 2002. **18 Suppl 1**: p. S241-8.
38. Perrin, B.E., et al., *Gene networks inference using dynamic Bayesian networks.* Bioinformatics, 2003. **19 Suppl 2**: p. ii138-48.
39. Zou, M. and S.D. Conzen, *A new dynamic Bayesian network (DBN) approach for*

*identifying gene regulatory networks from time course microarray data.* Bioinformatics, 2005. **21**(1): p. 71-9.

40. Opgen-Rhein, R. and K. Strimmer, *From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data.* BMC Syst Biol, 2007. **1**: p. 37.
41. Rhodes, D.R., et al., *Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.* Proc Natl Acad Sci U S A, 2004. **101**(25): p. 9309--9314.
42. Stevens, J.R. and R.W. Doerge, *Combining Affymetrix microarray results.* BMC Bioinformatics, 2005. **6**: p. 57.
43. Yang, I.V., et al., *Within the fold: assessing differential expression measures and reproducibility in microarray assays.* Genome Biol, 2002. **3**(11): p. research0062.
44. Irizarry, R.A., et al., *Summaries of Affymetrix GeneChip probe level data.* Nucleic Acids Res, 2003. **31**(4): p. e15.
45. Quackenbush, J., *Microarray data normalization and transformation.* Nat Genet, 2002. **32 Suppl**: p. 496-501.
46. Yang, Y.H., et al., *Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.* Nucleic Acids Res, 2002. **30**(4): p. e15.
47. Schadt, E.E., et al., *Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data.* J Cell Biochem Suppl, 2001. **Suppl 37**: p. 120-5.
48. Cleveland, W.S., *Robust locally weighted regression and smoothing scatterplots.* J. Amer. Stat. Assoc, 1979(74): p. 829–836.
49. Hoffmann, R., T. Seidl, and M. Dugas, *Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis.* Genome Biol, 2002. **3**(7): p. RESEARCH0033.
50. Tamayo, P., et al., *Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.* Proc Natl Acad Sci U S A, 1999. **96**(6): p. 2907-12.
51. Toronen, P., et al., *Analysis of gene expression data using self-organizing maps.* FEBS Lett, 1999. **451**(2): p. 142-6.
52. Wang, J., et al., *Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study.* BMC Bioinformatics, 2002. **3**: p. 36.
53. Herrero, J., A. Valencia, and J. Dopazo, *A hierarchical unsupervised growing neural network for clustering gene expression patterns.* Bioinformatics, 2001. **17**(2): p. 126-36.
54. Butte, A.J. and I.S. Kohane, *Unsupervised knowledge discovery in medical databases using relevance networks.* Proc AMIA Symp, 1999: p. 711-5.
55. Kim, S.K., et al., *A gene expression map for Caenorhabditis elegans.* Science, 2001. **293**(5537): p. 2087-92.
56. Raychaudhuri, S., J.M. Stuart, and R.B. Altman, *Principal components analysis to summarize microarray experiments: application to sporulation time series.* Pac Symp Biocomput, 2000: p. 455--466.
57. Weinstein, J.N., et al., *An information-intensive approach to the molecular*

*pharmacology of cancer.* Science, 1997. **275**(5298): p. 343-9.
58. Wen, X., et al., *Large-scale temporal gene expression mapping of central nervous system development.* Proc Natl Acad Sci U S A, 1998. **95**(1): p. 334-9.
59. Soukas, A., et al., *Leptin-specific patterns of gene expression in white adipose tissue.* Genes Dev, 2000. **14**(8): p. 963-80.
60. Aronow, B.J., et al., *Divergent transcriptional responses to independent genetic causes of cardiac hypertrophy.* Physiol Genomics, 2001. **6**(1): p. 19-28.
61. Fallah, S., D. Tritchler, and J. Beyene, *Estimating number of clusters based on a general similarity matrix with application to microarray data.* Stat Appl Genet Mol Biol, 2008. **7**(1): p. Article24.
62. Arima, C., et al., *Modified fuzzy gap statistic for estimating preferable number of clusters in fuzzy k-means clustering.* J Biosci Bioeng, 2008. **105**(3): p. 273-81.
63. Bolshakova, N. and F. Azuaje, *Estimating the number of clusters in DNA microarray data.* Methods Inf Med, 2006. **45**(2): p. 153-7.
64. Dudoit, S. and J. Fridlyand, *A prediction-based resampling method for estimating the number of clusters in a dataset.* Genome Biol, 2002. **3**(7): p. RESEARCH0036.
65. Costa, J.A. and M.L. Netto, *Estimating the number of clusters in multivariate data by self-organizing maps.* Int J Neural Syst, 1999. **9**(3): p. 195-202.
66. Bolshakova, N. and F. Azuaje, *Machaon CVE: cluster validation for gene expression data.* Bioinformatics, 2003. **19**(18): p. 2494-5.
67. Azuaje, F., *Clustering-based approaches to discovering and visualising microarray data patterns.* Brief Bioinform, 2003. **4**(1): p. 31-42.
68. Azuaje, F., *Genomic data sampling and its effect on classification performance assessment.* BMC Bioinformatics, 2003. **4**: p. 5.
69. Rousseeuw, P., *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.* Journal of Computational and Applied Mathematics, 1987. **20**(1): p. 53 -- 65.
70. Dunn, J., *Well separated clusters and optimal fuzzy partitions.* J.Cybernetics, 1974: p. 95-104.
71. Bezdek, J. and N. Pal, *Well separated clusters and optimal fuzzy partitions.* IEEE Transactions on Systems, 1998: p. 301-15.
72. Saal, L.H., et al., *Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity.* Proc Natl Acad Sci U S A, 2007. **104**(18): p. 7564-9.
73. Culhane, A.C., G. Perrière, and D.G. Higgins, *Cross-platform comparison and visualisation of gene expression data using co-inertia analysis.* BMC Bioinformatics, 2003. **4**: p. 59.
74. Fellenberg, K., et al., *Correspondence analysis applied to microarray data.* Proc Natl Acad Sci U S A, 2001. **98**(19): p. 10781--10786.
75. Neumann, J.V., *Theory of Self-Reproducing Automata*, ed. A.W. Burks. 1996, Champaign: University of Illinois Press.
76. Kauffman, S.A., *Metabolic stability and epigenesis in randomly constructed genetic nets.* J Theor Biol, 1969. **22**(3): p. 437-67.
77. I. Shmulevich, A.S., O. Yli-Harja, J. Astola. *Inference of genetic regulatory networks under the best-fit extension paradigm*. in *Proceedings of the IEEE EURASIP Workshop*

*on Nonlinear Signal and Image Proc*. 2001.

78. Thomas, R., D. Thieffry, and M. Kaufman, *Dynamical behaviour of biological regulatory networks--I. Biological role of feedback loops and practical use of the concept of the loop-characteristic state.* Bull Math Biol, 1995. **57**(2): p. 247-76.
79. R. Thomas and R. d'Ari, *Biological feedback*. 1990, FL, USA: CRC Press.
80. Huang, S., *Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery.* J Mol Med, 1999. **77**(6): p. 469-80.
81. Thieffry, D. and R. Thomas, *Qualitative analysis of gene networks.* Pac Symp Biocomput, 1998: p. 77-88.
82. Bornholdt, S., *Systems biology. Less is more in modeling large genetic networks.* Science, 2005. **310**(5747): p. 449-51.
83. Wagner, A., *Circuit topology and the evolution of robustness in two-gene circadian oscillators.* Proc Natl Acad Sci U S A, 2005. **102**(33): p. 11775-80.
84. Albert, R. and H.G. Othmer, *The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in Drosophila melanogaster.* J Theor Biol, 2003. **223**(1): p. 1-18.
85. Li, F., et al., *The yeast cell-cycle network is robustly designed.* Proc Natl Acad Sci U S A, 2004. **101**(14): p. 4781-6.
86. McAdams, H.H. and A. Arkin, *Stochastic mechanisms in gene expression.* Proc Natl Acad Sci U S A, 1997. **94**(3): p. 814-9.
87. Butte, A.J. and I.S. Kohane, *Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements.* Pac Symp Biocomput, 2000: p. 418-29.
88. Choi, J.K., et al., *Differential coexpression analysis using microarray data and its application to human cancer.* Bioinformatics, 2005. **21**(24): p. 4348-55.
89. Huang, Y., et al., *Systematic discovery of functional modules and context-specific functional annotation of human genome.* Bioinformatics, 2007. **23**(13): p. i222--i229.
90. Ucar, D., et al., *Construction of a reference gene association network from multiple profiling data: application to data analysis.* Bioinformatics, 2007. **23**(20): p. 2716-24.
91. Faith, J.J., et al., *Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles.* PLoS Biol, 2007. **5**(1): p. e8.
92. Basso, K., et al., *Reverse engineering of regulatory networks in human B cells.* Nat Genet, 2005. **37**(4): p. 382-90.
93. Margolin, A.A., et al., *ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.* BMC Bioinformatics, 2006. **7 Suppl 1**: p. S7.
94. Silkes, J.P., M.R. McNeil, and M. Drton, *Simulation of aphasic naming performance in non-brain-damaged adults.* J Speech Lang Hear Res, 2004. **47**(3): p. 610-23.
95. Ott, S., S. Imoto, and S. Miyano, *Finding optimal models for small gene networks.* Pac Symp Biocomput, 2004: p. 557-67.
96. de Jong, H., *Modeling and simulation of genetic regulatory systems: a literature review.* J Comput Biol, 2002. **9**(1): p. 67-103.
97. D. Heckerman, *A tutorial on learning with bayesian networks*. 1995, Redmond, WA, USA: Microsoft Research.

98. Friedman, N., et al., *Using Bayesian networks to analyze expression data.* J Comput Biol, 2000. **7**(3-4): p. 601--620.
99. Hasty, J., et al., *Computational studies of gene regulatory networks: in numero molecular biology.* Nat Rev Genet, 2001. **2**(4): p. 268-79.
100. Smolen, P., D.A. Baxter, and J.H. Byrne, *Modeling transcriptional control in gene networks--methods, recent results, and future directions.* Bull Math Biol, 2000. **62**(2): p. 247-92.
101. Friedman, N.M.K. and S. Russell, *Learning the structure of dynamic probabilistic networks*. 1998.
102. Yu, J., et al., *Advances to Bayesian network inference for generating causal networks from observational biological data.* Bioinformatics, 2004. **20**(18): p. 3594--3603.
103. Smith, V.A., E.D. Jarvis, and A.J. Hartemink, *Evaluating functional network inference using simulations of complex biological systems.* Bioinformatics, 2002. **18 Suppl 1**: p. S216--S224.
104. Bernard, A. and A.J. Hartemink, *Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data.* Pac Symp Biocomput, 2005: p. 459--470.
105. Khan, M.S., et al., *Sulfite reductase defines a newly discovered bottleneck for assimilatory sulfate reduction and is essential for growth and development in Arabidopsis thaliana.* Plant Cell, 2010. **22**(4): p. 1216-31.
106. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.
107. Takahashi, H., et al., *The roles of three functional sulphate transporters involved in uptake and translocation of sulphate in Arabidopsis thaliana.* Plant J, 2000. **23**(2): p. 171-82.
108. Takahashi, H. and K. Saito, *Molecular biology and functional genomics for identification of regulatory networks of plant sulfate uptake and assimilatory metabolism.* , in *Sulfur Metabolism in Phototrophic Organisms*, R. Hell, C. Dahl, and T. Leustek, Editors. 2008, The Netherlands: Springer: Dordrecht. p. 151-164.
109. Mugford, S.G., et al., *Disruption of adenosine-5'-phosphosulfate kinase in Arabidopsis reduces levels of sulfated secondary metabolites.* Plant Cell, 2009. **21**(3): p. 910-27.
110. Leustek, T., et al., *Pathways and Regulation of Sulfur Metabolism Revealed through Molecular and Genetic Studies.* Annu Rev Plant Physiol Plant Mol Biol, 2000. **51**: p. 141-165.
111. Vauclare, P., et al., *Flux control of sulphate assimilation in Arabidopsis thaliana: adenosine 5'-phosphosulphate reductase is more susceptible than ATP sulphurylase to negative control by thiols.* Plant J, 2002. **31**(6): p. 729-40.
112. Loudet, O., et al., *Natural variation for sulfate content in Arabidopsis thaliana is highly controlled by APR2.* Nat Genet, 2007. **39**(7): p. 896-900.
113. Nakayama, M., T. Akashi, and T. Hase, *Plant sulfite reductase: molecular structure, catalytic function and interaction with ferredoxin.* J Inorg Biochem, 2000. **82**(1-4): p. 27-32.
114. Yonekura-Sakakibara, K., et al., *Analysis of reductant supply systems for ferredoxin-dependent sulfite reductase in photosynthetic and nonphotosynthetic organs of maize.* Plant Physiol, 2000. **122**(3): p. 887-94.

115. Swamy, U., et al., *Structure of spinach nitrite reductase: implications for multi-electron reactions by the iron-sulfur:siroheme cofactor.* Biochemistry, 2005. **44**(49): p. 16054-63.
116. Patron, N.J., D.G. Durnford, and S. Kopriva, *Sulfate assimilation in eukaryotes: fusions, relocations and lateral transfers.* BMC Evol Biol, 2008. **8**: p. 39.
117. Kopriva, S., *Regulation of sulfate assimilation in Arabidopsis and beyond.* Ann Bot, 2006. **97**(4): p. 479-95.
118. Zimmermann, P., et al., *GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox.* Plant Physiol, 2004. **136**(1): p. 2621-32.
119. Brychkova, G., et al., *Sulfite oxidase protects plants against sulfur dioxide toxicity.* Plant J, 2007. **50**(4): p. 696-709.
120. Lang, C., et al., *Sulphite oxidase as key enzyme for protecting plants against sulphur dioxide.* Plant Cell Environ, 2007. **30**(4): p. 447-55.
121. Martin, M.N., et al., *The role of 5'-adenylylsulfate reductase in controlling sulfate reduction in plants.* Photosynth Res, 2005. **86**(3): p. 309-23.
122. Haas, F.H., et al., *Mitochondrial serine acetyltransferase functions as a pacemaker of cysteine synthesis in plant cells.* Plant Physiol, 2008. **148**(2): p. 1055-67.
123. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3**: p. Article3.
124. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* J. R. Stat. Soc. Ser. B, 1995. **57**: p. 289â€"300.
125. Kruse, C., et al., *Sulfur-enhanced defence: effects of sulfur metabolism, nitrogen supply, and pathogen lifestyle.* Plant Biol (Stuttg), 2007. **9**(5): p. 608-19.
126. Boettcher, M., et al., *Decoding pooled RNAi screens by means of barcode tiling arrays.* BMC Genomics, 2010. **11**(1): p. 7.
127. Sotiriou, C., et al., *Breast cancer classification and prognosis based on gene expression profiles from a population-based study.* Proc Natl Acad Sci U S A, 2003. **100**(18): p. 10393-8.
128. Silva, J.M., et al., *Profiling essential genes in human mammary cells by multiplex RNAi screening.* Science, 2008. **319**(5863): p. 617-20.
129. Berns, K., et al., *A large-scale RNAi screen in human cells identifies new components of the p53 pathway.* Nature, 2004. **428**(6981): p. 431-7.
130. Baum, M., et al., *Validation of a novel, fully integrated and flexible microarray benchtop facility for gene expression profiling.* Nucleic Acids Res, 2003. **31**(23): p. e151.
131. Parkinson, H., et al., *ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression.* Nucleic Acids Res, 2009. **37**(Database issue): p. D868--D872.
132. Moghaddas Gholami, A. and K. Fellenberg, *Cross-species common regulatory network inference without requirement for prior gene affiliation.* Bioinformatics, 2010. **26**(8): p. 1082-90.
133. Kapushesky, M., et al., *Gene expression atlas at the European bioinformatics institute.* Nucleic Acids Res, 2009. **38**(Database issue): p. D690-8.
134. Fortunel, N.O., et al., *Comment on " 'Stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature".* Science, 2003. **302**(5644):

p. 393; author reply 393.
135. Han, E.S. and M. Hickey, *Microarray evaluation of dietary restriction.* J Nutr, 2005. **135**(6): p. 1343-6.
136. Choi, J.K., et al., *Combining multiple microarray studies and modeling interstudy variation.* Bioinformatics, 2003. **19 Suppl 1**: p. i84--i90.
137. Su, A.I., et al., *A gene atlas of the mouse and human protein-encoding transcriptomes.* Proc Natl Acad Sci U S A, 2004. **101**(16): p. 6062-7.
138. Liao, B.Y. and J. Zhang, *Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution.* Mol Biol Evol, 2006. **23**(6): p. 1119-28.
139. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
140. Manoli, T., et al., *Group testing for pathway analysis improves comparability of different microarray datasets.* Bioinformatics, 2006. **22**(20): p. 2500-6.
141. Liu, Q., et al., *Comparative evaluation of gene-set analysis methods.* BMC Bioinformatics, 2007. **8**: p. 431.
142. Irizarry, R.A., et al., *Multiple-laboratory comparison of microarray platforms.* Nat Methods, 2005. **2**(5): p. 345-50.
143. Rifkin, S.A., J. Kim, and K.P. White, *Evolution of gene expression in the Drosophila melanogaster subgroup.* Nat Genet, 2003. **33**(2): p. 138-44.
144. Oshlack, A., et al., *Using DNA microarrays to study gene expression in closely related species.* Bioinformatics, 2007. **23**(10): p. 1235-42.
145. Nuzhdin, S.V., et al., *Common pattern of evolution of gene expression level and protein sequence in Drosophila.* Mol Biol Evol, 2004. **21**(7): p. 1308-17.
146. Whiteford, C.C., et al., *Credentialing preclinical pediatric xenograft models using gene expression and tissue microarray analysis.* Cancer Res, 2007. **67**(1): p. 32-40.
147. Gilad, Y., et al., *Expression profiling in primates reveals a rapid evolution of human transcription factors.* Nature, 2006. **440**(7081): p. 242-5.
148. Sartor, M.A., et al., *A new method to remove hybridization bias for interspecies comparison of global gene expression profiles uncovers an association between mRNA sequence divergence and differential gene expression in Xenopus.* Nucleic Acids Res, 2006. **34**(1): p. 185-200.
149. Gilad, Y., et al., *Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles.* Genome Res, 2005. **15**(5): p. 674-80.
150. Khaitovich, P., et al., *Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees.* Science, 2005. **309**(5742): p. 1850-4.
151. Vallee, M., et al., *Cross-species hybridizations on a multi-species cDNA microarray to identify evolutionarily conserved genes expressed in oocytes.* BMC Genomics, 2006. **7**: p. 113.
152. Whitfield, M.L., et al., *Identification of genes periodically expressed in the human cell cycle and their expression in tumors.* Mol Biol Cell, 2002. **13**(6): p. 1977-2000.
153. Menges, M., et al., *Cell cycle-regulated gene expression in Arabidopsis.* J Biol Chem, 2002. **277**(44): p. 41987-2002.
154. Laub, M.T., et al., *Global analysis of the genetic network controlling a bacterial cell*

*cycle.* Science, 2000. **290**(5499): p. 2144-8.

155. Alter, O., P.O. Brown, and D. Botstein, *Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms.* Proc Natl Acad Sci U S A, 2003. **100**(6): p. 3351--3356.
156. Lu, Y., R. Rosenfeld, and Z. Bar-Joseph, *Identifying cycling genes by combining sequence homology and expression data.* Bioinformatics, 2006. **22**(14): p. e314-22.
157. Dolédec, S. and D. Chessel, *Co-inertia analysis: an alternative method for studying species-environment relationships.* Freshwater Biology, 1994. **31**: p. 277-294.
158. Robert, P. and Y. Escoufier, *A unifying tool for linear multivariate statistical methods: the RV-coefficient.* Appl. Statist., 1976. **25**: p. 257–265.
159. Jeffery, I.B., et al., *Integrating transcription factor binding site information with gene expression datasets.* Bioinformatics, 2007. **23**(3): p. 298-305.
160. Kuhn, H.W., *The Hungarian method for the assignment problem.* Naval Research Logistics Quarterly, 1955. **2**: p. 83-87.
161. Bourgeois, F. and J.-C. Lassalle, *An extension of the Munkres algorithm for the assignment problem to rectangular matrices.* Communications of the ACM, 1971. **14**: p. 802-804.
162. Dray, S. and A.-B. Dufour, *The ade4 Package: Implementing the Duality Diagram for Ecologists.* Journal of Statistical Software., 2007. **22**: p. 1-20.
163. Dray, S., D. Chessel, and J. Thiolouse, *Co-inertia analysis and the linking of ecological data tables.* Ecology, 2003. **84**: p. 3078-3089.
164. Hartemink, A., et al. *Combining Location and Expression Data for Principled Discovery of Genetic Regulatory Networks*. in *Pacific Symposium on Biocomputing 2002 (PSB02)*. 2002: World Scientific: New Jersey.
165. Kanehisa, M., et al., *KEGG for linking genomes to life and the environment.* Nucleic Acids Res, 2008. **36**(Database issue): p. D480--D484.
166. Milo, R., et al., *Network motifs: simple building blocks of complex networks.* Science, 2002. **298**(5594): p. 824--827.
167. Wernicke, S., *Efficient detection of network motifs.* IEEE/ACM Trans Comput Biol Bioinform, 2006. **3**(4): p. 347--359.
168. Stossi, F., et al., *Transcriptional profiling of estrogen-regulated gene expression via estrogen receptor (ER) alpha or ERbeta in human osteosarcoma cells: distinct and common target genes for these receptors.* Endocrinology, 2004. **145**(7): p. 3473-86.
169. Moggs, J.G., et al., *Phenotypic anchoring of gene expression changes during estrogen-induced uterine growth.* Environ Health Perspect, 2004. **112**(16): p. 1589-606.
170. Cipollina, C., et al., *Revisiting the role of yeast Sfp1 in ribosome biogenesis and cell size control: a chemostat study.* Microbiology, 2008. **154**(Pt 1): p. 337-46.
171. Troyanskaya, O., et al., *Missing value estimation methods for DNA microarrays.* Bioinformatics, 2001. **17**(6): p. 520--525.
172. Bar-Joseph, Z., et al., *Continuous representations of time-series gene expression data.* J Comput Biol, 2003. **10**(3-4): p. 341--356.
173. Futschik, M.E. and H. Herzel, *Are we overestimating the number of cell-cycling genes? The impact of background models on time-series analysis.* Bioinformatics, 2008. **24**(8): p. 1063--1069.
174. Barrett, T., et al., *NCBI GEO: archive for high-throughput functional genomic data.*

Nucleic Acids Res, 2009. **37**(Database issue): p. D885--D890.

175. Huber, W., et al., *Variance stabilization applied to microarray data calibration and to the quantification of differential expression.* Bioinformatics, 2002. **18 Suppl 1**: p. S96--104.
176. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.
177. Santisteban, M.S., et al., *Histone octamer function in vivo: mutations in the dimer-tetramer interfaces disrupt both gene activation and repression.* EMBO J, 1997. **16**(9): p. 2493-506.
178. Rowland, B.D., et al., *Building sister chromatid cohesion: smc3 acetylation counteracts an antiestablishment activity.* Mol Cell, 2009. **33**(6): p. 763-74.
179. Unal, E., et al., *A molecular determinant for the establishment of sister chromatid cohesion.* Science, 2008. **321**(5888): p. 566-9.
180. Ben-Shahar, T.R., et al., *Eco1-dependent cohesin acetylation during establishment of sister chromatid cohesion.* Science, 2008. **321**(5888): p. 563-6.
181. Alon, U., *Network motifs: theory and experimental approaches.* Nat Rev Genet, 2007. **8**(6): p. 450-61.
182. Liu, C., et al., *Identification of the downstream targets of SIM1 and ARNT2, a pair of transcription factors essential for neuroendocrine cell differentiation.* J Biol Chem, 2003. **278**(45): p. 44857-67.
183. Harbison, C.T., et al., *Transcriptional regulatory code of a eukaryotic genome.* Nature, 2004. **431**(7004): p. 99-104.
184. Lee, T.I., et al., *Transcriptional regulatory networks in Saccharomyces cerevisiae.* Science, 2002. **298**(5594): p. 799-804.
185. Workman, C.T., et al., *A systems approach to mapping DNA damage response pathways.* Science, 2006. **312**(5776): p. 1054-9.
186. Zhu, G., et al., *Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth.* Nature, 2000. **406**(6791): p. 90-4.
187. Cipollina, C., et al., *Saccharomyces cerevisiae SFP1: at the crossroads of central metabolism and ribosome biogenesis.* Microbiology, 2008. **154**(Pt 6): p. 1686-99.
188. Chua, G., et al., *Identifying transcription factor functions and targets by phenotypic activation.* Proc Natl Acad Sci U S A, 2006. **103**(32): p. 12045-50.
189. Jorgensen, P., et al., *A dynamic transcriptional network communicates growth potential to ribosome synthesis and critical cell size.* Genes Dev, 2004. **18**(20): p. 2491-505.
190. Akache, B. and B. Turcotte, *New regulators of drug sensitivity in the family of yeast zinc cluster proteins.* J Biol Chem, 2002. **277**(24): p. 21254-60.
191. Gittins, R., *Canonical analysis, a review with applications in ecology.* Biomathematics, 1985. **12**.
192. Wold, H., *Multivariate Analysis*, ed. p. krishnaiah. 1966: Academic Press, New York.
193. Cao, K.-A.L., et al., *A sparse PLS for variable selection when integrating omics data.* Stat Appl Genet Mol Biol, 2008. **7**: p. Article 35.
194. Waaijenborg, S., P.C.V. de Witt Hamer, and A.H. Zwinderman, *Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis.* Stat Appl Genet Mol Biol, 2008. **7**(1): p. Article3.

195. Parkhomenko, E., D. Tritchler, and J. Beyene, *Sparse canonical correlation analysis with application to genomic data integration.* Stat Appl Genet Mol Biol, 2009. **8**: p. Article1.
196. Thioulouse, J. and J.R. Lobry, *Co-inertia analysis of amino-acid physico-chemical properties and protein composition with the ADE package.* Comput Appl Biosci, 1995. **11**(3): p. 321--329.
197. Dray, S., D. Chessel, and J. Thiolouse, *Procrustean co-inertia analysis for the linking of multivariate data sets.* Ecoscience, 2003. **10(1)**: p. 110-119.
198. Fagan, A.s., A.n.C. Culhane, and D.G. Higgins, *A multivariate analysis approach to the integration of proteomic and gene expression data.* Proteomics, 2007. **7**(13): p. 2162--2171.
199. Singh, A.V., K.B. Knudsen, and T.B. Knudsen, *Integrative analysis of the mouse embryonic transcriptome.* Bioinformation, 2007. **1**(10): p. 406-13.
200. Lage, K., et al., *A human phenome-interactome network of protein complexes implicated in genetic disorders.* Nat Biotechnol, 2007. **25**(3): p. 309--316.
201. Adie, E.A., et al., *SUSPECTS: enabling fast and effective prioritization of positional candidates.* Bioinformatics, 2006. **22**(6): p. 773--774.
202. Franke, L., et al., *Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes.* Am J Hum Genet, 2006. **78**(6): p. 1011--1025.
203. Brunner, H.G. and M.A. van Driel, *From syndrome families to functional genomics.* Nat Rev Genet, 2004. **5**(7): p. 545--551.
204. Smid, M., L.C.J. Dorssers, and G. Jenster, *Venn Mapping: clustering of heterologous microarray data based on the number of co-occurring differentially expressed genes.* Bioinformatics, 2003. **19**(16): p. 2065--2071.
205. Tsiporkova, E. and V. Boeva, *Fusing time series expression data through hybrid aggregation and hierarchical merge.* Bioinformatics, 2008. **24**(16): p. i63--i69.
206. Fierro, A.C., et al., *Meta Analysis of Gene Expression Data within and Across Species.* Curr Genomics, 2008. **9**(8): p. 525-34.
207. Smith, V.A., E.D. Jarvis, and A.J. Hartemink, *Influence of network topology and data collection on network inference.* Pac Symp Biocomput, 2003: p. 164--175.
208. Bansal, M., G.D. Gatta, and D. di Bernardo, *Inference of gene regulatory networks and compound mode of action from time course gene expression profiles.* Bioinformatics, 2006. **22**(7): p. 815-22.
209. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
210. Willis, R.C. and C.W. Hogue, *Searching, viewing, and visualizing data in the Biomolecular Interaction Network Database (BIND).* Curr Protoc Bioinformatics, 2006. **Chapter 8**: p. Unit 8 9.
211. Bader, G.D., D. Betel, and C.W.V. Hogue, *BIND: the Biomolecular Interaction Network Database.* Nucleic Acids Res, 2003. **31**(1): p. 248--250.
212. Bader, G.D. and C.W. Hogue, *BIND--a data specification for storing and describing biomolecular interactions, molecular complexes and pathways.* Bioinformatics, 2000. **16**(5): p. 465-77.
213. Breitkreutz, B.J., et al., *The BioGRID Interaction Database: 2008 update.* Nucleic Acids

Res, 2008. **36**(Database issue): p. D637-40.

214. Stark, C., et al., *BioGRID: a general repository for interaction datasets.* Nucleic Acids Res, 2006. **34**(Database issue): p. D535--D539.
215. Xenarios, I., et al., *DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions.* Nucleic Acids Res, 2002. **30**(1): p. 303-5.
216. Hermjakob, H., et al., *IntAct: an open source molecular interaction database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D452--D455.
217. Kerrien, S., et al., *IntAct--open source resource for molecular interaction data.* Nucleic Acids Res, 2007. **35**(Database issue): p. D561-5.
218. Aoki-Kinoshita, K.F. and M. Kanehisa, *Gene annotation and pathway mapping in KEGG.* Methods Mol Biol, 2007. **396**: p. 71-91.
219. Chatr-aryamontri, A., et al., *MINT: the Molecular INTeraction database.* Nucleic Acids Res, 2007. **35**(Database issue): p. D572-4.
220. Zanzoni, A., et al., *MINT: a Molecular INTeraction database.* FEBS Lett, 2002. **513**(1): p. 135--140.
221. Guldener, U., et al., *MPact: the MIPS protein interaction resource on yeast.* Nucleic Acids Res, 2006. **34**(Database issue): p. D436-41.
222. Mewes, H.W., et al., *MIPS: analysis and annotation of proteins from whole genomes in 2005.* Nucleic Acids Res, 2006. **34**(Database issue): p. D169-72.

# 8 APPENDICES

## 8.1 Banjo parameters

```
###-----------------------------------------------------
###  Search specifications
###-----------------------------------------------------
searcherChoice =                             SimAnneal
proposerChoice =                         AllLocalMoves
evaluatorChoice =             default to EvaluatorBDe
deciderChoice =       defaulted to DeciderMetropolis
statisticsChoice =                             default
###-----------------------------------------------------
###  Pre-processing options
###-----------------------------------------------------
discretizationPolicy =                              q3
createDiscretizationReport =        withMappedValues
###-----------------------------------------------------
### Search "problem domain" constraints
###-----------------------------------------------------
minMarkovLag =                                       1
maxMarkovLag =                                       1
dbnMandatoryIdentityLags =                           1
equivalentSampleSize =                             1.0
maxParentCount =                                     5
### Stopping criteria
###-----------------------------------------------------
maxTime =                                         10 m
minNetworksBeforeChecking =                       1000
###-----------------------------------------------------
### Parameters used by specific methods
###-----------------------------------------------------
### For simulated annealing:
```

```
initialTemperature =                           1000
maxAcceptedNetworksBeforeCooling =             1000
maxProposedNetworksBeforeCooling =            10000
minAcceptedNetworksBeforeReannealing =          200
reannealingTemperature =                        500
```

## 8.2 Supplementary Table S1

Table S1. List of pathway data sources used to explore expected network

| Abbr. | Name | Description | PubMed Articles | Content | |
|---|---|---|---|---|---|
| BIND | Biomolecular Interaction Network Database | Full descriptions of interactions, molecular complexes and pathways | [210-212] | Genes / Proteins:<br>Interactions / Reactions:<br>Experiments / PubMed IDs: | 57,971<br>198,905<br>23,010 |
| BioGRID | General Repository for Interaction Datasets | Protein-protein and genetic interaction networks | [213, 214] | Proteins:<br>Publications:<br>Organisms: | 529,018<br>21,268<br>22 |
| DIP | Database of Interacting Proteins | Experimentally determined interactions between proteins | [215] | Proteins:<br>Organisms:<br>Interactions:<br>Experiments describing an interaction:<br>Articles: | 20,728<br>274<br>57,683<br>64,952<br>3,915 |
| IntAct | IntAct | Protein-protein interactions maintained by the European Bioinformatics Institute (EBI). | [216, 217] | Proteins:<br>Interactions:<br>Experiments: | 59,971<br>201,094<br>10,900 |
| KEGG | Kyoto Encyclopedia of Genes and Genomes | Database of metabolic pathways from multiple organisms | [165, 218] | Genes / Proteins:<br>Small molecules:<br>Interactions / reactions:<br>Pathways: | 4,964,241<br>16,000<br>8,022<br>96,160 |
| MINT | Molecular Interaction Database | Database of molecular and protein-protein interactions | [219, 220] | Genes / Proteins:<br>Interactions / Reactions:<br>Experiments / PubMed IDs: | 29,718<br>83,210<br>3,141 |
| MIPS CYGD | MIPS Comprehensive Yeast Genome Database | Protein interactions, protein complexes and metabolic pathway diagrams for budding yeast. | [221, 222] | not available. | |

## 8.3 Publications

- **Moghaddas Gholami, A**. and K. Fellenberg, *Cross-species common regulatory network inference without requirement for prior gene affiliation.* Bioinformatics, 2010. **26**(8): p. 1082-90.

- Khan, M.S., F.H. Haas, A.A. Samami, **A. Moghaddas Gholami**, A. Bauer, K. Fellenberg, M. Reichelt, R. Hansch, R.R. Mendel, A.J. Meyer, M. Wirtz, and R. Hell, *Sulfite reductase defines a newly discovered bottleneck for assimilatory sulfate reduction and is essential for growth and development in Arabidopsis thaliana.* Plant Cell, 2010. **22**(4): p. 1216-31.

- Boettcher, M., J. Fredebohm, **A. Moghaddas Gholami**, Y. Hachmo, I. Dotan, D. Canaani, and J.D. Hoheisel, *Decoding pooled RNAi screens by means of barcode tiling arrays.* BMC Genomics, 2010. **11**(1): p. 7.

Publication not included in this thesis:

- Kumar Botla, S., M. Malekpour, **A. Moghaddas Gholami**, E. Moskalev, A. Aghajani, R. Omranipour, F. Malekpour, V. Bubnov, H. Najmabadi, J. Hoheisel, and Y. Riazalhosseini, *A breast-cancer DNA methylation signature associated with field defect.* (in press. *The Lancet Oncology*)

## 8.4 Curriculum vitae

### Personal details

Name: Amin Moghaddas Gholami

Date of Birth: 31/08/1976

Place of Birth: Tehran – Iran

### Educations

| | | |
|---|---|---|
| April 2007 – Present | PhD candidate in Bioinformatics | |
| | Technische Universität München | Germany |
| Dec 2006 | MSc in Bioinformatics | |
| | University of Leicester | UK |
| Sep 94 – Oct 2000 | BSc in Computer Engineering | |
| | Azad University of Tehran | Iran |

### Awards

| | | |
|---|---|---|
| Dec 2006 | The Pfizer Prize for Student Achievement | UK |
| May 2006 | The Novartis International Studentship | UK |

### Publications

- **Moghaddas Gholami, A.** and Fellenberg, K. (2010) Cross-species common regulatory network inference without requirement for prior gene affiliation, *Bioinformatics*, **26**, 1082-1090. PMID: 20200011.

- Khan, M.S., Haas, FH., Samami, AA., **Moghaddas Gholami, A.**, Bauer, A., Fellenberg, K., Reichelt, M., Hänsch, R., Mendel, R.R., Meyer, A.J., Wirtz, M., Hell, R. (2010) Sulfite Reductase Defines a Newly Discovered Bottleneck for Assimilatory Sulfate Reduction and Is Essential for Growth and Development in Arabidopsis thaliana. *Plant Cell.* PMID: 20424176.

- Boettcher, M., Fredebohm, J., **Moghaddas Gholami, A.**, Hachmo, Y., Dotan, I., Canaani, D. and Hoheisel, J.D. (2010) Decoding pooled RNAi screens by means of barcode tiling arrays, *BMC Genomics*, **11**, 7. PMID: 20051122.

- Kumar Botla, S., Malekpour, M., **Moghaddas Gholami, A.**, Moskalev, Aghajani, A., Omranipour, R., Malekpour, F., Bubnov, V., Najmabadi, H., Hoheisel, J.D., Riazalhosseini, Y. (2010) A breast-cancer DNA methylation signature associated with field defect. (in press, *The Lancet Oncology*)

**Refereed Conference publications**

- Zhixiang Wu, **Amin Moghaddas Gholami**, Kurt Fellenberg, Simone Lemeer, Bernhard Küster (2010) Comparison of label-free protein quantification approaches for chemical proteomics. 58th ASMS Conference on Mass Spectrometry and Allied Topics. Utah, USA. May 2010

- Yasser Riazalhosseini, **Amin Moghaddas Gholami**, Mahdi Malekpour, Sandeep Kumar Botla, Vladimir Bobnov, Azin Jahangiri, Hossein Najmabadi, Jörg Hoheisel (2009) Genome-wide DNA methylation screening in breast cancer revealed new candidate loci potentially involved in field defect. 2nd Annual Meeting of NGFN-Plus and NGFN-Transfer in the Program of Medical Genome Research. Berlin, Germany. Nov 2009

- **Moghaddas Gholami, A.** Hoheisel, J., Fellenberg, K. (2009) Application of a Hill-climbing Algorithm to Combine Information Across Different Microarray Studies. *International Conference on Systems Biology.* Stanford, California,USA. Sep 2009

- **Moghaddas Gholami, A.**, Schmitt, C., Hauser, N.C., Rupp, S., Hoheisel, J., Fellenberg, K.(2008) New M-CHiPS add-on features. International Workshop *"Transcriptome and Proteome Data Analysis and Warehousing towards Systems Biology"* Stuttgart, Germany. Jun 2008.

- **Moghaddas Gholami, A.**, Hauser, N.C., Hoheisel, J., Fellenberg, K. (2008) M-CHiPS, Microarray Data Warehouse and Analysis Software. *10 Jahre Statusseminar Chiptechnologien.* Frankfurt, Germany. Jan 2008

- **Moghaddas Gholami, A.**, Schmitt, C., Visvanathan M., Hauser, N.C., Rupp, S., Hoheisel, J., Fellenberg, K.(2007) Universal platform for Microarray Data Interpretation. *German Conference on Bioinformatics 2007 (GCB07).* Potsdam, Germany. Sep 2007

## 8.5 Erklärung

Ich erkläre an Eides statt, daß ich die vorliegende Arbeit am Lehrstuhl für Proteomik und Bioanalytik der Technischen Universität München unter Anleitung von Prof. Dr. Bernhard Küster ohne sonstige Hilfe erstellt und bei der Abfassung nur die gemäß § 6 Abs. 5 (Promotionsordnung) angegebenen Hilfsmittel benutzt habe.

Amin Moghaddas Gholami