Learning of Probabilistic Models to Infer Gene Regulatory Networks

Matthias Böck

TECHNISCHE UNIVERSITÄT MÜNCHEN

Fakultät für Informatik
Lehrstuhl für Bioinformatik

Learning of Probabilistic Models to Infer Gene Regulatory Networks

Matthias Böck

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. Martin Bichler
Prüfer der Dissertation:	
	1. Univ.-Prof. Dr. Burkhard Rost
	2. Univ.-Prof. Dr. Stefan Kramer Johannes Gutenberg Universität Mainz

Die Dissertation wurde am 27.07.2015 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 10.05.2016 angenommen.

Because we separate like ripples on a blank shore,
in rainbows.

Radiohead

Ich versichere, dass ich diese Dissertation selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

München, den 10.07.2015 Matthias Böck

Abstract

Since 2008, genomics data has been outpacing Moore's Law by a factor of four, and we are still only at the beginning of understanding biological systems as a whole. One of the most important regulatory levels in every process of life are the networks of directly or – via the corresponding proteins or RNAs – indirectly interacting genes. Although the learning of these gene regulatory networks from given biological data has been intensively studied in recent years, still only some smaller regulatory subnetworks are understood.

In this thesis, I examine three key objectives from the field of gene regulatory network inference. First, the analysis of gene expression time series data. This is done by both the integration of additional knowledge about biological network structures and already known interactions, as well as the identification of suitable data discretization procedures. Second, the development of an intuitive and easily adaptable analysis workflow to predict so-far unknown regulatory relations. Third, the integration and representation of heterogeneous biological data retrieved through experiments, as well as predicted data for the inference of gene regulatory networks.

I present two novel approaches for the analysis of time series on the first objective. The first approach uses the concept of Dynamic Time Warping on discretized time series profiles to learn undirected gene regulatory networks. The profiles are discretized based on the angle between two consecutive time points. Additionally, an alternative distance matrix was trained from the profiles of known interacting genes. Dynamic Time Warping uses this distance matrix and the time series profile for pairwise alignments and ranks the corresponding gene pairs according to their similarity. The second approach uses Bayesian learning on time series, discretized to a binary state space. The aim of this approach is to infer a directed network. Additionally, the known tendency of biological networks to form scale-free networks is integrated into the learning process. Scale-free networks are typical for most biological or social networks and imply that the majority of nodes have only a few connections, while a small number of nodes is linked to many other nodes. Both approaches were evaluated against synthetic and biological data sets and performed equally or better than state-of-the-art methods in several benchmarks.

I also present the development and application of a data driven and iterative analysis framework to identify new regulatory interactions in normal human urothelium (NHU) cell lines. This framework adopted a novel ranking approach, whereby changes in the amplitude of experiment and control time series were compared to identify common regulatory elements. The transcription factor (TF) ELF3 was identified and verified in the lab as an early regulator of NHU.

The last approach studies the integration of multiple heterogeneous data sources for four different eukaryotic organisms for the inference of gene regulatory networks. These data sources consist of predicted data, such as TF binding site predictions or text mining from PubMed abstracts, or experimental data, such as ChIP-seq or microarrays. I present a descriptive and predictive study of these data sources to infer their structure, their interdependencies, and their importance for possible predictions.

The presented results show that current gene regulatory prediction methods still can be improved and that the integration of additional data sources is an important step in this process. Furthermore, it is shown that methods have to be adapted and that an iterative approach towards a better understanding of the underlying processes and structures leads to new biological insights.

Zusammenfassung

Seit 2008 hat die Datenmenge aus dem Genomicsbereich Moore's Law bereits um einen vierfachen Faktor überholt und wir sind dennoch erst am Anfang davon biologische Netze in ihrer Gesamtheit zu verstehen. Eine der wichtigsten regulatorischen Ebenen in jeglichem Lebensprozess sind die Netzwerke von direkt oder – über die entsprechenden Proteine oder RNAs – indirekt interagierenden Genen. Trotz intensiver Studien über das Lernen von genregulatorischen Netzwerken aus gegebenen biologischen Daten in den letzten Jahren, können immer noch nur kleinere regulatorische Subnetzwerke beschrieben werden.

In dieser Arbeit untersuche ich drei Kernziele aus dem Bereich der genregulatorischen Netzwerkinferenz. Erstens die Analyse von Genexpressions-Zeitreihendaten. Dies wird sowohl über die Integration von zusätzlichem Wissen über biologische Netzwerkstrukturen, bereits bekannten Interaktionen, als auch der Findung geeigneter Datendiskretisierungsprozeduren durchgeführt. Zweitens die Entwicklung von intuitiven und leicht anzupassenden Analyseprozessen für die Vorhersage von bisher unbekannten regulatorischen Beziehungen. Drittens die Integration und Darstellung von heterogenen, biologischen Daten für die Ableitung von genregulatorischen Netzwerken. Die verwendeten Daten setzen sich sowohl aus Experimenten, als auch Vorhersagen zusammen.

Für das Erste der oben genannten Ziele beschreibe ich zwei neue Ansätze für die Analyse von Zeitreihendaten. Der erste Ansatz wendet das Konzept von Dynamic Time Warping auf diskretisierte Zeitreihenprofile an, um ungerichtete genregulatorische Netzwerke zu lernen. Die Profile sind diskretisiert durch die Steigungswinkel zwischen jeweils zwei aufeinander folgenden Punkten. Zusätzlich wurde eine alternative Distanzmatrix mit Hilfe der Profile von bereits bekannten interagierenden Genen trainiert. Dynamic Time Warping benutzt die Distanzmatrix und die Zeitreihenprofile für paarweise Alignments und bewertet die entsprechenden Genpaare nach ihrer Ähnlichkeit. Der zweite Ansatz wendet Bayessches Lernen auf Zeitreihen an, die in einem binären Zustandsraum diskretisiert wurden. Das Ziel dieses Ansatzes ist es ein gerichtetes Netzwerk abzuleiten. Zusätzlich wird die bekannte Eigenschaft von biologischen Netzwerken, skalenfreie Netzwerke zu formen, in den Lernprozess integriert. Skalenfreie Netzwerke sind typisch für die meisten biologischen oder sozialen Netzwerke und implizieren, dass die Mehrheit der Knoten nur eine geringe Anzahl von Verbindungen aufweisen, während eine kleine Anzahl mit vielen anderen Knoten verlinkt ist. Beide Ansätze wurden gegen synthetische und biologische Datensätze evaluiert und erreichten gleichwertige oder bessere Ergebnisse auf verschiedenen Benchmarks als andere aktuelle Methoden.

Darüber hinaus präsentiere ich auch die Anwendung eines entwickelten, datengetriebenen und iterativen Analyseverfahrens für die Identifizierung neuer regulatorischer Interaktionen in normalen, menschlichem Urothelium (NHU - Normal Human Urothelium) Zelllinien. Dieses Analyseverfahren beinhaltet einen neuen Rankingansatz, wodurch Veränderungen in den Amplituden von Experiment- und Kontrollzeitreihen miteinander verglichen werden um gemeinsame regulatorische Elemente zu finden. Der Transcriptionsfaktor (TF) ELF3 wurde identifiziert und verifiziert im Labor als ein früher Regulator in der Entwicklung von NHU.

Der letzte Ansatz beschäftigt sich mit der Integration von verschiedenen heterogenen Datenquellen für vier unterschiedliche eukaryotische Organismen zur Inferenz von genregulatorischen Netzwerken. Diese Datenquellen bestehen aus vorhergesagten Daten, sowie TF-Bindestellenvorhersagen, Text Mining aus PubMed Zusammenfasssungen oder experimentellen Daten, wie ChIP-seq oder Microarrays. Ich stelle eine deskriptive und prädiktive Studie für diese Datenquellen vor, um ihre Struktur und Interdependenzen im Detail zu beschreiben sowie ihrer Bedeutung für mögliche Vorhersagen herauszuarbeiten.

Die vorgestellten Ergebnisse zeigen, dass gegenwärtige genregulatorische Vorhersagemethoden immer noch verbessert werden müssen und dass die Integration von zusätzlichen Datenquellen ein wichtiger Schritt für diesen Prozess ist. Weiterhin wird gezeigt, dass Methoden an die speziellen Fragestellungen angepasst werden müssen und dass mit einem iterativen Ansatz ein besseres Verständnis der zugrundeliegenden Prozesse und Strukturen geschaffen werden und dies somit zu neuen biologischen Einsichten führen kann.

Acknowledgement

First of all, I would like to thank my supervisor Stefan Kramer who encouraged me to start the challenge of a PhD. He gave me the chance to join his lab and the RECESS graduate school program. This gave me the great opportunity to work on my thesis in a multi-disciplinary and highly international environment. Stefan not only introduced me to the manifold ways of machine learning and data mining, but also gave me the means and freedom to choose and follow my own paths.

During my thesis, I meet many interesting and wonderful persons, who I would like to thank for their inspiration, ideas and new perspectives. These few lines of text here can only describe my gratitude to a small extend and definitely not include everybody who should be mentioned here.

I want to thank Constanze Schmitt, with whom I not only shared the same room at the lab but also many exciting travelling adventures ranging from Gelendzhik to Moscow and St. Petersburg, but also to York and Paris. We also shared many hours of discussing and tackling the challenges of gene expression analysis. On the scientific trips to and through Russia also joined two other RECESS colleagues and friends, Johannes Raffler and Robert Pesch. We had many discussions about our projects and they often helped me reaching new ideas from the perspective of metabolomics and proteomics, which was mostly followed by a movie night.

I also thank my Russian friends, who I was fortunate to get to know during my visits at their lab in Moscow. Especially, Yerbol Kurmangaliyev and Sascha Favorov for many enlightening discussions ranging from scientific topics to dolphins and the open sky. I am also very glad to have met Prof. Gelfand, who not only introduced me to his lab but also to many interesting insights about Russia and science itself.

I thank Jenny Southgate for countless discussions on the best approach to find the regulatory factors in human urothelium and to help me getting a detailed understanding about the perspective of biologists and how to solve biological questions collaboratively.

I would like to thank my lab for having a few great years in Garching and in particular Jörg Wicker who still shows me the inspiring world of academic life. Hopefully, one day we manage to start finally working on the naked mole rat project. I also like to thank Tim Karl who does not only an amazing admin job, but also could brighten up the foggiest days in Garching. Thanks also to the rest of the lab, Marianne Müller, Jana

Schmidt, Andreas Hapfelmaier, Lothar Richter, Tobias Girschick and the lab-adopted, Simon Berger, for having an exciting time and the chance to learn from each other.

Last, I want to thank the persons closest to me, my parents and my little sister, who always encouraged me to follow my own way and supported me through all those years with their trust and fondness. Most of all I thank my wife Dorothea who supported me through all this time, with her strength, patience and love. The final thank you goes to my son, Janis, who has been enthusiastically commenting the last couple of weeks, my writing on this thesis.

Contents

CHAPTER 1

Introduction

1.1 Motivation

Probably one of the most frequently introduction slides used in talks on bioinformatics or systems biology is a graphic showing the exponential increase in available data versus the understanding of the system itself. The latter is usually depicted as a gently inclining linear function, in extreme contrast to the steep increase in the amount of data. With the rise of new computer technologies, the internet and more and more high throughput methods for the analysis, we have powerful tools at our disposal, which allow us to gain new insights into fields that were out of our reach before. The last two decades provided biologists with a tremendous amount of new possibilities to study biological processes on a detailed level and to analyze systems as a whole. Methods such as mass spectrometry, DNase footprinting, gene expression, methylation or copy number analysis via microarrays, and the more recent technology of next-generation sequencing — to name only a few of the most prominent examples. And even for the relatively young next-generation sequencing discipline, follow-up techniques already exist. So-called, next next generation sequencing methods exceed the methods currently used by several magnitudes with respect to speed and sequencing depth. While the first sequencing of the human genome took more than a decade and cost several billion dollars (Human Genome Project, finished in 2003), these new techniques speed up the genome sequencing process to days — if not hours (depending on the organism) — and reduce the costs to a few thousand dollars. All these continuously evolving techniques allow us to create an amount of data beyond human conceivability, and this is currently the greatest limitation. We are just beginning to understand what this data means, and we are still far from understanding whole biological systems that are more complex than the single-cell organism *E.coli*.

One elementary step towards understanding the various data sources is to develop methods which extract as much knowledge as possible from each data set. This thesis focuses on the field of gene regulatory networks and their inference from given time series data. Gene regulatory networks are one of the essential layers of controlling the cellular system, and if we gain a deeper insight into their functionality and structure, we also reach a better understanding of disease mechanisms and their cures. There is an obvious need for methods which can deal with large-scale data sets and even though there is a lot of data available, the dimension of the search space in data analysis is even bigger. The human genome is currently supposed to have approximately 20,000 protein coding genes, whose sequences only cover about 1.5% of the human genome. Not too long ago, the majority of the genome sequence was supposed to be junk DNA with no function at all, but this number has been turned upside down to a current estimate of 80% of DNA which also encodes regulatory elements. Focusing on the protein coding genes alone already leaves us with a search space of $20{,}000^2$ possible directed interactions (including self-regulations and feedback loops). This scales up further, since this is the search space for each measured condition of a cell type as well as for each cell type itself.

The second and even more challenging task is to bring together the various fields of study on biological systems and to integrate the gained knowledge to improve or even enable new gene regulatory network predictions. There were nearly one million new publications in PubMed only in 2011, covering novel approaches, data sets and insights into various biological processes, which can be of great value to new studies if their findings can be integrated. An impressive new resource are the results of the recently published ENCODE (Encyclopedia Of DNA Elements) project in Nature, Genome Biology and Genome Research on finding all functional elements in the human genome sequence. 440 scientists in 32 laboratories all over the world have participated in this project since 2003. They have generated about 15 trillion bytes of raw data and so far have used more than 300 years of computer time for their analysis. It may seem like a Herculean task to pick and combine the right data sets, but with every effort towards solving this task, a step towards a better understanding is made.

For these reasons, this thesis deals with the development of new methods to analyze time series experiments, their application to a real-world example and the integration of available data sources into the prediction process of gene regulatory networks.

1.2 Contributions

This thesis presents five main contributions for the task of inferring and understanding gene regulatory networks. As stated in the previous section, it is essential to get out as much use as possible from generated data sets. To this end, two newly developed methods are presented here, which focus on the learning of gene interactions from time series data.

The first method, described in Chapter 3.1, combines a *Dynamic Time Warping (DTW)* approach with discretized time series profiles to account for noise in the data. This approach is also able to deal with time shifts and different strengths and velocities of gene regulatory effects. Additionally, a semi-supervised method variant was developed which uses known gene regulatory interactions in the organism of interest—in the presented study *E. coli* or *S. cerevisiae*—to calibrate an organism specific distance metric for the DTW calculations [19]. Both methods performed equally well and in some cases superior compared to established methods. The methods are implemented in *R* and *C* and can deal with large-scale comparisons on a genome-wide level.

A second, more advanced method is described in Chapter 3.2. This method uses a Bayesian learning approach on gene expression time series, discretized to a binary state space [18]. This approach integrates the known tendency of biological networks to form scale-free networks into the network inference process. This particular network structure implies that the majority of the genes only interacts with a few other genes, while a small percentage of the genes functions as so-called *hub genes*, which have many interaction partners. The presented method outperforms other methods like ARACNE, Banjo or MRNet, especially for the task of finding hub genes. So far, there is no other method capable of integrating the hub gene structure, and which can be applied to large data sets.

Chapter 4.1 presents a data driven analysis framework which leads to the identification of a new transcription factor for the development of normal human urothelium (NHU) [17]. As described in Section 2.1, gaining insight into a given data set strongly depends on data quality, the choice of the methods, and the iterative exchange of hypotheses and results between biologists and data analysts. The outlined approach includes a detailed analysis pipeline which resulted in the better understanding of factors regulating the differentiation and proliferation of NHU. Furthermore, follow-up experiments were designed and executed following the analysis' findings.

One of the key findings of this thesis is the necessity of integrating additional, already existing knowledge into the learning process of gene regulatory networks. Chapter 5.1 gives an overview of the various available data sources for finding gene regulatory interactions and how these sources can be integrated and used for further predictions.

The fifth contribution of this thesis is the visualization of data sets. In particular, the interactive visualization of gene expression time series and the visualization of a large-scale heterogeneous gene interaction database. Usually, when starting to work on a new data set, one of the first steps is to visualize the data to get a better understanding of its structure and potential noise sources. In Appendix A.1 of Chapter 4.1, screenshots show the interactive gene expression time series visualization (available online), which allows to explore a given data set on gene level as well as on probe set level. Especially for large heterogeneous data collections it is important to get an overview of the different interdependencies and the overall structure of the data. The Circos plots in Figures 5.2,5.3, 5.4 and 5.5, presented in Chapter 5.1, give a detailed overview of the collected and predicted evidences for gene interactions.

1.3 Organization of this Thesis

Chapter 2 gives an overview of the biological background and of current gene regulatory network inference approaches. This thesis primarily focuses on gene interactions, which are briefly introduced as well as the structure and attributes of gene regulatory networks. This chapter also shows an exemplary process on how to work with real world data sets in the field of systems biology. In the following, the initial steps in data preprocessing, adapting algorithms, the difference between dynamic, and steady-state data are introduced. Additionally, the proper evaluation of predicted networks is discussed. The last part of this introduction describes different learning approaches, ranging from information theory models to ordinary differential equation models and gives a brief overview of the task of integrating several heterogeneous data sources to improve the predictions.

Chapter 3 presents two different approaches on gene expression time series data. The first one is based on Dynamic Time Warping and infers an undirected network. The learning process follows two main ideas. An angle based discretization to describe the time series and a dataset-specific distance matrix, which is calibrated with already known interacting genes. The second approach infers a directed network and integrates the knowledge about the topological features of biological networks into the learning process. A typical attribute of a biological network are so-called hubs, highly linked (interacting) genes. Only a few hubs exist, while the majority of genes has only a few interaction partners (scale-free structure). The approach also uses time series and integrates the scale-free property as prior knowledge into a Bayesian learning process.

An example of working with real world biological data sets is presented in Chapter 4, which describes a data driven analysis pipeline to analyze a microarray time series from normal human urothelium experiments. The approach shows how to overcome several quality issues, like missing data points, a possible discretization method, and how to iteratively reduce the set of candidate genes to a manageable size. The outcome of this project was the identification of the transcription factor ELF3 as an important regulator in human urothelium. Several other potentially interesting genes could be identified and additional time series experiments were designed and executed subsequently. Another result of this analysis was the interactive visualization platform of the data sets, which allows biologists to easily explore data sets.

Chapter 5 presents the application of several bioinformatics methods for the integration of heterogeneous data sources for the prediction of a gene regulatory network. For this task, several heterogeneous data sources on gene interactions were collected.

Among the collected data sets are text mining results, transcription factor binding site predictions, co-expression and protein-protein interaction networks, and ChIP-seq experiments. The aim of this data integration was to identify relevant data sources for generating a reliable database of prior knowledge for different organisms, like human and mouse. The integrated data sets where visualized for a first descriptive analysis and then more sophisticated methods, like Random Forests, were applied to identify feature importance and to predict a gene regulatory network.

The final Chapter 6 concludes the work. It discusses achieved results and possible future analyses, which tie in with the approaches and insights presented in this thesis.

CHAPTER 2

Inference of Gene Regulatory Networks

The following sections give a short overview of the state of the art in network inference. Prior to this the biological background of gene regulatory networks are briefly introduced. These first sections describe the concept of gene interactions and the specific structure and attributes of biological networks. Furthermore, the field of systems biology is introduced and an exemplary analysis workflow, which is based on multiple iterations of analyses between biologists and bioinformaticians. The second part of this chapter focuses on the data and evaluation of the analysis. In particular, the complexity of biological data and the various processing approaches and how synthetic data can be used to simulate experiments. Additionally, the aspect of proper evaluation of predictions and the differences of dynamic and static data sets are discussed. The last sections describe the different gene regulatory network inference models and the general concept of integration of heterogeneous data sources.

2.1 Gene Regulatory Networks

Our current understanding of a biological cell underlies the principle of different regulatory and interacting entities, which form a synergistic ensemble and result in the different observable phenotypes. These entities can be metabolites, proteins, the various forms of RNA (like mRNAs or ncRNAs), and form different intertwined layers of regulation. The overall aim would be to understand the full structure of these regulatory processes and to develop new cures for diseases. One level of insight towards this aim is the learning and analysis of gene regulatory networks (GRNs). GRNs focus on the interactions and interdependencies between genes in a cell and how signals are transferred and regulated between those genes.

Tremendous advances have been made in recent years in the various fields of biology, biochemistry and computer science, which allow to analyze biological systems in detail. However, the understanding of the GRNs as a whole is still limited to only the most prominent model organisms, such as *S. cerevisiae* or *E. coli*. Several insights have been gained especially for human, mouse and fruit fly, but mostly only single pathways or subnetworks are understood out of the actual system they are embedded in.

Mathematical modelling of biological systems can help us to better understand the underlying processes and to interpolate our current knowledge. Still, those models often underlie the limitation of prior knowledge, lack of data quantity and/or quality as well as the scientist's assumptions driving the analysis. The processes on the different levels of gene, cell, organism and environment, their dependencies and interactions are barely understood due to their complexity, and the more details we try to incorporate in our systems, the more complex our models become. For this reason, the field of systems biology emerged in recent years and a short introduction on this topic is given in the following sections.

The primary aim of this thesis lies on better understanding the level of transcription factor (TF) target (TG) interactions in gene regulatory networks, which play an essential role in systems biology. A TF is a protein which binds to a DNA sequence and thereby controls the transcription rate of another gene, its target. The DNA sequence is located in the so-called promoter region of a gene and each TF binds only to a specific sequence, its transcription factor binding site (TFBS). The majority of these TFBSs is supposed to be close (within 1000 or even 500 base pairs) upstream of the transcription start site (TSS) of the regulated gene. Nevertheless, TFBS can be also located several thousand base pairs distant from the TSS, both upstream and downstream.

The regulation of a gene's transcription rate by a TF can either increase (activating) or decrease (silencing). A TF can also form a protein complex with other TFs or proteins for the regulation process. TF-TG interactions are supposed to be the key regulators in cellular systems, but in recent years more and more additional factors of regulation (like ncRNAs or methylation) have been identified. When studying GRNs on the TF-TG level, one should be always aware of the additional perturbations not measured on the system.

2.1.1 Gene Interactions

As mentioned above, the interactions between two genes do not necessarily imply a physical interaction, like a TF which binds to the promoter region of its target gene, but might also refer to a indirect regulation, like on the protein or RNA level. When interpreting the results of GRN predictions it is important to note that some predicted interactions might be also caused by these indirect factors.

2.1.2 Structure and Attributes

The interactions between several genes can be represented as a network in which each node represents a gene and each link between the nodes a particular interaction. These interactions can be either directed if the causal effects are known, as for the relationship between TF and TG, or undirected. The latter can be used to describe co-expression networks and link pairs of genes which were found to have a strong correlation. A third option are partially directed graphs as a mixture of both. Networks can be used to describe either a particular state of the gene regulatory system or to give an overview of all possible interactions. Additionally, the edges in such a network can be labeled with a probability which defines the certainty for a connection between two genes. Biological network structures have some interesting properties, such as hub genes or that the small world principle applies, which are discussed in more detail in Chapter 3.2 of this thesis.

2.1.3 Systems Biology

The field of systems biology incorporates various disciplines, like transcriptomics, proteomics, metabolomics etc., and attempts to understand biological systems (organisms) as a whole. To achieve this goal it is important to connect the different disciplines and to develop mathematical models which allow to understand and simulate a system.

As already stated above, the focus of this thesis is on the transcriptomics part. Nevertheless, the concept of systems biology applies also to a single discipline and it is necessary to find suitable approaches and frameworks to handle the multidimensional complexity of biological systems and in particular GRNs.

A detailed overview of the different steps for a possible analysis framework of biological systems is given in Figure 2.1. This framework adapted from machine learning for a system and data from corresponding experiments follows three basic steps. These steps can be categorized as ideation (creative process to generate hypotheses), descriptive analysis (understanding the data structure and interdependencies) and prediction and interpretation. The strength of this process lies in its iterative nature, which allows and also requires feedback loops and a stepwise approach towards better understanding the system.

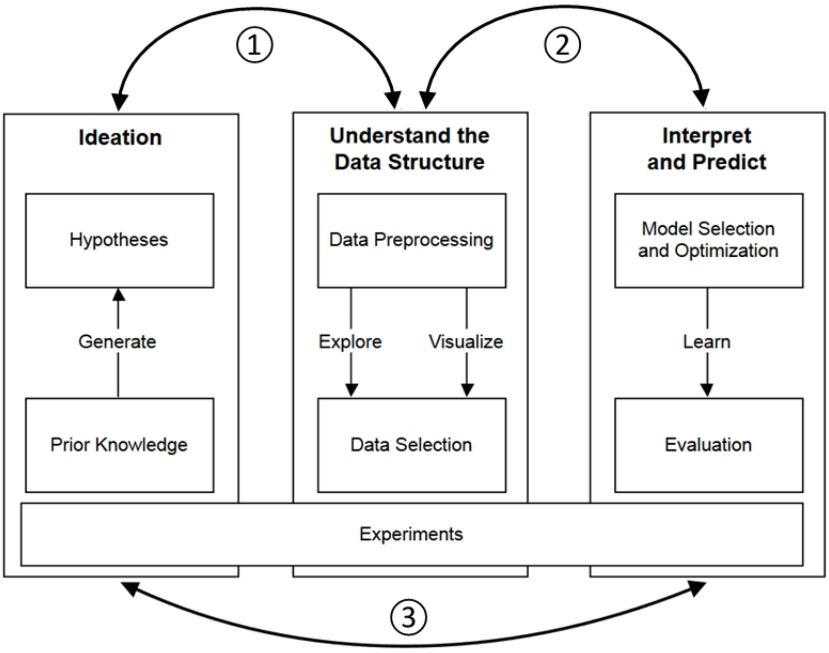


Figure 2.1: Iterative approach to analyze biological systems. The process of inferring a biological system or gaining a better understanding of single processes within a system is described here with the three domains, ideation, understand the data structure and interpret and predict. The overall process is an iterative process which requires continuously, exchange between the three domains and relies to a great extent on the underlying experiments. These experiments can be a newly created data set to follow a hypothesis, experiments from publicly available databases, or validation experiments to test intermediary results or the final predictions. The first domain on the left, ideation, deals with the generation of hypotheses about a system of interest. These hypotheses are derived from prior knowledge arising from literature, other experiments or even previous predictions which have to be evaluated in the lab. The result of this ideation process, is a hypothesis (or a list of hypotheses) and properly designed and executed experiments to deal with the resulting tasks. The bidirectional arrows (1) and (2) also imply that the design of these experiments has to be reconciled with the model selection and data processing steps. The domain in the middle, represents the second domain: understanding the data structure. It is crucial to understand the quality, structure and statistical significance of the available data to be able to assess the quality of any prediction in any of the later steps. Furthermore, it is of course important to have deeper knowledge about the analyzed system itself. Several data preprocessing and transformation steps might be necessary to bring the data in a proper format and to reduce noise in the data set. To gain a better understanding, the structure (distribution, ranges, means) has to be explored and the data itself visualized to compare the different results. This also allows to select a suitable subset of the data for the third analysis step in the right domain. This can be further refined or already validated with additional experiments in the lab. In the last domain, strongly depending on the two previous domains, it is important to make the proper choice of a predictive model which can give new insights into the system of interest. The predictions and interpretations of the data have to be thoroughly validated (again, also with experiments) and in an ideal circle, these newly gained insights, lead back to the ideation process (arrow (3)).

2.2 Data and Initial Considerations

2.2.1 Biological Data and Algorithm Adaptation

Over the last two decades an immense amount of publically available interaction databases were established. These databases cover fields like protein-protein interactions (STRING, DIP, BIND or HPRD), metabolic pathways (KEGG, ENZYME), TFs and interactions (JASPAR, TRANSFAC, RegulonDB), and many more [67]. Also more and more fully sequenced genomes are available (Ensembl, NCBI), which can be searched for transcription factor binding sites. In addition to these databases, further valuable insights can be extracted from the large amount of scientific literature. Several text mining tools have been developed which allow the automatic extraction of interaction data from the literature (see Chapter 4.1 for examples). Another source of interest can be functional annotations of genes. Gene Ontology [6] offers an hierarchical description of genes and gene product attributes for the domains, cellular component, molecular function and biological process. All these listed sources are possible sources of prior knowledge. This knowledge can be used to reduce the search space of possible interaction partners, improve or validate the predictions, or use the available knowledge itself to make new predictions about interdependencies in a GRN.

The main focus of this thesis lies on the level of interactions between genes and especially deriving new insight from gene expression time series data (microarray based). These interactions can be physical interactions (e.g., a TF binds to the promoter region of its TG) or indirect interactions (e.g., regulation on protein level).

Due to the fast changing landscape of available biological data, it is also an important task to adapt algorithms for the specific difficulties which arise with these multifaceted types of data. Each of these data sets comes with its specific strength but also drawbacks as usually experiments in the lab can strongly depend on their execution (e.g. scientists and environment). Additionally, the available funding can limit the statistical strength of the experimental setup. Further factors of influence can be new gene annotations, technical advances or newly gained insights into the gene regulatory system of interest. As a result of these constraints, hardly any algorithm is applicable out of the box to these data sets, rather, they have to be re-evaluated and modified for each analysis. Still, the major advantage to the discussed synthetic data sets in the next section of biological experiments is the possibility to actually verify the predictions.

2.2.2 Synthetic Data and Evaluation

One major challenge for each newly developed method is the unbiased evaluation against established methods. Biological data can only be of limited help to assess the performance of a method due to the huge amount of unknown and also uncontrollable factors which influence a biological system. Additionally, as mentioned in the previous section, most of the algorithms have to be adapted to the specific task. This might introduce a certain bias towards favoring a particular easy-to-use and modifiable algorithm.

Establishing a Synthetic Gold Standard

Several attempts have been made in the last couple of years to establish a synthetic standard which allows a reliable comparison platform between the different methods. The goal of these methods is to offer a data generation workbench which allows to control all factors of influence and to introduce perturbations on a gene regulatory system.

One of the best-known attempts to install common synthetic evaluation standards for systems biology models is the annual “Dialogue on Reverse-Engineering Assessment and Methods” (DREAM) which exists since 2006 (`http://www.the-dream-project.org/`). Every year, the DREAM project offers open challenges, mostly based on synthetic data in the area of systems biology, in which the participants can compare their predictions or for which the different predictions are combined in a wisdom of the crowd approach [94]. The data and the synthetic data generating tools are publically available at the end of the challenges. The GeneNetWeaver (`http://www.tschaffter.ch/projects/gnw/`) offers an intuitive open-source framework to generate benchmark *in silico* steady-state and time series expression data for the prediction of synthetic gene regulatory networks. This framework has been used for several DREAM challenges and offered valuable insights into the different network inference methods as well as into the requirements for a reliable evaluation [117, 95]. The gene expression data is simulated with a system of ordinary differential equations, which are supposed to adequately describe the dynamics of the mRNA and the protein concentration for each gene. These dynamics are regulated by a maximum transcription rate, a translation rate, as well as mRNA and protein degradation rates and an input function which computes the relative activation of a gene (active or inactive). Further details of this approach are described by Marbach *et al.* [95]. This method allows to generate steady-state data but also to introduce an external perturbation on the system (to simulate different experimental conditions) and to study the behavior of the system over time. The GeneNetWeaver workbench was also used for the evaluation of the presented method in Chapter 3.2.

In addition to the importance of finding suitable approaches which are capable of describing the dynamics of gene interactions, it is also crucial to apply these functions on realistic biological network structures. Hence, the generation of plausible networks for the data generation has to be addressed, too. The GeneNetWeaver approach uses known biological networks, like *S. cerevisiae* or *E. coli*, for the extraction of subnetworks.

Another synthetic data generation tool, which also uses subnetworks (but a different network sampling method), is SynTReN (`http://homes.esat.kuleuven.be/~kmarchal/SynTReN/index.html`). This generator simulates gene expression steady-state data based on Michaelis-Menten and Hill kinetics [143]. A more flexible software is GeNGe, which runs as a web application and allows to choose between various network as well as synthetic data generation options (`http://genge.molgen.mpg.de`). The gene expression algorithm is based on a non-linear differential equation system [62]. The data simulation tool Netsim (`http://www.medcomp.medicina.unipd.it/bioing/netsim/`) creates networks with a scale-free distribution and calculates the gene activity with differential equations, which also account for saturation effects (like GeneNetWeaver) [36].

Unfortunately, the last three mentioned approaches have not been further developed since their publication and are only to a limited extent intuitively applicable for the data generation due to outdated packages, software bugs and missing technical documentation. Currently, the DREAM project offers the most reliable synthetic – and for the recent challenges, also biological – data sets for inferring gene regulatory networks and assessing the performance of the applied methods. Still, it should be noted that this cannot replace the experimental validation of predictions in the lab and that of course a bias is always introduced in the synthetic data set by the choice of a certain algorithm for the data generation.

Finding the Adequate Evaluation Measure

Besides the necessity for a proper gold standard, the choice of a suitable evaluation measure is crucial for assessing the performance of a method with respect to predicting a reliable network structure.

The classical task consists of binary predictions, which have the goal to assess how well a method can predict if a connection between two genes exists or not. This corresponds to two classes for existing and not existing links. The comparison between the predictions of a method and a given gold standard can then give insights into the method's performance. Table 2.1 shows the possible outcomes of such a binary prediction.

Various possible metrics can be derived from this confusion matrix, which can give an

Table 2.1: Confusion Matrix

	Predicted Class	
	positive	negative
Actual Class		
positive	true positive (TP)	false negative (FN)
negative	false positive (FP)	true negative (TN)

Classifications of predicted links as true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN).

estimate how well a method performs on a certain task. The main difference between the metrics lies in the influence of each of the four outcomes from the confusion matrix in the chosen evaluation score. Simple metrics are the rates between the confusion matrix entries, like the *true positive rate* ($TPR = \frac{TP}{TP+FN}$), which is also called sensitivity or recall, or the corresponding *true negative rate* ($TNR = \frac{TN}{FP+TN}$), also referred to as specificity. Further rates are the *false positive rate*, the *false negative rate* or the *rate of positive predictions*. Another, well-known metric is the *precision* which is defined as $\frac{TP}{TP+FP}$. In addition to these rates, more complex metrics like the *Phi coefficient*, the *Mathews correlation coefficient* or the *F-score* exist. A more detailed discussion on the different metrics is given by Fawcett [40] and Küffner *et al.* [129].

However, reducing the prediction task to binary classes is a simplification since genes can up- and down-regulate their targets or be co- and anti-correlated. This already leads to a three class problem (up-regulation, down-regulation and no interaction). An even more challenging task is the highly unbalanced nature of the positive and negative (no interaction) classes, since the majority of genes only interacts with a few other genes and the predictions have to deal with sparse networks. When predicting on real world data sets and using given gold standards, one should always be aware that apparently not interacting genes might actually be interacting genes and of course this also influences the performance of the methods.

The above mentioned metrics are usually combined for the performance evaluation of a method, since they represent different aspects of a prediction. A commonly applied evaluation method for classification tasks are Receiver Operating Characteristic (ROC) curves. ROC curves combine the TPR with the FPR and are the resulting visualization if a step-wise increased cutoff value (or threshold) is applied to a prediction and the TPR and FPR are plotted against each other for each step. This curve can then be used to assess the method's performance with respect to sensitivity (TPR) and 1-specificity

(FPR) and to weigh these two measures against each other for the choice of a suitable threshold. Fawcett *et. al.* [40] give a very detailed introduction to this topic as well as the area under the ROC curve (AUROC) and the precision-recall (PR) curves, which are briefly introduced in the following. The AUROC summarizes the results of ROC curves in one value and the larger this value, the better the prediction is supposed to be. For a two class problem an AUROC of 0.5 refers to randomly guessing the classification and an AUROC of 1 to a perfect classifier. Another common evaluation curve are PR curves, which plot the precision as a function of the recall (sensitivity) and also apply a step-wise increased threshold on the prediction. In contrast to ROC curves, it is insufficient to linearly interpolate between the points of the PR plot [34] to calculate the area under the precision recall curve (AUC-PR). The calculation of precision and recall curves is more sensitive than the ROC curves towards the difference in class distributions of positive and negative examples in the data (also class imbalance or skew). This is also the case for each point of the PR plot and if the class distribution strongly varies for different thresholds (local class skew), the linear interpolation leads to an overly-optimistic curve [34]. Davis and Goadrich show in [34] how to properly interpolate the PR curve with the use of the analogous ROC convex hull. Additionally, it has been shown that a curve dominates in ROC space if and only if it dominates also in the PR space [34]. Vice versa, it is not necessarily true that a method optimizing the ROC curve also optimizes the PR curve.

In general, ROC curves have the advantage of being less sensitive to the class distribution and are hence a valid measure to compare the AUROC performance of different methods on different networks. Still, the different PR curves should be additionally used to assess the performance details for different thresholds. For the task of predicting a network for a specific organism or to rank the TF-TG interactions, it is of more importance to find the best tradeoff between precision and recall (finding as many as possible true interactions while keeping the number of false interactions low). Hence, PR curves are the more appropriate choice for an evaluation metric, but it can be useful to also include ROC curves for specific tasks [129]. There are several additional factors of influence for the evaluation of network inference methods, like the unknown number of false negatives in the data set which have yet not been discovered, or the split of TFs and their TGs over the different sets used for the cross-validation (or training and test sets). An introduction to this topic is given by Küffner *et al.* [129]. In case of the cross-validation on predicted interactions it is suggested to divide the predictions into four groups in which the prediction method has either seen both nodes (genes), only one node (TF or TG) of the interaction, or none of the nodes. The performance of the predictions can strongly vary due to these groups. Another important topic, which will not be discussed in more detail here, is the difficulty of distinguishing direct from indirect regulations in a predicted gene regulatory network. Several methods have been proposed to overcome this issue, like the data processing inequality [97], partial correlation [25], two-way ANOVA [85] or network deconvolution [41].

For this thesis, the different presented methods are evaluated with both AUROC as well as AUC-PR to follow the suggestions above. In Chapter 3.2, the predictions extend the binary prediction task also to a three-class prediction.

2.2.3 Data Pre- and Post-Processing

The data pre- and post-processing steps can have a large influence on the performance of a method and should be applied carefully. Different data sets require different procedures and often the processing steps have to be adapted specifically for a data set. Hence, the processing steps are also discussed in Chapters 4.1, 3.2, 5.1 and 3.1, for each of the specific tasks.

In general, four different pre-processing steps should be considered, data selection, transformation or discretization, sampling and feature selection. For the first step it is important to assess the data quality and structure and to select a suitable data subset which is used for the further analysis. Second, some kinds of data require to transform the data, like microarray data which has to be normalized to be able to compare the different genes over different experiments. Examples of microarray pre-processing can be found in Chapters 4.1 and 5.1. Additionally, the performance of the prediction might be increased by using discrete rather than continuous values to represent gene expression (or also any other data). This can decrease noise and complexity of the data (see Chapters 3.1 and 3.2). Possible discretization methods range from simple mean, median or quantiles (unsupervised; equal-width or equal-frequency techniques) to IR, ID3 or ChiMerge (supervised techniques) [93]. A more detailed overview of the different methods is given by Garcia *et al.* [52] and a survey of discretization methods on microarray data by Li *et al.* [90].

Third, if the number of data points is too large for applying the algorithm of choice, the use representative (stratified) subsamples of the data set should be considered. Furthermore, it is usually the case that for network inference problems, the number of linked genes is magnitudes lower than the number of unlinked genes. Many algorithms cannot account for this class skew (already discussed in the previous section), and either a cost sensitive classifier is applied or samples of the classes are used for the training of a predictor. Several sampling procedures have been suggested. Most of these approaches either oversample the minority class (linked genes) or undersample the majority class (unlinked genes) to be of comparable size to the other class. The undersampling procedure SMOTE is described in Chapter 5.1. Feature selection is the fourth option which can be either data or knowledge driven. The overall goal is to identify a smaller subset of relevant features (either attributes or variables) which can be used in the construction of a predictive model. The feature selection can be either on gene level or on the different samples or experiments, which are available for a particular gene. Again, various methods exist for data driven feature selection and an example of this procedure is given in Chapter 5.1. A partly knowledge driven feature selection of possible genes of interest is presented in Chapter 4.1. For this approach, prior knowledge is used to reduce the number of genes. This prior knowledge can be for example from literature research or publicly available databases.

After the method of choice is applied on the pre-processed data set, it is sometimes helpful to further refine the results. This can be the simple application of a threshold, which removes low scoring results, as described in the following section, or the integration of additional knowledge to rank more or less likely results.

2.2.4 Time Series versus Steady-States

Steady-state gene expression data sets can be seen as *snapshot* of the cellular system at a certain time with certain conditions. A practical use of this kind of data would be the study of the differentially expressed genes in cell lines from different populations or conditions, like healthy and cancer patients. The shortcoming of this approach is that it does not capture the dynamics of the underlying gene regulatory network, which might not only have one steady-state but several. Time series are a possible approach to study the dynamics of regulatory networks over time, but they are usually more complex in their experimental setup.

An ongoing debate is how and if it is possible to infer not only correlation but also causality from experiments. The dynamics of time series experiments can give valuable insight into the interdependencies between different genes, but still, the observed cause and effect might be the result of an external perturbation of the system which we are not capable of measuring. The same holds for steady-state experiments, where for example, the differences between a regulatory system and its perturbated version can be analyzed. In general, a gene expression time series or steady-state experiment will not be enough to answer the questions of causality, and additional knowledge about the data-generating process is needed. An often quoted phrase on this topic is “correlation does not imply causation”. A detailed overview of causal inference in statistic and in particular the distinction between association and causation is given by Pearl [113].

Steady-state experiments can give a snapshot of the state of a cell and have the advantage over time series experiments that the factor of time does not bring additional limitations to the experimental setup. For time series experiments it is important to control as precisely the synchronization of the experiments as possible, especially for the control cell lines, to be able to perform statistical tests on the resulting data.

Time series experiments allow to study the dynamic effects of the perturbation of a gene regulatory system. Possible perturbations can be single gene overexpression or knockdown experiments. Usually, only a single factor is changed, and the remaining system kept under the same conditions to limit the amount of variables of influence.

Some algorithms, like ARACNE or Banjo can be applied to both types of data, but usually the focus of a method lies in one or the other type [7].

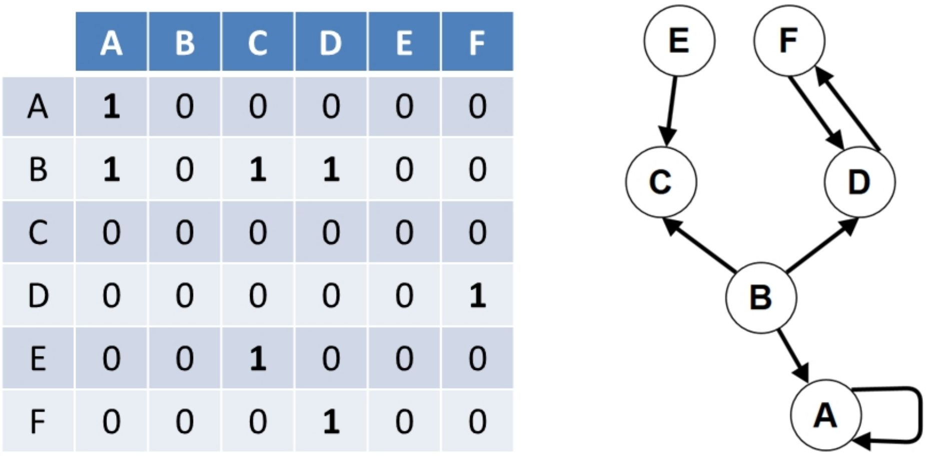
2.3 Overview of GRN Inference Models

A detailed overview of the different GRN inference models is given by Bansal *et al.* [7] and Hecker *et al.* [67]. In the following, the most important concepts and their categories are briefly introduced.

The aim in this context, given a set of gene expression data from an experiment of interest, is usually either the network inference or the ranking of possible targets (differentially expressed, pathways, TFs, etc.) of interest for a given set of data. Both tasks demand different approaches and have their own challenges. An example for the ranking task is given in Chapter 4.1.

In general, most network prediction algorithms can be described with the following three categories of attributes [67]. First, the representation of the activity level of a node (gene or protein) in the network, which can be Boolean (active or inactive) or denoted by other (discrete) categorical or continuous values. Second, the network model can be based on a stochastic or deterministic, static or dynamic type of model. The third attribute describes the interactions between the nodes in the network which can be directed or undirected, linear or non-linear. Figure 2.3 gives an additional general overview of the different network inference models which will be also described in more detail in the following sections.

Another additional group of prediction methods are so-called meta-predictors, which combine several approaches of the list above. One key finding of the DREAM challenge



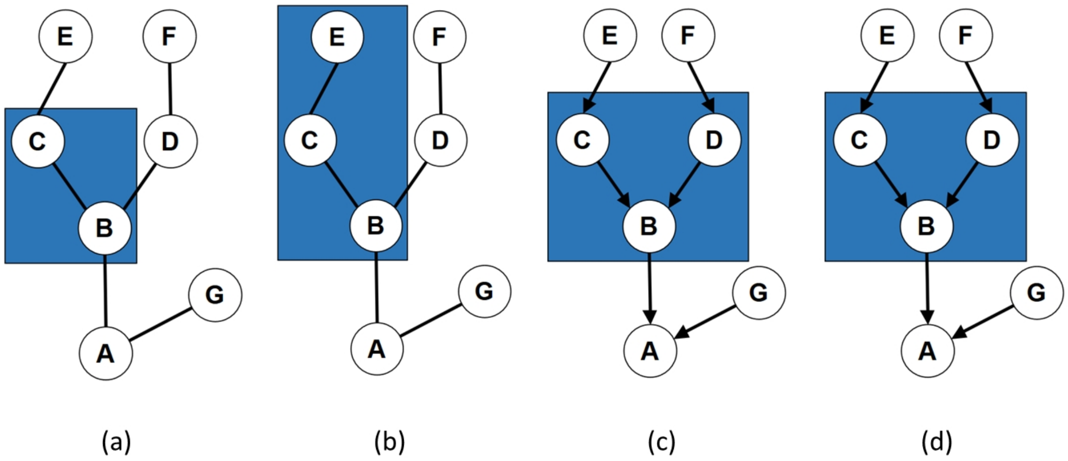
	A	B	C	D	E	F
A	**1**	0	0	0	0	0
B	**1**	0	**1**	**1**	0	0
C	0	0	0	0	0	0
D	0	0	0	0	0	**1**
E	0	0	**1**	0	0	0
F	0	0	0	**1**	0	0

Figure 2.2: Matrix and graph representation of a network. A network can be represented as a matrix where each entry is a directed link between two genes. Each row represents all outgoing links from a particular gene, and each column the ingoing links. In case of an undirected network the matrix is symmetric. The corresponding graph is shown on the right and includes also self-regulations (node A) and feedback loops (D and F). It should be noted that the matrix, also called adjacency matrix, does not necessarily imply Boolean states and that each cell entry can also be filled with a continuous value which represents the certainty or strength of this edge (interaction). In case of a directed network, the relation of nodes can be termed with child and parents. In the example above, E is the parent of its child node C.

analysis on network inference was that these wisdom of the crowd approaches outperform the individual methods (especially when only the top ranked methods are combined) and that it usually improves the predictions if the best performing methods are combined by the average rank of their combined edges [95].

2.3.1 Information Theory Models

One of the simplest network learning approaches use correlation, which is either Pearson's or Spearman's correlation coefficient. This approach is rather a clustering approach based on the idea that genes with similar expression profiles can be grouped together [38]. The result of this correlation analysis is a co-expression network with undirected interactions between pairs of genes (an example of such an analysis is given in Chapter 4.1). The strength of the interaction is quantified by the correlation coefficient. An empirically chosen threshold can then be applied to prune low scoring interactions with



$$\begin{aligned} corr(B,C) &> TH \\ corr(B,G) &\leq TH \end{aligned} \tag{a}$$

$$\begin{aligned} MI(A,H) &= 0 \\ MI(B,C) &> 0 \\ 0 < MI(B,E) &\leq min(MI(B,C),MI(C,D)) \end{aligned} \tag{b}$$

$$P(B|C,D,E,F) = P(B|C,D) \tag{c}$$

$$dB/dt = \sigma_1 B + \sigma_2 C + \sigma_3 D \tag{d}$$

Figure 2.3: Overview of inference models. Figure a) and b) represent undirected networks. The first figure shows a simple correlation network, where only pairs of genes (nodes) are compared and low scoring interactions (edges) are pruned with a predefined threshold (TH). The second undirected network represents information theoretic approaches, combined with pruning rules, like the Data Processing Inequality (IDP). Figure c) and d) represent directed networks based on Bayesian models and ordinary differential equations (ODEs). Their structures are the same but their underlying models differ. In both models B only depends on its direct predecessors (parent nodes C and D). For the Bayesian network in Figure c), B is conditionally independent of E and F given C and D. In case of ODEs the expression of B is described as a function of the expression rates of C and D and itself. All of these four model types do not allow for cycles (i.e. no feedback loops). An example for networks with cycles are Dynamic Bayesian Networks.

an absolute correlation coefficient below this threshold.

Information-theoretic models are based on the measure of entropy. In general, this measure is used to quantify the amount of information ($I(X_i; X_j)$) in common between two random variables, X_i and X_j, with the entropy of a variable X_i defined as:

$$H(X_i) = - \sum_{k_i \in X_i} p(x_{k_i}) \log p(x_{k_i}) \tag{2.1}$$

The probability of each discrete state (value), is described by $p(x_{k_i})$. The distribution of p is generally unknown and has to be estimated. Several different entropy estimators exist, like the empirical, Miller-Madow, Pearson or Spearman correlation [112]. The latter two can be applied directly to continuous variables, such as gene expression data. An example for information-theoretic models is the Mutual Information (MI) score, which is a generalization of the pairwise correlation coefficient between the gene expression profiles. *Relevance networks* calculate this score for all gene pairs and infer their relation if the MI score is above a certain threshold [23].

Other more advanced MI-based examples are ARACNe (Algorithm for the Reconstruction of Accurate Cellular Networks) [97], CLR (Context Likelihood of Relatedness) [39] or MRNET (Minimum Redundancy NETwork) [114]. All these methods calculate in a first step the MI matrix for all pairwise genes and apply in a second step different algorithms to rerank the edges of the MI matrix.

Besides the simple filtering of results with a threshold, some methods apply additional processing steps, like the earlier mentioned data inequality processing (DPI) by ARACNe, to remove indirect or redundant interactions. The DPI is computed by finding the minimal MI between three variables (X_1,X_2,X_3):

$$I(X_1; X_3) \leq min(I(X_1; X_2), I(X_2; X_3)) \tag{2.2}$$

The edge with the lowest value is interpreted as indirect edge and if this value is above a given threshold, removed from the network. The CLR method, extends the relevance network approach by additionally integrating the mean (μ) and the standard deviation (σ) of the empirical distribution of $I(X_i; X_j)$ into a score $w_{ij} = \sqrt{z_i^2 + z_j^2}$ with:

$$z_i = max \left\{ 0, \frac{I(X_i; X_j) - \mu_i}{\sigma_i} \right\} \tag{2.3}$$

The MRNET approach applies a maximum relevance/minimum redundancy feature selection technique [114]. This feature forward selection chooses at each step, among the least redundant variables, the one having the highest MI with the given target. A backward selection version has also been introduced recently by Meyer *et al.* [103] as well as a comparison of the different methods on synthetic data sets.

Olsen *et al.* also thoroughly compared the different MI approaches and suggested to use MRNET combined with Spearman correlation for complete and accurate data, and the CLR method combined with the Pearson correlation for data with noisy and missing values [112].

A major advantage of these types of models are their simplicity, the relatively low computational costs (maximum complexity of $O(n^3)$) and the low number of required samples, since only bivariate distributions are estimated [112]. This allows to apply them also on large-scale data sets to study regulatory systems as a whole. On the downside, these approaches only take into account pairwise interactions without considering multiple possible regulators of a gene. The resulting network is only an undirected network.

2.3.2 Boolean Models

Boolean models are a special case of a sequential dynamic system, where time and states are discrete. They consist of two components, a directed graph as representation and Boolean functions and states of genes to describe the relationships between the nodes in the graph [80]. The Boolean network (BN) also allows loops, and each node represents a gene, which can either be active or inactive (Boolean states {1,0}). Node relationship can be described with transition functions between the nodes of the network. These functions – in the simplest case built with the logical operators AND, OR, NOT – are defined for each node and are used to determine the state of each node for each of the discrete time points. Usually, these updates are made synchronously in each step for the whole network. Synchronously means that all edges have the same duration and hence nodes are updated at multiples of a fixed time step [3].

Due to the structure of BNs, there exist only 2^N possible states for a network with N nodes. This means that at a certain time point the network is always going to fall into an attractor, which is either a single state (steady-state) or a repeating sequence of states (cycle). The inference of the Boolean networks is done from the list of observed states (transition tables).

Two well-known examples for the inference of Boolean network models are REVEAL [91] and the Akutsu algorithm [2]. The REVEAL approach uses information theoretical principles for the inference of the network from a given transition table.

A possible extension of the BNs are probabilistic BNs, which can handle uncertainty and allow the quantification of the relative strength of an interaction between a TF and its TG [130].

In summary, BNs are a useful qualitative approach to cope with the complexity and high dimensionality of biological systems. Albert *et al.* present several interesting case studies for the representation of biological systems with BNs [3]. Still, the level of abstraction, introduced with BNs has several disadvantages, since gene regulation not only follows a switching behavior, but signals can also be amplified, subtracted or added over time, or even have a negative feedback control. Another complex topic is the choice of a suitable discretization procedure for gene expression data (already discussed in the previous Section 2.2.3).

2.3.3 Bayesian Models

Bayesian models are graphical models which encode the probabilistic relationships among a set of random variables X_i. These relationships are usually described by a directed acyclic graph (DAG) G. The vertices of this graph represent the random variable X_i, and the links (interactions) between the variables are described by a joint probability distribution:

$$P(X_i,...,X_n) = \prod_{t=1}^{N} P(X_i = x_i | X_j = x_j,...,X_{j+p} = x_{j+p}) \tag{2.4}$$

Applied to a GRN, the set X (vertices) represents the genes, with X_j being the regulators (or parent nodes) of X_i (child node). A more detailed introduction on the structure and use of Bayesian networks is given in Section 3.2.

The three main tasks for Bayesian networks are a) inferring the parameters if the structure of the network is known, b) learning the structure if the parameters are known, and c) inferring both, structure and parameters. The latter is usually the case for GRNs since neither the exact gene regulatory function nor the gene regulatory network is known for the majority of the systems. The advantage of this type of model are its capability to handle noisy and incomplete data, as well as its ability to avoid overfitting. Additionally, due to the probabilistic nature of Bayesian networks, it is possible to simply add prior knowledge into the equations.

A disadvantage of Bayesian networks is the limitation to DAG structures, which cannot represent feedback loops, which are an important regulatory pattern. Dynamic Bayesian networks overcome this issue and are of interest especially for time series data and the dynamic changes of the underlying network. It should furthermore be noted that the parents of a node in Bayesian networks do not necessarily imply a direct causal effect.

Banjo is one of the best-known gene network inference tools and uses both Bayesian and Dynamic Bayesian networks [155]. In Section 3.2, Banjo is also used as a benchmark for the predictions. Banjo can be applied to steady-state as well as time series data.

2.3.4 Ordinary Differential Equation Models

In this thesis, the focus of the methods lies on the three previously mentioned sections. Hence, the use of ordinary differential equations (ODEs) will be just briefly introduced. In contrast to Bayesian or information theory models, ODEs do not rely on conditional probabilities and are a deterministic approach. A gene regulatory system is described with one ODE per gene, where each ODE describes the gene regulations as a function of other genes [7]:

$$\frac{dx}{dt} = f(x,p,u,t) \tag{2.5}$$

where $x(t) = (x_1(t),...,x_n(t))$ represents the gene expression vector for the genes 1 to n at time t. The function describes the rate of change for each of the genes (x_i) in dependence of the model parameters p and the external perturbation signals u [67].

ODEs have the advantage to be able to simulate different perturbations (i.e. gene knock downs, drugs of interest, etc.) on the inferred system and can be used to predict the reactions to these perturbations. The DREAM project also used a set of ODEs for the generation of the synthetic gene expression data sets.

NIR, MNI and TSNI are examples for ODE-based algorithms and have been successfully applied and extended over the past years [53, 14, 8].

Besides the ODEs, there are several more complex models, like stochastic differential equation models [111].

2.4 Data Integration of Various Sources

As stated earlier, the focus of this work lies on the gene interaction networks. However, relying on only one type of source in the era of omics, would be ignoring the tremendous amount of potential prior knowledge. On the transcriptomics level, data sources such as microarrays, TFBS predictions and prior knowledge from literature research. Recently, focus has shifted to more advanced high-throughput approaches like large-scale text mining, next generation sequencing (NGS), ChIP-seq or DNase footprintings. The “Encyclopedia of DNA Elements” (ENCODE) published in 2012, started in 2003 to systematically map regions of transcription, TF association, chromatin structure and histone modification and enabled scientists to assign biochemical functions to 80% of the human genome [139]. The published data sets offer a great opportunity to enrich the database of GRN analyses.

Not restricted to GRNs but a general review about the current and future challenges of data integration throughout the different omics disciplines is given by Gomez-Cabreror *et al.* [59]. They additionally used a websurvey to identify current opinions and needs on this topic from the research community. The three key findings of this review are the use of prior knowledge and its efficient storage, the development of methods which are capable of analyzing heterogeneous data sets and the creation of data exploration tools that incorporate useful statistics and visualizations. These are all topics also covered by the recently wide-spread discussion around "Big Data". Besides the need for data management systems, efficient and reliable machine learning and visualization approaches and user-friendly tools to give a broader audience access to this new dimension of possible insights across heterogeneous data sources, I would add the need to establish reliable evaluation benchmarks for algorithms and to improve the available processes for data collection. The latter problem of data standardization is also addressed by Gomez-Cabreror *et al.* and the lack of easily accessible large collections of data, also limits the potential of generating a database of prior knowledge for the prediction of biological systems. Additionally, more and more advanced high-throughput techniques allow to generate faster and more and more detailed experimental data, while the problem of efficiently storing the "old" data sets has not even been solved. The field of data integration applies to many disciplines and faces many exciting challenges. Chapter 5.1 describes an approach which combines various data sources, as well as how to make use of this combined knowledge base and to visualize the data and results.

A thorough overview of the field of data integration for the inference of gene regulatory networks is given by Hecker *et al.* [67]. The number of different data sources and experimental measurements needed to model a GRN is difficult to define but depends on the size of the search space, which in turn consists of the model parameters (number of genes N and weights $w_{i,j}$ describing the interactions between the genes) for the GRN model. Increasing the number of data sources usually also increases the complexity of the GRN model and increasing the number of measurements per data source comes usually at the cost of additional calculation time. A possible solution for the first of these two problems can be feature selection and mapping. Hecker *et al.* describe three possible data- and knowledge-driven approaches, which reduce the number of genes in the GRN by removing genes which are only poorly represented in the available data (either due to experimental errors or because of inactive genes) or by combining similarly behaving genes into functional entities [67]. Additionally, when using different sources for learning of GRN, a feature selection on these data sources can help to understand the impact of a single sources for predicting the GRN of interest.

A list of different GRN learning approaches is also given by Hecker *et al.* and five different approaches are discussed in Chapter 5.1. All these methods have to decide on a suitable data set and a set of reliable data sources, where each data set has to be preprocessed with respect to data quality, normalization, discretization and selection. Furthermore, a scoring function has to be defined which is capable of integrating and weighting the data or different data sources to learn the parameters of a GRN and to predict the future behavior of genes of interest.

2.5 Summary

In this chapter, an overview of the basic concepts of working with gene expression data has been given. These concepts include general knowledge about the domain of interest, here TF-TG interactions and gene regulatory networks. A more detailed description is out of the scope of this introduction, but it should be noted that every analysis approach is limited by the understanding of the studied domain. Furthermore, the concepts of data selection, pre-processing and result evaluation should be understood. In the following sections, newly developed and applied methods from the introduced fields above, such as time series analysis, large-scale data integration, Bayesian learning and integration of prior knowledge into the learning process, are going to be presented.

CHAPTER 3

Time Series Based Methods for the Inference of Gene Regulatory Relationships

The analysis of time series data is still one of the most challenging fields and occurs in many scientific disciplines. Steady-state data can only give a snapshot of the actual dynamics, while time series allow to study the processes over time and to capture the dependencies between the forces and protagonists. It should be noted that the topic of causality is a large philosophical field by itself, and it can still be argued, whether the observations are cause or effect of an certain outcome. In the following, two approaches which deal with the inference of regulatory relationships from time series data are presented. The first focuses on the transformations and shifts of gene expression time series and how to learn undirected interactions [19]. The second approach uses Bayesian learning on Boolean discretized gene expression time series along with a prior on the inferred network structure, which consists of directed interactions [18].

3.1 Dynamic Time Warping for the Inference of Undirected Gene Regulatory Relationships

This section focuses on learning of undirected interactions and dependencies from given gene expression time series data. A novel angle-based discretization of given microarray data time series and a new alignment approach, combining the ideas of *Dynamic Time Warping* (*DTW*) with *Stochastic Local Search* (*SLS*) are introduced.

The building of alignments of discretized profiles is supposed to be robust against noisy data and to overcome the assumption of strictly linear relationships between two interacting genes. A basic assumption for the alignment of time series is that co-regulated genes also show similar expression behavior over time and hence similar amplitudes which can be aligned with suitable transformations. Testing and evaluation of the approach has been done with one synthetic data set as well as four biological data sets.

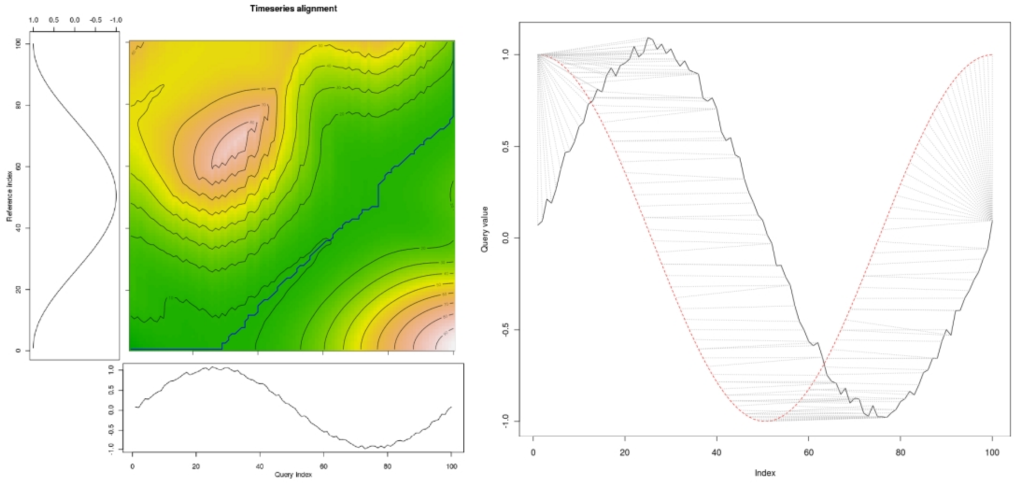
In the following, different variants of *Dynamic Time Warping* (*DTW*) for the inference of gene regulatory relationships are assessed. A positive influence of the distance optimization on the performance of the alignments of gene expression profiles could not firmly be established. However, the results show that discretization can be important to the outcome of the alignments. The discretization is not only able to keep the important features of the time series, it is also able to perform better than regular *DTW* on the original data.

3.1.1 Dynamic Time Warping

Dynamic time warping (*DTW*) was introduced in the 1960s [12] and has been intensively used for speech recognition and other fields, like handwriting recognition systems, gesture recognition, signal processing and gene expression time series clustering [1]. The basic idea of this unsupervised learning approach is that a suitable distance measure, which is most generally the Euclidean distance, allows the algorithm to stretch (or compress) the time and expression rate axis to find the most suitable fit of two given time series. The *DTW* algorithm will be described briefly in the following. Consider two given sequences $S = s_1,...,s_n$ and $T = t_1,...,t_m$ and a given distance function $\delta(s_i,t_j)$ with $1 \leq i \leq n$ and $1 \leq j \leq m$, *DTW* tries then to minimize with the given δ over all possible warping paths between the two given sequences based on the cumulative distance for each path. This is solved by a recursive dynamic programming approach for each $i \in [1,...,n]$ and $j \in [1,...,m]$:

$$DTW(i,j) = \begin{cases} 0 & \text{for } i = j = 0 \\ min \begin{cases} DTW_{i-1,j-1} + \delta(s_i,t_j) \\ DTW_{i-1,j} + \delta(s_i,t_j) \\ DTW_{i,j-1} + \delta(s_i,t_j) \end{cases} & \text{for } i,j > 0 \\ \infty & \text{otherwise} \end{cases} \tag{3.1}$$

$DTW[n,m]$ is the total distance $DTW(S,T)$ and can be calculated in $O(nm)$. The traceback through the matrix D gives the optimal warping of the aligned sequences. In the following, the symmetrical version of DTW which is supposed to perform better on equally sampled time series is used [123].



(a) Scheme of DTW Alignment

(b) Effect of Alignment

Figure 3.1: DTW Alignment of two given sequences. Alignment with DTW of a cosine for reference and a noisy sine wave as query with traceback through the DTW matrix (replaced for a better overview with the corresponding density levels). The plots and calculations were done using the dtw R package [57]. In the right plot the warping effect of the alignment from Figure a can be seen. The grey dotted lines indicate aligned time points and show the mapping of the corresponding time points on the two time series.

In contrast to other existing methods, the approach deals also with anti-correlated time series and uses a supervised method to infer a data specific distance matrix for the alignment. The result is a scoring matrix for the pairwise distances between the measured genes.

Four different gene expression time series of different lengths are used for the evaluation. An overview of the networks and data sets will be given in the next section. All time series are centered around the x-axis by applying a z-score transformation to account for scale inconsistencies of the microarray experiments. Cubic smoothing splines are used to interpolate the time series for missing values and smoothen out smaller fluctuations from experimental or biological noise.

The discretization of each time series for each gene is done according to the steepness of the expression change $\delta\ exp$ between two consecutive time steps. This is done by calculating the angle: $\alpha = \textbf{atan}\ \delta\ exp \cdot \frac{180}{\pi}$. The angles are then discretized into positive and negative (increasing or decreasing) integer values according to a predefined threshold. Defining the threshold is done by dividing the largest found angle for increases or decreases for each time series for a gene, into equally sized subsectors. Consider a maximum found angle of 180 degrees, which should be split into n subsectors, resulting a range of $\frac{180}{n}$ degrees each. Each of this sectors represents a possible range of angles for the increase or decrease between two consecutive time points and has assigned a discrete value. For n sectors the range of these values would be $[-\frac{n}{2}, -\frac{n}{2}+1, ..., \frac{n}{2}-1, \frac{n}{2}]$. To account for noise in the data, the two sectors which are neighboring the x-axis (one in the positive and one in the negative direction) are combined into one sector with the discrete value zero. See Figure 3.2 for an example of the discretization.

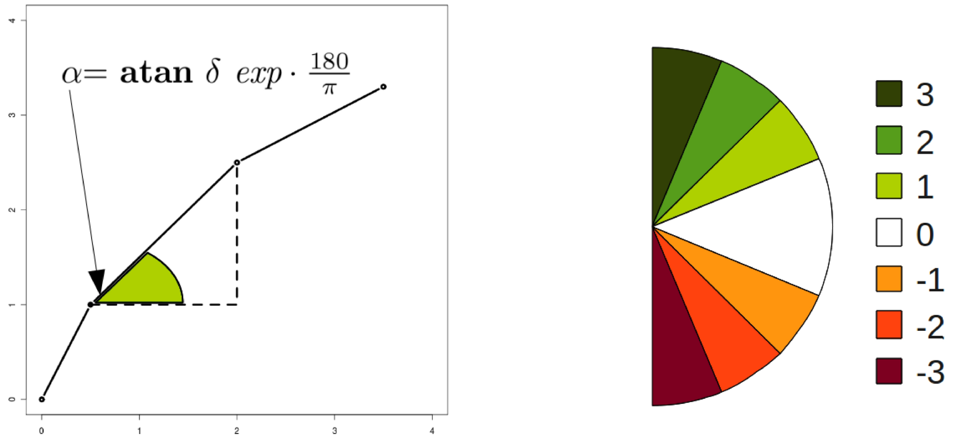


Figure 3.2: Example of the Angle-Based Discretization: The left figure shows a time course for one gene with four measured time points. For the discretization in this example n is set to three, and the resulting possible sectors are shown in the right colored graphic. According to this setting, the second time point of the time course would be discretized to one.

A crucial point for the quality of the alignments is the choice of a suitable distance matrix which defines the distances between the discretized values of the time series.

This motivates our supervised approach to use a set of already known interacting genes I to infer the distance matrix δ. These gene pairs are chosen randomly from a given gold standard network along with a further randomly chosen set of not interacting genes N. The size of the latter is set if possible to twice the size of I. From this larger set N between successive iterations of the distance calibration process new subsets are resampled to prevent accidentally chosen existing interaction partners between I and N genes to distort the result.

The resulting δ should minimize the distance for I and maximize the distance for N. Since *DTW* is not differentiable, a combination of *Stochastic Local Search* (*SLS*) and simulated annealing for the stepwise improvement of δ is applied. For a more detailed introduction to *SLS*, see the work of Hoos and Stützle [61].

Three constraints on the step-wise altering of the distance matrix δ are imposed to reduce the search space and to keep the basic distance structure between different bins of angles: $\delta(i,j) = 0$ for $i = j$, $\delta(i,j) = \delta(j,i)$ and $\delta(i,j) < \delta(i,j-1)$.

The resulting distance matrix is then used for the calculations of the alignments and the score defines the distance between each pair of genes. Additional alignments are done for each comparison with flipped signs for one of the time series to find anti-correlated pairs. All calculations were done in R except for the alignment matrix calculations, which were done for runtime efficiency in C.

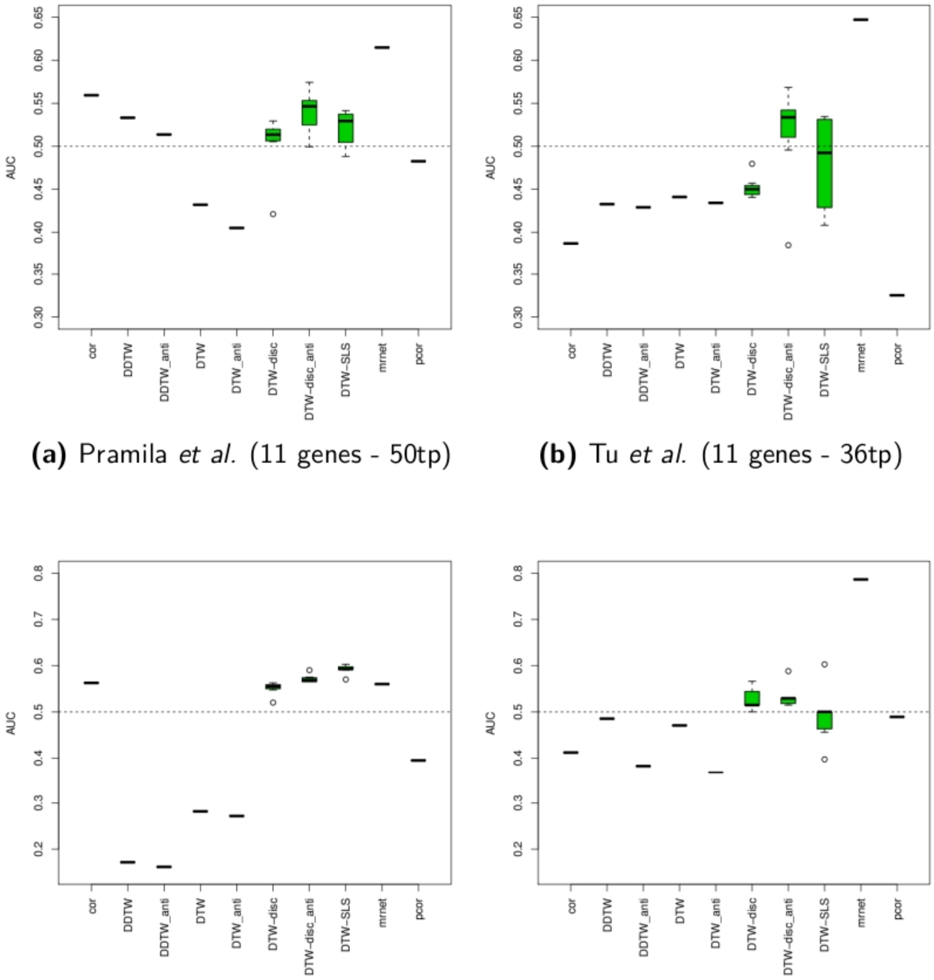
3.1.2 Performances of Different DTW Variants

The evaluation is done on five differently sized networks, a synthetic five gene network of yeast, called IRMA [24], the SOS signaling pathway in *E. coli* consisting of eight genes [120], a 11 cell cycle regulating network derived by Li *et al.* from the literature [89], and a full set of cell regulating genes, consisting of 1129 genes published by Rowicka *et al.* [121]. Gold standard and time series for the IRMA and SOS signaling pathway were taken from the R package *TDARACNE* [157] and consist of 16 and 14 measurements. The 11 cell cycle network by Li *et al.* as well as the suggested set of Rowicka *et al.* are tested with two time series experiments by Pramila *et al.* [116] and Tu *et al.* [142]. These sets follow several full cell cycles and include 50 and 36 time points. Genes of the large-scale network which were not found in the experimental sets were left out. This resulted in gene sets of size 961 and 944 for Pramila *et al.* and Tu *et al.*. As a benchmark network for the large-scale cell cycle analysis, the protein-protein interaction network from the STRING database (v8.3) [73] is used. STRING calculates for each interaction a score based on the evidence from various sources like experiments, interaction databases or abstract text mining. A cutoff of 0.8 was applied to select only interactions with high confidence. It is clear that the PPI network is only able to cover part of the gene regulatory processes but still, observations on this level can provide insight into the performance of the methods. STRING is also considering pairs derived from co-expression analysis and might therefore be more suitable than other PPI databases. Self-regulations were excluded from all data sets

The performance on the data sets are compared to the results with simple correlation, partial correlation, MRNET (mutual information) [102], DTW and $DDTW$ (a modification of DTW which uses for the discretization the first derivative for each point) [83]. DTW_{disc} applies our discretization method with different numbers of sectors (n) and calculates the alignments with DTW. DTW_{SLS} additionally applies the distance calibration before the calculations. Methods ending with *_anti* also consider anti-correlated time series in the calculations. The evaluation is done based on ROC curves and the AUC. Interactions are undirected and hence only a two class problem considered, interaction predicted or not.

The results from the small networks in Figure 3.3 show that MRNET performs well even on shorter time series. Our methods perform only on the *E.coli* data better but are the second best performer on this task compared to the established methods. The discretized version performs in all cases, except for the Tu *et al.* data, better than guessing and outperforms DTW and $DDTW$, except for the Pramila *et al.* data, where $DDTW$ performs equally well. Including anti-correlation into the calculations improves in all cases of the discretized method the performance but has no positive effect for the regular DTW and $DDTW$. On the large-scale network evaluation in Figure 3.4, the use of only correlated genes performs significantly better than with anti-correlation.

In general, the different DTW approaches perform better on the large-scale data sets than correlation or MRNET, except in the case of regular DTW and $DDTW$ on the Pramila *et al.* data. The results of DTW_{disc} show that the discretization keeps the important features and performs well even with a small number of sectors. The approach of DTW_{SLS} seems, to this date, not to be able to improve the distance measure and achieves slightly smaller AUC values. The discretization method outperforms DTW and $DDTW$ on the Pramila *et al.* data and performs only slightly worse on the other data set.



(a) Pramila *et al.* (11 genes - 50tp)

(b) Tu *et al.* (11 genes - 36tp)

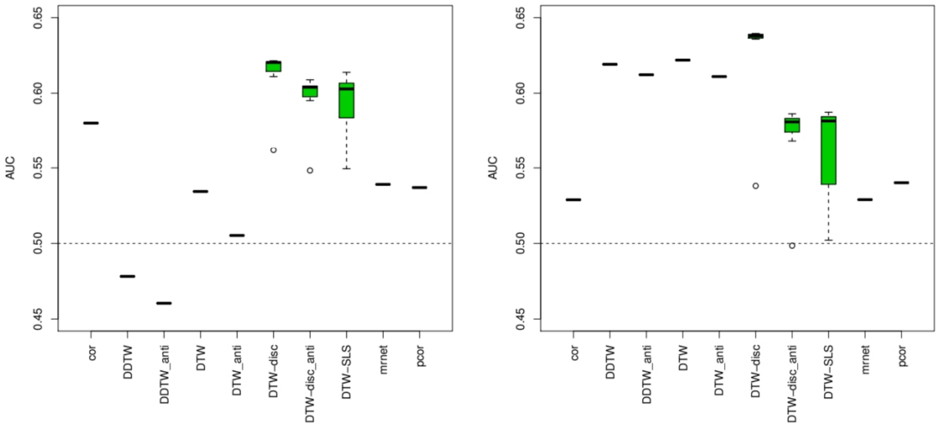
(c) *E.coli* SOS (8 genes - 14tp)

(d) Synth. yeast (5 genes - 16tp)

Figure 3.3: Comparison of the performances on three small networks. The dotted line indicates guessing. The number of sectors ranges from 1 to 8. Knowledge size for calibration was set to 2. MRNET performs best on all yeast data sets and only slightly worse than our proposed method on *E.coli*. The discretized version of *DTW* performs, except for case b), always better than guessing and best for the *E.coli* data set. Including anti-correlation improves in all cases the performance of DTW_{disc}. The other *DTW* versions perform quite differently on the data sets and in most cases even worse than guessing, especially in c).

3.1.3 Conclusion

Several variants of *Dynamic Time Warping* for the detection of gene regulatory relationships were investigated in detail. While the supervised optimization of the distance



(a) Pramila *et al.* (961 genes - 50tp)

(b) Tu *et al.* (944 genes - 36tp)

Figure 3.4: Comparison of the performances on a large-scale network. The dotted line indicates guessing. Number of sectors ranges from 1 to 8. Knowledge size for calibration was set to 10. Gold Standard: STRING DB. Our approach performs significantly better in a) and only slightly worse in b). $DDTW$ and DTW perform best on the Tu *et al.* set but the influence of the anti-correlation is only small. DTW_{disc} performs much better in b) without the anti-correlation.

matrix did not lead to improvements, a novel discretization approach seems, even with a small number of defined sectors, able to keep the main features and appears as a suitable qualitative transformation for time series alignments. On the biological data sets, our approach seems to be more stable compared to DTW and $DDTW$. In contrast to correlation-based methods, DTW is also able to infer the orientation of the time shift through the traceback and hence able to hint at possible causalities. A next step would be to make use of this information and to further evaluate the robustness of the discretization method compared to DTW and $DDTW$.

3.2 Hub-Centered Gene Network Reconstruction using Automatic Relevance Determination

3.2.1 Introduction

With the development of large-scale experimental platforms for the acquisition of genome-wide data, massive amounts of experimental data describing complex cellular processes are becoming widely available. The extraction of knowledge and development of models from such data remains a major challenge. Manual model development is constrained to small models involving a few dozen components, and requires extensive prior biological knowledge. The alternative is to use automated machine learning approaches to infer models directly from data, as reviewed by Kaderali and Radde [79].

For small models involving only a few dozen genes, detailed quantitative network inference approaches using *nonlinear differential equations* can be employed [100]. Such approaches fail for larger networks due to computational limitations and practical non-identifiability of model parameters. *Boolean network models* have been proposed as an alternative, neglecting the quantitative detail and assuming genes to be in only one of two states, active or inactive [80, 91]. Updates of the states are then done using logical rules, either synchronously for all genes or using asynchronous update rules [66]. Further extensions are based on fuzzy logic [152] or probabilistic Boolean networks, which basically use alternative sets of Boolean update rules that are stochastically employed [130].

Bayesian networks on the other hand are stochastic models that use conditional probabilities to describe dependencies between genes in a network [134, 48, 64, 49]. These conditional distributions can be discrete or continuous, and are used to compute the likelihood of given data. Using Bayes' theorem, this is then used to compute the posterior distribution over alternative models given the data.

For large-scale network inference involving thousands of genes, *relevance network approaches* are often used. They consider the similarity or dissimilarity between pairs of genes in a network, for example using pairwise correlation or mutual information, and use the "guilt by association" principle to reconstruct the underlying network. ARACNE is a representative approach of this type, it uses Gaussian kernel estimators to compute the mutual information between two genes, and then filters the resulting networks using different criteria [97].

Main challenges in automated network reconstruction arise from (1) The exponential growth of possible model topologies for increasing network size, (2) the high level of biological and experimental variability in measured data with often low signal to noise ratios, and (3) the frequently large number of different components that are measured, combined with an – in comparison – small number of different observations under changing conditions, e.g. number of time points or perturbations of the biological system. Together these problems lead to non-identifiability and overfitting of models. Regularization methods are therefore widely employed to penalize overly complex models.

The most commonly used regularization assumption in gene regulatory network reconstruction is that the inferred models should be sparse: There are typically only a low number of regulators acting on each gene [5, 27, 144, 60]. Some studies furthermore indicate that the degree distribution in biological networks often follows a power law distribution, with only few highly-connected genes, and most genes having only a low number of interaction partners [74]. While there is ongoing debate about the statistical support of this claim [84, 92], it is widely believed that central hubs do exist in gene regulatory networks. This is usually incorporated into network inference approaches only indirectly, by limiting the number of regulators in the network [26, 21].

We here propose to use Bayesian networks with a Boolean state space to reconstruct transcriptional networks from gene expression time series data. We furthermore introduce a hierarchical prior distribution on the edge-weights in the network, which not only leads to sparse networks, but explicitly aims for the identification of central hub genes in the network, and centers the network reconstruction around these hubs.

We show results with the proposed approach on simulated as well as real experimental data sets of different sizes. Specifically, we present inference results on the genetic regulatory network controlling progression through the yeast cell cycle, based on three published genome-wide microarray studies. A first interesting result of our study indicates that large-scale network inference on this dataset is a very difficult problem, where none of the published methods we employed was able to significantly outperform random guessing. However, using the hierarchical prior presented in this work, key regulators could correctly be identified. Focusing our analysis on a smaller sub-network, we were able to reconstruct a core network regulating progression through the cell cycle. Our findings confirm that MCM1/SFF, CLB5/6 and CLN3 are key regulators in the yeast cell cycle network.

3.2.2 Bayesian Network Model

We describe the activity of genes in a transcriptional network of n genes using discrete variables $x_i \in \{\pm 1\}$, $i = 1, ..., n$, where $x_i(t) = 1$ means that gene i is active at time

t, and $x_i(t) = -1$ means the gene is inactive. We furthermore assume discrete time $t = 0, 1, ..., T$, and model the time-invariant probability for each gene $x_i(t)$ to be active at time t, conditional on the states of all genes at the previous time point, $\mathbf{x}(t-1) = (x_1(t-1), x_2(t-1), ..., x_n(t-1))$ using the probability distribution

$$p\left(x_i(t)|\mathbf{x}(t-1), \mathbf{W}\right) = \frac{1}{1 + e^{-x_i(t)\sum_{j=1}^{n} \mathbf{W}_{j,i} x_j(t-1)}}. \tag{3.2}$$

$\mathbf{W} \in \mathbb{R}^{n \times n}$ is a weight matrix and describes the strength of regulation between all genes. In case of an activation of gene i by gene j, $\mathbf{W}_{j,i} > 0$, in case of an inhibition, $\mathbf{W}_{j,i} < 0$, and $\mathbf{W}_{j,i} = 0$ if there is no effect of gene j on gene i.

Equation (3.2) describes a sigmoid function over the weighted sum of incoming regulations on a given gene x_i. If the sum $\sum_{j=1}^{n} \mathbf{W}_{j,i} x_j(t-1)$ is positive, the probability that $x_i(t) = 1$ will be larger than the probability that $x_i(t) = -1$, if the sum is negative, gene i will more likely be inactive than active.

Summarizing the logarithm of the likelihood (3.2) over all genes and all time points, the log-likelihood of given data D can be written as

$$\ln p(D|\mathbf{W}) = \sum_{t=1}^{T} \sum_{i=1}^{n} \ln p(x_i(t)|\mathbf{x}(t-1), \mathbf{W}), \tag{3.3}$$

where $D = \{\mathbf{x}(0), ..., \mathbf{x}(T)\}$ is the data, and $\mathbf{x}(t) \in \{-1, +1\}^n$ is the state vector of the system at time t.

We have previously used a similar model to reconstruct small signaling networks from RNAi perturbation data, see [78]. We are here extending this model for gene expression data, and use a hierarchical prior distribution to enable the hub-centered reconstruction of large-scale gene regulatory networks.

3.2.3 Prior Distribution

For this purpose, we employ a hierarchical prior distribution on the regulation strengths $\mathbf{W}$ to regularize the network reconstruction. As first level prior, independent normal distributions with variance σ_j^2 are used as prior on the weights $\mathbf{W_{j,i}}$, where the same variance σ_j^2 is used for all prior distributions over weights emanating from the same node j:

$$p(\mathbf{W}|\text{œ}) = \prod_{i,j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{1}{2}\frac{\mathbf{W}_{j,i}^2}{\sigma_j^2}}. \tag{3.4}$$

The variance serves as hyperparameter, and determines the strength of the regulatory effect a given node j can have on all other nodes. Therefore, **œ** describes the respective regulatory strength for each node in the network.

We furthermore use a second-level prior on the hyperparameter σ. Since a standard deviation needs to be positive by definition, and should neither become too large nor too small, we use a gamma distribution on the σ_j, thus

$$p(\boldsymbol{\text{œ}}|a,r) = \prod_{j=1}^{n} \frac{a^r \sigma_j^{r-1}}{\mathit{\Gamma}(r)} e^{-a\sigma_j}, \tag{3.5}$$

with positive shape and rate parameters r and a, respectively, and gamma function $\mathit{\Gamma}(r)$.

Importantly now, the same value of σ_j is used for all regulations exhibited by the same gene, i.e., for all *outgoing* edges for gene j. *Incoming* edges for a particular gene can have different values of σ. The combined effect of these two priors is that genes that receive a large weight also get a larger variance hyperparameter, and are more likely to attract further large edges in future inference steps, making the gene a hub. Correspondingly, genes with small weights get a small variance parameter σ, and it becomes increasingly difficult for these genes to attract large edges. Such a hub formation can not be achieved with ordinary sparseness priors such as L1 regression.

The shape and rate parameters r and a of the second-level prior ultimately control how large the weights of edges emanating from a particular node in the network can become. The choice of gamma distribution implies that most genes in the network have small variance hyperparameter σ. Only few genes receive large values of σ, and hence, larger values for the weights on their outgoing edges. Pruning edges with small values would then directly lead to sparse networks, where edges are concentrated around central hub genes.

If we would allow different values of σ for each edge (i.e., σ is a property of the edge, not of the gene), we would still obtain sparse networks where most edges have small values and only few edges receive large values, but the “large” edges would not center around hub genes anymore, but would be evenly distributed over the network.

We note that a similar automatic relevance determination (ARD) model has successfully been used in pattern recognition using neural networks by Neal [106], but the approach has not been used so far for genetic regulatory network reconstruction. Other related ARD approaches include Bayesian principal component analysis [15] and ARD-nonnegative matrix factorization [137].

The proper choice of prior hyperparameters (the shape and rate parameters a and r)

is critical to obtain optimal performance of the method. The values of a and r indirectly control how many hub genes there are. The regularization through the prior distribution should be sufficiently strong to learn hub genes and avoid overfitting, but regularization should not be too strong to completely dominate the learning from the data. “Good” values for a and r hence depend not only on the size of the network, but also the amount of experimental data available, the expected number of hubs in the data, and the level of noise in the data. The choice of parameters is hence a difficult issue, that – as with other Bayesian approaches and regularization parameters in general – requires a lot of experience and skill. We discuss this issue further at the end of the results section.

3.2.4 Optimization of the Posterior Distribution

Given $\mathbf{œ}$, we can write the log-posterior distribution over $\mathbf{W}$ using Bayes’ theorem as

$$\ln p(\mathbf{W}|\mathbf{œ}, D) = \ln P(D|\mathbf{W}) + \ln P(\mathbf{W}|\mathbf{œ}) - C_1, \tag{3.6}$$

where $C_1 = \ln p(D|\mathbf{œ})$ is independent of $\mathbf{W}$ and can be neglected. Similarly, given $\mathbf{W}$, again using Bayes’ rule, we can write the log-posterior distribution over $\mathbf{œ}$ as

$$\ln p(\mathbf{œ}|\mathbf{W}, D) = \ln p(\mathbf{W}|\mathbf{œ}) + \ln p(\mathbf{œ}|a, r) - C_2, \tag{3.7}$$

where again $C_2 = \ln p(\mathbf{W})$ is independent of $\mathbf{œ}$ and can be neglected.

We now iteratively optimize equation (3.6) with respect to $\mathbf{W}$ and equation (3.7) with respect to $\mathbf{œ}$, until the optimization converges. The idea here is that the optimization with respect to $\mathbf{W}$ serves to reconstruct the network, whereas the optimization with respect to $\mathbf{œ}$ controls the magnitudes of the outgoing edge weights any given node j can have. If a node j receives an outgoing weight with large value $\mathbf{W}_{j,i}$, its hyperparameter $\mathbf{œ}_j$ will increase in the next iteration, thus increasing the likelihood that other edges emanating from j will also receive larger weights, making j a hub gene. We note that the shape and rate parameters r and a of the second level prior (3.5) indirectly control the expected number of hub genes.

The choice of starting point for optimization algorithms such as gradient descent is an important issue, depending on which different local or global optima can be identified in the optimization. We expect resulting networks to be sparse, and therefore, most of the weights $\mathbf{W}$ should be close to zero. We therefore suggest to start the gradient descent with respect to equation (3.6) at or in the vicinity of the origin, with a fairly large starting value of σ to initially avoid a strongly peaked prior distribution $P(\mathbf{W}|\mathbf{œ})$.

The major disadvantage of gradient based optimization is that only a single maximum a posteriori estimate of $\mathbf{W}$ and **œ** is returned. However, multiple different networks might explain given data, corresponding to different modes of the posterior distribution. Although we expect the resulting network to be sparse, starting the gradient descent at the origin for $\mathbf{W}$ may results in getting stuck in a suboptimal local optimum. As an alternative for small networks, we therefore sample from the posterior distribution using the Hybrid Monte Carlo algorithm, a Markov chain Monte Carlo sampler that was originally proposed by Duane [37], see also Neal [106] and Kaderali [77]. The basic idea for our application is in each step of the Markov chain to randomly decide whether to sample from equation (3.6) or from equation (3.7) using hybrid Monte Carlo. These results can not only be used to validate the gradient based computations, but furthermore allow it to study the full posterior distribution over networks and model parameters, given the data. This is of particular value in case of multimodal distributions, when several different network topologies or sets of model parameters are consistent with the observed data.

3.2.5 Evaluation of Networks

We use Receiver Operator Characteristic and Precision-Recall analysis to evaluate results of the network reconstruction. In our model, the **œ**$_j$ provide information on the importance of individual genes in the network, the $\mathbf{W}$ describe the inferred network topology. To assess the quality of reconstructed networks, we evaluated precision (fraction of true positives in all predicted regulations), sensitivity (=recall, fraction of true positives in all actual positives) and specificity (fraction of true negatives in all actual negatives) of our approach. For this purpose, a variable threshold c on the absolute value of the weights W is introduced, edges with weights below the threshold are pruned from the network, and precision, sensitivity and specificity of edge recognition are then computed. Receiver Operator Characteristic (ROC) and Precision to Recall (PR) curves can then be plotted by varying the threshold c and plotting the resulting sensitivity over specificity, or precision over sensitivity (recall), respectively. Each value of c results in a specific point in these plots, the ROC and PR curves arise by varying c continuously and connecting the resulting points. ROC graphs nicely describe the overall relationship of positive to negative instances in the predicted model, and have the advantage to be insensitive to changes in the class distribution. On the other hand, precision to recall curves consider only the correctly inferred positive instances amongst all predicted links, and are therefore particularly useful for sparse networks. PR and ROC curves are then summarized further using the area under the curve (AUC), which is a value between 0 and 1. The closer this value is to one, the better is the reconstructed network. We compute the AUC for both ROC and PR curves.

We note here that the computation of sensitivity, specificity and precision usually requires two-class problems. In our context, three classes are possible for each edge – a positive regulation, an inhibition, or no regulation between two given genes. The assignment of predicted links to the four possible outcomes *true positive (TP), false positive (FP), true negative (TN) and false negative (FN)* used for the computation of sensitivity, specificity and precision is shown in Table 3.1.

Table 3.1: Evaluation of Predicted Networks

	Predicted Regulation		
	Activation	Inhibition	No Regulation
Actual Regulation			
Activation	TP	FP	FN
Inhibition	FP	TP	FN
No Regulation	FP	FP	TN

Classifications of predicted links as true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). The assignment given here is used in the three-class classification problem to compute sensitivity, specificity and precision.

Importantly, the three classes imply that guessing a network will on average not result in an AUC value of 0.5 anymore, but values smaller than 0.5, depending on the number of activations, inhibitions and nonexistent edges in the true network. For a technical proof, see Mazur *et al.* [100].

We performed a statistical test to assess the significance of the difference of the obtained AUC values from AUCs for randomly generated Networks. The null hypothesis is that the AUC of the ROC curve is not different from the AUC for guessing. Since our ROC curves are based on a three-class problem, we can not apply out of the box solutions for the calculation of the p-value. Therefore, we extended the methods of the R package pROC developed by Xavier *et al.* [119], which employs the method by DeLong *et al.* [35]. Briefly, the method of DeLong employs the mathematical equivalence of the AUC to the Mann-Whitney U-statistic. ROC curves can then be compared by evaluating the difference of the AUCs, which is asymptotically normal. To compare two AUC values, the method uses the covariance matrix for each of the ROC curves and finally does a two-sided t-test on the score of this comparison. To be be able to apply DeLong's method, we extended the Mann-Whitney kernel implementation of the pROC package as follows:

$$\theta(x,y,z) = \begin{cases} 1 & \text{if } z = 1 \text{ and } y < x \\ 1/2 & \text{if } z = 1 \text{ and } y = x \\ 0 & \text{if } z = 1 \text{ and } y > x \\ 1 & \text{if } z = -1 \text{ and } y > x \\ 1/2 & \text{if } z = -1 \text{ and } y = x \\ 0 & \text{if } z = -1 \text{ and } y < x \end{cases} \tag{3.8}$$

with x being the cases or TN, y being the controls or TP and z being the signs of the edges in the gold standard, either 1 or -1.

A further problem in the evaluation of reconstruction performance on real data arises due to the lack of a "gold standard" network. Hence, to evaluate the hub gene identification on real data, we extracted protein networks from the STRING database [136]. STRING calculates for each interaction a score based on the evidence from various sources like experiments, interaction databases or abstract text mining. It is clear that the PPI network reflects only a part of the gene regulatory processes but still, observations at this level can provide insight into the performance of the methods. STRING is also considering pairs derived from co-expression analysis and might therefore be more suitable than other PPI databases. We then computed the degree d_i of each gene i in the STRING network, and assessed correlations between d_i and the network inference hyperparameter σ_i. We then again used receiver operator characteristic analysis to study the predictive strength of σ to identify hub genes, by varying a threshold on σ for a fixed threshold on the degree d, and computing sensitivity and specificity. ROC curves were summarized using the AUC, and AUC was plotted for continuously varied d.

3.2.6 Overview of Results

We implemented our method in C++, using the gnu gcc compiler under the Linux operating system. All computations reported were carried out on a 3 GHz 64 bit Intel processor using a single processor core (no parallel processing). For a systematic evaluation of the approach, we used different simulated datasets, as well as real, publicly available microarray data.

Simulated data has the advantage that the real network underlying the data is known, and can be used to evaluate the performance of the network reconstruction and hub identification. We therefore discuss simulated data first. More specifically, we start by showing results using data that was simulated with the Boolean model used also in the inference method, using three different network sizes (11 genes, 100 genes, and 1000 genes), and using different dataset sizes generated from these networks for the inference task (20, 40 and 200 time points). This simulated dataset allows it to study the effect of network size and dataset size on performance of the network inference. To evaluate whether the choice of prior introduces artificial hubs even on random networks where no hubs are present, we furthermore simulated data for a 1000 gene Erdös-Rényi [107] random network, again with different numbers of time points (20, 40 and 200 time points).

We next proceed by using a further simulated dataset, that was simulated with a realistic kinetic model for gene regulation, implemented in the GeneNetWeaver (GNW) package [96]. GNW uses systems of differential equations for simulation, data hence need to be discretized before they can be used in the network inference. GNW allows the simulation of time course data using a realistic model of noise for microarray data, this dataset hence allows to study the effect of noise on the network reconstruction.

We finally applied our network inference method to three different publicly available microarray gene expression data sets regarding the yeast cell cycle, published by Spellman [133], Cho [29] and Pramila [116]. These three datasets were pooled, and network inference done on the ensemble dataset. We start by showing results on a small subset of the genes in this pooled dataset, representing a core network of 11 genes known to be involved in the yeast cell cycle. Thereafter, we present results on the reconstruction of a relatively large yeast transcriptional network comprising almost 800 genes. On this dataset, we compare results of our approach with results obtained using the relevance-network approaches ARACNE [97] and MRNet [99], as well as the Bayesian approach implemented in Banjo [13].

All analyses done and results achieved on simulated and real data are summarized in Tables 3.2 and 3.3.

3.2.7 Synthetic Data Simulation with Boolean Model

To systematically evaluate our network reconstruction approach, we simulated data for three different network topologies, with different numbers of genes. The smallest network contained 11 genes, and is the yeast cell cycle core network described by Li and coauthors [89], as shown in Figure 3.5. We furthermore used the *CenturySF* network topology comprising 100 genes, and the *JumboSF* network topology comprising 1000 genes, proposed by Mendes [101]. These topologies include desired properties such as regulatory loops, hub genes, and are sparse.

Table 3.2: Overview of Analyses on Network Inference

Simulated Data

Network	**Nodes**	**Edges**	**TP**	**Gradient Descent**			**MCMC**		
				ROC	**p-val**	**PR**	**ROC**	**p-val**	**PR**
Boolean Model									
Yeast Cell Cycle (CC)	11	34	20	0.74	0.0035	0.37	0.68	0.024	0.37
Core (Simulated)			40	0.76	0.0029	0.48	0.74	0.0045	0.48
			200	0.93	7.78e-08	0.67	0.91	6.4e-08	0.77
Mendes CenturySF	100	200	20	0.64	< 1e-8	0.13	0.43	0.0007	0.005
			40	0.75	< 1e-8	0.3	0.52	< 1e-8	0.04
			200	0.90	< 1e-8	0.66	0.67	< 1e-8	0.19
Mendes JumboSF	1000	999	20	0.68	< 1e-8	0.05			
			40	0.77	< 1e-8	0.26			
			200	0.88	< 1e-8	0.62			
Random Network	1000	5000	20	0.42	-	0.003			
			40	0.62	-	0.09			
			200	0.79	-	0.4			
GeneNetWeaver					-				
No noise	100	532	25	0.53	-	0.053			
	250	1317	25	0.50	-	0.020			
	500	2150	25	0.50	-	0.008			
With noise	100	532	25	0.51	-	0.054			
	250	1317	25	0.50	-	0.021			
	500	2150	25	0.50	-	0.009			

Real Data

Network	**Nodes**	**Edges**	**TP**	**Gradient Descent**			**MCMC**		
				ROC	**p-val**	**PR**	**ROC**	**p-val**	**PR**
Li *et al.* Yeast CC Core	11	34	98	0.56	-	0.27	0.59	-	0.26
Large Yeast CC Network	781	unk.	98	0.52	-	0.01			

Overview of all results on the simulated and biological datasets, using the approach presented in this manuscript. See the main text for comparison with other methods. Shown are results for the full network reconstruction task; table 3 shows corresponding results for hub identification. Each row in the table corresponds to one dataset. Nodes, edges and TP gives the number of genes, regulations and time points in the respective dataset. ROC and PR are the area under the curve values (AUC) of the Receiver Operator Characteristic (ROC) and Precision-Recall (PR) analysis, respectively. P-values were computed to test the null hypothesis of a significant deviation from random guessing for the AUC ROC values. Due to runtime limitations, MCMC results were calculated only for small networks, and p-values only for the synthetic networks with AUC values >0.5. unk.: True number of edges for Yeast CC Network is unknown.

Table 3.3: Overview of Analyses on Hub Identification

Simulated Data					
Network	**Nodes**	**Edges**	**TP**	**AUC Hub**	
				Top 10 Hubs	Overall
GeneNetWeaver					
No noise	100	532	25	0.76	0.77
	250	1317	25	0.31	0.56
	500	2150	25	0.74	0.85
With noise	100	532	25	0.29	0.45
	250	1317	25	0.83	0.83
	500	2150	25	0.92	0.92
Real World Data					
Network	**Nodes**	**Edges**	**TP**	**AUC Hub**	
				Top 10 Hubs	Overall
Large Yeast CC Network	781	unk.	98	0.94	0.97

Overview of all hub identification results on the simulated and biological datasets. Hub AUCs were only calculated for the large networks since they are only of little relevance for small networks. Each row in the table corresponds to one dataset. Nodes, edges and TP gives the number of genes, regulations and time points in the respective dataset. AUC Hub is the AUC value computed for hub identification, shown are AUC values for the top 10 hub genes and maximum overall AUC values. A value of 0.5 corresponds to random guessing, values between 0.5 and 1 measure the hub identification performance. unk.: True number of edges for Yeast CC Network is unknown.

Weights for given topology were uniformly randomly generated between -2 and $+2$, a starting state was randomly chosen, and time courses were simulated with 20, 40 and 200 time points, using the stochastic model described by equation (3.2). Weights in this range correspond to a moderate level of noise in the experimental data, due to the probabilistic model employed to simulate the data.

We then took the simulated data, and used our gradient descent and Markov chain approaches to reconstruct the underlying networks from the data alone. Shape and rate parameters of the gamma prior (3.5) were set to $r = 4.8$ and $a = 0.4$ for these computations.

3.2.8 Results with Gradient Descent

Iterative gradient descent on the equations (3.6) and (3.7) was carried out as described in the methods section, until convergence was reached. For the 11 gene network, computation finished in a few seconds. For 100 genes, computation time varied between 17s for

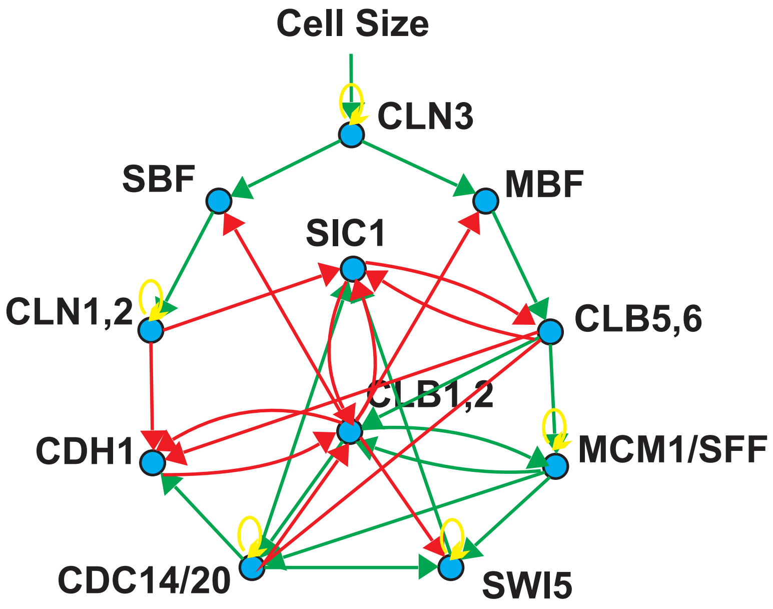


Figure 3.5: Yeast Cell Cycle Core Network. Core yeast cell cycle network, as derived by [89] from literature. There is one external checkpoint, cell size, which initiates progression through the cell cycle. Activations are shown in green, inhibitions in red, and self-regulations in yellow.

20 time points, up to 4.1 min for 200 time points. On the 1000 gene network, gradient descent required 12.3 min, for 20 time points, 90.5 min for the 40 time point data set, and 447.33 min and roughly 7 1/2 hours on the 200 time point data set.

Figure 3.6 shows ROC and PR curves for the networks reconstructed from the data, in dependence of network size and number of time points available. As expected, for small network sizes (11 genes) and many (200) time points, the network reconstruction performs very well, and performance decreases with increasing network size and decreasing number of time points. Corresponding AUC values together with p-vales to assess the significance of the results (H0: AUC values are not superior to guessing, see methods for details) are shown in Table 3.2. To put these results further into perspective, we generated 1000 random “reconstructed” networks with 11, 100 and 1000 genes each, by drawing weights from a standard normal distribution, and computed the AUC for these networks. For the 11, 100 and 1000 gene networks this yields an average AUC_{ROC}^{guess} of 0.34, 0.38 and 0.39, respectively, and an average AUC_{PR}^{guess} of 0.14, 0.009 and 0.0005. This reconfirms that results are significantly better than random.

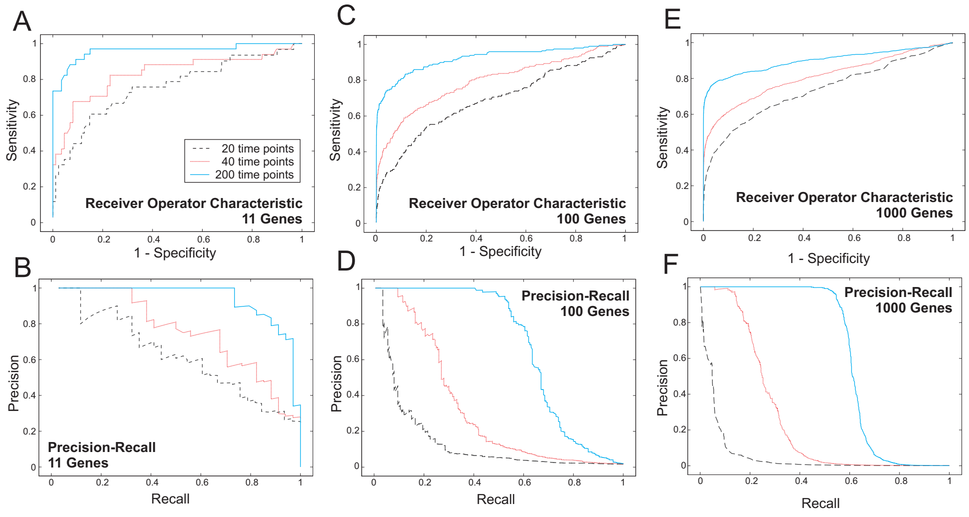


Figure 3.6: ROC and PR Results on Simulated Data. The Figure shows receiver operator characteristic (ROC) and precision to recall curves (PR) for network reconstruction on simulated data, for different network sizes and different numbers of time points. A, B: ROC and PR curves for the network with 11 genes, C,D: ROC and PR curves for network with 100 genes, E.F: ROC and PR curves, respectively, for network with 1000 genes. Black: 20 time points used for network reconstruction, red: 40 time points, blue: 200 time points. It can clearly be seen how performance deteriorates with increasing network size and decreasing number of different time points. We note that, due to the three-class classification problem underlying the graphs, random guessing of network topologies would not yield a diagonal line in the ROC plots, but a significantly lower line with an area under the curve of approximately 0.33.

3.2.9 Results with MCMC

We next repeated the computation using the Markov chain Monte Carlo sampling approach. Due to the high running time, an evaluation was done only for the 11 and 100 gene networks, by iteratively sampling from the distributions (3.6) and (3.7). 1 million sampling steps were done for the 11 gene network. Due to runtime constraints, only 800,000 steps were done on the 100 gene network. Running times for 20, 40 and 200 time points were 116, 207 and 929 minutes for the 11 gene network, and 10, 20 and 80 days for the 100 gene network, respectively. We note that computations were done using a single processor thread, and significant speed-ups can clearly be expected from parallelization of the sampler.

To simplify analysis of the reconstructed networks, we summarized the different values sampled for each parameter by the mean. This clearly is a crude intervention, and disregards much of the additional information contained in the distribution, for example, in case of a bimodal distribution. More sophisticated methods such as cluster analysis, and the consideration of higher order moments, can be used here. In spite of this simplification, results for the 11 gene network were completely equivalent to results for the gradient descent method (see Table 3.2), indicating that in this simulated example, only one set of parameters corresponding to one network topology is consistent with the experimental data, and is recovered using both gradient descent and Markov chain. Results of the 100 Gene Network obtained using the MCMC sampler were still significantly better than guessing, but inferior to results obtained from gradient descent, compare Table 3.2. This is likely due to multiple local optima of the posterior distribution. In this situation, averaging over multiple modes leads to an average result with low posterior probability, and thus suboptimal results. Furthermore, the number of sampling steps carried out (800,000) may not be sufficient to achieve adequate sampling from the stationary distribution, but this was a limiting factor due to runtime.

3.2.10 Results on a Non-Hub Network

To test our approach for biases towards inferring a scale-free structure also if no such structure is present in the gold standard network, we tested the gradient descent method on a random (Erdös-Rényi) 1000 gene network (generated with igraph [33]) with 5000 interactions. The data set size is the same as for the scale-free networks, we simulated each of 20, 40 and 200 time points as described above. Network reconstruction was then done using the same settings as above, with conjugate gradient descent.

We focused our analysis of the results on the question if the network inference learns artificial hubs from the data, although none are present. Correspondingly, we evaluated the degree distribution of the reconstructed networks. Figure 3.7 shows the resulting degree distribution, for the 1000 gene scale-free network above (JumboSF, Figure 3.7 left plot), as well as the Erdös-Rényi random network (Figure 3.7, right plot). The results clearly show that the approach does not identify artificial hubs, provided sufficient amounts of experimental data are available. In case of the data set with 20 time points, a distortion of the random network result can be seen. This data set is too sparse and the method cannot infer the right topology from it. In this situation the prior starts dominating the obtained results.

3.2.11 Synthetic Data Simulation with GeneNetWeaver

As a further test of the method, we next simulated data using a realistic kinetic model, implemented in the GeneNetWeaver (GNW) package [96]. We subsampled networks of

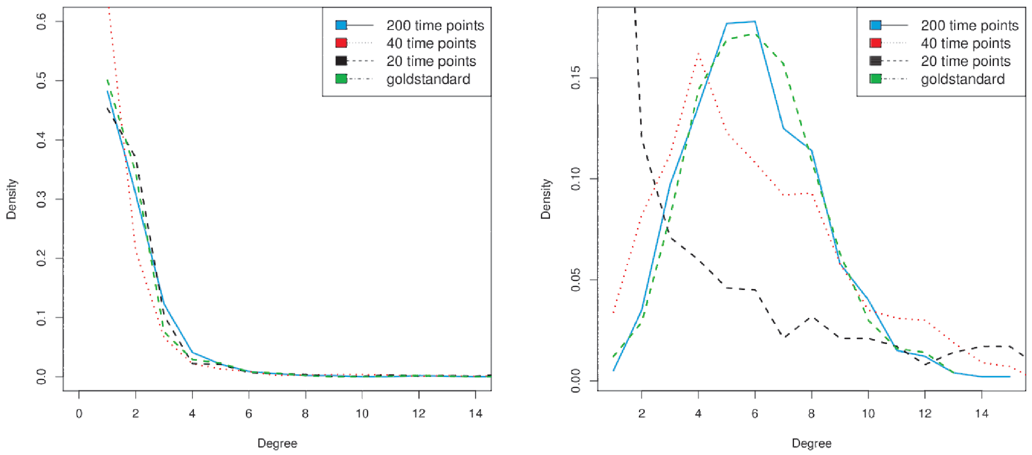


Figure 3.7: Inferred Degree Density Distribution on Scale-free and Random Networks. To test whether artificial hubs are generated in network inference due to their used prior distribution, we performed a comparative analysis on two different 1000 gene networks. The first network is the JumboSF network, a large scale-free network with central hub genes. The second network is a random Erdös-Rényi network, which does not contain any hubs. Network inference was performed using identical parameter values for the hyperparameters on both data sets. The figure shows the degree distribution of the inferred networks, in dependence of the number of time points used for network inference (left: JumboSF, right: random network). The plot shows that, provided sufficient data is available, the prior distribution does not lead to artificial hubs. On the other hand, if only little data is used for network inference, the prior starts dominating the results, as one would expect.

size 100, 250 and 500 from the Yeast transcriptional network implemented in GNW, and generated 25 time points using the ordinary differential equation model, with settings as for the DREAM challenge (see GeneNetWeaver documentation). Data were simulated without noise and with noise using the DREAM microarray noise model implemented in GNW. Data were discretized to Boolean states using a threshold of 50% on the maximum of the simulated gene activity levels. We then used the gradient descent approach with the hierarchical ARD prior presented, as well as a standard L1 sparseness prior to reconstruct the underlying networks from the data. Results of the network reconstruction were summarized by computing the area under the ROC curve for the reconstructed edges, as well as the area under the ROC curve for the hub identification.

Due to the low number of time points simulated, overall performance of the network reconstruction was not significantly superior to guessing in all runs. However, both the L1 prior as well as the hierarchical prior led to a successful identification of hub genes, with AUC values as shown in Table 3.4, and in all but one case superior performance of

Table 3.4: AUC Results for Network Reconstruction and Hub Identification on Simulated Data:

Network Size (Genes)	no noise		noise	
	L1	**ARD**	**L1**	**ARD**
	Network Reconstruction			
100	0.508	0.527	0.510	0.511
250	0.499	0.499	0.504	0.504
500	0.504	0.497	0.499	0.496
	Hub Identification			
100	0.526	0.767	0.449	0.453
250	0.789	0.563	0.755	0.827
500	0.698	0.849	0.859	0.924

Data was simulated using the GeneNetWeaver package, subsampling networks of size 100, 250 and 500 from the yeast transcriptional network. Simulation was done using an ordinary differential equation model, with and without experimental noise added to the data. Network reconstruction was carried out using the model described, using an L1 and a hierarchical automatic relevance determination (ARD) prior, respectively. Shown are the area under the ROC curve values for the correct identification of edges (top) and hub identification (bottom). A value of 0.5 is equivalent to guessing, a value of 1 corresponds to perfect identification of hub genes.

the hierarchical ARD prior. Interestingly, in case of the smallest network simulated, the addition of noise was so detrimental that no successful hub identification was feasible using either method. This probably reflects the situation that when more genes and hence more edges are present in a network, the influence of noise on the hub identification is less severe simply due to more edges contributing information on an individual hub gene. Overall, the results indicate that in the simulated data, information content seems not sufficient to reconstruct the full network, but it is still possible to identify key regulatory genes. Together, these observation motivate the use of hub-centered methods in particular on larger networks, where full reconstruction of a network is very difficult or even fails completely, but still some information on hubs can be extracted.

3.2.12 Results using Microarray Data on the Core Network of the Yeast Cell Cycle

We next evaluated our reverse engineering approach using publicly available microarray data regarding the yeast cell cycle. Data were pooled from the studies by Spellman [133], Cho [29] and Pramila [116]. We discarded the CDC15-synchronized data from the Spellman data set, due to previous reports of quality problems [42]. Experimental measurements were interpolated using smoothing splines, and binarized using the median of each gene as threshold. Missing values were interpolated with the mean of the preceding and the following time point. This discretization of the data into binary (Boolean) states can lead to several consecutive time points without any changes in all genes, such time points were then collapsed into a single time point, i.e. repetitive states after the binarization were removed. Network inference was performed using all time series simultaneously.

As reference network to evaluate the performance of our reconstruction, we used the 11 gene yeast cell cycle model proposed by Li *et al.* [89], see Figure 3.5. This network was carefully constructed from the literature, and we constrained our further analysis on reconstructing the interaction network between the 11 genes contained in this core network.

Network reconstruction was done using gradient descent, with shape parameter $r = 4.6$ and rate parameter $a = 0.2$. Precision, sensitivity and specificity for reconstructed networks were computed as described in methods, and used to plot receiver operator characteristic and precision to recall curves. The area under the curve was then calculated, resulting in $AUC_{ROC} = 0.56$ and $AUC_{PR} = 0.27$. As it has been done for the synthetic networks, we generated 100 random networks and computed the AUC for these networks. This yields an average AUC_{ROC}^{guess} of 0.35 and an average AUC_{PR}^{guess} of 0.13, indicating that our approach performs significantly better than guessing.

To furthermore study the effect of the choice of starting point for the gradient descent, we performed computations with different starting values, results are summarized in Figure 3.8. These results support the choice of the origin as starting point for the gradient descent, which seems to give good results. The rationale here is that we expect sparse networks, hence most edges should have weights equal to or close to zero. Apparently, if largely distinct values are chosen, the optimization tends to get stuck in local optima corresponding to overly complex, non-sparse networks.

We next repeated the network reconstruction using the Monte Carlo sampler, using 800,000 iterations and a burn-in phase of 50,000 steps. Computation time was 264 minutes, or 4 hours and 24 minutes. To check for convergence of the Markov Chains, several chains were run with different starting points, length, and random seed, and results were compared, indicating good convergence of the chains to the stationary distribution. We summarized values sampled for each model parameter by the mean, and used this to evaluate the reconstruction performance. Results overall were very similar to the ones obtained using gradient descent, with $AUC_{ROC}^{\text{MCMC}} = 0.59$ and $AUC_{PR}^{\text{MCMC}} = 0.26$, again significantly outperforming guessing.

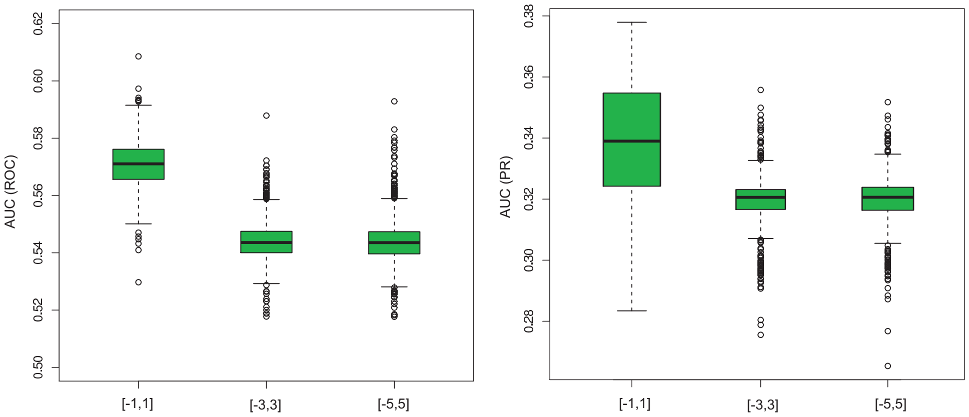
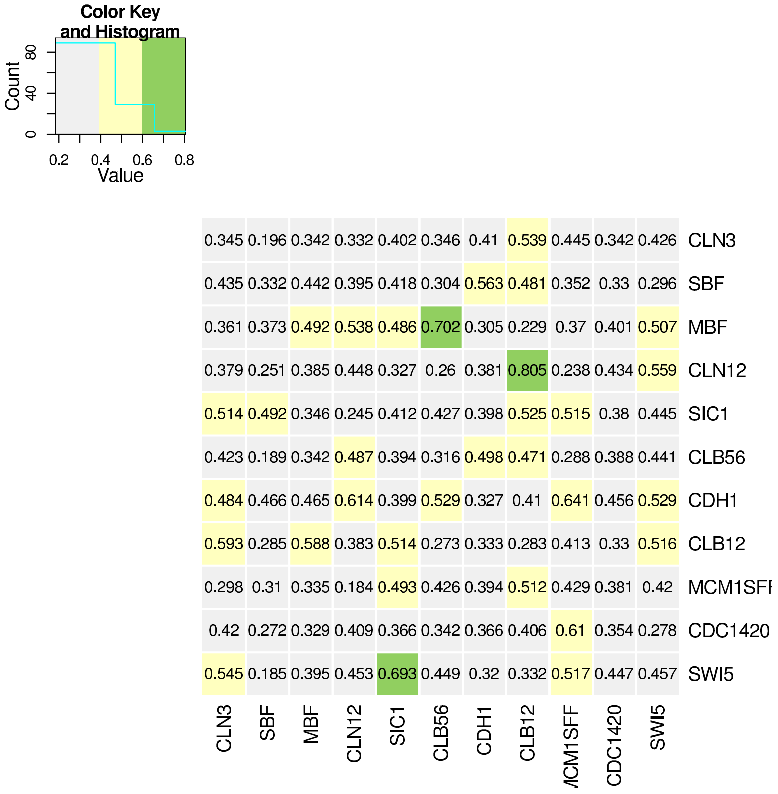


Figure 3.8: Effect of Starting Point on obtained AUC values. Shown are the distribution of AUC values (left: ROC, right: PR) of 1000 gradient descent runs, for randomly chosen starting values for **W**, on the yeast core network. For the parameter vector **W**, randomly chosen values within ranges of [−1,1], [−3,3] and [−5,5] were used as a starting points for the calculations with CG. This was done for each of the suggested ranges 1000 times, and AUC ROC and AUC PR values were computed. The boxplots show the comparison between the different AUC values for these calculations. It can be clearly seen, that randomly sampled start values close to zero allow the approach to obtain better results for the optimal values of w. If the range of initial values for **W** is too large, the optimization ends in suboptimal local optima corresponding to overly complex networks with many non-zero edges.

Interestingly, obtained values for the hyperparameter σ were very similar for all genes, both for the Markov chain Monte Carlo and the gradient descent approach. This probably reflects the fact that on such small networks, consisting of only 11 genes, the definition of hub genes is not or only marginally useful, and does not significantly influence network reconstruction. Still, largest hyperparameter values were attained by MCM1/SFF, CLB5/6, SBF and CLN3 which are key genes in the cell cycle network. For example, CLN3 initiates the cell cycle, or the transcription factor MCM1/SFF controls downstream genes like CLB2, CDC20 and SWI5.

It is clear that an analysis based on the mean of all values sampled for each parameter is a major simplification, and will actually yield inferior results in case of multimodal distributions. We have sampled 750,000 different values for each edge from the posterior distribution over model parameters, given the data, and clearly, this data can not only be used to provide confidence intervals on parameter estimates, but might also point to alternative topologies consistent with the data. To gain a better picture of the landsacpe of different modes and thus possible alternative topologies, we used the Dip test of unimodality on the Markov chains. This test, suggested by Hartigan and Hartigan

(1985) [65], measures the departure of an empirical distribution from the best fitting unimodal distribution. The smaller this Dip-score statistic becomes, the more likely the distribution is unimodal. Due to the large sample size used in our Markov chains, the Dip test would reject the null hypothesis of unimodality for all edges in our network. We hence directly use the Dip value as a measure of the "deviation from unimodal". Figure 3.9 shows the average Dip values for three prior settings ($a = 0.2$ and $r = 4.6$, $a = 1.6$ and $r = 1.6$, $a = 4.6$ and $r = 0.2$), indicating that several of the edges show clear multimodal distributions. These edges could now be characterized further experimentally, to assess the true underlying network. A cell might have different interaction paths for different states, like proliferation and differentiation, and also fallback paths exists to avoid single point of failures.



CLN3	SBF	MBF	CLN12	SIC1	CLB56	CDH1	CLB12	MCM1SFF	CDC1420	SWI5	
0.345	0.196	0.342	0.332	0.402	0.346	0.41	0.539	0.445	0.342	0.426	CLN3
0.435	0.332	0.442	0.395	0.418	0.304	0.563	0.481	0.352	0.33	0.296	SBF
0.361	0.373	0.492	0.538	0.486	0.702	0.305	0.229	0.37	0.401	0.507	MBF
0.379	0.251	0.385	0.448	0.327	0.26	0.381	0.805	0.238	0.434	0.559	CLN12
0.514	0.492	0.346	0.245	0.412	0.427	0.398	0.525	0.515	0.38	0.445	SIC1
0.423	0.189	0.342	0.487	0.394	0.316	0.498	0.471	0.288	0.388	0.441	CLB56
0.484	0.466	0.465	0.614	0.399	0.529	0.327	0.41	0.641	0.456	0.529	CDH1
0.593	0.285	0.588	0.383	0.514	0.273	0.333	0.283	0.413	0.33	0.516	CLB12
0.298	0.31	0.335	0.184	0.493	0.426	0.394	0.512	0.429	0.381	0.42	MCM1SFF
0.42	0.272	0.329	0.409	0.366	0.342	0.366	0.406	0.61	0.354	0.278	CDC1420
0.545	0.185	0.395	0.453	0.693	0.449	0.32	0.332	0.517	0.447	0.457	SWI5

Figure 3.9: Multimodal Distributions in the Yeast Cell Cycle Core Network. Shown are Dip scores for the distribution of sampled edge weigths from the Markov chain. The Dip value measures the departure of an empirical distribution from the best fitting unimodal distribution. Large scores indicate a stronger deviation from unimodality. Rows in the diagram represent source (regulating) genes for edges, columns the target (regulated) genes. Colors have been used to indicate the magnitude of the deviation from unimodality.

3.2.13 Hub Genes in the Yeast Transcriptional Regulation

The previous example on the core cell cycle network regards a relatively small network. For such small networks, the definition of hub genes is not so useful, and accordingly, the parameters σ_i describing the importance of individual genes in the network were all similar, and essentially peaked at the mode of the prior distribution (3.5). To evaluate hub genes in larger networks, we took the set of 800 cell cycle regulated genes reported by Spellman et al. [133], and intersected this gene set with the genes in the Pramila data set [116], resulting in a set of 781 genes. Data were preprocessed as described above, network reconstruction was carried out using gradient descent. Shape and rate parameters of the prior were set to $r = 3$ and $a = 7.5$, posterior optimization took 145 minutes. Computation with the Markov chain sampler is not feasible for this large network due to excessive running time. We furthermore used ARACNE, MRNet and Banjo for comparison, and furthermore repeated the computation with the model (3.3) using a standard L1 sparseness prior. ARACNE and MRNet results were computed using the R package minet [99]. ARACNE results were computed using default parameters in the minet implementation. Since minet uses additive tolerance instead of multiplicative tolerance, we furthermore used parmigene [124] to compute ARACNE results, and applied DPI thresholding at three different thresholds from 0.01, 0.05 and 0.15. MRNet results were computed using the Spearman estimator, the number of bins was set to $\sqrt{N}$ with N being the number of samples, as suggested in the documentation. Banjo was run with default parameters.

Since our simulation study indicates that at least 200 time points are required to successfully reconstruct a network of the given size, it is clear that individual edges predicted in our inferred network must be interpreted with great caution and need further experimental validation. In fact, we directly evaluated the reconstructed networks by comparison with the STRING database, using only experimentally verified or all interactions. We computed sensitivity/specificity and precision/recall of the reconstructed networks, and plotted ROC and precision-recall curves. None of the methods was able to perform better than guessing on this pooled dataset (Area under the curve for ROC [PR] analysis: Hierarchical Prior 0.515 [0.0106], L1-Prior 0.49744 [0.059], L2-Prior 0.496 [0.058], ARACNE 0.4993 [0.0099], Banjo 0.500 [0.06], MRNET 0.498 [0.059]).

We will therefore concentrate our analysis of this large reconstructed network on the identification of central hubs predicted.

Figure 3.10 shows a histogram of the reconstructed regulation strengths for the $781^2 = 609{,}961$ possible regulations between all pairs of the 781 genes. Negative weights

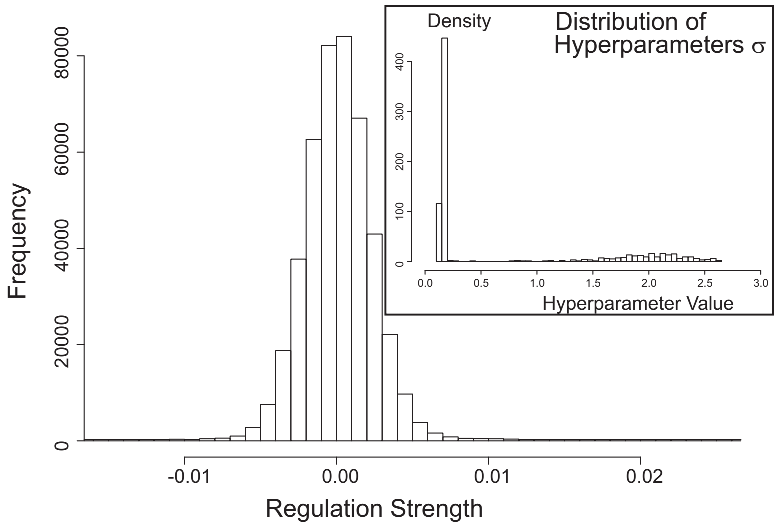


Figure 3.10: Hub Genes in the Yeast Cell Cycle. Histogram of reconstructed regulation strength for the full yeast cell cycle dataset. Negative weights correspond to inhibitions, positive weights to activations. Weights in the vicinity of zero indicate no regulation between two genes. The plot shows the distribution of regulation strengths between any two genes, showing clearly that only few genes exhibit strong regulations. The inset shows a histogram of the corresponding hyperparameters σ (equation 3.7), controlling the magnitude of the regulations exhibited by a particular gene. As can clearly be seen, most genes have only small importance corresponding to low values of σ, and only few genes are assigned large values of σ and correspondingly large weights on their outgoing connections.

correspond to inhibitions, positive weights to activations, and weights in the vicinity of zero indicate no regulation between two genes. The inset in the figure shows the distribution of hyperparameters σ for the 781 genes, providing a direct measure of the importance of individual genes. A large value of σ_i for a gene i indicates that the gene has strong (positive or negative) effects on other genes. For example, 114 genes (14.5%) have a hyperparameter of $\sigma_i > 2$ and 209 genes (26.7%) have $\sigma_i > 1$, predicting that these genes play important roles in the yeast gene regulatory network.

Since predicted regulation strengths are continuous, we pruned all weights with absolute value < 0.75 from the network. This yields a network with average out-degree 2.65. A plot of the correlation between the number of other genes regulated by a gene and σ shows a good linear correlation (Pearson $\rho = 0.699$, plot not shown), reconfirming that σ appropriately summarizes the genes importance in the reconstructed network. 114 genes have a hyperparameter value $\sigma_i > 2$, they on average are predicted to regulate 16.8 other genes, whereas an average gene in the full network regulates only 2.65 other genes.

We next evaluated in more detail genes identified as “hubs” in the transcriptional network. We retrieved interactions between the 781 genes in our dataset from the STRING database, using all interaction types. We then computed the degree d_i of each gene i in the STRING network, and assessed correlations between d_i and the network inference hyperparameter σ_i.

Pearson correlation between σ_i and d_i was only weak ($\rho = 0.0497$), probably due to the large number of non-hub genes contributing significant noise to the correlation coefficient, and possibly also influenced by false positives in the database network. Accordingly, correlation improves to $\rho = 0.127$ if the top 25%, $\rho = 0.334$ if the top 5%, and $\rho = 0.711$ if only the top 1% predicted hub genes are used.

We then used receiver operator characteristic analysis to study the predictive strength of σ to identify hub genes, by varying a threshold on σ for a fixed threshold on the degree d, and computing sensitivity and specificity. ROC curves were summarized using the AUC, and AUC was plotted over different thresholds on the degree d, as shown in Figure 3.11. To compare results obtained using our approach with other methods, we reconstructed networks using ARACNE [97], MRNet [99] and Banjo [13], using the same input data. Importance values σ_i were then computed for each gene from the reconstructed edge weights as described above, and we then computed ROC and AUC values. We furthermore compared these results with a reconstruction using equation (3.3) with a normal and a L1 prior distribution, to study the effect of the hierarchical prior distribution used.

Figure 3.11 summarizes the AUC values obtained with these different approaches, in dependence of the STRING degree of the underlying genes. The dashed grey line in Figure 3.11 corresponds to the expected AUC for random guessing, the solid red curve shows the AUC for our ARD approach using the full posterior distribution. The dotted brown curve shows results using a L1 sparsity prior, the dotted black curve was obtained using a Normal distribution as prior. In comparison, the green and the pink dot-dashed curves were obtained using the relevance network approaches ARACNE and MRNet, respectively, whereas the dashed blue line shows results of the Bayesian method Banjo. While the Bayesian ARD approach performs only slightly better than guessing for low-degree genes ($AUC \approx 0.55$), it makes excellent predictions for highly connected genes, which it identifies as hub-genes with high area under the ROC curve, and thus with high sensitivity and specificity. A comparison with the same model using an L1 and a normal prior shows clearly how the prior distribution used helps identify hub genes. Interestingly, at least on this dataset, the relevance network approaches ARACNE and

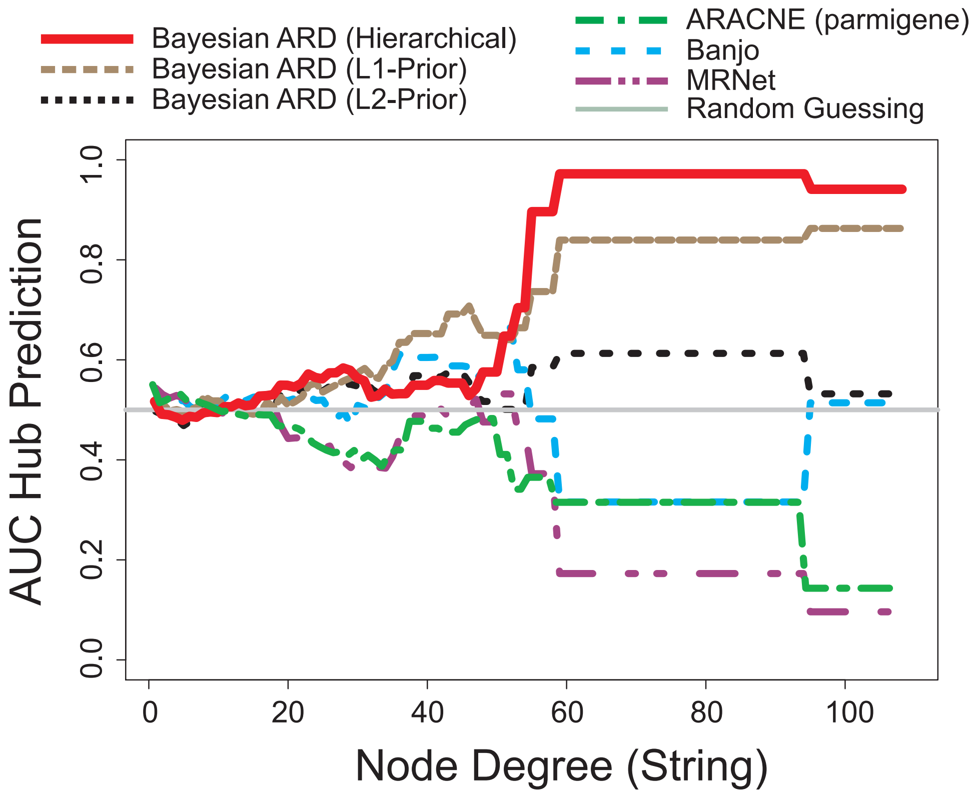


Figure 3.11: Receiver Operator Characteristic Analysis for the Prediction of Hub Genes in the Yeast Cell Cycle. Genes were split in two groups "hub" and "non-hub" based on a threshold θ on the degree of the gene in the literature derived network, and ROC curves were computed by then varying the threshold on σ. ROC curves were summarized for each θ using the area under the curve. The plot shows $AUC(\theta)$ over θ. The red curve shows results for the inferred network using the method presented, the black dotted line shows results using the method with a Normal prior, the brown dashed line using a L1 "sparseness" prior distribution. The dashed blue line was obtained using Banjo, the dot-dashed green lines shows results of ARACNE, the dot-dashed pink line represents results of MRNet. The grey dashed line corresponds to the expected value for randomly guessing a network. Larger AUC values indicates better performance.

MRNet performed worst, and actually make hub predictions that are inferior to guessing.

3.2.14 Choice of Prior Hyperparameters

A critical issue is the choice of hyperparameter values a and r for the ARD prior. Optimal values for a and r depend on the size of the network, the number of experimental data points, level of noise in the data, and expected number of hub genes. Some theoretical insight on the effect of changing a and r can be gained from a marginalization of the prior over σ:

$$p(w|a,r) = \int_{0}^{\infty} p(w|\sigma)p(\sigma|a,r)\, d\sigma, \tag{3.9}$$

which can be solved and analyzed numerically. By plotting $p(w|a,r)$ over w for different values of a and r, one can see that choosing smaller values of r corresponds to a more "peaked" prior, i.e. a stronger "sparsity" of the inferred networks, whereas smaller values of a cause the overall importance of the prior to decrease. Hence, for larger networks and in case of small amounts of data, smaller values of r and larger values of a should be preferred, whereas in case of excellent and large amounts of data and small networks, r should be chosen larger and a smaller, to decrease the influence of the prior distribution.

Due to the difficulty in manually choosing these parameters, we performed a sensitivity analysis to assess the sensitivity of results with respect to choices for a and r. On the simulated data (synthetic 11 gene, 100 gene and 1000 gene data sets), we modified parameters a and r in the range from 0.2 to 4.8, performed the network inference for each combination using gradient descent using 20 and 200 time points from the data, and computed resulting AUC_{ROC} values. Results are shown in the heatmaps in Figure 3.12. The plots show that results are relatively insensitive over a large range of parameters. Smaller values of the hyperparameter r correspond to a more peaked prior distribution, resulting in "sparser" networks. Correspondingly, the figure shows that smaller values of r should be chosen for larger networks. In comparison, the correct choice of a seems less important.

On the experimental data regarding the hub genes in the yeast cell cycle, we also performed a similar analysis. We note that parameters chosen for this analysis ($a = 7.5$, $r = 3$) result in a significantly narrower distribution of σ than the hyperparameter values used on the synthetic data, corresponding to much stronger regularization – in line with expected larger levels of noise in the data. We modified both parameters individually and together by up to ±50%, reran the network inference, and computed average AUC values for the reconstructed networks. Figure 3.13 shows the resulting AUC values, and clearly shows that in spite of considerable variation of the hyperparameter values over a wide range, performance is again only marginally affected.

3.2.15 Discussion

In this paper, we present a novel approach to reconstruct gene regulatory networks from microarray gene expression time series data, which employs the concept of *hub genes* for regularization. Our evaluation on the simulated data shows that the method precisely

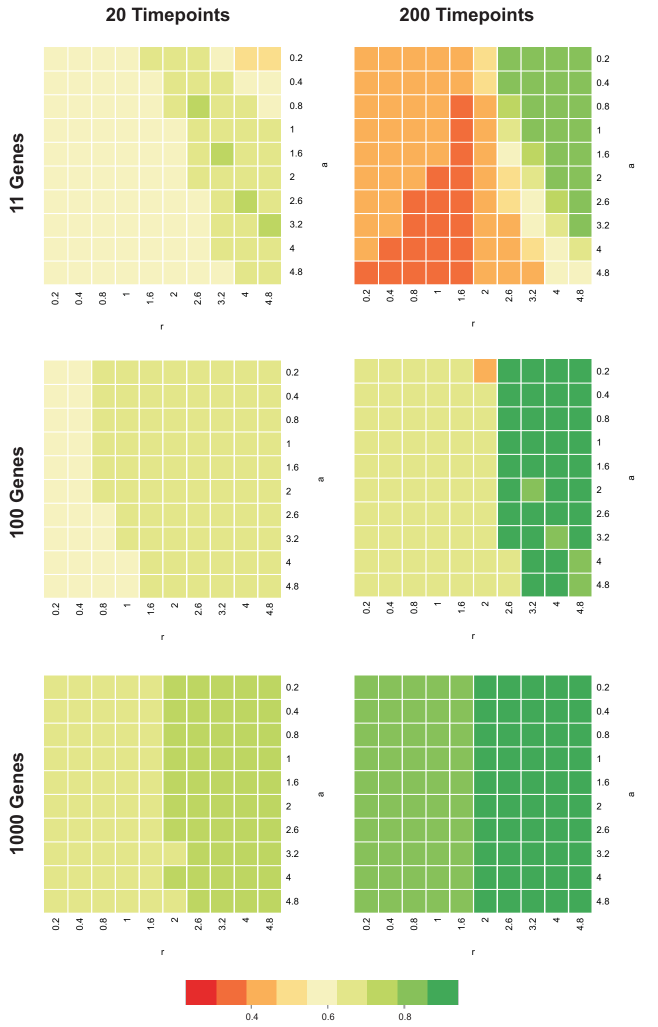


Figure 3.12: Sensitivity Analysis for the Network Inference Performance on Synthetic Data with Respect to Parameters a and r. Plots comparing distributions of AUC values for ROC graphs for different a and r settings (x- and y-axis), for the synthetic networks of sizes 11, 100 and 1000, using data sets with 20 and 200 time points, respectively. The plots show that results are relatively insensitive over a large range of parameters. Smaller values of the hyperparameter r correspond to a more peaked prior distribution, resulting in "sparser" networks. Correspondingly, the figure shows that smaller values of r should be chosen for larger networks. Although the effect of changing a seems not as pronounced, larger values of a correspond to a narrower prior distribution, and should therefore be used if fewer data are available to avoid overfitting.

retrieves the original network from the data, provided sufficient time points are available. Furthermore, the approach can help identify hub genes in regulatory networks, and we have shown an application to a large biological dataset on yeast, where we successfully

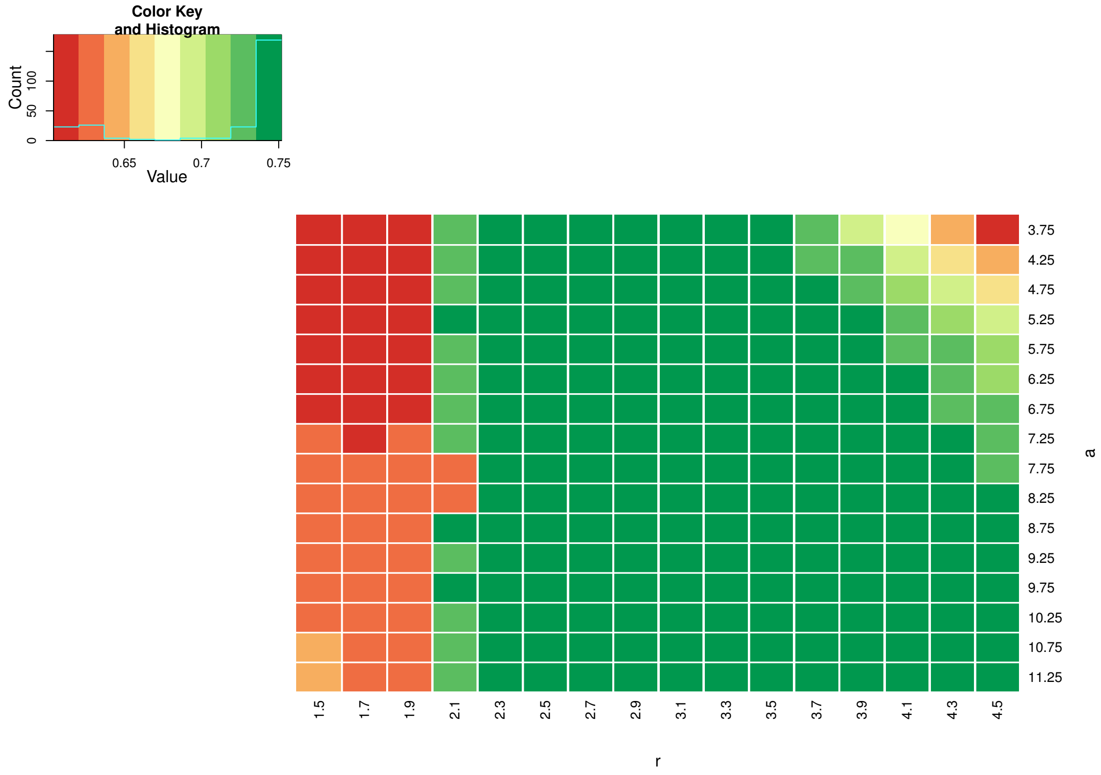


Figure 3.13: Sensitivity Analysis for the Prediction of Hub Genes in the Yeast Cell Cycle with Respect to Parameters a and r. To assess the effect of changes of model parameters a and r, both parameters were varied individually and together by up to ±50 percent. Network reconstruction was restarted for each combination of values for a and r, and average AUC values were computed for the reconstructed networks in comparison to the STRING network. The figure shows the resulting AUC values over a, r, indicating that results are relatively insensitive over a wide range of parameter values.

identified several important hub genes.

While a considerable number of approaches to reconstruct networks from data have been published to date, to our knowledge, this is the first method that simultaneously identifies hubs in the regulatory network and centers the network reconstruction around these hub genes, by using a hierarchical Bayesian prior distribution on the edge weights. While clearly other network reconstruction approaches can also be used to identify hubs *in retrospect*, our approach specifically centers the reconstruction of the network around central hub genes. In particular on large, noisy datasets, this may be a major advantage over other approaches that may only identify clusters of correlated genes, but not necessarily induce a hierarchical structure. This is shown on the yeast cell cycle network, where ARACNE, Banjo and MRNet all failed to correctly identify the highest-degree hubs in the network. We therefore believe our approach to have high potential for the identification of hubs in unknown regulatory networks, around which further experimental effort should be centered in elucidating the respective network. Ultimately, this could be highly useful for an iterative procedure of network reconstruction, experiment design, further biological experiments, and feeding the results back into network reconstruction, of particular interest for large networks. Indeed, if certain hubs in a network are already known, this can even be integrated into the network inference by choosing a different prior over σ for the known hub genes.

We have shown two different approaches to evaluate the posterior distribution over models given the data. On the one hand, we used a Markov chain Monte Carlo approach to sample from the posterior distribution. The advantage of this is that full distributions are evaluated, hinting to possible different, alternative network topologies, yielding additional information on confidence in results. The disadvantage of this method is the computational burden involved, making it infeasible for networks involving more than a few dozen genes, at least without further parallelization of the sampler. On the other hand, we use gradient based optimization to maximize the posterior, yielding a single optimal network topology. This can be computed considerably faster and is feasible for networks with several hundred to thousands of genes, but does not provide any information on alternative, high-probability networks, and no confidence intervals are available on model parameters.

We showed results on simulated data, indicating that even with only moderate noise, for a network of approximately 1000 genes, at least 200 time points are needed for reliable network reconstruction. Hence, while the size of the used yeast data set clearly is not sufficient for a precise reconstruction of the whole network structure, we could identify important hubs in the regulatory network, which were validated using the STRING database. An interesting result from our point of view is that all published approaches that we tried, including our own, failed to reconstruct a yeast transcriptional network from the microarray data, at least in comparison to the gold standard network from the STRING database. This may be due to low quality of the experimental data and the lack of targeted interventions, but these results are in line with findings in recent results of the DREAM competition, where also even the best submitted methods showed surprisingly weak performance, and most methods did not perform better than guessing [135, 117]. Under conditions of high noise and limited amounts of experimental data, for large-scale network reconstruction, a method that centers on hubs may therefore be of value to concentrate further experimental efforts and network reconstruction attempts around these hub genes.

Interestingly, the HUB prediction for the yeast dataset shows good performance with high AUC values if a fairly strict definition of hub genes is enforced, by requiring a hub to have a large number (> 50) of interaction partners in the STRING network. If this threshold is relaxed, AUC values drop rapidly. We offer two explanations for this behavior: On the one hand, genes with low connectivity in the network probably contribute significant noise to the network reconstruction, simply due to their large number. On the other hand, false positives in the STRING dataset will affect genes with few interaction partners more than genes with a large number of partners, since a gene with say 100 interaction partners would still be considered a hub, even if 20 of the interactions are false. It is somewhat surprising that the transition occurs so rapidly around a value of 50 interaction partners, one would expect a more smooth transition where AUC gradually increases with increasing degree. To study this further and exclude the possibility that this is an artifact of the method employed, we additionally performed the same computation on the 500 gene simulated network with noise from GeneNetWeaver, where indeed a smooth increase of the AUC values is observed. We therefore speculate that the rapid transition in the Yeast dataset is not due to the method we used, but rather an artifact of the data set.

A difficulty in using our approach, that all Bayesian methods share, is the need to select parameters for the prior distributions. In some cases, these can significantly influence results, and the choice of parameters a and r in our method is not straightforward. Optimal values depend on the size of the network, the amount of available experimental data, the level of noise in the data, and the expected number of hub genes. Importantly, our sensitivity analysis of the yeast cell cycle network reconstruction with respect to parameters a and r shows that results are relatively insensitive over a wide range of parameter choices. Still, considerable experience is required in tuning these parameters. Methods to assist finding sensible choices, such as empirical Bayes approaches or careful cross-validation, could be used to address these issues.

The main assumption we make in our model is the binarization of state space – each gene is assumed to be either active or inactive. This implies a loss of detailed expression levels, but allows us to tremendously reduce model complexity and computation time, and hence, to explore biological networks of a much larger scale. This discretization of the data may furthermore have advantages in case of microarray data as used in this study, in particular if the data is more of a qualitative than of a quantitative nature due to inherent noise, or if data from different platforms or different studies shall be integrated. Furthermore, in contrast to co-expression based approaches, the underlying Boolean model allows causative inferences, hence edges between genes are directed and can be interpreted not only as correlation or co-expression, but causality.

A difficulty associated with the use of a Boolean model is the requirement to discretize the experimental data. We have used smoothing cubic splines in this work to smooth out smaller fluctuations in the experimental data, and thus take care of some of the noise in the data. Data were then discretized for each gene separately by using the median of the respective gene as threshold. For the small 11 gene network, we have manually checked the resulting data, and the discretized values were compared with the raw data to assure that the interpolation and discretization has produced reasonable results. However, this is clearly not feasible for large-scale network inference with hundreds to thousands of genes, and discretization can then become a difficult issue, in particular since it will clearly have a considerable effect on results of the network inference. Already using the mean instead of the median as discretization threshold can lead to a completely different data set, if the time course for a particular gene has a single large outlier.

The spline interpolation itself requires the choice of a smoothing factor, and clearly, also other interpolation functions could be employed (for example linear, polynomial, etc.). Kaderali *et al.* have previously proposed an iterative procedure between spline interpolation and network inference for a model using ordinary differential equations [100]. In this work, model predictions are fed back into the interpolation, to adaptively choose parameters for the interpolation. It is not immediately evident how such a procedure can be used with a Boolean model, but this might be an interesting question for future work.

Overall, our results show that the approach presented may be a valuable tool for large-scale network reconstruction, and may guide experimental efforts to characterize identified hubs in more detail. The Boolean discretization used in principle allows the reconstruction of larger networks and may in fact be an advantage in case of noisy data, but our results also clearly indicate that an accurate reconstruction of a large network is not feasible with present limited data sets containing at most a few dozen time points or different conditions. In addition to larger experimental data sets, a key to overcome these challenges will be the integration of as much biological knowledge as is available. Our method contributes to this aim by providing a general framework for reconstructing sparse networks with small world properties.

CHAPTER 4

Analysis of Regulatory Relationships in Real-World Domains

In the previous chapter two new methods were presented for the analysis of gene expression time series data. In this chapter the focus lies on the analysis of a completely new gene expression data set, which is also a time series. This time series describes the behavior of human cell lines for two experimental setups and their controls. The evaluation of any predictions in human is difficult, since no gold standard exists to which the results could be compared. Furthermore, due to the costs of measuring experiments, the data set is a short time series, which consists only of four time points. Further limitations to the analysis were experimental noise and time shifts between the time points. Nevertheless, this is not an untypical task in this field since most experiments have only limited funding and have to focus on the time points which are supposed to be the most interesting ones. It is a challenging task to deal with these preliminaries. Further obstacle are batch effects from the experiments, choosing the proper normalization method, finding a suitable ranking of genes of interest and to identify genes of interest for follow-up studies. In the following sections an approach is presented, which is capable of dealing with all of these topics. A general workflow from the raw analysis to the identification of genes of interest will be outlined. The application of this workflow resulted in the identification of the TF ELF3 as a new regulatory factor of human urothelium [17]. This analysis has been done in close collaboration with the Jack Birch Unit for Molecular Carcinogenesis from the University of York, where all the lab work was done.

4.1 Identification of ELF3 as an Early Transcriptional Regulator of Human Urothelium

Despite major advances in high-throughput and computational modelling techniques, understanding of the mechanisms regulating tissue specification and differentiation in higher eukaryotes, particularly man, remains limited. Microarray technology has been explored exhaustively in recent years and several standard approaches have been established to analyze the resultant datasets on a genome-wide scale. Gene expression time series offer a valuable opportunity to define temporal hierarchies and gain insight into the regulatory relationships of biological processes. However, unless datasets are exactly synchronous, time points cannot be compared directly. Here we present a data-driven analysis of regulatory elements from a microarray time series that tracked the differentiation of non-immortalised normal human urothelial (NHU) cells grown in culture. The datasets were obtained by harvesting differentiating and control cultures from finite bladder- and ureter-derived NHU cell lines at different time points using two previously validated, independent differentiation-inducing protocols. Due to the asynchronous nature of the data, a novel ranking analysis approach was adopted whereby we compared changes in the amplitude of experiment and control time series to identify common regulatory elements. Our approach offers a simple, fast and effective ranking method for genes that can be applied to other time series. The analysis identified ELF3 as a candidate transcriptional regulator involved in human urothelial cytodifferentiation. Differentiation-associated expression of ELF3 was confirmed in cell culture experiments and by immunohistochemical demonstration in situ. The importance of ELF3 in urothelial differentiation was verified by knockdown in NHU cells, which led to reduced expression of FOXA1 and GRHL3 transcription factors in response to PPARγ activation. The consequences of this were seen in the repressed expression of late/terminal differentiation-associated uroplakin 3a gene expression and in the compromised development and regeneration of urothelial barrier function.

4.1.1 Introduction

The bladder and associated lower urinary tract is lined by urothelium, a transitional epithelium that functions as a permeability barrier to limit exposure to urinary toxins and to minimise alterations in urine and blood composition ([43], reviewed [88]). The maintenance of this vital urinary barrier is supported by an exceptional regenerative capacity, whereby the urothelium switches from a mitotically-quiescent to a highly proliferative

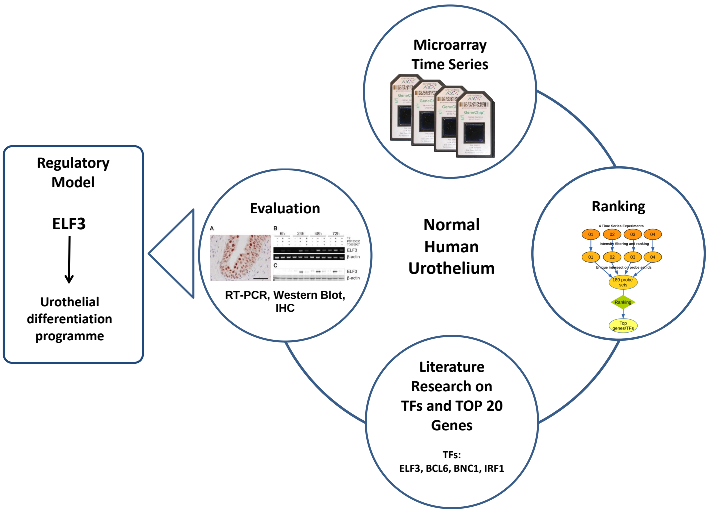


Figure 4.1: Graphical abstract of the presented workflow.

state in response to damage [87].

Urothelium shows an increase in morphological complexity between basal, intermediate and superficial cell zones. The lumen-facing superficial cells are uniquely specialised to provide urinary barrier function. With well-developed tight junctions limiting paracellular permeability [146], the major transurothelial barrier is provided by thickened plaques of asymmetric unit membrane (AUM) decorating the apical membrane of the superficial cells [68]. The AUM is constituted in the Golgi as a result of precise uroplakin protein interactions (reviewed [153]), the disruption of which has devastating consequences for urothelial permeability and urinary tract development. Thus, the targeted disruption of the uroplakin UPK3a gene in mice resulted in distinctive structural and functional abnormalities of the urothelium and high grade vesicoureteric reflux [69]. Although a few examples of minor uroplakin gene anomalies associated with urinary tract malformations have since been found in man [72], there is no common association [54, 56, 75, 82], indicating that major disruption of urothelial differentiation during human development is likely non-viable.

Although morphological and molecular features of urothelial differentiation are well characterised, relatively little is known of the transcriptional mechanisms underpinning this process. In mice, Kruppel-like factor 5 (KLF5) has been shown to be involved in the embryological development and differentiation of bladder urothelium [11]. In the same study, KLF5-null fetal urothelium was shown to be deficient for expression of PPARγ, GRHL3, ELF3 and OVOL1, supporting the participation of these factors in a hierarchical transcriptional network regulating urothelial development. The authors further showed in transient transfection assays that KLF5 regulated expression of the mGRHL3 promoter. In mouse development, GRHL3 plays an essential role in epidermal morphogenesis, with GRHL3-deficient mice exhibiting failed skin barrier formation, defective wound repair and loss of eyelid fusion [20]. GRLH3 has also been shown to be critical to urothelial differentiation during mouse development [156]. Although it is assumed that these factors and relationships are conserved from mouse to man, there is a lack of experimental approach to enable these developmental relationships to be assessed in human cells.

We have developed a robust experimental system for the propagation and differentiation of normal human urothelial (NHU) cells in vitro. When isolated from the tissue and cultured in a serum-free low calcium medium as finite cell lines, NHU cells subsume a basal squamous (CK14+) phenotype, are highly proliferative and do not show spontaneous differentiation even at confluence [131]. Nevertheless, cultured NHU cells retain the capacity to differentiate to form a functional barrier urothelium, as shown by subculture in medium containing bovine serum and physiological [Ca^{2+}], where transepithelial electrical resistances of $>$3,000 Ω.cm^2 are routinely attained [31]. In addition, pharmacological activation of the nuclear receptor peroxisome proliferator activated receptor gamma (PPARγ) initiates the urothelial differentiation gene expression programme in individual cells, but without self-organisation into a barrier urothelium [148, 146, 149, 147]. Thus, the outcomes of these two differentiation-inducing protocols are not identical, yet both result in development of a differentiated urothelial cell phenotype. We reasoned that comparison of the two protocols to identify common changes in gene expression over time would help limit method-dependent artefacts and hence was a strategy that could help identify key regulatory genes involved in determining urothelial differentiation. To these ends, we performed a gene array series at different time-points following differentiation induction using the two differentiation-inducing protocols performed in parallel. Our aim was to perform an unbiased analysis to identify common regulatory features and in the following, we describe the quality assessment of the data, the overcoming of synchronization issues by performing a qualitative ranking approach, the identification of a set of significant genes involved in transcriptional regulation and the experimental validation of a previously unidentified regulator of human urothelial cytodifferentiation.

4.1.2 Materials and Methods

Troglitazone (TZ) was obtained from Sigma-Aldrich (Dorset, UK) and the EGF receptor tyrosine kinase inhibitor PD153035 was obtained from Merck Millipore (Darmstadt, Germany). The PPARγ-specific antagonist, T0070907 was obtained from Cambridge Bioscience (Cambridge, UK). Rabbit anti-ELF3 antibody (ab97310) was obtained from Abcam, Cambridge, UK. Mouse anti-β-actin (clone AC-15) was obtained from Sigma-Aldrich (Dorset, UK).

Tissue Samples

Human urothelial tissue samples were sourced ethically with informed written consent from patients and approval for use in research from Leeds (East) and York Research Ethics Committees. The surgical specimens were collected from patients with no history of urothelial cancer and were processed for histology or used to establish urothelial cell cultures. Samples taken for (immuno)histology were fixed for 16h in 10% (v/v) formalin, dehydrated and processed into paraffin wax.

Cell Culture

Finite NHU cell lines were established as detailed elsewhere [132]. For routine propagation, cultures were maintained as monolayers in low calcium [0.09 mM] Keratinocyte Serum Free Medium containing bovine pituitary extract and EGF (Invitrogen) and further supplemented with cholera toxin (KSFMc). Cultures were sub-cultured by trypsinisation at just-confluence and used for experiments between passages 3 to 5.

To induce differentiation, two previously described methods were used. In the first (referred to as TZ/PD): NHU cell cultures were treated with 1μM TZ with concurrent 1μM PD153035 to block EGFR activation and induce individual cell differentiation [149]. In the second protocol (referred to as ABS/Ca^{2+}): cultures were pre-treated with 5% adult bovine serum (ABS, Harlan Sera-Lab) for 3 days before subculture (time point T=0 h) into KSFMc supplemented with 5% ABS and 2 mM $CaCl_2$, leading to generation of a differentiated, tight barrier epithelium as described [31]. Vehicle control non-differentiated cultures were maintained in parallel in KSFMc and used at the same time points (between 24 to 144 hours). Cultures were lysed in situ with TRIzol® to prepare RNA by the manufacturer's recommended protocol (Invitrogen). RNA samples were treated with a DNA-free kit (Ambion) and quantified by UV spectrophotometry.

Microarray Experiments and Preprocessing

Time series experiments were performed on two independent donor NHU cell lines (Y579 and Y676) using the TZ/PD and ABS/Ca^{2+} differentiation-inducing protocols described above. For each arm of the experiment, parallel non-differentiated control cultures were included and RNA was extracted at 6, 24, 72 and 144 hours, where t=0 coincided with the treatment to induce differentiation in the TZ/PD cultures. The nature of the differentiation process meant that (as described above and Cross *et al.* [31]) the ABS/Ca^{2+} arm had a serum pre-treatment stage, which affected absolute synchronization of the two arms of the experiment. Following RNA extraction, induction of differentiation in the experimental arms was verified by assessing the expression of UPK2 transcript by quantitative real time PCR (not shown).

For the arrays, mRNA was converted to cDNA and then to biotin-labelled cRNA before hybridising to HG-U133 Plus 2.0 arrays (Affymetrix). The array chips were washed and scanned at 560nm using an Affymetrix GeneChip Scanner. Quality assessment of the microarrays was performed with the *arrayQualityMetrics* package [81] and two samples were discarded due to quality issues (one control sample: 72h ABS/Ca^{2+} from cell line Y676; and one experimental sample: 6h TZ/PD from cell line Y676). The experiment thus yielded 30 arrays (2 cell lines x (2 experimental and 2 control arms) x 4 time points) - 2 discarded arrays).

All calculations were performed in R. The microarray data were RMA normalized [71] using the Bioconductor package *affy* [55]. To reduce the dimensionality of the data and to filter out insignificant signals, an intensity filter was applied, which retained probe sets for which at least 25 per cent of all time points for the experimental time series had an intensity above $log_2(100)$. 25,461 probe sets remained after the filtering step, corresponding to about 13,500 genes. Since the time series began at 6h, the first control time point was used as a zero time point to account for changes of the expression level between the starting point of the control cell lines and the experiments. This left three time series with five time points and one with four because of the discarded sample (the 6h time point in the TZ/PD Y676 experiment). The quality analysis and a principal component analysis of the different samples are provided in the Appendix A. An interactive visualisation of the results can be accessed from: `http://www.informatik.uni-mainz.de/groups/information-systems/research/timeseries-visualisation`

All calculations were performed in R. Microarrays were RMA normalized [71] using the Bioconductor package *affy* [55]. To reduce the dimensionality of the data and to filter out insignificant signals, an intensity filter was applied, which retained probe sets for which at least 25 per cent of all time points for the experimental time series had an intensity above $log_2(100)$.

Filtering and Ranking

We developed an approach that focused on the expression changes over time on a qualitative level to overcome the lack of precise synchronization of events. We searched only for the maximum changing event per gene and experiment and globally compared changes over time from control to experiment, thus avoiding direct comparison of time points.

We calculate for each probe set time series a $\mathit{\Delta X}$, the difference between the maximum intensity of all time points $x_i \in X$ to the minimum intensity of all time points:

$$\mathit{\Delta X} = max(x_i \in X) - min(x_i \in X) \tag{4.1}$$

This difference $\mathit{\Delta X}$, describes the maximum change over time per probe set for each time series. In a second step, we calculate the difference of the $\mathit{\Delta X}_{exp}$ for the experiment and the $\mathit{\Delta X}_{ctrl}$ for the control time series. We refer to this as amplitude log fold change (ALFC):

$$ALFC = \mathit{\Delta X}_{exp} - \mathit{\Delta X}_{ctrl} \tag{4.2}$$

This value defines if a probe set changes significantly over time compared to its control without having to compare the individual time points. A positive ALFC means the change is larger for the experiment and a negative value vice versa, indicating a silenced gene. Probe sets were then ordered in decreasing order from their ALFC value.

Validation Experiments

In validation experiments, NHU cell cultures (independent cell lines from arrays) were induced to differentiate by co-treatment with 1μM troglitazone (TZ as PPARγ agonist) and 1μM PD153035 (EGFR inhibitor) as described above in [149]. To further define the role of PPARγ, parallel cultures were pretreated for 3 hours with 5μM T0070907 as a specific PPARγ antagonist prior to induction of differentiation. Replicate cultures were harvested at 6, 24, 48 and 72 hours and used to extract DNA-free RNA for analysis of transcript expression (see below). Parallel cultures were lysed and processed for immunoblotting and probed with rabbit anti-ELF3 antibody, which was also used to assess ELF3 localisation in paraffin wax-embedded tissue sections of human urothelium by immunoperoxidase histochemistry (see below). Appropriate vehicle (DMSO), loading (β-actin for RT-PCR and immunoblotting) and specificity (RT-negative and irrelevant antibody) controls were included in all experiments.

Reverse-transcribed (RT) and real-time quantitative (RTq) PCR

Cell cultures were solubilised in TRIzol® and the RNA was isolated by chloroform extraction and iso-propanol precipitation, according to the manufacturer's protocol (Life Technologies, Paisley, UK). The RNA was treated with DNase I (DNA-freeTM kit from Ambion, Huntingdon, UK). cDNA was synthesised from 1μg of total RNA using the Superscript first-strand synthesis system (Life Technologies, Paisley, UK). RT-PCR was performed using Go Taq® Hotstart Polymerase (Promega, Southampton, UK) with primer sets designed to amplify specific human products (Table 4.1). RT negative and no template (water) controls were always included. For semi-quantitative analysis, template cDNA was mixed with SYBR® Green PCR Master Mix (Applied Biosystems) and 300nM of each forward and reverse target gene primers (Table 4.1) and analyzed on an ABI StepOnePlus™ Real Time PCR System. The thermal profile was: 20 sec hold at 95°C, followed by 40 cycles of denaturation at 95°C (3 sec) and elongation at 60°C (30 sec). Dissociation curves were performed to confirm the presence of a single amplification product and the absence of primer dimers for each primer set. Assay efficiency, validated using the CT slope method prior to use, confirmed that both the test and endogenous assays were of equivalent efficiency (within tolerance range). SYBR® Green results were expressed as relative quantification (RQ) values (Applied Biosystems).

Immunohistochemistry

De-waxed 5μm tissue sections were blocked for endogenous peroxidase activity with 3% (v/v) hydrogen peroxide for 10 minutes. Antigen retrieval was performed by microwave boiling of tissue sections in a 10mM citric acid buffer (pH 6.0) for 10 minutes, followed by 10 minutes cooling on ice. Tissue sections were treated with an Avidin/Biotin blocking kit (Vector labs, Peterborough, UK), before applying 10% goat serum for 5 minutes to prevent non-specific binding of the secondary antibody. Rabbit anti-ELF3 antibody (2μg/ml) was applied, followed by biotinylated goat anti-rabbit secondary antibody (1/800, Dako Cytomation Ltd, Ely, UK) – each for 15 minutes at ambient temperature, with washing in between. Antibody binding was detected using a tyramide-based signal amplification system (CSA system, Dako Cytomation Ltd, Ely, UK), as described in the manufacturer's protocol. Tissues were lightly counterstained in Mayer's haematoxylin,

Table 4.1: Primer sequences used for RTPCR and RTqPCR

PCR Product	Forward Primer (5’-3’)
(gene name)	Reverse Primer (5’-3’)
ELF3 (RTPCR)	GTTCATCCGGGACATCCTC
	GCTCAGCTTCTCGTAGGTC
ELF3 (RTqPCR)	TCAACGAGGGCCTCATGAA
	TCGGAGCGCAGGAACTTG
GRHL3 (RTqPCR)	TGGAATATGAGACGGACCTCACT
	CAGACACGTTCTCTGTCAGGAATT
FOXA1 (RTqPCR)	CAAGAGTTGCTTGACCGAAAGTT
	TGTTCCCAGGGCCATCTGT
UPK3a (RTPCR)	CGGAGGCATGATCGTCATC
	CAGCAAAACCCACAAGTAGAAAGA
UPK2 (RTqPCR)	CAGTGCCTCACCTTCCAACA
	TGGTAAAATGGGAGGAAAGTCAA
CLDN7(RTqPCR)	GCAGTGGCAGATGAGCTCCTAT
	CATCCACAGCCCCTTGTACA
β-actin (RTPCR)	ATCATGTTTGAGACCTTCAA
	CATCTCTTGCTCGAAGTC
GAPDH (RTqPCR)	CAAGGTCATCCATGACAACTTTG
	GGGCCATCCACAGTCTTCTG

This list gives an overview of the primer sequences for the different marker genes.

dehydrated through ethanol to xylene and mounted in DPX (Sigma Aldrich).

Immunoblotting

Culture lysates were resolved on 4-12% gradient bis-Tris acrylamide NuPAGE® gels (Life Technologies, Paisley, UK) and electrotransferred onto 0.45-μm PVDF-FL membranes (Merck Millipore, Darmstadt, Germany). Membranes were probed with anti-ELF3 (1gμ/ml) and β-actin (1/250,000) antibodies for 16 hours at 4°C; bound antibody was detected with goat anti-rabbit Ig conjugated to IRDye® 800 (50ng/ml; Rockland Immunochemicals; supplied by Tebu-bio, Peterborough, UK) or anti-mouse immunoglobulins conjugated to Alexa Fluor® 680 (200ng/ml; Life Technologies), as appropriate. Immunolabelled protein bands were visualized and relative quantifications generated using an Odyssey infrared imaging system (LiCor, Cambridge, UK).

Knock-down of ELF3 by retroviral-mediated shRNA interference

For RNA interference experiments, siRNA oligos were designed to target the ELF3 coding sequence before adding the hairpin loop, restriction overhangs for directional cloning and an Mlu1 restriction site (to verify cloned inserts) in to generate the following ELF3 sense shRNA sequences including hairpin loop:

shRNA1:	GATCCGCTACCAAGTGGAGAAGAACATTCAAGAGA
	TGTTCTTCTCCACTTGGTAGCTTTTTACGCGTG
shRNA2:	GATCCGCTCTTCTGATGAGCTCAGTTTTCAAGAGA
	AACTGAGCTCATCAGAAGAGCTTTTTACGCGTG
shRNA3:	GATCCGCTCAGTTGGATCATTGAGCTTTCAAGAGA
	AGCTCAATGATCCAACTGAGCTTTTTACGCGTG

A scrambled control shRNA was also prepared. Oligonucleotides were annealed and cloned into the RNAi-Ready pSIREN-RetroQ retroviral expression vector (Clontech) and transfected into the PT67 packaging cell line using the manufacturer's protocols. Following antibiotic selection, conditioned virion-containing medium was harvested from confluent PT67 cultures and filtered through a 0.45μm low protein binding Tuffryn® filter to remove cell debris. NHU cells were transduced for 6hr with 8ml of conditioned medium supplemented with 8μg/ml Polybrene (Hexadimethrine Bromide, Sigma), after which virus-containing medium was removed and cultures replenished with KSFMc. Transduced NHU cell cultures were selected with 1μg/ml puromycin and screened for ELF3 protein expression by immunoblotting following induction of differentiation.

Trans-epithelial electrical resistance (TER) studies

Differentiating urothelial cell cultures were established on Snapwell™ membranes using the ABS/Ca^{2+} differentiation-inducing protocol and with medium changed on alternative days. TER readings were taken daily using a portable EVOM™ Epithelial Voltohmmeter (World Precision Instruments), as described [122]. After stabilisation of the TER readings, cultures were scratched to create a wound of 250μm wide. Further TER measurements were taken at regular intervals over a period of 61h during wound closure.

4.1.3 Results

Candidate Gene Selection

The time series microarray data was pre-processed and filtered as described in the Methods section. The first intensity filtering left 25,461 probe sets (about 13,500 genes). The initial goal was to prioritize and group genes with respect to their possible importance in regulating the differentiation or proliferation process. The main difficulty for the analysis of the time series data arose due to inherent a) variability between the two biological replicates (independent human donor cell lines) studied and b) differences in the pre-treatment and hence precise timing of the two procedures (ABS/Ca^{2+} and TZ/PD) used to induce differentiation. Together, these confounded synchronization of the experimental arms and impaired direct comparison of control and experimental time points. To overcome these issues, we developed an approach that focused on the expression changes over time on a qualitative level, which we refer to as the ALFC. Thus, instead of performing pairwise comparisons between each time point of control and experiment, we looked for the maximum expression burst in the time series for each gene and compared the global expression change over time between the control and experiment. By discretizing the global differences between experimental and control arms, we effectively circumvented any problems related to shifts in timing between or within experiments. The result was that we were able to avoid direct comparison of the time points, whilst still considering the time information. An overview of the resulting sets and their intersections is shown in Figure 4.2.

The calculated ALFC was then used for further filtering of the data sets. We used a set of 25 pre-defined marker genes selected as implicated in urothelial differentiation or proliferation (Table 4.2) in order to select the size of sets for further investigation. Figure 4.3 shows the stair-step plots with the number of markers included by applying different thresholds for each of the four experimental arms (see Table 4.3 for an overview of the included markers). Based on this analysis, we selected the top 1000 ranked probe sets associated with each experimental arm. We proceeded to analyze the different intersections between the four selected top 1000 probe sets (see Figure 4.2A and B for an overview). By definition, the overlap between all four experimental arms (two biological replicates and two differentiation-inducing procedures) identified common (method-independent) genes, whereas the overlap exclusive to either of the two differentiation-inducing procedures (ABS/Ca^{2+} or TZ/PD) revealed method-dependent factors.

Identification of common regulatory genes implicated in urothelial differentiation.

For the intersection of all four sets, the ALFC was normalized to one for each experimental arm and the absolute sum of the ALFC was calculated and used to prioritize the list of probe sets. The overlap of all four experiments resulted in 189 probe sets and

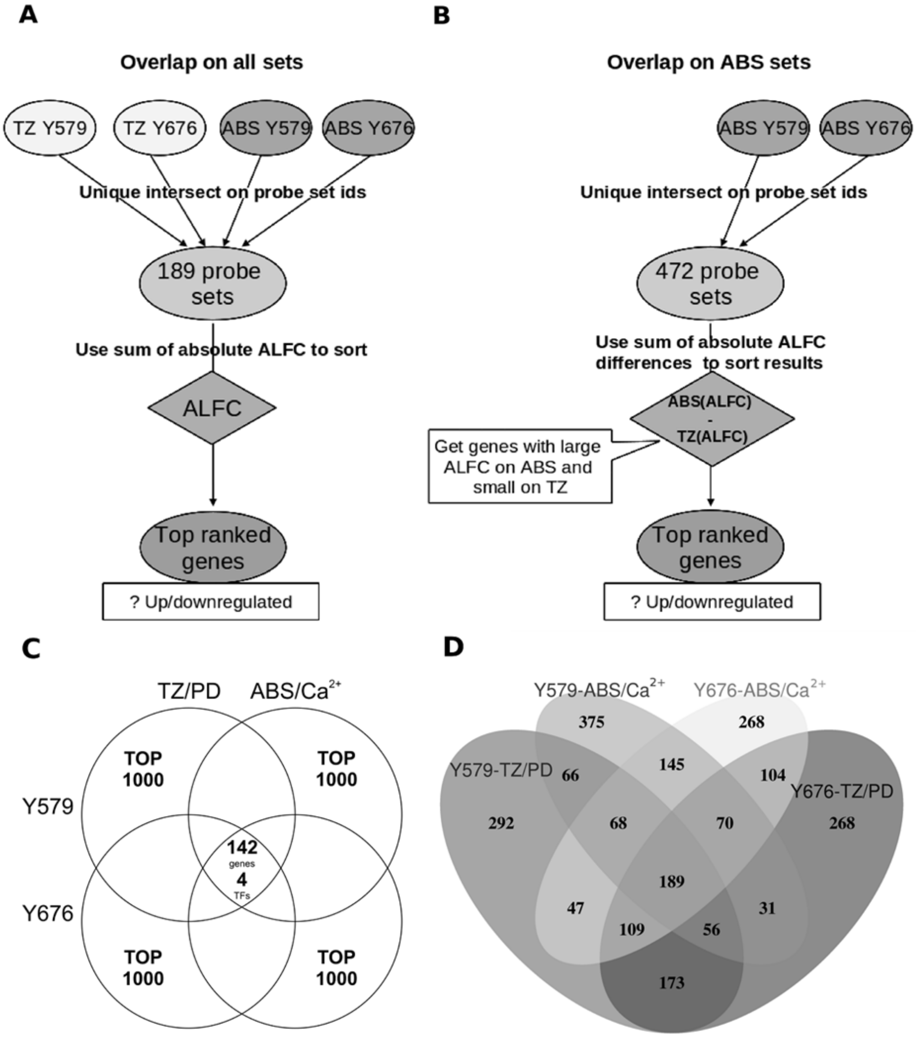


Figure 4.2: Overview of the probe set overlaps between the experiments. Figures A and B give an overview of how the top ranked probe sets were combined. The intersection was built either on all four filtered 1000 top ranked sets or the two ABS/Ca^{2+} experimental arms. The ALFC for each probe set was then normalized to one and the set sorted according to the sum of absolute ALFC over all experiments. The intersection of the ABS/Ca^{2+} sets was ranked from the difference between the absolute sum of ALFC on the ABS/Ca^{2+} arm and the absolute sum of the ALFC on the TZ/PD arm. Figures C and D display Venn diagrams which show the intersection and overlaps of the four differentiation series (TZ/PD and ABS/Ca^{2+}, on two independent donor cell lines Y579 and Y676). The overlaps were generated using the top 1000 probe set lists from each experiment. The Venn diagram in panel C shows the number of genes present at the intersection of all four data sets. 142 up-/downregulated genes were found within the intersection. Within this subset, four TFs were found (ELF3, BCL6, BNC1, IRF1). Figure D shows the same overlaps on the probe set level. The 189 probe sets in the intersection of all sets map to the 142 genes from Figure C. The overlap between the ABS/Ca^{2+} experimental arms contains 472 probe sets with 13 identified as TFs (see Table 4.2 for details). From these 13 TFs, nine factors (GRHL1, GRHL3, FOXC1, ID2, SMAD3, FOXN2, ETS1, MITF and FOXD1) were unique to the ABS/Ca^{2+} arm and not included in the TZ/PD lists.

Table 4.2: List of markers for probe set filtering

Expected behavior versus control	Marker genes
Expressed, but not changing	KRT7, KRT19, GAPDH, ACTB, SMAD2
Downregulated	KRT5, KRT6A, KRT6B, SMURF2, SMAD7, MYBL2, BUB1, PLK1, CREB1
Upregulated early	FOXA1, GATA3, IRF1, IRF2
Upregulated late	KRT13, CLDN4, UPK2, UPK3A, UPK1A, UPK1B, UPK3B

This list gives an overview of the proliferation and differentiation-associated transcribed gene markers used to define the filtering thresholds.

Table 4.3: Overview of top ranked probe sets and their associated marker genes

Experiment Cell Line	Probe Sets (Genes)	pos. +	neg. -	Marker genes
TZ/PD Y579	1000 (707)	955	45	UPK1A, UPK1B, UPK2, UPK3A, IRF1, FOXA1, CLDN4, KRT5/6A/6B, SMAD7
TZ/PD Y676	1000 (695)	997	3	UPK1A, UPK1B, UPK2, UPK3A, UPK3B, ACTB, KRT5/6A/6B/13, IRF1, FOXA1, GATA3
ABS/Ca^{2+} Y579	1000 (731)	953	47	UPK1B, KRT5/6A/6B/13, IRF1, CLDN4
ABS/Ca^{2+} Y676	1000 (755)	996	4	UPK1A, UPK1B, UPK3A, UPK3B, IRF1, SMAD7, SMURF2, KRT5/6A/6B/13, CLDN4, FOXA1, GATA3

This Table gives an overview of the four experiments with the top 1000 ranked probe sets (number of genes represented) according to ALFC and showing the associated marker genes (from Table 4.2). Genes were assigned as positive (+) or negative (-) depending on the ALFC value. A positive ALFC value indicates that over the time series, the gene expression of the probe set for the experimental arm changed more than for the control arm.

the overlap of the ABS/Ca^{2+} arms resulted in 427 probe sets (see Figure 4.2C and D). The 20 top ranked genes on these lists are shown in Table 4. We next used the dragon TcoF-DB [127] to find transcription factors (TFs) of interest within the two sets. TcoF-DB currently consists of 1365 TFs which were manually curated [145] or are contained in TRANSFAC, where they have passed a manual curation step.

Four TFs (ELF3, BNC1, BCL6 and IRF1) were found in the overlap of all four ex-

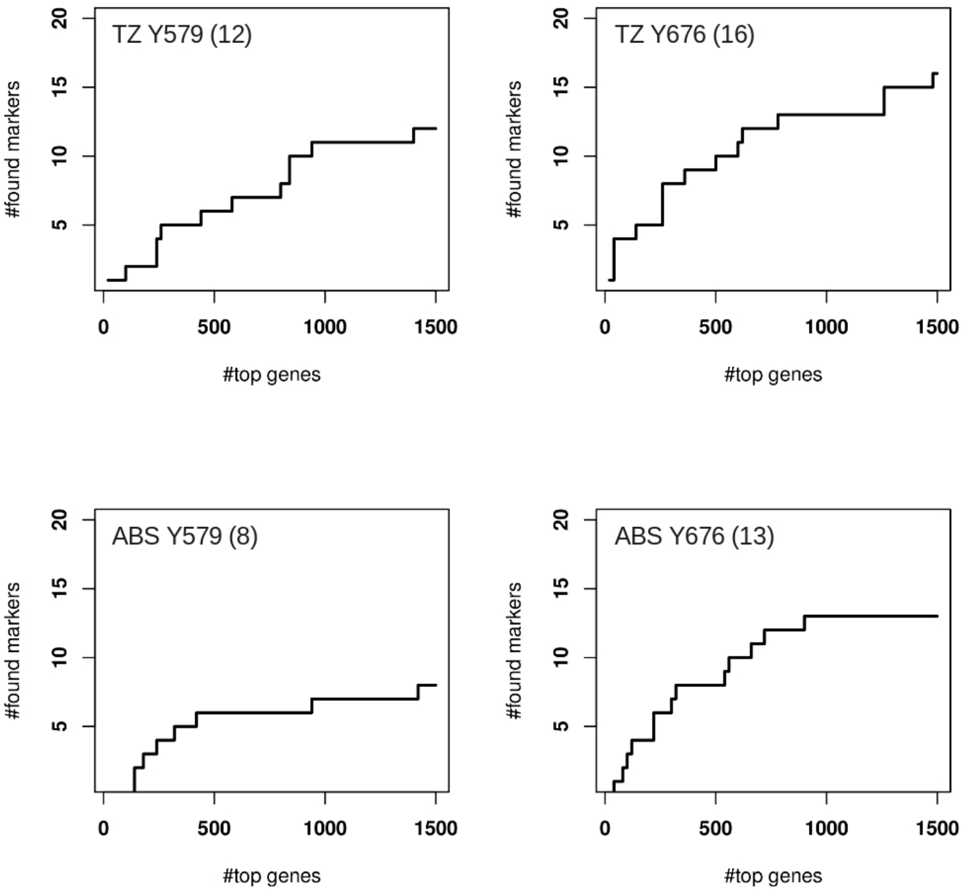


Figure 4.3: Defining the ALFC threshold. Definition of a suitable ALFC threshold: Stair-step plots for the four experiments showing the number of rediscovered markers in the data set for differently sized sets of top ranked probe sets (up to 1500). The ranking is done using the calculated ALFC and each bar represents the number of found markers for the particular amount of chosen genes. A plateau around 1000 is reached in all four data sets and hence a threshold of choosing the top 1000 probe sets is applied for the second filtering step. The maximum amount of found markers is shown in brackets next to the experiment identifier in the left corner of each chart.

periments and 13 TFs were associated with the ABS/Ca^{2+} overlap, of which nine TFs (GRHL1, GRHL3, FOXC1, ID2, SMAD3, FOXN2, ETS1, MITF and FOXD1) remained after the TZ/PD arm was filtered out. It should be noted that the gene list for the ABS/Ca^{2+} overlap does not necessarily exclude a gene from being also relevant to the TZ/PD model: the list contains similar genes but the specific ranking is based only on the ABS/Ca^{2+} arm.

Table 4.4: Overview of TFs and 20 top ranked genes within the overlaps of the experiments

Overlap of all sets	
TFs	Top 20
ELF3, BNC1, BCL6, IRF1	RARRES1, LIMCH1, TRIM31, UBD, SPINK1, PDE10A, TMPRSS2, CP, PIGR, C10orf116, HPGD, CX3CL1, GKN1, IGFBP3, ELF3, RHOU, SYTL5, TFF1, MUC20, RARRES3

Overlap of ABS/Ca^{2+} experiments	
TFs	Top 20
GRHL1, GRHL3, FOXC1, ID2, BNC1, SMAD3, ELF3, BCL6, FOXN2, ETS1, IRF1, MITF, FOXD1	MUC4, CLIC5, CXCL6, BCL2A1, DUOX2, SERPINB9 SERPINA3, PIGR, MAP3K8 IDO1, IL1RL1*, ZG16B IL8, PDZRN3, RARRES3 MMP7, GABRP*, PPBP* RARRES1

This table shows the lists of top ranked genes in the overlaps/intersections of all four experimental arms and for the two ABS/Ca^{2+} specific arms (see Figure 4.2C and D). The genes are ordered according to their ALFC. We observe in general for the TZ/PD model the largest gene expression burst between 24h and 72h. For the ABS/Ca^{2+} model, the burst occurs earlier, between 0h and 6h (but note that this protocol involves a priming pre-treatment). *The majority of the genes are upregulated, except the genes IL1RL1, GAPRP and PPBP from the overlap of the ABS/Ca^{2+} arms.

ELF3 was the top ranked TF (overlap of all sets) with a larger ALFC value than most other genes in all four differentiated datasets (see Figure 4.4). The 95% quantile of all probe sets (in the overlap of all sets) lay at an ALFC value of 0.6, while the three probe sets for ELF3 represented on the HG-U133 Plus 2.0 chip (210827_s_at, 229842_at, 201510_at) reached values of 1.57, 2.06 and 1.72, respectively. A two-tailed p-value test for all three probe sets was significantly smaller than 0.05 (between 2.6e-10 and 5.3e-18). The main expression burst for ELF3 occurred early, between 0h to 24h, and increased in case of the TZ/PD arm until 72h. As ELF3 has not been previously implicated in human urothelial cytodifferentiation, we validated its expression in human urothelium in situ and during differentiation of normal human urothelial cells in vitro.

In addition, we examined how the expression of ELF3 was regulated in relationship to PPARγ activation.

Validation of ELF3 expression by human urothelium.

By immunohistochemistry, ELF3 localised specifically to the urothelium in sections of human ureter. The localization pattern was exclusively nuclear and the intensity of expression increased from basal to the most differentiated superficial cells (Figure 4.5A).

In cultures of NHU cells induced to differentiate by activation of PPARγ (through co-treatment with TZ and PD153035), ELF3 transcript expression was induced within 24 hours and remained high at 72 hours. Inhibition of PPARγ activation by pretreatment of cells with the specific PPARγ antagonist, T0070907, resulted in inhibition of ELF3 induction (Figure 4.5B).

The ELF3 transcript results were confirmed by immunoblotting for ELF3 protein, which revealed upregulation of ELF3 protein at 24 hours. Inhibition by pretreatment with T0070907 confirmed that this was a PPARγ-mediated process (Figure 4.5C).

Immunoblotting revealed that shRNAs designed to three regions of the ELF3 protein coding sequence were successful at inhibiting ELF3 protein expression following induction of differentiation by PPARγ activation (Appendix A.4). Sequence shRNA1 showed

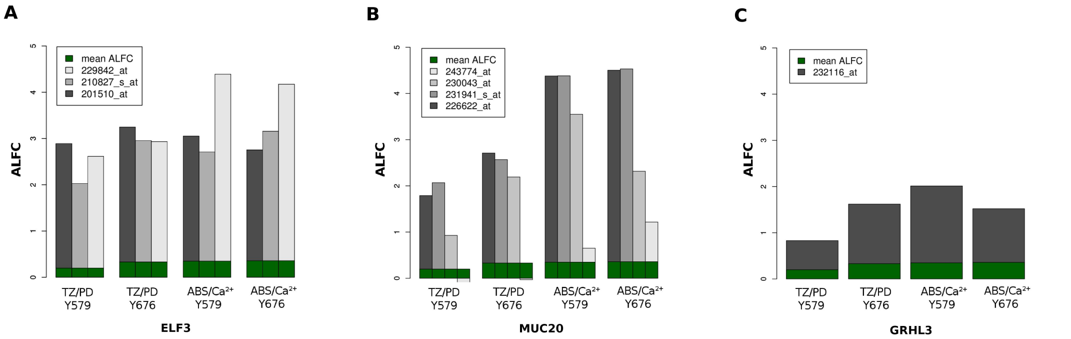


Figure 4.4: ALFC values for ELF3, MUC20 and GRHL3. ALFC values for the probe sets of three genes of interest, ELF3, MUC20 and GRHL3. The mean ALFC of all probe sets is shown by the green bars. For all three genes the ALFC is significantly larger than for the average probe set. The only exception is the 243774 at probe set for MUC20 in case of the TZ/PD model. In this case, the ALFC is slightly negative. According to GeneAnnot, this probe set is the least sensitive of the four probe sets, which can be mapped to MUC20.

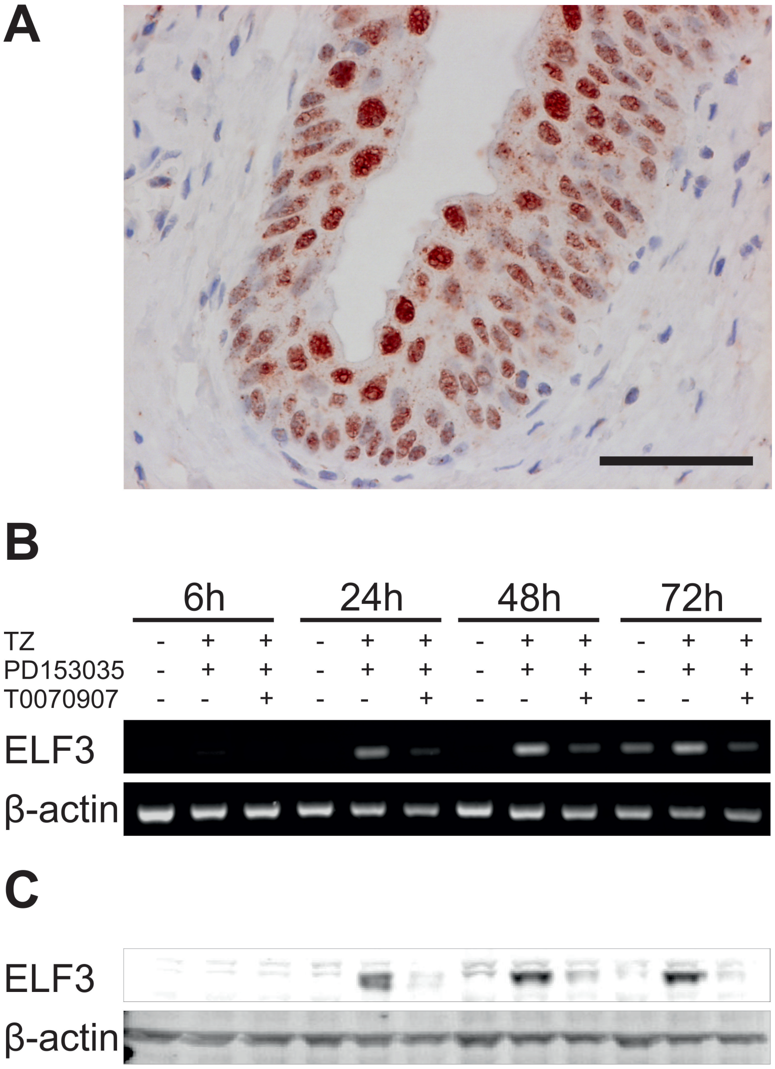

A
B
6h
24h
48h
72h
TZ - + + - + + - + + - + +
PD153035 - + + - + + - + + - + +
T0070907 - - + - - + - - + - - +
ELF3
β-actin
C
ELF3
β-actin

Figure 4.5: Experimental validation of ELF3. ELF3 immunolocalisation in human urothelium on a section of normal ureter (A). Note nuclear localisation and increased intensity in the most highly differentiated superficial cells (Scale bar = 50μM). Expression of ELF3 transcript (B) and protein (C) was examined in NHU cells induced to differentiate in response to the TZ/PD protocol. The antagonist T0070907 was used to confirm specific involvement of PPARγ. Cells were pre-treated with 5μM T0070907 (or vehicle control) for 3h prior to addition of 1μM troglitazone (TZ) and 1μM PD153035. RNA extractions and whole-cell protein lysates were collected for analysis at 6h, 24h, 48h and 72h post-treatment. At each time point a DMSO vehicle control was included. For transcript analysis (B), ELF3 gene expression was analyzed by RT-PCR and β-actin was included as a normalisation control. Whole cell lysates were processed for western blotting (C) and labelled with antibodies against ELF3 or β-actin as loading control.

the most efficient knockdown and was used in subsequent experiments.

Following induction of differentiation in response to PPARγ activation (TZ/PD protocol), cells stably transduced with ELF3 shRNA showed reduced induction of ELF3 transcript and of the terminal differentiation-associated UPK3a gene, compared to the transduced scrambled control cells (Figure 4.6A). Quantitative analysis by real-time PCR of the ELF3 knock-down cultures showed reduced differentiation-induced expression of CLDN7 and of the transcription factors FOXA1 and GRHL3 (Figure 4.6B).

To assess the consequence of ELF3 knockdown on differentiated urothelial barrier function, TER was used to monitor barrier development post-induction of differentiation using the ABS/Ca^{2+} protocol. A measurable barrier first became apparent at four days post-induction of differentiation in both the control and ELF3 knockdown cells, but the ultimate TER attained was significantly reduced in the latter (control versus ELF3k/d: 3,420±443 versus 2,06±176 $\Omega.\mathrm{cm}^2$; $p<0.001$ ±sd; n=6; Figure 4.6C). Following scratch wounding of the cultures, initiation of barrier repair was less efficient in the knockdown cells and the final barrier attained was reduced compared to the transduced scrambled control (Figure 4.6D).

4.1.4 Discussion

Much has been learned about transcriptional regulation during tissue development from null expression studies in transgenic mice, but translation to human systems is more challenging. Here we demonstrate how bioinformatics analysis of a normal human differentiating cell culture time series can identify key transcriptional regulators involved in human tissue-specific determination and differentiation, and provide insight into the relationships and hierarchies of the transcriptional networks.

We have described a qualitative approach for the analysis of asynchronous time series data and applied it to four gene expression time series representing two differentiation-inducing protocols in bladder and ureter derived finite human urothelial cell lines. Unlike other methods for analysing short time series [30, 157], our method is directly applicable to analysing non-synchronous short time series and overcomes issues of missing time points or replicates. The result of our analysis is a ranked list that offers an intuitive pipeline for successive iterations between data analysis and biological evaluation to attain a manageable set of candidate genes. The same approach can be used to look for specific differences between different experimental arms to identify genes upregulated by one protocol only. For example, such analyses could provide insight as to why the ABS/Ca^{2+} protocol generates a differentiated multi-layered and functional barrier ur-

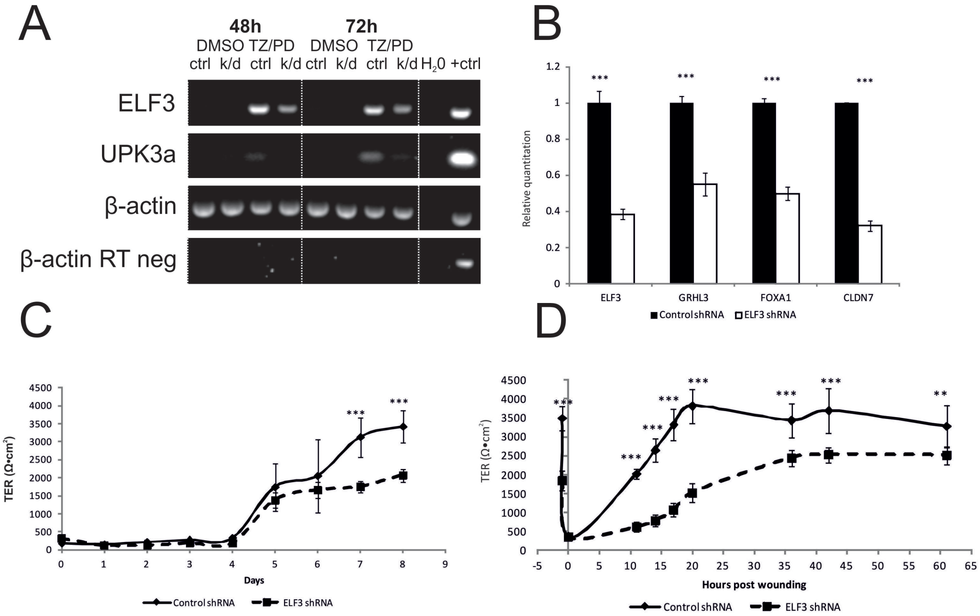


Figure 4.6: Analysis of ELF3 knock-down on NHU cytodifferentiation and barrier repair. Expression of urothelial differentiation-associated genes was examined in ELF3 versus scrambled shRNA transduced cells. A) Scrambled shRNA control (ctrl) and ELF3 knockdown (k/d) cells cultures were exposed to vehicle only (0.1% DMSO) or differentiated by co-treatment with TZ and PD153035 (TZ/PD) and the expression of ELF3 and late differentiation-associated UPK3a was analyzed by RT-PCR at 48h and 72h post-treatment. B) The expression of ELF3, GRHL3, FOXA1 and claudin 7 was analyzed by RTqPCR in cells transduced with scrambled versus ELF3 shRNA at 48h post differentiation with TZ/PD. In A and B, the reaction controls included an RT-negative for each RNA sample, a no-template (H2O) control, a β-actin normalisation control and a genomic DNA positive control. C) Barrier function in ELF3 versus scrambled shRNA transduced cultures was followed over an 8 day period following differentiation in ABS/Ca^{2+} by measurement of the transepithelial electrical resistance (TER). D) The same cultures were then wounded and the restoration of barrier function was observed over the subsequent 61 hours. Statistical analysis was calculated by ANOVA with Bonferroni Multiple Comparisons post-test (***P<0.001, **P<0.01).

othelium, compared to the generation of differentiated but non-organised monolayer cell cultures that result from the TZ/PD protocol.

In this report, we concentrated on identifying common (method-independent) genes involved in regulating the development of a differentiated phenotype from normal human urothelial cells. The culture system is such that it maintains NHU cells in a proliferative squamous basal phenotype, characterised by CK14+/CK13- expression [131] and we have shown that PPARγ activation [146, 149, 147] or subculture in serum [31] can switch cells into a differentiating CK14-/CK13+ transitional epithelial programme. Our analysis has identified the epithelium-specific Ets domain transcription factor ELF3 as a differentiation-associated gene whose expression is regulated downstream of PPARγ. In mice, ELF3 has been shown to be induced in the urothelium following infection with uropathogenic *E.coli* (UPEC) [105].

Infection of the bladder epithelium of mice with UPEC triggers a response in which bacterial-laden superficial cells are exfoliated and the urothelium is reconstituted through differentiation of underlying basal and intermediate cells [105]. One of the early transcriptional responses to attachment of the UPECs is thought to be upregulation of the transcription factor ELF3, which has also been implicated in keratinocyte terminal differentiation [4, 110]. Mysorekar and colleagues hypothesised that ELF3 has a dual role in regulating urothelial differentiation and mediating host defense through transactivation of iNOS [105].

Targeted disruption of ELF3 in the mouse resulted in 30% lethality, with the remaining offspring reported as showing disrupted morphological and cellular differentiation of the small intestinal epithelium [108]. ELF3-deficient enterocytes expressed markedly reduced levels of the transforming growth factor type II receptor (TGFβRII) and could be genetically rescued by introduction of a human TGFβRII transgene, demonstrating that ELF3 is the critical upstream regulator of TGFβRII in the mouse small intestinal epithelium [45].

Transcriptional reprogramming of the TGFβR pathway, including downregulation of TGFβRII, has been documented in NHU cytodifferentiation [44]. However, TGFβR signalling was not associated in urothelial differentiation itself and instead was implicated in priming an autocrine tissue repair programme [44].

A GRHL3-null mouse embryo model was used to demonstrate that the transcriptional regulator GRHL3 is required for formation of normal superficial cells and terminal differentiation of bladder urothelium [156]. The gene and protein expression of the uroplakins was significantly downregulated in bladders of GRHL3-null mice and a functional GRHL3 binding site was identified on the UPK2 gene promoter. More recently it has been proposed that the transcriptional regulator KLF5 is required for urothelial maturation and differentiation [11]. Thus, in mice with KLF5 deficient bladder epithelium, the urothelium fails to stratify and there was reduced expression of terminal differentiation markers, including uroplakins and claudins. Eleven transcription factors were downregulated in KLF5-deficient bladder urothelium including PPARγ, ELF3, FOXA1 and GRHL3. The murine GRHL3 gene has been shown to be a downstream target of KLF5 and it has therefore been proposed that PPARγ and GRHL3 participate in a

KLF5-dependent transcriptional network regulating urothelial differentiation [11].

Southgate *et al.* have previously shown a specific role for PPARγ in the induction of differentiation in normal human urothelial cell cultures, which depends both on the suppression of EGFR activity and availability of activating ligand [149]. PPARγ activation leads to de novo expression of intermediary transcription factors including FOXA1 and IRF1 that act directly as transcription factors for inducing the de novo expression of uroplakins and other genes associated with urothelial differentiation [148]. ELF3 lies downstream of PPARγ: it has predicted PPAR response elements in the promoter and is induced specifically by PPARγ activation. We have now demonstrated that ELF3 influences urothelial differentiation, as the induction of UPK3a expression and the ultimate acquisition of barrier function were both inhibited by ELF3 knockdown. The effect on ELF3 on UPK3a transcription must be indirect as no Ets binding sites are predicted in the UPK gene promoters (not shown), although as shown by knockdown, ELF3 does influence expression of other implicated transcriptional regulators including FOXA1 and GRHL3.

In conclusion, we propose a hierarchy in the specification of human urothelium that has ELF3 downstream of PPARγ, but upstream of GRHL3 and FOXA1. We suggest that our strategy of studying the temporal development of a differentiated phenotype in vitro can help unravel the hierarchical relationships between candidate transcription factors and their individual roles in the differentiation programme and we have provided a new approach for extracting this information from gene array studies.

CHAPTER 5

Large-Scale Integration of Heterogeneous Data

The previous chapters have introduced different methods to analyze gene expression time series data. Time series offer, as already mentioned, a great opportunity to explore possible causalities in the data but are also challenging to use for the inference of gene regulatory networks. The previous sections have also shown that relying on a set of gene expression data alone is not enough. As outlined in Chapter 4, for the analysis of time shifted human urothelium, it can be seen that for identifying gene interactions or network structures, the incorporation of additional prior knowledge can be of great value. In the last couple of years, a tremendous amount of experimental data has been collected and stored in publicly available databases. This data covers various levels of cellular systems, such as gene expression data from microarrays (and more recently from Next Generation Sequencing), protein-protein interaction data, DNase footprinting, chip on chip or mass spectrometry data. Furthermore, various prediction methods have been published, tested and improved over the last years, which offers a great opportunity to mine these data sets for new insights into the underlying systems. The major challenge is to combine the different sources into one integrated database and to make use of this data as prior knowledge for learning gene regulatory networks. The next sections describe the curation of a heterogeneous gene interaction data set for different organisms, as well as an approach for combining these different data sets for the prediction of gene regulatory networks. The data generation and analysis was done in collaboration with Robert Pesch, who provided the text mining, ChIP-seq and gold standard data within the TUM and LMU graduate school RECESS. A part of this, data collection for the analysis of conserved regulatory networks in eukaryotes has been published in Pesch *et al.* [115].

5.1 Introduction

Regulatory networks consisting of transcription factor and target genes are the core of biological systems. A transcription factor (TF) binds to a target gene (TG) to enhance or suppress the expression rate of this gene to response to environmental changes, intercellular signals or to control the cell cycle. Unfortunately, little is known about the details of regulatory networks and a reliable structure is only known for *E. coli* and to some extent for *S. cerevisiae.* For more complex eukaryotes like *D. melanogaster*, *M. musculus* and of course *H. sapiens*, we are still only able to describe small fragments of the regulatory mechanisms.

The predictive performance of gene regulatory network analyses often relies on the combination of different data sources and how well these sources can be integrated into the learning process. In the simplest case, this is done by just using the knowledge about TFs. Genes are then grouped into TFs and TGs, and only directed interactions are added from TFs to TGs. This reduces the amount of possible interactions tremendously. In the following, five different approaches dealing with the integration of additional data sources for GRN learning for eukaryotes will be introduced. Their common challenges are the network size, data quality and pre-processing, feature generation and selection, model complexity and performance evaluation. Each approach offers different solutions, but has also shortcomings regarding the mentioned challenges.

A possible data integration framework, which uses inductive logic programming (ILP) for static data for *S. cerevisiae* was introduced by Fröhler *et al.* [50]. The data consisted of discretized gene expression data (microarrays) for 173 different experimental conditions, TFBS data, PPI data (physical and genetic interactions) and functional categories. The different data sets were represented with predicates and relations with the goal to learn the corresponding predicate (consisting of GeneID, ConditionID and discretized expression value) for each gene. Learning was done with the Tilde tool [16] and the set of genes reduced to 1,411 TG and 53 TFs with high gene expression variance over all conditions. An advantage of the ILP approach is its ability to easily integrate additional constraints and data sources. The field of logic programming has not yet received as much attention for GRN inference as the classical fields mentioned in Chapter 2. Recent promising approaches, such as from Brouard *et al.* [22], come from the field of Markov logic networks (MLNs) [118]. Besides the integration of heterogeneous data sources, an attractive aspect of MLNs is the more intuitive interpretability compared to other black-box models like SVMs. Furthermore, they are capable of modeling cycles, such as regulatory feedback loops in the network. A drawback of MLNs is their need for a larger amount of observations and also their computational complexity.

An approach using Bayesian networks to combine expression data with multiple sources of prior knowledge was presented by Werhli *et al.* [150]. They also used data sets from yeast and additionally, the RAF signaling pathway. Each integrated source is encoded with an energy function and the sampling of the prior and posterior probabilities was done with Markov chain Monte Carlo (MCMC), similar to the procedure described in Chapter 3.2. The main focus of this paper lies on the algorithm and less on the biological aspects of data integration. They integrated two additional sources and applied their method on a smaller subset of genes, 25 in yeast and 11 genes for the RAF signaling pathway. Besides the computational complexity of this method, Bayesian networks do not allow for feedback loops, which are also an important regulatory structure. Still, their method was able to infer the GRNs and outperformed other methods such as Graphical Gaussian Models, Bayesian networks with and without prior and prior knowledge only as baseline. It also offers a mathematical framework to add additional data sources and train weights for them individually.

A biology driven framework, illustrated on data from *C.elegans*, has been published by Cheng *et al.* [28]. They based their approach on the three main regulation categories: TF to gene, TF to miRNA and miRNA to gene. They used ChIP-seq, RNA-seq, known protein-protein and TF-TF interactions, and a set of predicted miRNA target sites at the 3'UTRs. Positive and negative signs for the regulating TFs were assigned according to the RNA-seq expression profiles and their correlations. Their integrated GRN was simply generated by keeping those genes, which had both a TF and a miRNA predicted binding site. The network was then enriched with known TF-TF interactions. Due to the integration rule, their approach is limited to only those interactions which are covered by all data sources. For example the set of *C.elegans* was reduced to 22 TFs out of 393 TFs. They also applied their approach to smaller data sets of human and mouse and analyzed the hierarchical structure and motif enrichment for all three results. The advantage of this approach lies in restricting the analysis to only qualitative, preprocessed data sources, which of course also strictly limits this approach to small data sets and also cannot account for noise or uncertainties in the data. Also, there is no standard evaluation possible, except testing in the lab.

Another genome-wide machine learning approach on learning a GRN in human is the regulatory interaction predictor (RIP) [10]. The interesting concept of this approach is the feature generation from the given data sets. These data sources consist of the network topology from the gold standard, TFBS predictions and the correlation of gene expression microarray data. The features include the number of correlation neighbors to a gene, which are regulated by the corresponding TF, or the number of genes regulated by a TF, or the number of correlating neighbors to a gene, which have a significant PWM hit of the corresponding TF. In total, ten different features were created and an ensemble classifier of SVMs (with Gaussian kernels) with 20 x 100-fold stratified crossvalidation was employed for the training. The gold standard consisted of 2,896 true regulatory interactions (derived from TRANSFAC) and 284,651 non-interacting pairs. The evaluation was done with precision-recall curves on the average values of the 20 ensemble classifiers, as well as the predicted interactions on a microarray study.

Marbach *et al.* applied supervised and unsupervised machine learning methods on a diverse functional genomics data set [94]. For *D. melanogaster*, they predicted about 300,000 regulatory edges, given 600 TFs with 12,000 TGs. As input data they used conserved TFBS motifs and ChIP binding of TFs (both referred to as physical features), as well as chromatin marks and gene expression data (both referred to as functional features). For their unsupervised method they used the mean of the scores for the integrated features and kept the top 2% edges ranked by the highest scores. For the supervised method they used logistic regression and accounted for the class imbalance problem by using a stratified crossvalidation, where positive and negative examples were balanced. Additionally, the weights of positive instances were scaled to be balanced between negative and positive examples. They evaluated their predictions on different data sets, such as tissue-specific expression patterns. They were also able to use a regression model on their integrated regulatory network to predict the expression of ~17% of the genes correctly from the expression level of TFs in a new experiment. One of their findings was that physical information is the most informative feature when evaluating against their gold standard, but has almost no predictive power for the gene expression levels. Another interesting outcome was that the supervised and unsupervised approach performed comparably. A reason for this might be the small number of available edges for the training set of the supervised approach.

All the presented methods focus more or less on the biological part of the analysis, scale only to small networks with very granular details or to large networks, which allow to explore the topological properties and to identify functional modules. Furthermore, they use different evaluation procedures, which in most cases do not allow comparison of the results to newly developed methods. Due to more reliable gold standards and complete data sets, most of the approaches focus on the more well-known model organisms, like *S. cerevisiae*, *D. melanogaster* or *C. elegans*. Only the RIP approach attempts to generate newly combined features from the data sources and only Marbach *et al.* assess the predictive contribution of each data source. None of these methods offers a global visualization of the data sources and their overlapping features and only Cheng *et al.* also compared their results on smaller networks between different organisms.

In the following, the focus will be on the description and analysis of more complex organisms (like mouse and human) and how well predictions can be made with the currently available data. In particular, three major points are addressed. First, the quantitative and qualitative description and visualization of the available data for those organisms and how well the different sources agree on the interactions. Second, the predictive power of the data and newly generated features to classify TF-TG interactions and third, how the different data sources contribute to the predictions.

The database created for this thesis consists of five different information sources: relations extraction from textual documents, transcription factor binding site (TFBS) predictions, protein-protein interaction (PPI), ChIP-seq and gene expression data. Considered individually, all of these data sources have different strengths and weaknesses.

In the next sections the individual methods will be described, as well as their different predictive capabilities. This is followed by using the data of two of the organisms – mouse and human – to classify TF-TG pairs with state of the art machine learning techniques and finally, assessing the importance of each data source for the predictions.

The results of this analysis provide the following insights. First, as expected, the overlap between the integrated data sources is very limited but at least for half of the known interactions for each organism, two or more data sources describe the same interaction between two genes. Second, the integration of features, which are generated by combining the different data sources or their network properties (such as hub score or shortest paths), can improve the predictions. Third, the collected data sets can be used to correctly classify TF-TG pairs from already known interactions. Fourth, across all data sets ChIP-seq experiments show the proportionally largest overlap with the known interactions and also a larger overlap with TFBS predictions. Additionally, this analysis shows that a combination of the different sources contributes best to the predictive performance.

5.2 Data Sources and Applied Methods

Four different standard organisms were chosen for the integration into the database, *S. cerevisiae*, *D. melanogaster*, *M. musculus* and *H. sapiens*, of which the main focus will be on mouse and human, since these are the most complex and least understood organisms on this list. All of these organisms have been heavily explored with respect to different aspects over the last decades and offer the most reliable possibility to evaluate our predictions with already established biological knowledge about the gene regulatory interactions. The five main pillars of the here presented approach are text mining, co-expression, PPI networks, ChIP-seq experiments and TFBS predictions. The data sources and the general data processing steps will be described in the following. This is followed by the definition of the gold standards for the different organisms to assess the performance of the predictions. The last sections describe the architecture of the MySQL database in which the data is stored, the feature generation and the applied classification algorithms.

5.2.1 Text Mining

For text mining data, a set of abstracts from PubMed (20,766,340 abstracts) and full text Publications from the PubMed central open access subset (389,322 articles) was used to search for descriptions of potentially interacting genes. The gene names were identified with syngrep [32], which is a dictionary-based gene name identification tool. Dictionaries were compiled for the different species by combining the gene names, aliases and synonyms for each gene from UniProt, Ensembl, HGNC, MGI and FlyBase. The relations between the genes were identified with a simple tri-occurrence approach, RelEx [51] and a shallow linguistics (SL) kernel [58] on a sentence base. A brief description of the methods is given in the following (the generation of the text mining data set is described in more detail in Pesch *et al.* [115]):

Tri-occurrence: A list of keywords indicating regulatory interactions was created. For each sentence where at least one keyword was found, a regulatory relation between the respective genes was assumed.

RelEx: A rule-based relations extracting tool using dependency parse trees.

SL: A SVM kernel for identifying a relation using the local and global context of the sentence.

From RelEx all Tri-occurrence relations stating a regulatory relation were used. For SL, the simple margin active learning approach [141] was used to create a model for the identification of regulatory sentences. For this purpose a set of 450 identified regulatory relations found by RelEx was manually corrected. This set was used to train an initial model. The model was refined by applying the learned predictor and 10,000 randomly selected relations found by the Tri-occurrence approach und selecting the 100 instances, which were closest to the separating margin of the SVM. This set of relations was then manually annotated and included in the training set. The model was refined with this approach until no further performance improvement on a control set could be observed.

5.2.2 Chromatin Immunoprecipitation Data (ChIP-seq)

The use of high-throughput sequencing (seq) in combination with chromatin immunoprecipitation (ChIP) allows to study genome-wide protein bindings. ChIP allows to selectively enrich DNA fragments, which are bound by a particular protein. The DNA fragments can then be sequenced. ChIP-seq makes it possible to analyze the physical interactions between TFs and the corresponding DNA sizes. The data was taken from the ENCODE project. The ChIP-seq experiments were downloaded for mouse from `http://genome-euro.ucsc.edu/ENCODE/downloadsMouse.html` and the other data sets from ENCODE `http://genome-euro.ucsc.edu/ENCODE/downloads.html`.

5.2.3 Co-expression Network Generation

The data sets and their sources are described in Table 5.1. The samples were taken from different experiments and in case of the larger organisms, also from different cell lines, and combined into single compendia. All samples are based on the Affymetrix platform and the probe sets were mapped to Ensembl gene identifiers. The mapping was necessary for combining the different data sources in the following analyses. This mapping was done partly with the available mapping files from the data sources, available Bioconductor annotation packages and the DAVID gene ID conversion tool [70]. Some of the probe sets or given identifiers were ambiguous or outdated and could not be mapped onto Ensembl identifiers. Those probe sets were discarded from our final data set. Probe sets, which referred to the same Ensembl identifier were integrated by using the one with the maximum interquartile range (IQR). This ensures to keep the probe set with significant changes across the different samples. The drawback of this approach is the possibility that a large IQR might be also caused by strong noise in the measurement. Still, using the mean of all matching probe sets could lead to discarding interesting signals from single probe sets.

Three methods were used to assess the co-expression patterns in the different data sets, correlation (Pearson), partial correlation and mutual information. All calculations were done in R with the default parameters. In contrast to regular correlation, partial correlation not only examines the relationship between two variables, but also subtracts the possible effect of the other variables on this particular correlation. The GeneNet [128] package was used to calculate the partial correlation. The third method, mutual information, has its origin in information theory and measures the mutual dependence of two variables. ARACNe is a reverse engineering approach, particularly designed for cellular networks, which uses mutual information in combination with the *Data Processing Inequality* (DPI) [97]. The DPI is used to remove the weakest interaction (edge) between any three genes in the resulting network. For the calculations, the R package parmigene [124] was used and the default setting for the DPI of 0.5 was applied. It should be noted that including the different conditions from the different experiments could give additional insights, but for this analysis only the overall correlating genes were considered. Furthermore, gene co-expression can only capture the regulation on mRNA level and not on the protein level. A possible extension of the co-expression analysis could be to either use the conditions or to cluster the given samples and to calculate the correlation of the genes within those clusters.

5.2.4 PPI Network Curation

For the PPI network data, the current dump of the PPI database STRING (version 9.0) [73] was downloaded and the protein identifier mapped with the given protein aliases file to the corresponding Ensembl gene IDs. Table 5.2 gives an overview of the different PPI networks. STRING integrates four different sources of knowledge in its PPI network: genomic context, high-throughput experiments, (conserved) co-expression and previous knowledge (from publications and other databases). From these sources a confidence score between 0 and 1000 is calculated, which reflects the reliability of an interaction. Scores equal or above 700 are considered to be highly reliable (this is an empirically chosen threshold). The largest network exists for human, with nearly twice the number of interactions as in the mouse and fruit fly PPI networks. The number of proteins is similar for human and mouse, as well as the proportion of PPIs above the threshold of 700. In addition to the interactions and their scores, for all pairs of proteins in the networks the shortest path (unweighted) and the number of shared neighbors, which have a direct interaction with the two proteins, were calculated. This was done with the igraph package in R [33]. The number of neighbors per protein in mouse is significantly smaller than for human or fruit fly. As already discussed in Section 3.2, biological networks tend to be scale-free, which means that the majority of nodes has only a few interactions while only a few, so-called hubs, have many interaction partners [9]. This structure ensures topological robustness against single deletions of nodes in the network, as well as quick information transfer over a few connections from one point in the network to another. The nearly equal average path lengths for all organisms in Table 5.2 are an indicator for

Table 5.1: Overview of curated gene expression data

Organism	Samples	Mapped Ensembl Genes (All Known Genes/ Protein Coding)	Gene Mapping	Sample Source
S. cerevisiae	904	6054 (7126/6692)	M3D Annotations and David	M3D
D. melanogaster	1102	12597 (15246/13940)	Given Mapping and David	COXPRESdb
M. musculus	2226	16267 (37681/22705)	Given Mapping and David Bioconductor: mouse4302.db	COXPRESdb
H. sapiens	4401	18206 (54843/21160)	Bioconductor: Hgu133plus2.db	COXPRESdb

This table gives an overview of all curated gene expression data sets. All genes were mapped to Ensembl gene IDs to be able to integrate the different sources. All samples are Affymetrix arrays and were downloaded from the *Co-expressed Gene Database* (COXPRESdb - `http://coxpresdb.jp/top_download.shtml`) [109]. The mapping of the Affymetrix IDs was done with a combination of given mapping files, Bioconductor annotations and the DAVID gene id conversion tool. The column *Ensembl Genes* lists only the number of genes, which could be mapped from the given samples. In brackets are the actual total number of all known genes for this organism and the ones which are protein coding (the numbers were also taken from Ensembl).

this structure. Additionally, Kleinberg's hub score (implemented in the igraph package) was calculated for each gene in the network and used as an additional feature (described in Section 5.4.1).

5.2.5 TFBS Prediction

The extraction of the promoter region for each organism was done with the RSAT workbench [140]. For all organisms, the same promoter size was chosen, 500bp upstream and 500bp downstream of the transcription start site (TSS). It should be noted that TFBS can be found also in far more distant regions, in human regions up to 5000bp (and even further distant regions) are considered to contain binding sites. Additionally, in some organisms, TFBS can be found even far downstream of the gene of interest. Still, the majority of the TFBS are considered to be close to the TSS and increasing the search space for the TFBS prediction introduces additional noise to the results. Open reading frames (ORFs) from other genes were excluded from the promoter regions. If available, the sequences were directly extracted from Ensembl (version 66) with RSAT, otherwise from the RSAT database. The lists of all possible TFs for each organism were extracted from the *Transcription factor prediction database* (DBD) [86]. For some of these predicted factors, position weight matrices (PWMs) are available from the transcription factor binding profile databases JASPAR [125] and TRANSFAC (version 9.3) [98]. The PWMs were used to search the extracted promoter sequences for significant TFBS. An overview of all sequences and PWMS are available in Table 5.3. A parallel version of the TFBS prediction tool cureos [151] in R has been developed and was applied with default parameters for the predictions.

5.2.6 Database Evidence for Regulatory Relations

Regulatory relations where extracted from the multi-species curated databases TRANSFAC [98] and ORegAnno [104]. Furthermore species-specific direct relations were manually extracted from YeastRact [138], RedFly [63] and the pathway databases Biocarta and NCI-Pathway [126]. In the following the collected sets of known interactions are referred to as database evidences and the other predicted and collected evidences as regulatory evidences. Table 5.4 gives an overview of this.

Table 5.2: Overview of curated PPI data

Organism	Proteins	PPIs	PPIs $\geq$ 700	Avg. Path	Avg. Neighbors
S. cerevisiae	6065	979,869	211,802	2.34	8.54
D. melanogaster	10,427	1,841,622	187,830	2.86	6.26
M. musculus	15,086	2,173,740	331,910	2.75	2.6
H. sapiens	17,328	4,035,506	550,604	2.53	5.83

In this table, the number of mapped proteins and interactions from the extracted STRING data is shown. *PPIs* $\geq$ *700* shows the number of interactions with high significance for each organism. The last two columns show the average undirected, shortest path length and number of shared direct neighbors between any two genes in the PPI network.

Table 5.3: Overview of curated transcription factor and promoter sequence data

Organism	DBD TFs	PWMs		Promoter Size (upstream)
		JASPAR	TRANSFAC	(All Seqs/Unique Seqs)
S. cerevisiae	177	177(-)	38(1)	1000
		170 unique genes		
D. melanogaster	1102	125(-)	51(-)	1000
		139 unique genes		
M. musculus	1385	53(1)	374(7)	1000
		276 unique genes		
H. sapiens	1336	76(3)	420(9)	1000
		300 unique genes		

This table shows the collected data for the TFBS prediction. The genes were mapped with the David conversion tool to Ensembl identifiers and the promoter sequences were extracted and pre-processed with the RSAT workbench. I chose the promoter region to be 500 base pairs up- and downstream of the transcription start site and to exclude ORFs from other genes from these regions. The lists of all possible TF for each organism were extracted from the DBD. The number of available TFs with PWMs are shown in columns two and three. The PWMs were used to search the extracted promoter sequences for significant TFBS.

Table 5.4: Overview of curated database evidences for each organism

Organism	Interactions
S. cerevisiae	4310
D. melanogaster	461
M. musculus	928
H. sapiens	3222

This table shows the number of directed interactions between pairs of genes derived from the different databases. As expected, most of the known regulatory interactions were found for *S. cerevisiae.* Even though a similar number of interactions was found for *S. cerevisiae* and human, the number of the interactions for the latter only cover a very small proportion of the total number.

5.3 Descriptive Data Analysis and Visualization

The following sections give an overview of the different collected data sources, visualize their shared predictions and assesses how well those regulatory evidences are potentially able to describe the known interactions (database evidences).

5.3.1 Storing the Data and Overview of Regulatory Evidences

All the different data sources are stored into a MySQL database and all found regulatory relations with their data sources and evidences were combined into one table. Unfortunately, considering all possible gene pairs for an organism with n genes would lead to $n^2/2$ (including self-regulations and with undirected interactions) entries in the database. This kind of dimensionality is not convenient to handle — even with a database — and the amount of information is still sparse for many interactions where only a correlation value exists. Therefore, the search space was limited to gene pairs with evidence from at least one of the of the following methods: text mining, PPI (STRING), ChIP-seq, database evidence or TFBS prediction. Table 5.5 shows an overview of the resulting database entries.

5.3.2 Overlaps of Regulatory Evidences

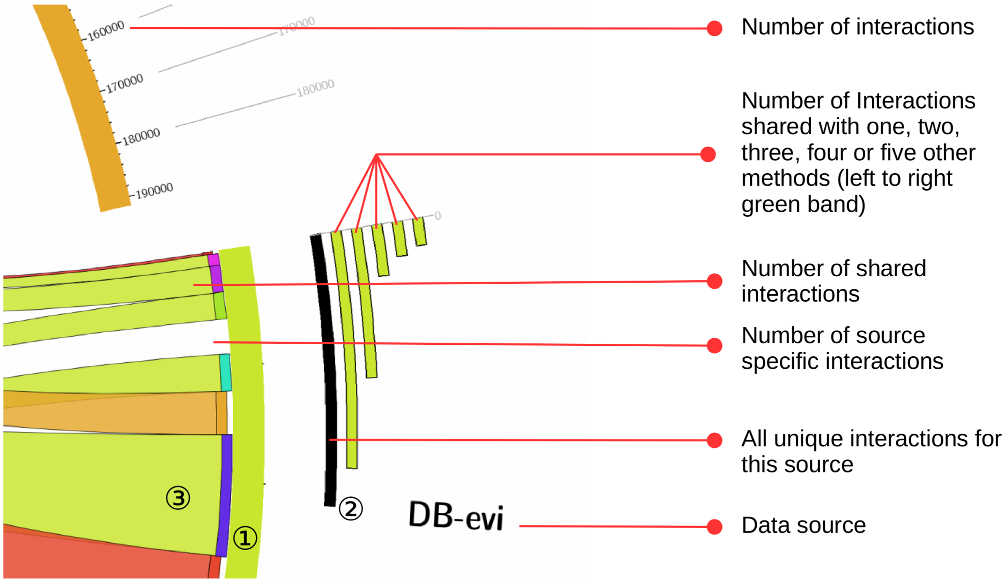
In the following, the shared and unique predictions of the different methods for the different organisms are visualized with Circos plots (Figures 5.2, 5.3, 5.4 and 5.5). This kind of plot arranges the different methods in a circular manner and allows to compare more methods than common Venn diagrams, which are readable only up to four or five

Table 5.5: Number of relations in the database

Organism	**STRING**	**TFBS**	**Cor. (ARACNe)**	**Co-occ.**	**Tri-occ.**	**RelEx**	**SL**	**ChIP-seq**
S. cerevisiae	489,891	110,201	623,241 (17,509)	137,031	31,053	4353	31,040	28,762
D. melanogaster	847,360	165,686	1,048,808 (36,327)	230,764	70,640	15,342	70,545	64,865
M. musculus	1,004,612	1,834,034	2,243,840 (540,135)	848,456	275,577	49,747	275,373	284,659
H. sapiens	1,806,229	4,717,964	4,479,719 (263,167)	1,479,470	451,311	75,044	432,917	1,687,564

This table gives an overview of the the relation available from the different experimental and predicted sources in the database. The number of interactions with calculated partial correlation is the same as for correlation. ARACNe has an additional filtering step, the DPI, and predicts a much smaller amount of relevant gene pairs.

different methods. See Figure 5.1 for a detailed description of the Circos plots.



Number of interactions

Number of Interactions shared with one, two, three, four or five other methods (left to right green band)

Number of shared interactions

Number of source specific interactions

All unique interactions for this source

Data source

① Segments represent each method
② Highlight tracks show unique interactions
③ Ribbons link two methods

Figure 5.1: Description of different data tracks in the Circos plots. Circos plots offer the possibility to visualize dependencies on a large-scale between different objects. Originally used for the display of chromosomes, they have been recently applied for various other tasks. Here, the plot visualizes the shared interactions between each possible pairing of the data sources in the database. Each source is represented by a segment and a color (see ①). The different segments are connected with ribbons (see ③), which represent the number of shared predicted TF-TG pairs between the two sources. The thicker these ribbons are, the more predicted interactions are shared between the two sources. Parts of the segment without ribbons mean predictions, which are solely made by this particular source. The segment size might also contain multiple connections for the same TF-TG pair, if more than one other source predicts the same pairing, and can be, because of this, longer than the real number of predicted TF-TG pairs. The additional highlight tracks (see ②) show the real number of unique TF-TG interactions for this source in black and the additional highlight bands in green show how many other methods (at least one to at most five, from left to right) share these predictions.

For all four organisms, the strongest connected sources are of course the text mining approaches, since these are partly subsets of each other. It can be also seen that TFBS predictions come with a large number of TF-TG predictions, which are not confirmed by any other method and are possibly noise. For the Circos plots there was also no filtering on the prediction scores and the number might be reduced by setting a threshold. It should be noticed that the TFBS predictions are also limited to the number of available PWMs. In case of *D. melanogaster*, only 139 PWMs are used, which is only about one tenth of the total number of TFs (Figure 5.4). The best coverage is for yeast with nearly a PWM for each TF. Figure 5.5 shows that this may come with a lot of false positive predictions. The only data source not shown are the correlations, since, without applying a threshold, an interaction for each pair of genes would be shown. Hence, the TFBS predictions could be partially confirmed by stronger correlations.

Another interesting insight is that the ChIP-seq experiments, which were initially supposed to reflect mostly true regulatory interactions (given a sufficient sequencing depth [76]), share only very few interactions with other regulatory evidences and have a large number of specific regulatory evidences for this source (ChIP-seq only). An explanation could be that many regulatory interactions are so far unannotated or low data quality. One source of influence on this could be the number of replicates needed for each experiment, which was set from the ENCODE consortium to be at least two, while a recent publication claimed that this might not be sufficient [154]. Nevertheless, considering the smaller total amount of ChIP-seq experiments with the other sources and the proportional overlap with database evidences, ChIP-seq still seems to be a highly reliable resource. The largest ChIP-seq overlaps are found with the TFBS predictions, and it could be interesting to combine these regulatory evidences.

More than half of each of the collected database evidences, which were used as gold standard for the predictions are confirmed by at least two different methods. For the data representation it is important to note that a part of these confirmations might be the result of the strong overlaps between the text mining approaches.

One obvious statement from these four model organisms is scarcity of data on confirmed TF-TG interactions. This makes it even more challenging to make predictions for these organisms. In addition to the small number of known interactions for a gold standard, there exists a huge search space of possible interactions. Several interesting questions arise from this, like to which extent the different methods are suitable for even finding new TF-TG pairs and if the different regulatory evidences should be weighted according to their accuracy. It is also interesting to which extent we can predict interactions now, especially in higher eukaryotes. These are standard organisms and a tremendous amount of experiments and work has been put into exploring their regulatory networks in the last couple of years, and consequently, there is already a lot of data available, which is stored across several databases. In a simple way, this collected regulatory evidences could be used to compare results and predictions from single experiments to the given evidences and to rank genes of interest accordingly.

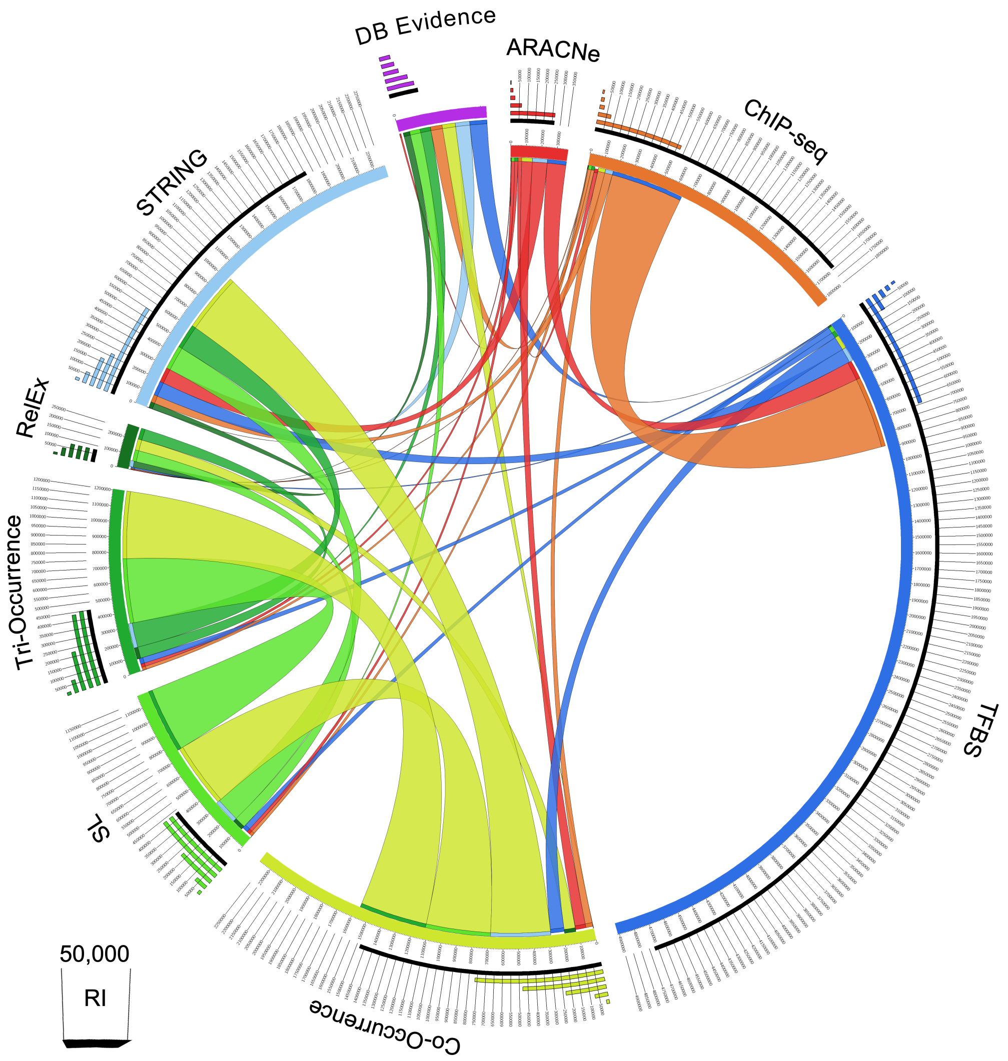


Figure 5.2: Circos plot for intersections on all collected regulatory evidences on human. This overview shows the shared predicted and experimental verified interactions between TF-TG pairs in human. Correlation predictions are left out since these exist for nearly every TF-TG pair shown here. As to be expected, the strongest connections are between the three text mining predictions since tri-occurrence is also a subset of co-occurrence. There is also a stronger connection between ChIP-seq and TFBS prediction, which includes about 500,000 TF-TG pairings. STRING and Co-occurrence shares about 100,000 interactions with the TFBS predictions and ARACNe even slightly more. Nevertheless, there are only about 100,000 interactions for the TFBS predictions, where at least two are confirmed by two other sources. Nearly all interactions from the database evidences are covered by at least one other method. Also notice that this region is zoomed, and the scaling is different to the other methods. In total, this represents 3222 interactions, where nearly half of these interactions are at least confirmed by three other methods.

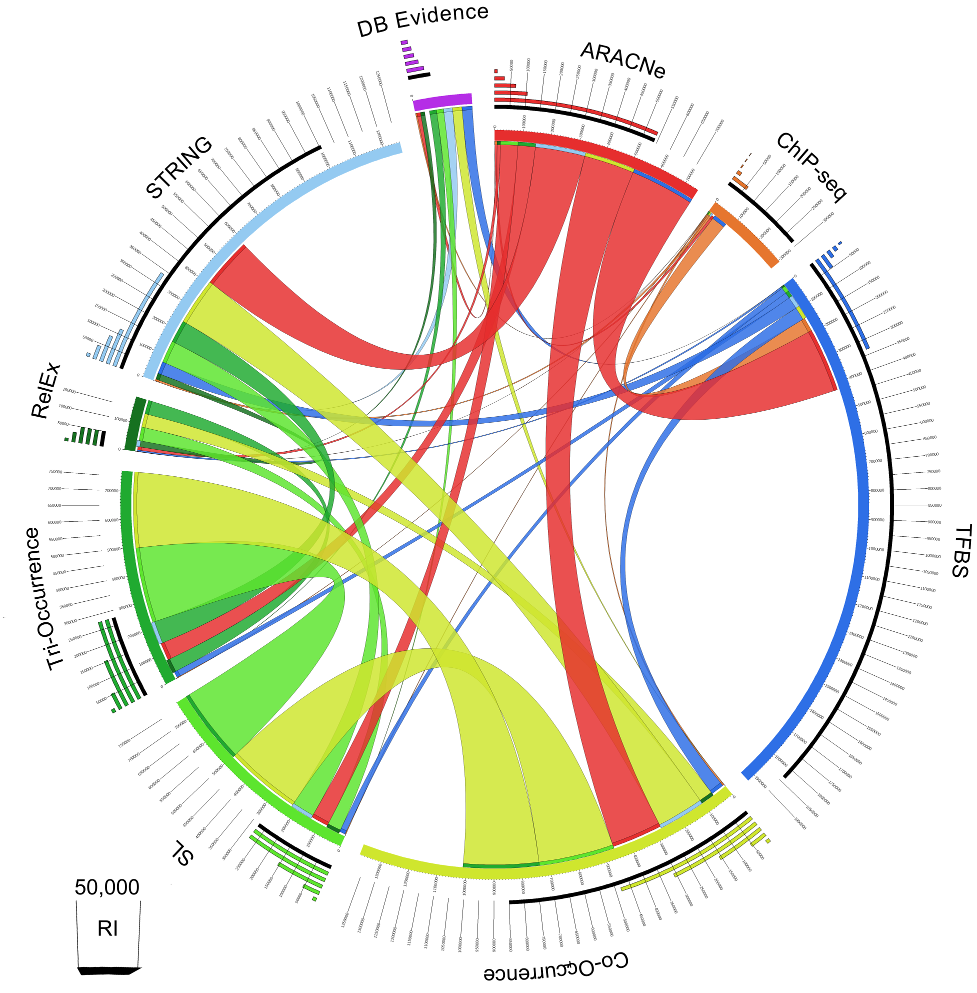


Figure 5.3: Circos plot for intersections on all collected regulatory evidences on mouse. Similar to the overview of the human data set, also STRING has the most overlaps across all other sources. Again, the text mining approaches share many predictions and also the TFBS predictions have the most predictions with no other source confirming them. The predictions with ARACNe seem to filter out substantially less interactions and keep more than twice the number of interactions as in human. Considering experimental evidences from ChIP-seq, human has more than five times as many regulatory evidences as mouse. The overlap of the ChIP-seq experiments with the database evidences is also very small and compared to the number of experiments, the smallest of all here collected organisms.

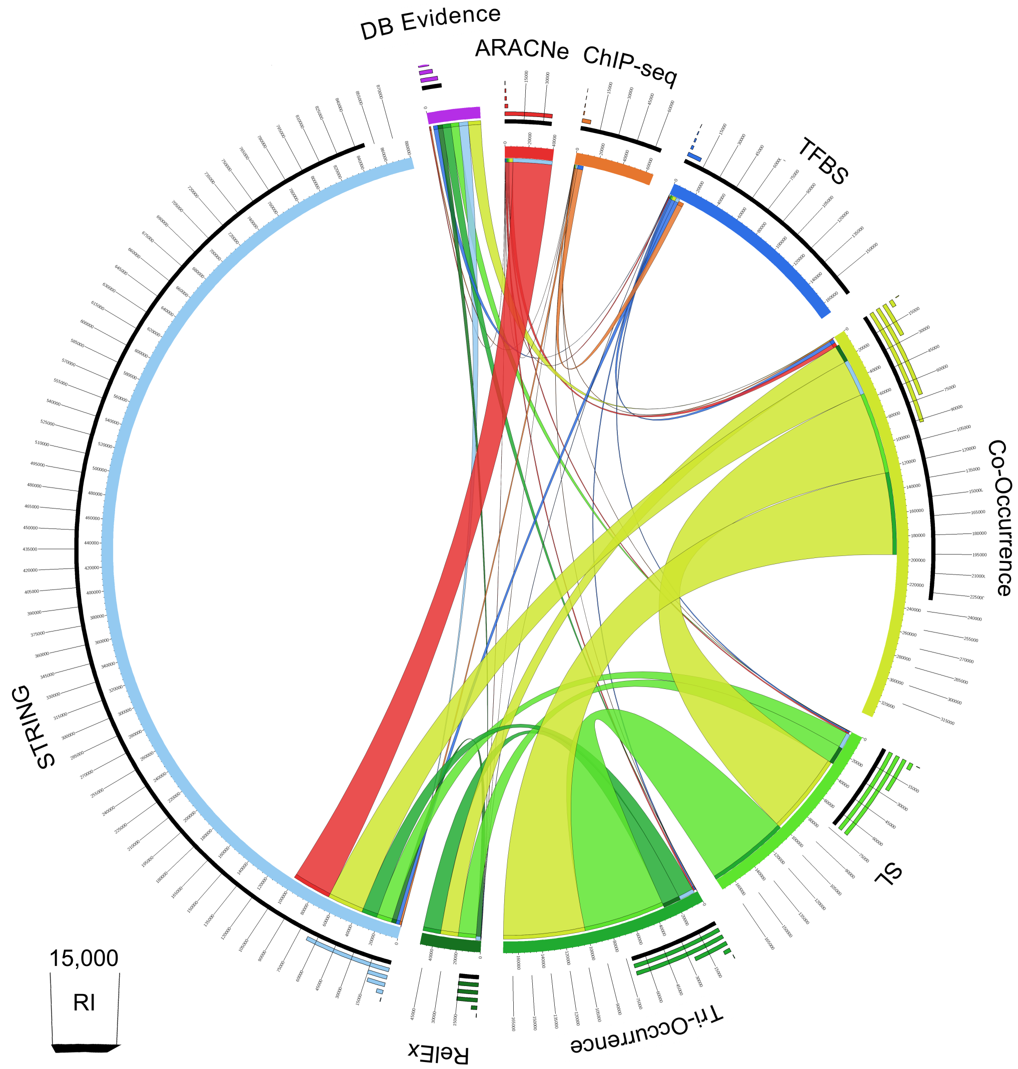


Figure 5.4: Circos plot for intersections on all collected regulatory evidences on *D. melanogaster*. This data set is an interesting outlier compared to human and mouse, since the STRING data clearly outnumbers all other methods combined. This data set seems to be more similar to the yeast data (Figure 5.5) regarding the distribution of regulatory evidences. The overlap of TFBS predictions and database evidences is also very small for this data set. This could be an evidence for generally low performing PWMs, poor choice of the promoter regions, or of course for, many so far unannotated regulatory interactions.

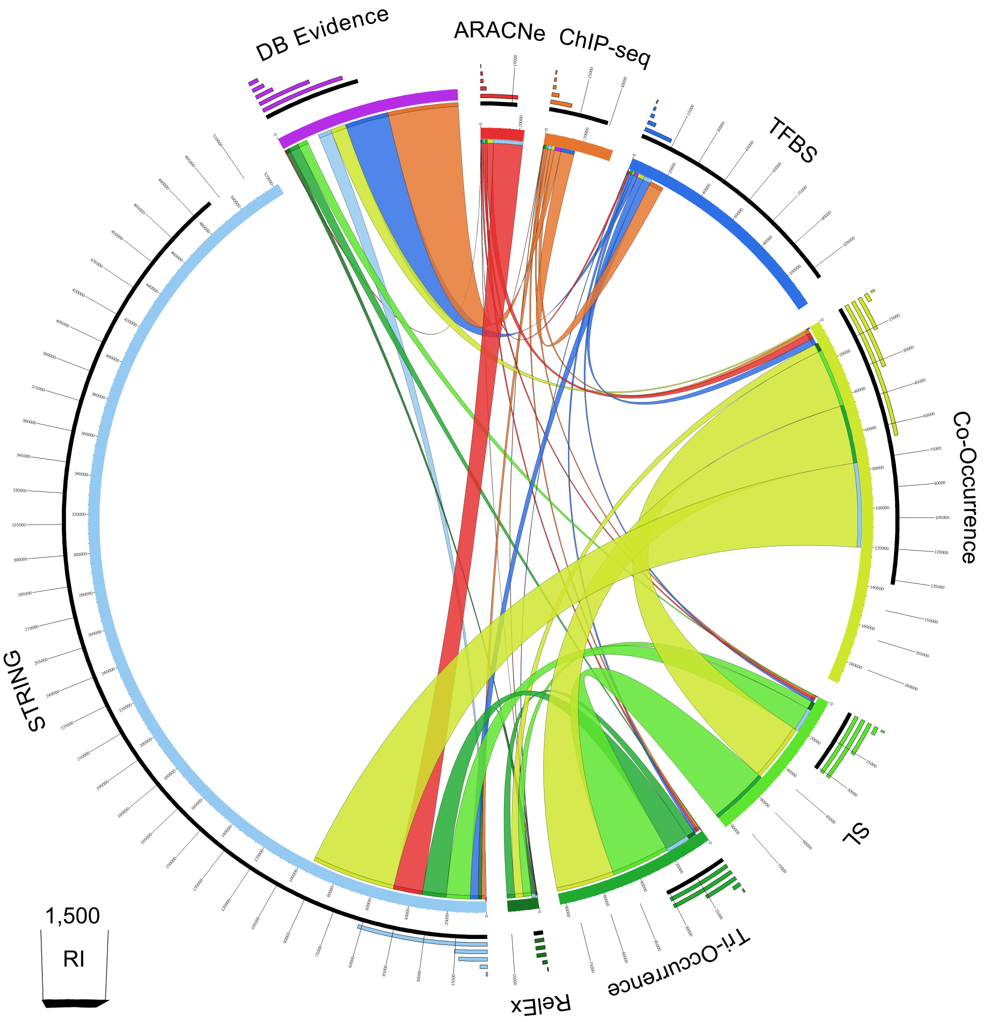


Figure 5.5: Circos plot for intersections on all collected regulatory evidences on *S. cerevisiae*. Even for yeast, not all database evidences also have a regulatory evidence. The majority has less than two evidences.

5.4 Predictive Analysis on *M. musculus* and *H. sapiens*

As stated earlier, the focus of this data integration study lies on the more complex and less well-tested organisms. Hence, the analysis will be illustrated on *M. musculus* and *H. sapiens.* Following the study of Marbach *et al.* [94], the gold standard – consisting of database evidences – was used to generate a training and test set for the predictions. All genes included in the gold standard were extracted from the full data set, along with all possible interactions between those genes and the corresponding regulatory evidences. For *M. musculus* this resulted in a set with 532 genes and 39,811 possible interactions, and for *H. sapiens* in a set with 1995 genes and 371,341 possible interactions. Interactions having only the correlation score as evidence were discarded. Even for this small number of genes, the search is huge, considering the comparably small number of regulatory evidences.

It is clear that this set can only represent a snapshot of the full regulatory network. Nevertheless, this subset offers the opportunity to create a benchmark of known interactions, which allows to evaluate the predictive importance of each data source. Several approaches for evaluation benchmarks and difficulties have already been introduced in Section 2.2.2. The predictions in the following are evaluated against this generated gold standard. The performance is measured with AUROC and precision-recall curves. The 10-fold crossvalidation is done on regular samples instead of sampled subnetworks.

5.4.1 Feature Generation and Selection

Figure 5.6 gives an overview of the features used for the prediction of regulatory interactions. Basically, features can be divided into interaction and gene specific features. In general, the interaction specific features are supposed to contribute more to the learning process. Gene specific features can give hints regarding the tendency of genes to interact with other genes, like the calculated hub score from the PPI data. All interaction-specific features, except the ones from text mining and ChIP-seq, have an associated score. The other features are Boolean. A simplification of the learning process, described in the next section, is made by only considering the interactions as undirected, which means that most of the features refer only to the interaction and not specifically to one of the genes.

Additionally, features were generated from the base sets above. These features were extracted from the PPI networks (STRING) and the combination of partial correlation and TFBS predictions. The STRING networks offers additional insight into another

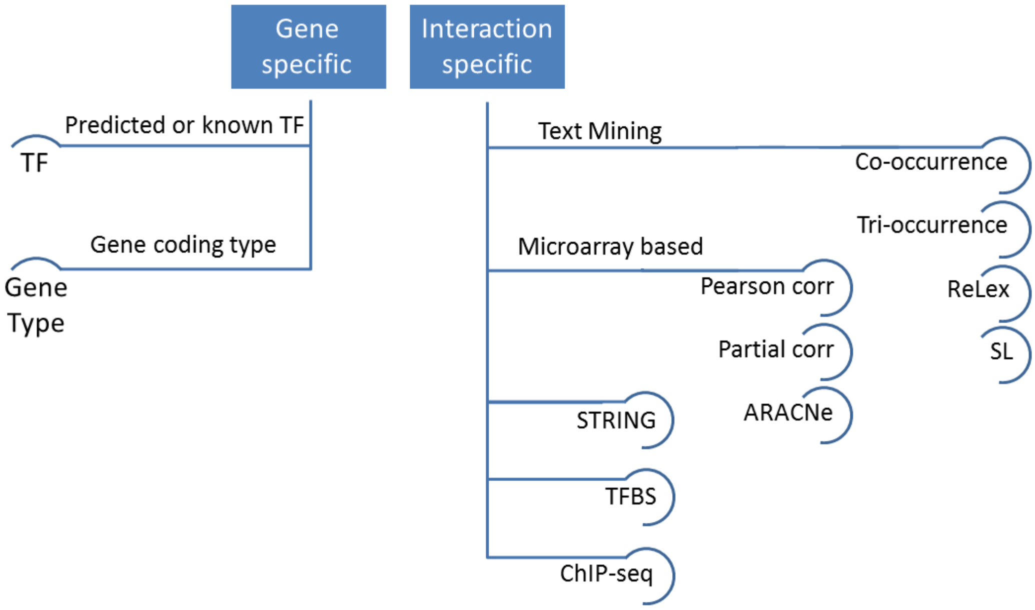


Figure 5.6: Overview of generated features. This figure shows the different generated features from the regulatory evidences. These features can be divided into two main groups: the gene specific and the interaction specific features.

level of regulation on the basis of PPI. Even though it is known that gene regulatory networks and PPI networks can be compared only to a very small extent, the structure and the distances in this network might predict if two genes are more or less likely to interact. The shortest paths between any two proteins in the network were therefore computed (edge count), as well as the hub score of each protein and the number of shared neighbors between two proteins. Calculations were performed using the *igraph* R package. Hubs are of particular interest, since they control a large amount of other genes or proteins and play an essential role in the regulation process. A threshold of 700 was applied to the STRING network to exclude PPIs with a low confidence score. This resulted in the following features:

1. Shortest path in all data
2. Shortest path in data above 700
3. STRING hub score
4. STRING hub score in data above 700

5. STRING shared neighbors
6. STRING shared neighbors in data above 700

The following features were generated for the combination of partial correlation and TFBS predictions:

1. TG (neighbors +TF) pVal: Calculated *p*-value for TF-TG pairs where the TG has correlated neighbors with also significant hit of the same TF. The *p*-value was calculated by getting all neighbors for the respective TG with a partial correlation above the 90% quantile (on all possible interactions) and by sampling the same amount of neighbors 100 times randomly. A two-tailed test was applied to compute the statistical significance.
2. TG (neighbors +TF): Number of TF-TG correlated neighbors (with TG) with a hit of the same TF
3. TG (neighbors) Number of correlation neighbors
4. TFBS PWM length: Number of positions described by PWMs

The combination of the different sources of regulatory evidences can be used to improve the predictions and additionally for testing for suitable extensions of the data sources.

5.4.2 Feature Imputation

Due to the different preprocessing steps and initially missing data points, not all features (regulatory evidences) have a value and unfortunately not all prediction methods are capable of dealing with missing values. The categorical and numerical values were filled using the GBM (Gradient Boosting Machine) function from the *imputation* R package [46]. The imputation uses boosted trees, where each column (feature) is treated as a regression problem. For each of the columns i, boosted regression trees are applied to predict i by using all other columns except i. Should the predictor variables also contain missing data, the GBM function uses surrogate variables as substitutes for the predictors. Additionally, a simple majority class imputation was used, where the most frequent value is used for imputing missing values. A comparison of the effect of the two imputation methods with respect to AUC is given in the following prediction section. Figure 5.10 gives an overview of the results when applied to the mouse data set.

5.4.3 Correlation of Regulatory Evidences and their Features

Despite the overlaps of the data sources, it is also of interest how well the created features correlate to each other in the training data. Figures 5.7 and 5.8 give an overview of these correlations.

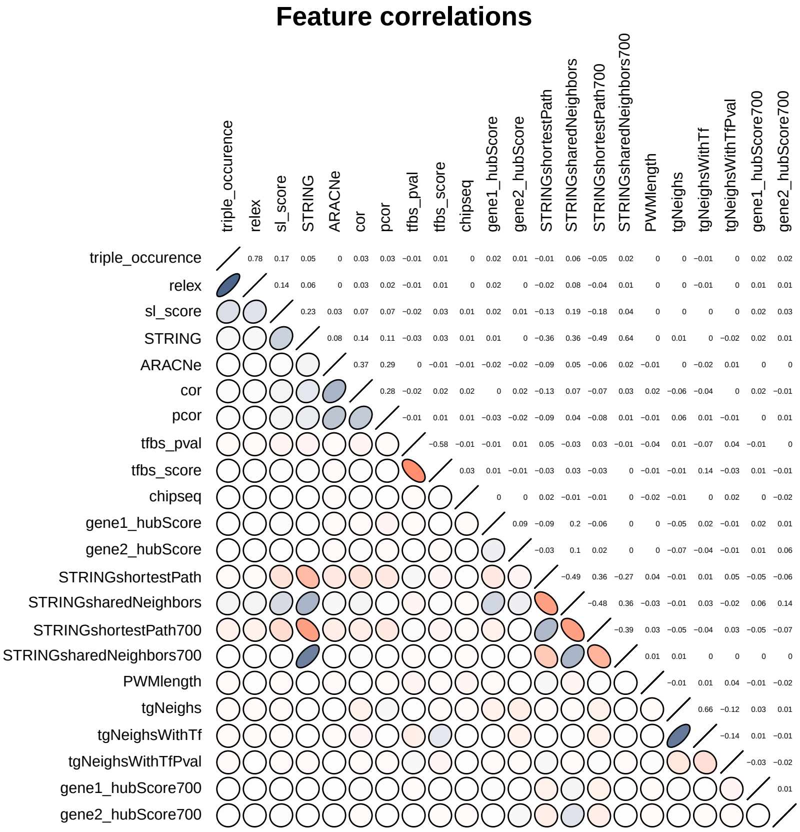


Figure 5.7: Numerical feature correlation on the mouse data set. This figure shows an all against all correlation comparison of the generated features. The color code ranges from red via white to blue. Red indicates anti-correlation and blue correlation. The flatter the shape of the circle, the stronger the absolute correlation. As expected, the strongest correlations occur between similar methods and the newly generated features within those methods. The microarray-based data sets for correlation, partial correlation and ARACNe have a strong positive correlation, as have the text mining methods (co-occurrence has been left out, since it would dominate the other methods, due to its low specificity) and the STRING based features. The generated features with the threshold of 700 on the STRING data set seem to be significantly different compared to the full data sets (shortest part, shared neighbors).

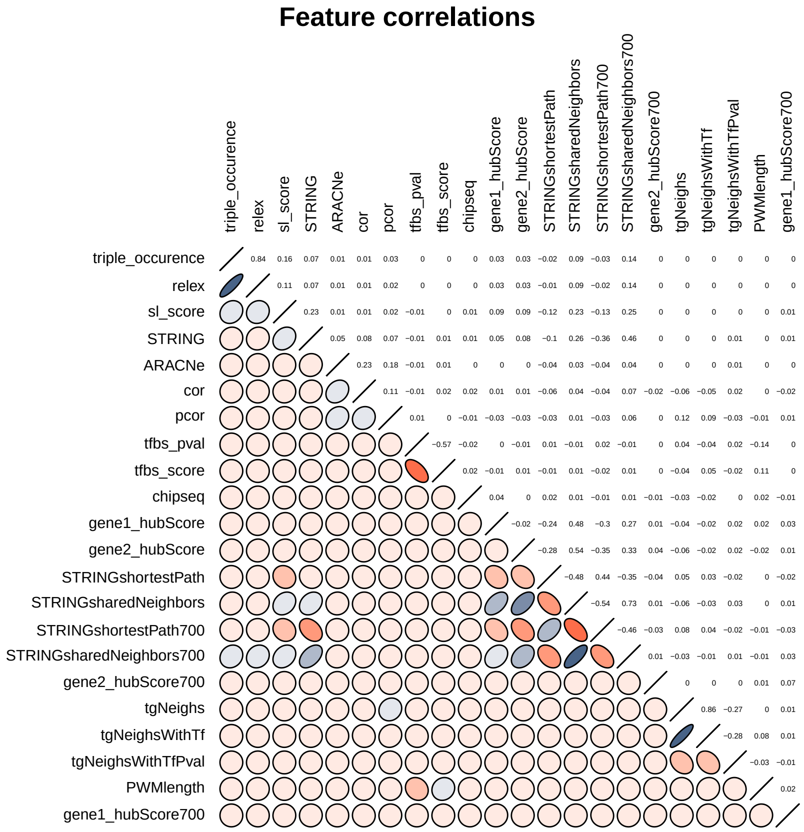


Figure 5.8: Numerical feature correlation on the human data set. The correlations are quite similar to the ones in Figure 5.7. Only some features correlate with stronger intensity, like the shortest path and shared neighbors features from STRING with a cutoff of 700 correlate stronger positively and negatively with the hub scores. A large hub score for at least one of the two observed genes possibly indicates shorter paths between those two genes. The SL text mining score correlates with the STRING score, while tri-occurrence and ReLex only correlate with the shared neighbors from STRING with a cutoff. Again, the microarray-based features do not seem to have a clear tendency towards another source of regulatory interactions.

In conclusion of the comparisons of the feature correlations, most of the features correlate with little overlap to each other, except for the closely related ones, which originate from the same data source. The pattern of correlations is very similar between the mouse and the human data set. For the predictions, it is expected that the features can all contribute very differently to the predictive performance and might also reflect

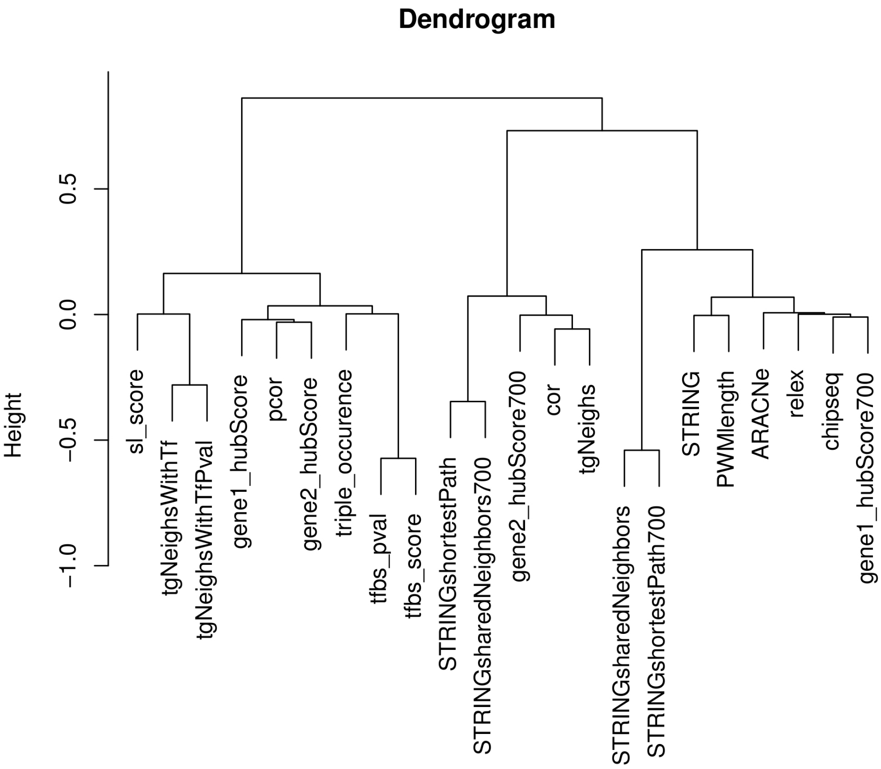


Figure 5.9: Hierarchical clustering of feature correlation on the human data set. In addition to the correlation analysis in Figure 5.8, a hierarchical clustering was applied to the correlation values. The clustering identified three major groups. The left group contains both text mining methods (SL and tri-occurrence), the original TFBS prediction scores and the gene specific hub scores from STRING. The group in the middle, mostly STRING-based features, except for the correlation, and the right group is the most diverse group, containing the STRING score, a text mining method (ReLex), ChIP-seq and also a microarray based feature (ARACNe). The latter ones do not seem to group specifically with each other.

different perspectives on the regulatory interactions. This could result in a lower coverage of single interactions by evidences from multiple data sources. In the previous section, it has already been stated that there exists only a limited overlap of multiple regulatory evidences for the interactions from the gold standard (database evidence). The clustering in Figure 5.9 shows three potential groups, but should be used with care due to the low overall correlations. The microarray-based features are split over all three groups, while the text mining approaches cluster together with the binding site predictions and the hub scores (left group). The STRING, ChIP-seq and ARACNe results form another group (right group) with other, mostly STRING-based features.

The following section describes how the different features can be used to make predictions of regulatory interactions.

5.4.4 Predictions of Regulatory Evidences and Performance Evaluation

The predictions were done with random forests using the *caret* R package [47]. The number of random features for the prediction was set to ten and the number of trees to 500. 10-fold crossvalidation was used on the data set to assess the performance (AUROC) of the prediction. Three different pre-processing methods were applied to the data set to overcome the unevenly distributed classes and the missing values (described in the feature imputation section 5.4.2).

Figure 5.10 shows a comparison between the random forest predictions for the mouse data set with and without down-sampling of the majority class, as well as the effect of GBM feature imputation and imputation with the majority class. Additionally, four other approaches and random guessing of the interactions was evaluated against the random forest performance. Those approaches were neural networks (NN), naive Bayes (NB), k-nearest neighbors (KNN) and an unsupervised approach (SUM). For the latter, the scaled sum of all features was calculated for each regulatory interaction in the data set. Random forests in general also perform well on data with unevenly distributed classes. This is due to the random selection of attributes for each decision tree and the use of samples that are drawn by bootstrapping (n samples with replacement). Because of the much larger size of the human data set, down-sampling of the data was applied to account for computational complexity. For the mouse data set, it was also tested if down-sampling reduces the predictive performance (see green bar in Figure 5.10). Sampling was done in ten iterations with sampling a class distribution of 1200 to 800 (negative to positive) examples. The performance was similar to the predictions on the full data set with random forests and therefore, sampling was applied for the calculations on the human data set. Additionally, the originally two Boolean features denoting if the interacting genes are a TF or not, were combined into one Boolean feature (true if at least one feature is a TF and false otherwise). The results from the feature importance analysis on mouse in the next section (see Figure 5.12) indicated that the two single features already had a stronger impact on the result.

The calculations on the mouse data set reached an AUROC of 0.87 and of 0.95 on human. The corresponding ROC and precision-recall curves are shown in Figures 5.11. The predictions show that the features can be used to train a predictive model. They also give a good example for the importance of carefully choosing the evaluation metrics. AUROC values offer a good possibility to compare different methods, but especially for the case of human and mouse, it can be seen that even though the predictions in human have a larger AUROC, the performance is lower with respect to precision and recall. The reason for this is the larger data set for human, which has of course an order of magnitudes more possible interactions. The more of these mostly not existing interactions the prediction method manages to classify as TN, the larger the AUROC. For the case of a great majority of not existing interactions, classifying an interaction as TN is quite easy.

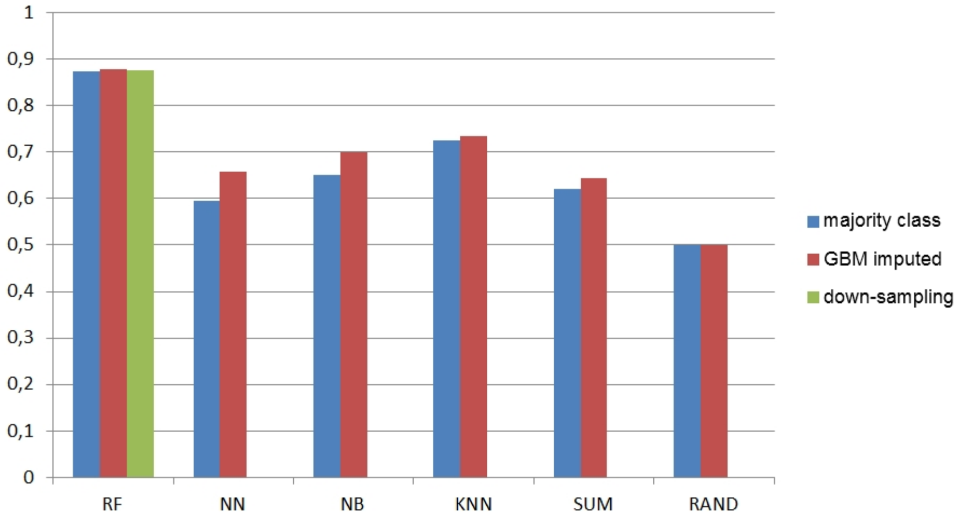


Figure 5.10: Comparison of different prediction and imputation methods on the mouse data set. This barplot shows the different outcomes of the imputation methods along with different prediction methods. The performance is measured with AUC. The applied methods from left to right are random forest, neural networks, naive bayes, k-nearest neighbors, unsupervised approach, which uses the sum of the scaled features and random guessing of interactions. On all methods, the GBM imputation (red bars) increases the performance. Random forests outperform the other prediction methods regardless of the imputation method. Additionally, the effect of down-sampling the majority class (green bar) was evaluated, since calculations on a larger data set, like human, is computationally expensive. The performance was comparable to the calculations on the full data set.

5.4.5 Feature Importance

In the following, the importance of the single features is evaluated. This is supposed to give insights as to which features should be generated for future predictions, or which

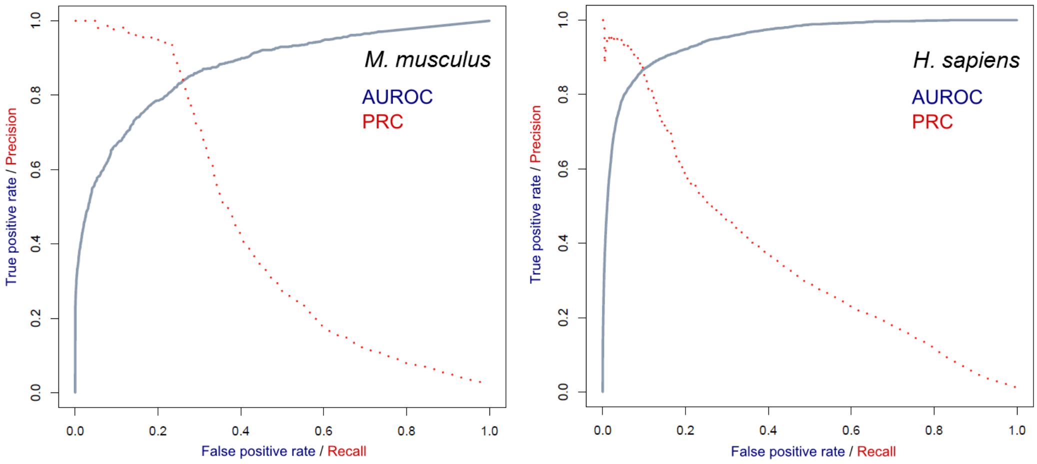


Figure 5.11: Comparison of AUROC and precision and recall curves. This plot gives an overview of the results of the predictions with random forest on the two data sets. The blue line indicates the AUROC curve with the true positive rate (also sensitivity) on the y-axis and the false positive rate (1-specificity) on the x-axis. The red dotted line refers to the corresponding precision-recall (PR) curve. Considering only the AUROC curve, the prediction on the human data reaches a larger AUC value than the prediction on the mouse data. Still, the results of the PR curve show that both methods perform nearly equally well and that the prediction on mouse is superior for the regions of lower recall.

features are obsolete, especially if they are computationally exhaustive (like the TG (neighbors + TF) pVal feature). The *randomForest* R package also allows to calculate the feature importance by calculating the class-specific contribution of each feature with a mean raw importance score and additionally the overall mean decreased accuracy. The larger the value, the more important the particular feature for both values. The mean decreased accuracy measures the effect of including a particular feature in the model and how much this reduces the classification error. The same applies to the raw importance score, just only calculated separately for each class (positive and negative or interacting and not interacting). The results were sorted by the mean decreased accuracy and are shown for the top 23 features in Figures 5.12 and 5.13.

An obvious insight from these calculations is the importance of the TF information. The combined TF feature in human is ranked for both classes at the top. In contrast to this, the information about the gene type had nearly no influence on the predictions. The ChIP-seq feature is of high importance, especially for the identification of the negative class. The microarray based features are inconclusive and strongly vary between the predictions on the human and mouse data. For both data sets, one of these features is ranked on the fourth position for identifying the positive class. The different positions of ARACNe could indicate that it depends on the filtering or on the included experiments for these features. The text mining approaches also performed quite differently and a more thorough study on these results should be done. A possible explanation could be the stronger overlap with STRING (see Figures 5.3 and 5.2), which could cause the algorithms to favor the impact of evidences from STRING over text mining. This could also explain the differences in the importance of the STRING results. Especially for human, the STRING based features are ranked at the top. The hub score seems to play an important role and it could be also interesting for the following calculations to combine this score into one feature. For the mouse data set, the hub score seems to be of greater importance for predicting the negative class. The TFBS score also plays an important role for this data set. On human, it is still important but ranked only on the 10th position. A reason could be the low specificity of some of the TFBS and of course the potentially large number of falsely predicted binding sites.

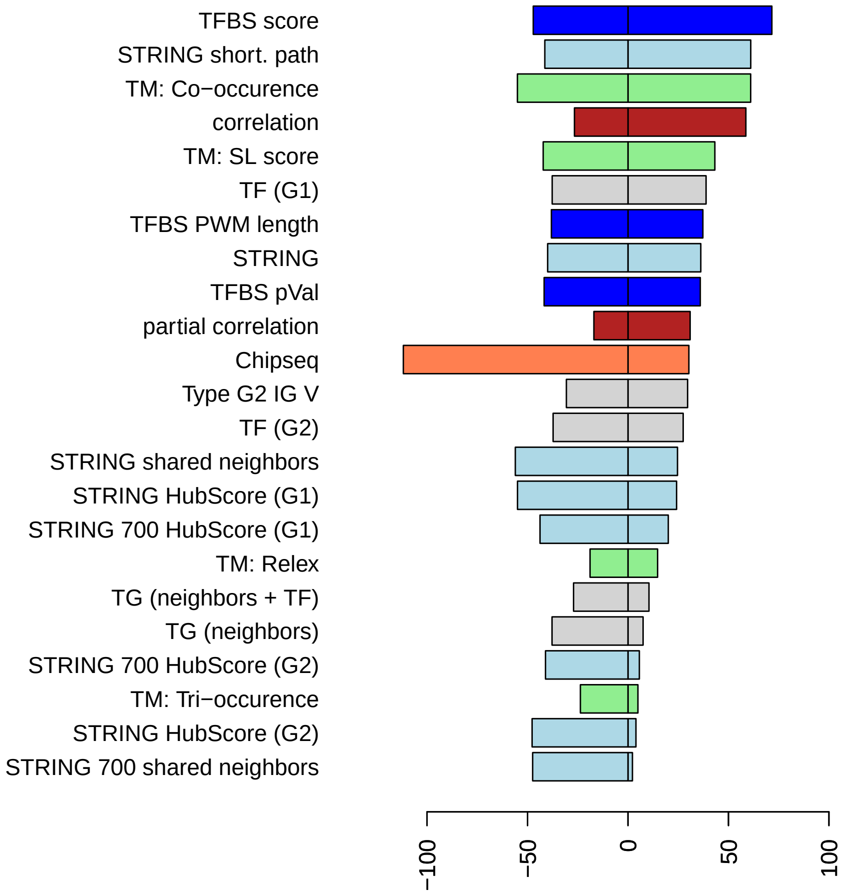


Figure 5.12: Feature importance for random forest predictions on the mouse data set. To assess the contribution of each feature to the predictions we used the *randomForest* R package to calculate the mean decrease in accuracy for each class-specific feature. The larger the value of a feature, the more important its contribution to the prediction. The right side of the plot shows the values for the positive class (interaction) and the left side represents the negative class (no interaction). Text mining features are colored in green, TFBS based features in blue, STRING based features in light blue, microarray-based features like correlation are red, ChIP-seq feature in orange and the remaining features, which were either extracted from databases or are a mixture of other features (like TFBS and partial correlation) are colored light grey. A mixture of all different data sources is ranked among the top eleven features (considering the positive class). ChIP-seq is particularly important for identifying the negative class. The information if a gene is a TF has a high rank for both classes. Most of the generated STRING features are ranked low for identifying the positive class but are at the same time a good indicator for the negative class. The TFBS predictions and the text mining approach are performing well for both classes.

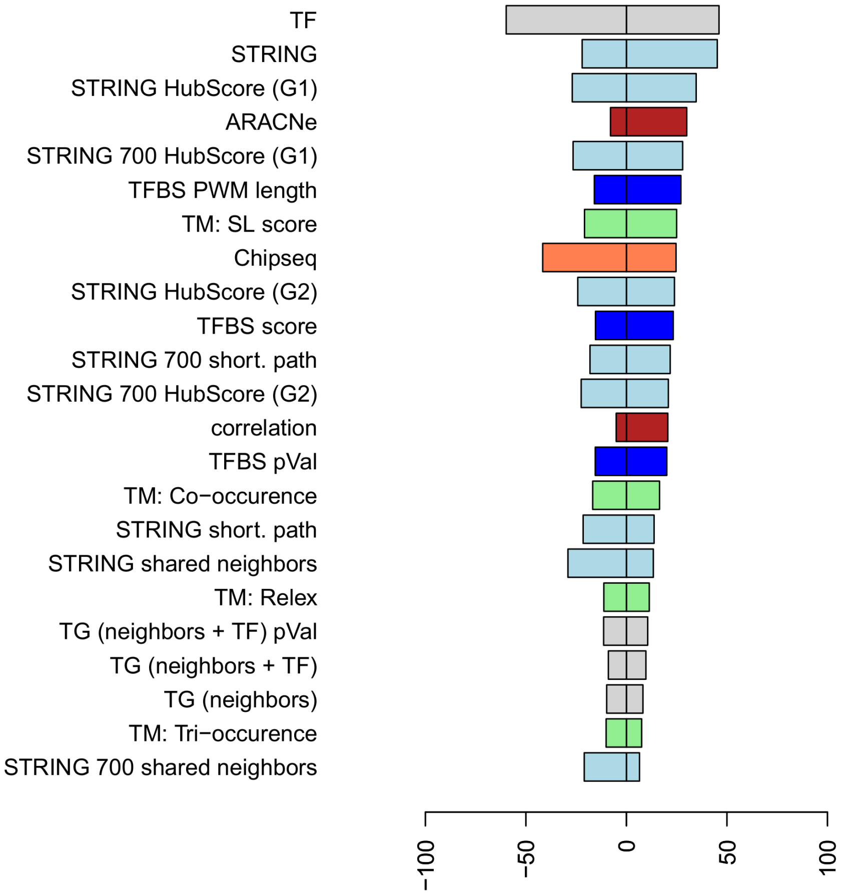


Figure 5.13: Feature importance for random forest predictions on the human data set. The merged feature “TF” was ranked first, for the positive and negative class. This is a quite intuitive result, since in interacting gene pairs usually one of the genes is a TF. In contrast to the predictions on the mouse data, the STRING features perform better and the text mining results seem to have less influence on the result. The ChIP-seq feature again plays an important role for identifying the negative class. Also, in contrast to the other prediction, the ARACNe feature is ranked at the top. Comparing the ARACNe feature in the Circos plots 5.3 and 5.2, the results for mouse seem to be much less restrictive and maybe a higher DPI for this data set would have removed less reliable results. For both predictions the TFBS PMW length is an important feature.

5.5 Discussion

The analysis of the different heterogeneous data sources was done from different perspectives starting with the simple descriptive visualization of the data. The visualization indicated that TFBS predictions come with a potentially huge amount of false positives and that they are of course limited to the known PWMs for the TFs. This is not a surprising insight but from the overall overview, it can be seen in which dimension the number of false positives might be compared to other prediction methods. Another insight is the high correlation of the text mining results to each other, with simple co-occurrence being the least restrictive. The latter data set has the largest overlaps to ARACNe and STRING. Furthermore, only a comparably small number of known interactions (database evidence) is available for the training of machine learning approaches. Still, for most of the collected data sets more than two sources of regulatory evidences exist for about half of the database evidences. The ChIP-seq data is supposed to be the most reliable source but, as mentioned earlier, also depends on the experimental setup. This data source has proportionally the largest overlap with the database evidences and additionally a larger overlap with the TFBS predictions. It might be of interest to further evaluate the origin of the large number of regulatory evidences, which are measured only by ChIP-seq. Another insight is that STRING offers the strongest overlap across all sources, but these overlaps are in most cases only specific to one other source and not confirmed by a third. Comparing the different organisms to each other with respect to their distribution of features, mouse and human are similar to each other, as are yeast and fruit fly. The PPI network from STRING seems to be proportionally much bigger for the latter two organisms compared to the other available data sources.

Following the approach by Bauer *et al.* [10], additional features were generated, which should combine different base features and also statistical tests. The p-value describing for each pair of TF-TGs the probability that the same result would have been also achieved for a set of TF-TGs with their strongly correlated neighbors, which have also a hit from the same TF, is unfortunately of limited use for the predictions. The combination of TFBS predictions and co-expression was supposed to reduce the noise in the TFBS predictions and also in the microarray data. In contrast to this, the generated STRING features performed well according to their feature importance. The shortest path and the hub score seem to be supporting indicators for predicting interactions. These two features also strongly correlate and probably relate to each other. A gene with a large hub score, hence, many interactions, probably has also a shorter path to all other genes. Short paths and highly connected genes might also originate in the same pathway and, thus, share a similar regulation. Another surprisingly well performing feature was the TFBS PWM length feature, which seems to reduce the noise of the TFBS predictions and is a relevant feature for the positive and the negative class. A longer PWM binding site is usually more specific and less likely to be hit by chance in another promoter sequence.

The predictions on mouse and human show promising results and are capable of inferring the TF-TG interactions from the gold standard. Random forests outperformed the other methods on the mouse data set and also the GBM feature imputation could improve the predictions. For the mouse data set it was also shown that even with down-sampling, the majority class the predictions performed equally well. For computational reasons down-sampling was also applied to the human data calculations.

5.6 Outlook

The previous sections introduced a framework for analyzing and collecting a heterogeneous data set for multiple organisms. Each data source has its own important preprocessing steps and offers different perspectives on the GRN of an organism. The framework presented here should encourage analysis of more complex data sources and the generation of new features from the single data sets and their combinations.

Additionally, the following suggestions and insights arise from this analysis. The integration of different data sources is already possible with state of the art approaches and for different organisms these data sources contribute differently to the predictions and should be handled accordingly. The presented framework offers an initial descriptive visualization, which allows to explore the collected data sets and also to evaluate the possibilities to combine or further refine features. An example for the latter are the hub score and the shortest path feature from the STRING data. Both features contributed to the predictive performance positively.

The combination of the data sources into more complex features did not improve the predictions, but several other combinations are suggested from the analysis, like the combination of TFBS predictions with ChIP-seq or text mining and STRING.

Next steps should include further assessing the performance of each data source and studying single overlaps, like TFBS predictions and ChIP-seq or the overlap of text mining results and STRING. Furthermore, predictions on the full data sets should be evaluated and could be used to find new functional annotations for related genes. Another interesting extension of this approach might be to compare the conserved genes between organisms and their local network structures or to predict the next state of a

GRN network, given a new experiment.

The already available data sets offer an amazing variety of possible analyses and opportunities for the development of new methods and frameworks. Still, the most challenging task lies in the data collection and one major goal for the future is the automatization of data integration processes.

CHAPTER 6

Conclusion

The here presented work contributes to both the field of development of new methods for the inference of GRNs, as well as the application on real-world data sets by gaining insight into biological processes in human urothelium cells. A key statement of this thesis is the importance of combining the technical understanding, such as machine learning and data mining methods, and deeper domain knowledge for the inference of gene regulatory networks.

Achieving this goal is not a process of following a classical water fall model, where everything can be defined and planned in the beginning. It is rather an iterative and agile process which requires close collaboration between biologists and bioinformaticians. The insights of Chapter 4 are the result of an iterative approach towards the understanding of the interdependencies in the normal human urothelium time series experiments. These iterations consist of several steps like the quality assessment, a descriptive as well as predictive exploration and visualization of the data and the successive evaluation of the intermediate results with the biologists as well as updates on new insights from the lab. The choice for this analysis felt on a rather simple, yet effective, data driven method which could help to identify ELF3 as a new regulation factor for human urothelium. The predictions were validated in the lab and the basis for a set of new microarray time series experiments. These experiments are the logical subsequent step in the iterative process and aim to extend the role of the ELF3 regulation in more detail. The contributions of this analysis are the adaptation of an iterative and agile analysis approach and the development of a novel data driven ranking procedure, which is capable of working also on short and noisy time series data (e.g. asynchronous or missing time points). Another contribution is the implementation of an interactive time series visualization and of course the identification of the new regulator gene in human urothelium. The visualization helped to discuss the potentially interesting gene sets and to further refine the final set of genes, which were supposed to be highly relevant for the development of human urothelium.

Despite the data and process driven considerations, it is also important to improve or to develop new methods for the inference of GRNs. Gene expression time series offer a great opportunity to analyze gene regulatory systems but of course it is necessary to gain as much detailed insight as possible for answering the question of interest from the given data. Questions such as the search for a regulatory factor driving the observed process, a subset of interacting genes, the ranking of possible targeted genes or even the whole GRN structure. My contributions to this field are outlined in detail in Chapter 3. Two novel GRN inference approaches are presented there, which deal with the data pre-processing, discretization, the developing of new algorithms, as well as their evaluation on synthetic and real world data sets. The first approach deals with the task of learning undirected, also called co-expression, networks. I adapted the concept of Dynamic Time Warping (DTW) to align pairs of gene expression time series profiles and to account for asynchronous time series. The time shifts can be caused by slower or faster evolving processes for different genes or by quality issues from the data generation itself. I developed an approach, which combines an angle based time series discretization and DTW for the alignment of pairs of gene expression profiles. Additionally, I developed a Stochastic Local Search based supervised extension of this approach to infer an organism specific distance matrix for the alignments. The training of this specific distance matrix is done by using known interacting genes and their profiles to find a distance matrix, which optimizes the alignment for the corresponding gene pairs. Both methods performed equally and even better than the compared methods, but the results with the standard distance matrix performs more stable over all benchmarks. The two versions of the DTW based approach run in parallel and can be applied to complete genomes. The second contribution to this chapter is an Bayesian approach, which integrates prior knowledge into the learning process and infers a directed GRN. As described in Section 3.2, biological networks tend to be scale-free and this tendency is used as prior for the Bayesian network inference. Each interaction in the inferred GRN has a score, which reflects the certainty for this prediction. The prior calculates an additional score for each gene which reflects the tendency to have more or less interactions partners. The parameters for the GRN were inferred by applying a Hybrid-Monte-Carlo sampling procedure to explore also different possible states of the GRN and additionally, conjugate gradient optimization was implemented for the optimization on large-scale networks. The performance was evaluated for the inference of GRNs from given synthetic and gene expression time series experiments. Additionally, the capability of identifying hubs was evaluated against other GRN inference methods, which it also outperformed.

The various properties and conditions of biological data sets can provide different limitations and challenges to the analysis. Gaining new knowledge about gene regulatory systems with experiments is an important task, but with the standardization of data generation like microarrays or next generation sequencing techniques, the need to get as much as possible of information from each experiments requires to advance the methods but also to find an efficient way of integrating the immense amount of already existing data and knowledge from public databases. Chapter 5 reviews five different concepts for the integration of different data sources into the learning process of GRNs. To study the different aspects of large-scale data integration, various sources for regulatory predictions and experiments were collected for four different organisms (and additionally *E. coli* and *A. thaliana*, which are not presented in this thesis). The generation of these multi-source data sets is also one of the contributions. Each source has to be pre-processed and integrated, or even predicted, like it is the case for TFBS. Additionally, a set of different features was created, such as the hub score for proteins in the STRING network or the p-value, which combines the TFBS predictions with the correlation values from gene expression microarrays. The second contribution of this analysis is the descriptive visualization and analysis of the collected data, which allows to compare the complete data set and also the different organisms to each other. Finally, the data sets for two eukaryotic organisms – human and mouse – were used to train a random forest model. The GRNs of both of these organisms have so-far been rarely analyzed on a large-scale due to the lack of data or gold standards. The performance evaluation showed that the trained model is capable of predicting gene regulatory interactions and allowed to weigh the importance of the single data sources and generated features. One insight of this analysis was the importance of the information about hub genes, which might not only be important TFs, but also influence regulation on the PPI level.

In summary all four approaches contribute to each other. Improving the methods helps to make more reliable predictions. Defining a suitable workflow for the analysis helps to understand and integrate new insights into the already known structure of GRNs. Using different sources of data helps to analyze the GRN from different perspectives and is a valuable extension to the time series analysis methods.

CHAPTER 7

Outlook

As mentioned in the previous chapter the analysis of the human urothelium cell lines, resulted in valuable insights into the underlying regulatory process. Following the suggested analysis workflow, follow-up experiments were designed and executed. These experiments are again microarray time series experiments and work with cell lines with knockdowns of the identified TF ELF3. Furthermore, the analysis on normal human urothelium is intended to be combined with another project in collaboration with the Institut Curie in Paris. This project focuses on the corresponding NHU cancer cell lines. The interactive visualizations have already proven to be a valuable tool for the discussions of potentially interesting genes. It is intended to extend this tool with additional features, such as gene ontology information or the integration of the predictions on the different gene expression data sets, or a more generic framework, which allows to upload own gene expression time series for further analysis.

Concerning the novel approaches, the DTW based approach could be also extended to infer direct interactions, if the direction of the shifts from the alignment matrices were also considered. Additionally, the effects of the organism specific distance matrix should be explored in more detail and maybe restricted to experiment specific matrices. Another approach could aim to find clusters of gene expression profiles and to compare their available annotated pathways to each other.

The Bayesian inference approach could be extended to also work on continuous variables and the impact of different discretization methods studied in more detail. Additionally, the approach offers a great opportunity to be extended towards the integration of different data sources. A possible application for this approach could be the data from Chapter 5. The posterior distribution could again use a sigmoid function to combine the regulatory evidences to train the weight for the directed interactions between any two genes. The prior could infer an additional data source or feature specific weight. This weight can be used to rank the overall importance of a feature to the GRN inference process. The posterior could again follow a normal distribution and the prior a Gamma distribution. The latter expects to have some very important features and a larger majority of the features, which contribute only little to the identification of gene regulatory interactions. These considerations already extended the functions from Section 3.2, but the different scale types, ordinal and nominal variables, have to be dealt with first.

The analysis of Chapter 5 shows first promising results concerning the integration of heterogeneous data sources. A next step could be to apply the trained random forest models on the full data sets of human and mouse. These predictions could then be used to help developing new experiments or at least to prioritize a set of potentially interesting genes. Additionally, the inferred GRN structure could be assessed for known models and of course compared to the insights of Chapter 4. Part of the database has already been visualized and it is possible to interactively explore the data [115], and another next step could be to visualize the whole data along with the new predictions. Finally, the most obvious extensions of this approach would be to integrate further data sets, such as gene ontology information, more recent gene expression data sets or DNase footprinting experiments.

List of Figures

List of Tables

Bibliography

[1] Aach, J., and Church, G. Aligning gene expression time series with time warping algorithms. *Bioinformatics 17*, 6 (2001), 495–508.

[2] Akutsu, T., Miyano, S., and Kuhara, S. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Proc. Pac. Symp. Biocomput. 4* (1999), 17–28.

[3] Albert, I., Thakar, J., Li, S., Zhang, R., and Albert, R. Boolean network simulations for life scientists. *Source Code Biol. Med. 3*, 16 (2008).

[4] Andreoli, J. M., Jang, S., Chung, E., Coticchia, C. M., Steinert, P. M., and Markova, N. G. The expression of a novel, epithelium-specific ets transcription factor is restricted to the most differentiated layers in the epidermis. *Nucleic Acids Res. 25* (1997), 4287–4295.

[5] Arnone, M., and Davidson, E. The hardwiring of development: Organization and function of genomic regulatory systems. *Development 124* (1997), 1851–1864.

[6] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M.and Rubin, G. M., and Sherlock, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet. 25*, 1 (2000), 25–29.

[7] Bansal, M., Belcastro, V., Ambesi-Impiombato1, A., and di Bernardo, D. How to infer gene networks from expression profiles. *Mol. Syst. Biol. 3*, 78 (2007), 1–10.

[8] Bansal, M., Della Gatta, G., and di Bernardo, D. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics 22* (2006).

[9] Barabasi, A., and Oltvai, Z. Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet. 5*, 101–113.

[10] Bauer, T., Eils, R., and König, R. RIP: the regulatory interaction predictor - a machine learning-based approach for predicting target genes of transcription factors. *Bioinformatics 27*, 16 (2011).

[11] Bell, S. M., Zhang, L., Mendell, A., Xu, Y., Haitchi, H. M., Lessard, J. L., and Whitsett, J. A. Kruppel-like factor 5 is required for formation and differentiation of the bladder urothelium. *Dev. Biol. 358*, 1 (2011), 79–90.

[12] Bellman, R., and Kalaba, R. On adaptive control processes. *IRE Trans. Autom. Control 4*, 2 (1959), 1–9.

[13] Bernard, A., and Hartemink, A. Informative Structure Priors: Joint Learning of Dynamic Regulatory Networks from Multiple Types of Data. In *In Proc. Pac. Symp. Biocomput. (PSB05)* (New Jersey, 2005), R. Altman, A. Dunker, L. Hunter, T. Jung, and T. Klein, Eds., World Scientific, pp. 459–470.

[14] Bernardo, D., Thompson, M., Gardner, T., Chobot, S., and Eastwood, E. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene. *Nat Biotechnol 23* (2005).

[15] Bishop, C. M. Bayesian PCA. In *Advances in Neural Information Processing Systems* (1998), vol. 11, MIT Press, pp. 382–388.

[16] Blockeel, H., and Raedt, L. D. Top-down induction of first-order logical decision trees. *Artif. Intell. 101*, 1–2 (1998).

[17] Böck, M., Hinley, J., Schmitt, C., Wahlicht, T., Kramer, S., and Southgate, J. Hub-Centered Gene Network Reconstruction using Automatic Relevance Determination. *Dev. Biol. 386*, 2 (2014), 321–330.

[18] Böck, M., Ogishima, S., Tanaka, H., Kramer, S., and Kaderali, L. Hub-Centered Gene Network Reconstruction using Automatic Relevance Determination. *PLoS ONE 7*, 5 (2012), e35077.

[19] Böck, M., Schmitt, C., and Kramer, S. A Study of Dynamic Time Warping for the Inference of Gene Regulatory Relationships. *In: Proceedings of the German Conference on Bioinformatics (GCB11)* (2011).

[20] Boglev, Y., Wilanowski, T., Caddy, J., Parekh, V., Auden, A., Darido, C., Hislop, N. R., Cangkrama, M., Ting, S. B., and Jane, S. M. The unique and cooperative roles of the Grainy head-like transcription factors in epidermal development reflect unexpected target gene specificity. *Dev. Biol. 349*, 2 (2011), 512–522.

[21] Bonneau, R., Reiss, D., Shannon, P., Facciotti, M., Hood, L., Baliga, N., and Thorsson, V. The Inferelator: An Algorithm for Learning Parsimonious Regulatory Networks from Systems-Biology Data Sets de novo. *Genome Biol. 7* (2006), 36.

[22] Brouard, C., Vrain, C., Dubois, J., Castel, D., Debily, M.-A., and d'Alche Buc, F. Learning a Markov Logic network for supervised gene regulatory network inference. *BMC Bioinf. 14* (2013), 273.

[23] Butte, A. J., and Kohane, I. S. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Proc. Pac. Symp. Biocomput.* (2000), 418–429.

[24] Cantone, I., Marucci, L., Iorio, F., Ricci, M., Belcastro, V., Bansal, M., Santini, S., di Bernardo, M., di Bernardo, D., and Cosma, M. A Yeast Synthetic Network for In Vivo Assessment of Reverse-Engineering and Modeling Approaches. *Cell*, 137 (2009), 172–181.

[25] Castelo, R., and Roverato, A. Reverse engineering molecular regulatory networks from microarray data with qp-graphs. *Comput. Biol. 16* (2009), 213–227.

[26] Chen, T., Filkov, V., and Skiena, S. Identifying Gene Regulatory Networks from Experimental Data. *Parallel Comput. 27* (2001), 141–162.

[27] Chen, T., He, H., and Church, G. Modeling gene expression with differential equations. *Proc. Pac. Symp. Biocomput. 4* (1999), 29–40.

[28] Cheng, C., Yan, K.-K., Hwang, W., Qian, J., N., B., Rozowsky1, J., Lu, Z. J., Niu, W., Alves, P., Kato, M., Snyder, M., and Gerstein,

M. Construction and Analysis of an Integrated Regulatory Network Derived from High-Throughput Sequencing Data. *PLoS Comput. Biol. 7*, 11 (2011).

[29] Cho, R., Campbell, M., Winzeler, E., and Steinmetz, L. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell 2*, 1 (1998), 65–75.

[30] Conesa, A., Nueda, M. J., Ferrer, A., and Talon, M. maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics 22* (2006), 1096–1102.

[31] Cross, W. R., Eardley, I., Leese, H. J., and Southgate, J. A biomimetic tissue from cultured normal human urothelial cells: analysis of physiological function. *Am. J. Physiol-Renal 289*, 2 (2005), F459–68.

[32] Csaba, G. syngrep - Fast synonym-based named entity recognition, Personal communication. *LMU, Munich* (2008).

[33] Csardi, G., and Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Systems* (2006), 1695.

[34] Davis, J., and Goadrich, M. The relationship between precision-recall and ROC curves. *In: Proceedings of the 23rd international conference on Machine learning 36* (2006), 223–240.

[35] DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics 44*, 3 (1988), 837–845.

[36] Di Camillo, B., Toffolo, G., and Cobelli, C. A gene network simulator to assess reverse engineering algorithms. *Ann. N.Y. Acad. Sci. 1158* (2009), 125–142.

[37] Duane, S., Kennedy, A., Pendleton, B., and Roweth, D. Hybrid Monte Carlo. *Phys. Lett. B 195* (1987), 216–222.

[38] Eisen, M. B., Spellman, P. T., Brown, P. O., and D., B. Cluster analysis and display of genome-wide expression patterns. *PNAS 95*, 25 (1998), 14863–14868.

[39] Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., and Gardner, T. S. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol. 5*, 1 (2007), e8.

[40] Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett. 27* (2006), 861–874.

[41] Feizi, S., Marbach, D., Medard, M., and Kellis, M. Network deconvolution as a general method to distinguish direct dependencies in networks. *Nat. Biotechnol. 31*, 8 (2013).

[42] Fellenberg, K., Hauser, N., Brors, B., Neutzner, A., Hoheisel, J., and Vingron, M. Correspondence analysis applied to microarray data. *Proc. Natl. Acad. Sci. U.S.A. 98* (2001), 10781–10786.

[43] Fellows, G. J., and Marshall, D. H. The permeability of human bladder epithelium to water and sodium. *Invest. Urol. 9* (1972), 339–344.

[44] Fleming, J. M., Shabir, S., Varley, C. L., Kirkwood, L. A., White, A., Holder, J., Trejdosiewicz, L. K., and Southgate, J. Differentiation-associated reprogramming of the transforming growth factor beta receptor pathway establishes the circuitry for epithelial autocrine/paracrine repair. *PLoS One 7* (2012), e51404.

[45] Flentjar, N., Chu, P. Y., Ng, A., Johnstone, C. N., Heath, J. K., Ernst, M., Hertzog, P. J., and Pritchard, M. A. TGF-betaRII rescues development of small intestinal epithelial cells in Elf3-deficient mice. *Gastroenterology 132* (2007), 1410–1419.

[46] Friedman, J. H. Stochastic Gradient Boosting. *Comput. Stat. Data Anal. 38*, 4 (2002), 367–378.

[47] Friedman, J. H. Building Predictive Models in R Using the caret Package. *J. Stat. Softw. 28*, 5 (2008), 367–378.

[48] Friedman, N. Using Bayesian Networks to Analyze Expression Data. *J. Comput. Biol. 7* (2000), 601–620.

[49] Friedman, N. Inferring cellular networks using probabilistic graphical modes. *Science 303* (2004), 799–805.

[50] Fröhler, S., and Kramer, S. Inductive logic programming for gene regulation prediction. *Mach. Learn. 70* (2007).

[51] Fundel, K., Küffner, R., and Zimmer, R. RelEx - relation extraction using dependency parse trees. *Bioinformatics 23*, 3 (2007).

[52] Garcia, S., Luengo, J., Saez, J. A., Lopez, V., and Herrera, F. A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning. *IEEE Trans. Knowl. Data Eng. 25*, 4 (2013).

[53] Gardner, T., di Bernardo, D., Lorenz, D., and Collins, J. Inferring genetic networks and identifying compound mode of action via expression. *Science 301* (2003).

[54] Garthwaite, M., Thomas, D., Subramaniam, R., Stahlschmidt, J., Eardley, I., and Southgate, J. Urothelial differentiation in vesicoureteric reflux and other urological disorders of childhood: a comparative study. *Eur. Urol. 49* (2006), 154–160.

[55] Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. affy – analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics 20*, 3 (2004), 307–315.

[56] Giltay, J., van de Meerakker, J., van Amstel, H., and de Jong, T. No pathogenic mutations in the uroplakin III gene of 25 patients with primary vesicoureteral reflux. *J. Urol. 171* (2004), 931–932.

[57] Giorgino, T. dtw: Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. *Journal of Statistical Software*, 31(7) (2009), 1–24.

[58] Giuliano, C., Lavelli, A., and Romano, L. Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. *In Proc. EACL* (2006).

[59] Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A., and Tegner, J. Data integration in the era of omics: current and future challenges. *BMC Sys. Biol. 8* (2014).

[60] Guthke, R., Möller, U., Hoffman, M., Thies, F., and Töpfer, S. Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection. *Bioinformatics 21* (2005), 1626–1634.

[61] H. Hoos, H., and Stützle, T. *Stochastic Local Search - Foundation and Applications.* Elsevier, 2005.

[62] Hache, H., Wierling, C., Lehrach, H., and Herwig, R. GeNGe: systematic generation of gene regulatory networks. *Bioinformatics 25*, 9 (2009), 1205–1207.

[63] Halfon, M. S., Gallo, S. M., and Bergman, C. M. REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila. *Nucleic Acids Res. 36*, Database issue (2008).

[64] Hartemink, A., Gifford, D., Jaakkola, T., and Young, R. Bayesian Methods for Elucidating Genetic Regulatory Networks. *IEEE Intell. Syst. 17* (2002), 37–43.

[65] Hartigan, J. A., and Hartigan, P. M. The Dip Test of Unimodality. *The Annals of Statistics 13*, 1 (1985), 70–84.

[66] Harvey, I., and Bossomaier, T. Time out of joint: Attractors in asynchronous random Boolean networks. In *Proceedings of the Fourth European Conference on Artificial Life (ECAL97)* (1997), P. Husbands and I. Harvey, Eds., MIT Press, pp. 67–75.

[67] Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., and Guthke, R. Gene regulatory network inference: data integration in dynamic models - a review. *Biosystems 96*, 1 (2009), 86–103.

[68] Hicks, R. M. The fine structure of the transitional epithelium of rat ureter. *J. Cell Biol. 26* (1965), 25–48.

[69] Hu, P., Meyers, S., Liang, F. X., Deng, F. M., Kachar, B., Zeidel, M. L., and Sun, T. T. Role of membrane proteins in permeability barrier function: uroplakin ablation elevates urothelial permeability. *Am. J. Physiol. Renal Physiol. 283* (2002), F1200–1207.

[70] Huang, D. W., Sherman, B. T., and Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc. 4*, 1 (2009).

[71] Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics 4*, 2 (2003), 249–264.

[72] Jenkins, D., Bitner-Glindzicz, M., Malcolm, S., Hu, C. C., Allison, J., Winyard, P. J., Gullett, A. M., Thomas, D. F., Belk, R. A., Feather, S. A., Sun, T. T., and Woolf, A. S. De novo Uroplakin IIIa heterozygous mutations cause human renal adysplasia leading to severe kidney failure. *J. Am. Soc. Nephrol. 16* (2005), 2141–2149.

[73] Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., and von Mering, C. STRING 8 : a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res. 37*, suppl 1 (2009), D412–D416.

[74] Jeong, H., Tomber, B., Albert, R., Oltvai, Z., and Barabasi, A. The large-scale organization of metabolic networks. *Nature 407* (2000), 651–654.

[75] Jiang, S., Gitlin, J., Deng, F. M., Liang, F. X., Lee, A., Atala, A., Bauer, S. B., Ehrlich, G. D., Feather, S. A., Goldberg, J. D., Goodship, J. A., Goodship, T. H., Hermanns, M., Hu, F. Z., Jones, K. E., Malcolm, S., Mendelsohn, C., Preston, R. A., Retik, A. B., Schneck, F. X., Wright, V., Ye, X. Y., Woolf, A. S., Wu, X. R., Ostrer, H., Shapiro, E., Yu, J., and Sun, T. T. Lack of major involvement of human uroplakin genes in vesicoureteral reflux: implications for disease heterogeneity. *Kidney Int. 66* (2004), 10–19.

[76] Jung, Y. L., Luquette, L. J., Ho, J. W. K., Ferrari, F., Tolstorukov, M., Minoda, A., Issner, R., Epstein, C. B., Karpen, G. H., Kuroda, M. I., and Park, P. J. Impact of sequencing depth in ChIP-seq experiments. *Nucleic Acids Res.* (2014).

[77] Kaderali, L. *A Hierarchical Bayesian Approach to Regression and its Application to Predicting Survival Times in Cancer.* Shaker Verlag, Aachen, 2006. ISBN 978-3-8322-5216-8.

[78] Kaderali, L., Dazert, E., Zeuge, U., Frese, M., and Bartenschlager, R. Recontructing Signaling Pathways from RNAi Data using Probabilistic Boolean Threshold Networks. *Bioinformatics 25(17)* (2009), 2229–2235. doi:10.1093/bioinformatics/btp375.

[79] Kaderali, L., and Radde, N. Inferring Gene Regulatory Networks from Expression Data. In *Computational Intelligence in Bioinformatics*, A. Kelemen,

A. Abraham, and Y. Chen, Eds., vol. 94 of *Studies in Computational Intelligence*. Springer-Verlag, Berlin, 2008, pp. 33–74.

[80] Kauffman, S. Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol. 22* (1969), 437–467.

[81] Kauffmann, A., Gentleman, R., and Huber, W. arrayQualityMetrics – a Bioconductor package for quality assessment of microarray data. *Bioinformatics 25*, 3 (2009), 415–416.

[82] Kelly, H., Ennis, S., Yoneda, A., Bermingham, C., Shields, D. C., Molony, C., Green, A. J., Puri, P., and Barton, D. E. Uroplakin III is not a major candidate gene for primary vesicoureteral reflux. *Eur. J. Hum. Genet. 13* (2005), 500–502.

[83] Keogh, E. J., and Pazzani, M. J. Derivative Dynamic Time Warping. *Proceedings of the SIAM International Conference on Data Mining* (2001), 1–11.

[84] Khanin, R., and Wit, E. How scale-free are biological networks. *J. Comput. Biol. 13* (2006), 810–818.

[85] Küffner, R., Petri, T., Tavakkolkhah, P., Windhager, L., and Zimmer, R. Inferring gene regulatory networks by ANOVA. *Bioinformatics 28*, 10 (2012), 1376–1382.

[86] Kummerfeld, S. K., and Teichmann, S. A. DBD: a transcription factor prediction database. *Nucleic Acids Res. 34* (2006), D74–D81.

[87] Lavelle, J., Meyers, S., Ramage, R., Bastacky, S., Doty, D.and Apodaca, G., and Zeidel, M. Bladder permeability barrier: recovery from selective injury of surface epithelial cells. *Am. J. Physiol. Renal Physiol. 283* (2002), F242–253.

[88] Lewis, S. A. Everything you wanted to know about the bladder epithelium but were afraid to ask. *Am. J. Physiol. Renal Physiol. 278* (2000), F867–874.

[89] Li, F., Long, T., and Tang, C. The yeast cell-cycle network is robustly designed. *PNAS 101*, 14 (2004), 4781–4786.

[90] Li, Y., Liu, L., Bai, X., Cai, H., Ji, W., Guo, D., and Zhu, Y. Comparative study of discretization methods of microarray data for inferring transcriptional regulatory networks. *BMC Bioinf. 11*, 1 (2010).

[91] Liang, S., Fuhrman, S., and Somogyi, R. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. *Proc. Pac. Symp. Biocomput. 3* (1998), 18–29.

[92] Lima-Mendez, G., and van Helden, J. The powerful law of the power law and other myths in network biology. *Mol. Biosyst. 5* (2009), 1482–1493.

[93] Liu, H., Hussain, F., Tan, C. L., and Dash, M. Discretization: An Enabling Technique. *Data Mining and Knowledge Discovery 6* (2002).

[94] Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., and Allison, K. R. Wisdom of crowds for robust gene network inference. *Nat. Methods 9* (2012), 796–804.

[95] Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. Revealing strengths and weaknesses of methods for gene network inference. *PNAS 107*, 14 (2010), 6286–6291.

[96] Marbach, D., Schaffter, T., Mattiussi, C., Floreano, D., Mayer, E., Lafitte, F., and Bontempi, G. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *J. Comp. Biol. 16*, 2 (2009), 229–239.

[97] Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla-Favera, R., and Califano, A. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinf. 7* (2006), S7.

[98] Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., and Wingender, E. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res. 34*, Database issue (2006).

[99] Mayer, P. E., Lafitte, F., and Bontempi, G. minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information. *BMC Bioinf. 9* (2008), 461.

[100] Mazur, J., Ritter, D., Reinelt, G., and Kaderali, L. Reconstructing Nonlinear Dynamic Models of Gene Regulation using Stochastic Sampling. *BMC Bioinf. 10* (2009), 448.

[101] Mendes, P., Sha, W., and Ye, K. Artificial gene netwokrs for objective comparison of analysis algorithms. *Bioinformatics 19* (2003), ii122–ii129.

[102] Meyer, P. E., Kontos, K., Lafitte, F., and Bontempi, G. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J. Bioinf. Syst. Biol.* (2007).

[103] Meyer, P. E., Marbach, D., Roy, S., and Kellis, M. Information-Theoretic Inference of Gene Networks Using Backward Elimination. *BIOCOMP 1* (2010), 700–705.

[104] Montgomery, S. B., Griffith, O. L., Sleumer, M. C.and Bergman, C. M., Bilenky, M., Pleasance, E. D., Prychyna, Y., Zhang, X., and Jones, S. J. ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics 22*, 5 (2006).

[105] Mysorekar, I. U., Mulvey, M. A., Hultgren, S. J., and Gordon, J. I. Molecular Regulation of Urothelial Renewal and Host Defenses during Infection with Uropathogenic Escherichia coli. *J. Biol. Chem. 277*, 9 (2002), 7412–7419.

[106] Neal, R. *Bayesian Learning for Neural Networks*, vol. 118 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1996.

[107] Newman, M. The structure and function of complex networks. *SIAM Rev. 45* (2003), 167.

[108] Ng, A. Y., Waring, P., Ristevski, S., Wang, C., Wilson, T., Pritchard, M., Hertzog, P., and Kola, I. Inactivation of the transcription factor Elf3 in mice results in dysmorphogenesis and altered differentiation of intestinal epithelium. *Gastroenterology 122* (2002), 1455–1466.

[109] Obayashi, T., Okamura, Y., Ito, S., Tadaka, S., Motoike, I. N., and Kinoshita, K. COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals. *Nucleic Acids Res. 41* (2013), D1014–20.

[110] Oettgen, P., Alani, R. M., Barcinski, M. A., Brown, L., Akbarali, Y., Boltax, J., Kunsch, C., Munger, K., and Libermann, T. A. Isolation and characterization of a novel epithelium-specific transcription factor, ESE-1, a member of the ets family. *Mol. Cell Biol. 17* (1997), 4419–4433.

[111] Oksendal, B. *Stochastic differential equations: an introduction with applications.* Springer Science & Business Media, 2013.

[112] Olsen, C., Meyer, P. E., and Bontempi, G. On the Impact of Entropy Estimation on Transcriptional Regulatory Network Inference Based on Mutual Information. *EURASIP J. Bioinform. and Syst. Biol. 1* (2009), 308959.

[113] Pearl, J. Causal inference in statistics: An overview. *Statistics Surveys 3*, R-350 (2009), 96–146.

[114] Peng, H., Long, F., and Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell. 27*, 8 (2005), 1226–1238.

[115] Pesch, R., Böck, M., and Zimmer, R. ConReg: Analysis and Visualization of Conserved Regulatory Networks in Eukaryotes. *German Conference on Bioinformatics 2012 26* (2012).

[116] Pramila, T., Wu, W., Miles, S., Nobe, W., and Breeden, L. The forkhead transcription factor HCM1 regulates chromosome segregation and fills the s-phase gap in the transcriptional citcuitry of the cell cycle. *Genes Dev. 20* (2006), 2266–2278.

[117] Prill, R. J., Marbach, D., Saez-Rodriguez, J., Sorger, P. K., Alexopoulos, L. G., Xue, X., Clarke, N. D., Altan-Bonnet, G., and Stolovitzky, G. Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges. *PLoS ONE 5* (2010), e9202.

[118] Richardson, M., and Domingos, P. Markov Logic Networks. *Mach. Learn. 62*, 1–2 (2006).

[119] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinf. 12*, 1 (2011), 77.

[120] Ronen, M., Rosenberg, R., Shraiman, B. I., and Alon, U. Assigning Numbers to the Arrows: Parameterizing a Gene Regulation Network by Using Accurate Expression Kinetics. *PNAS 99*, 16 (2002), 10555–10560.

[121] Rowicka, M., Kudlicki, A., Tu, B. P., and Otwinowski, Z. High-resolution timing of cell cycle-regulated gene expression. *PNAS 104*, 43 (2007), 16892–16897.

[122] Rubenwolf, P., and Southgate, J. Permeability of differentiated human urothelium in vitro. *Methods Mol. Biol. 763* (2011), 207–222.

[123] Sakoe, H., and Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech Signal Process. 26* (1978), 43–49.

[124] Sales, G., and Romualdi, C. parmigene - a parallel R package for mutual information estimation and gene network reconstruction. *Bioinformatics 27*, 13 (2011), 1876–1877.

[125] Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., and Lenhard, B. JASPAR: an openâaccess database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res. 32*, Issue suppl 1 (2004), D91–D94.

[126] Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K. H. Pathway Interaction Database. *Nucleic Acids Res. 37*, Database issue (2009).

[127] Schaefer, U., Schmeier, S., and Bajic, V. B. TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins. *Nucleic Acids Res. 39*, Suppl 1 (2010), D106–D110.

[128] Schäfer, J., Opgen-Rhein, R., and Strimmer, K. Reverse engineering genetic networks using the GeneNet package. *R News 6*, 5 (2006), 50–53.

[129] Schrynemackers, M., Küffner, R., and Geurts, P. On protocols and measures for the validation of supervised methods for the inference of biological networks. *Front. Genet. 4*, 262 (2013).

[130] Shmulevich, I., Dougherty, E., and Zhang, W. From Boolean to probabilistic Boolean networks as models of genetic regulatory network. *Proceedings of the IEEE 90* (2002), 1778–1792.

[131] Southgate, J., Hutton, K. A., Thomas, D. F., and Trejdosiewicz, L. K. Normal human urothelial cells in vitro: proliferation and induction of stratification. *Lab. Invest. 71*, 4 (1994), 583–94.

[132] Southgate, J.and Masters, J. R., and Trejdosiewicz, L. K. Culture of human urothelium. *Culture of Epithelial Cells 2nd* (2002), 381–400.

[133] Spellman, P., Sherlock, G., and Zhang, M. Conprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell 9* (1998), 3273–3297.

[134] Spiegelhalter, D., Thomas, A., and Best, N. Computation on Bayesian graphical models. *Bayesian Statistics 5* (1996), 407–425.

[135] Stolovitzky, G., Prill, R. J., and Califano, A. Lessons from the DREAM2 Challenge. *Ann. N.Y. Acad. Sci. 1158* (2009), 159–195.

[136] Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., M., S., Muller, J., Bork, P., Jensen, L. J., and von Mering, C. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res. 39* (2011), D561–D568.

[137] Tan, V. Y. F., and Févotte, C. Automatic Relevance Determination for Nonnegative Matrix Factorization. In *SPARS09 - Signal Processing with Adaptive Sparse Structured Representations* (2009), pp. 1–19.

[138] Teixeira, M. C., Monteiro, P., Jain, P., Tenreiro, S., Fernandes, A. R., Mira, N. P., Alenquer, M., Freitas, A. T., Oliveira, A. L., and Sa-Correia, I. The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Res. 34*, Database issue (2006), D446–D451.

[139] The. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature 489* (2012).

[140] Thomas-Chollier, M., Defrance, M., Medina-Rivera, A., Sand, O., Herrmann, C., Thieffry, D., and van Helden, J. RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Res. 39*, Web Server issue (2011), W86–91.

[141] Tong, S., and Koller, D. Support Vector Machine Active Learning with Applications to Text Classification. *J. Mach. Learn. Res. 2* (2002).

[142] Tu, B. P., Kudlicki, A., Rowicka, M., and McKnight, S. L. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science 310* (2005), 1152–1158.

[143] Van den Bulcke, T., Van Leemput, K., Naudts, B., van Remortel, P., Ma, H., Verschoren, A., De Moor, B., and Marchal, K. SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinf. 7*, 1 (2006), 43.

[144] Van Someren, E., Wessels, L., Reinders, M., and Backer, E. Searching for limited connectivity in genetic network models. In *Proc. 2nd Intl. Conf. Sys. Biol.* (Pasadena, California, 2001), pp. 222–239.

[145] Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet. 10*, 4 (2009), 252–263.

[146] Varley, C., Garthwaite, M., Cross, W., Hinley, J., Trejdosiewicz, L., and Southgate, J. PPARgamma-regulated tight junction development during human urothelial cytodifferentiation. *J. Cell Physiol. 208* (2006), 407–417.

[147] Varley, C., Stahlschmidt, J., Smith, B., Stower, M., and Southgate, J. Activation of peroxisome proliferator-activated receptor-gamma reverses squamous metaplasia and induces transitional differentiation in normal human urothelial cells. *Am. J. Pathol. 164* (2004), 1789–1798.

[148] Varley, C. L., Bacon, E. J., Holder, J. C., and Southgate, J. FOXA1 and IRF-1 intermediary transcriptional regulators of PPARgamma-induced urothelial cytodifferentiation. *Cell Death Differ. 16*, 1 (2009), 103–114.

[149] Varley, C. L., Stahlschmidt, J., Lee, W. C., Holder, J., Diggle, C., Selby, P. J., Trejdosiewicz, L. K., and Southgate, J. Role of PPARgamma and EGFR signalling in the urothelial terminal differentiation programme. *J. Cell Sci. 117* (2004), 2029–2036.

[150] Werhli, A. V., and Husmeier, D. Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Multiple Sources of Prior Knowledge. *Stat. Appl. Genet. Mol. Biol. 6*, 1 (2007).

[151] Westermann, F., Muth, D., Benner, A., Bauer, T., Henrich, K. O., Oberthuer, A., Brors, B., Beissbarth, T., Vandesompele, J., Pattyn, F., Hero, B., König, R., Fischer, M., and Schwab, M. Distinct transcriptional MYCN/c-MYC activities are associated with spontaneous regression or malignant progression in neuroblastomas. *Genome Biol. 9*, 10 (2008), R150.

[152] Woolf, P., and Wang, Y. A fuzzy logic approach to analyzing gene expression data. *Physiol. Genomics 3* (2000), 9–15.

[153] Wu, X. R., Kong, X. P., Pellicer, A., Kreibich, G., and Sun, T. T. Uroplakins in urothelial biology, function, and disease. *Kidney Int. 75* (2009), 1153–1165.

[154] Yang, Y., Fear, J., Hu, J., Haecker, I., Zhou, L., Renne, R., Bloom, D., and McIntyre, L. M. Leveraging biological replicates to improve analysis in ChIP-seq experiments. *Comput. Struct. Biotechnol. J. eCollection 2014* (2014).

[155] Yu, J., Smith, V. A., Wang, P. P., Hartemink, A. J., and Jarvis, E. D. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics 20* (2004).

[156] Yu, Z., Mannik, J., Soto, A., Lin, K. K., and Andersen, B. The epidermal differentiation-associated Grainyhead gene Get1/Grhl3 also regulates urothelial differentiation. *EMBO J. 28*, 13 (2009), 1890–1903.

[157] Zoppoli, P., Morganella, S., and Ceccarelli, M. TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. *BMC Bioinf. 11*, 1 (2010), 154.

APPENDIX A

Appendix

A.1 Supplementary Material: Identification of ELF3 as an Early Transcriptional Regulator of Human Urothelium

A.1.1 Additional tables and plots

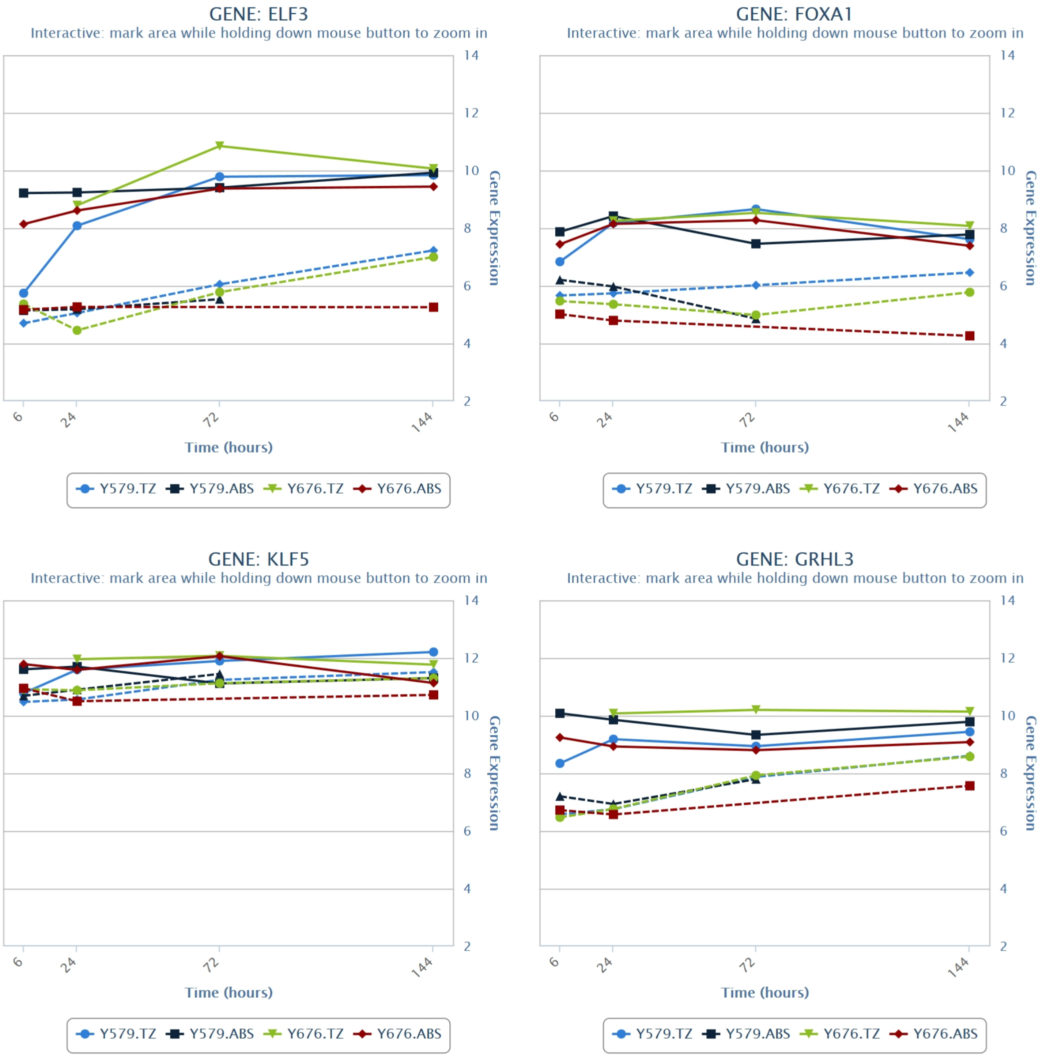


Figure A.1

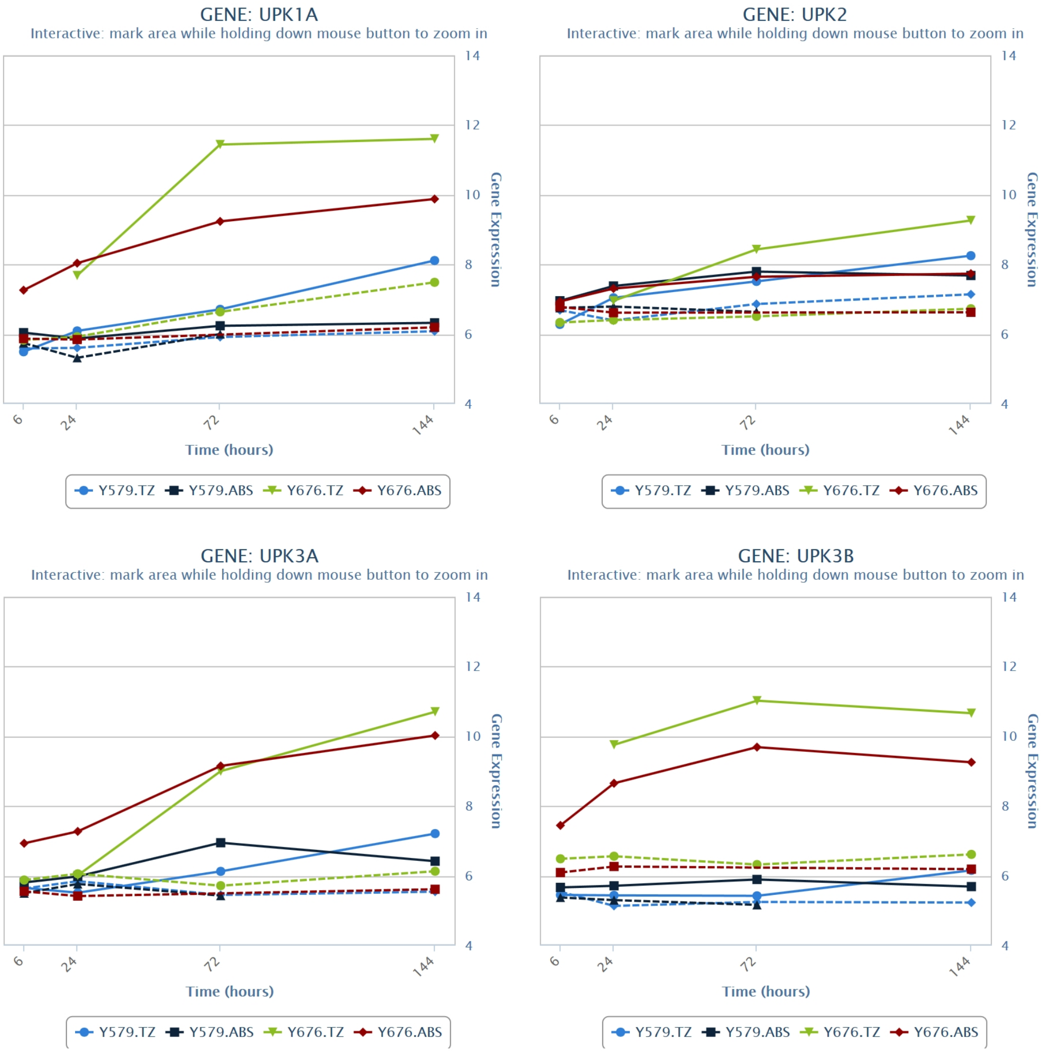
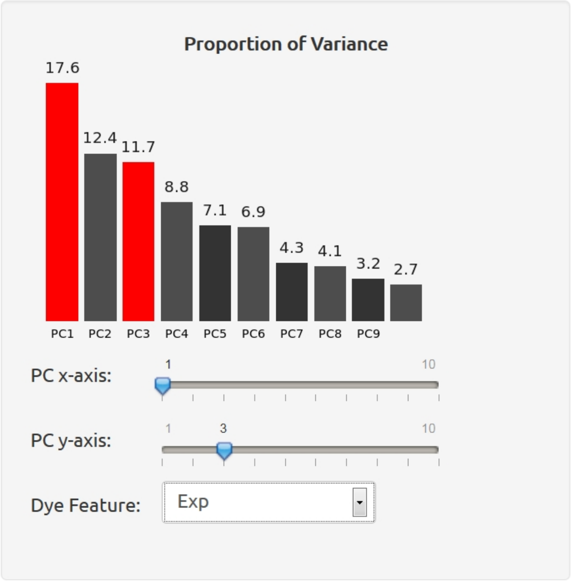


Figure A.1: Gene expression time series plots for a set of possible key genes. These plots give an overview of the transcript expression profiles of the following key genes: UPK2, UPK1A, UPK3A, UPK3B, ELF3, GRHL3, KLF5 and FOXA1. For genes with multiple probe sets the probe set with the largest inter quartile range (IQR) has been selected as representative.



PC x-axis: 1

PC y-axis: 3

Dye Feature: Exp

Figure A.2: Principal Component Analysis (PCA) of the available samples. PCA was performed on the samples to assess how well the different conditions could be separated. Unfortunately, up to the first ten components are needed to explain nearly 80% of the variability of the data. Figure A.3 gives an example of one possible choice for the principal components (PC) and how these two components separate experiment from control.

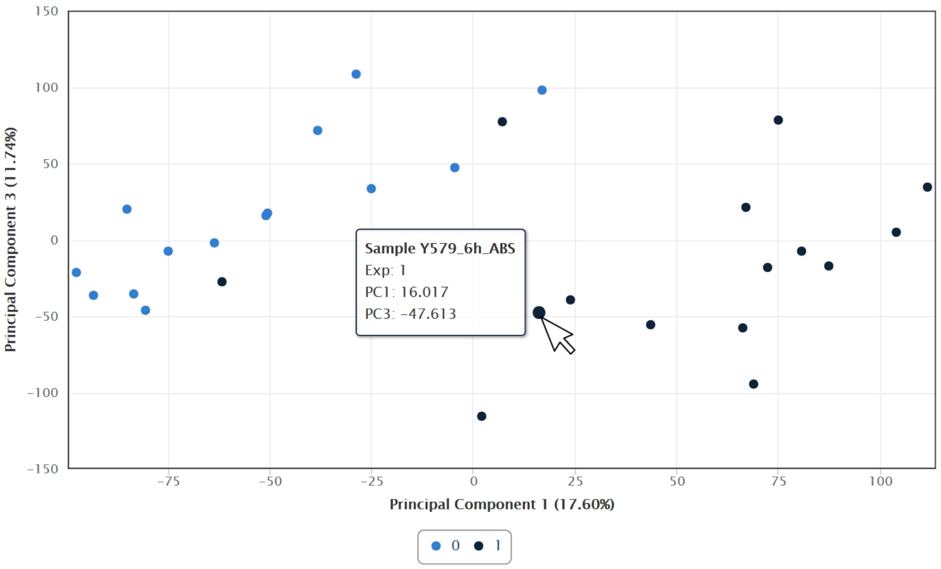


Figure A.3: Principal Component Analysis (PCA): Rotation of the samples with PC1 and PC3. As described in Figure A.2, principal components (PCs) one and three were used to separate experimental and control samples. Experiments are in dark blue (1) and controls in light blue (0). Depending on the choice of the PCs, the samples (microarrays) can be separated for different conditions. In general, samples which are close to each other in this plot tend to be also close in the timing of the corresponding experiments to each other (6h, 24h, 72h and 144h). For more details we refer to the project website where different features can be visualized and various settings of the PC combinations can be tested.

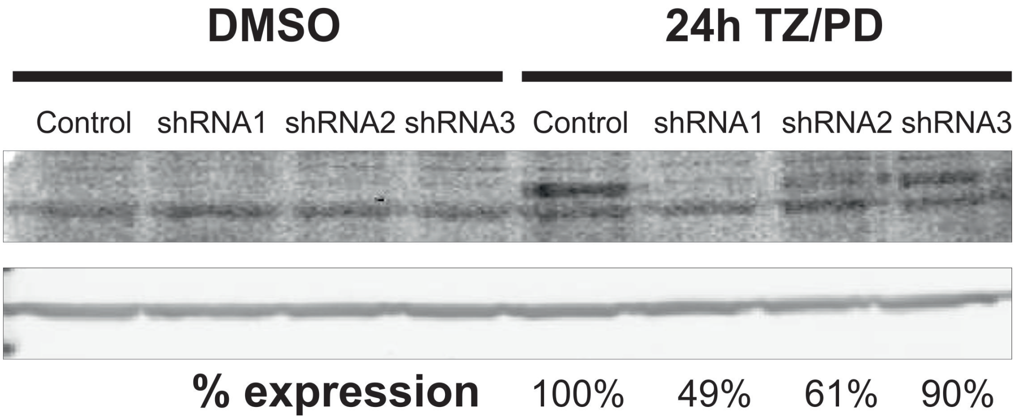


Figure A.4: Immunoblot: Analysis of ELF3 knock-down. For ELF3 knock-down, three ELF3 shRNA and a control shRNA were transduced into NHU cells and selected with puromycin. Cultures were induced to differentiate with 1μM troglitazone (TZ) and 1μM PD153035 or DMSO vehicle control for 24h prior to immunoblotting for ELF3. Band intensity was determined by densitometry analysis and normalised to β-actin; the efficiency of knockdown is expressed as a percentage of the scrambled control.