Institut für Informatik
der Technischen Universität München

Inference Modeling of Gene Regulatory Networks

Mathäus Dejori

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

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Dr. Jürgen Schmidhuber
Prüfer der Dissertation:		
	1.	Univ.-Prof. Dr. Dr. h.c. mult. Wilfried Brauer
	2.	Hon.-Prof. Bernd Schürmann, Ph.D. (Univ. of Cape Town) (Johann-Wolfgang-Goethe-Universität Frankfurt am Main)

Die Dissertation wurde am 20.04.2005 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 17.10.2005 angenommen.

Dedicated to my family

Abstract

The characterization of genetic mechanisms underlying normal cellular function, pathogenesis or the effect of drug treatment is one of the most challenging topics for todays lifesciences. The basis upon which the vast complexity and flexibility of life processes emerges is constituted by the interaction of various types of molecules such as proteins, mRNA, DNA or metabolites. The perhaps most important part of this complex signaling network is based on the interaction of proteins with the genome - the gene regulatory network. Thus, genetic network inference forms an enormous drive towards understanding, both the principles and details of the machinery which underlies the operation of living cells and systemic disease mechanisms such as cancer or diabetes.

The invention of high throughput screening techniques, such as DNA microarrays, and the resulting growth of biological data let data driven methods become very popular in the fields of genetic network inference. This thesis is concerned with elucidating gene regulatory network features by means of machine learning methods, more precisely, by means of graphical models. Most of the here presented approaches are related to Bayesian networks which fit the natural factorization of cellular processes into regular relationships between molecules and which are able to model causal relationships. One general problem, that data driven approaches are facing, is the noise and sparseness of given data sets. We therefore investigate measures which help to extract reliable features from Bayesian networks trained on microarray data with the goal to gain a more accurate interpretation of learned networks. The topology of such learned networks is the basis of a second approach, aiming to estimate genes that play a key role in controlling the state of regulatory genetic networks. By introducing new topological features we are able to estimate the effect of genes on the network stability finding those ones that represent the Achilles Heel of a molecular interaction network.

Existing methods for inferring gene regulatory networks from genome wide expression profiles provide important information about gene interactions and regulatory relationships. However, these methods do not provide information about the impact of possible interventions or changes on such regulatory networks to study cause-effect relationships at a systems biology level. We therefore propose a data driven method called ”generative inverse modeling” which simulates the effect of local genetic changes on the global cellular state, as reflected by an altered genome wide expression profile. The method can be used to estimate the relevance of genes regarding disease specific genetic mechanisms and to simulate local genetic changes on a global scale. Another strength of Bayesian networks is their probabilistic nature including the opportunity to make use of prior belief. We investigate the use of additional biological information as a probabilistic prior to guide structure learning. These approaches present two ways towards more robust estimates and equips Bayesian networks with further biological knowledge. The final chapter deals with an alternative graphical model approach to represent molecular networks. Instead of focusing on causal relationships between molecules, as done with Bayesian networks, decomposable models describe the underlying densities in terms of modules. This approach accounts for another fundamental principle of the molecular interplay, namely the union of several molecules into biological functional modules, to accomplish a certain task.

Acknowledgments

Since almost 3 years I am working at Siemens Corporate Technology and from the first day on I appreciated the atmosphere in which basic research and industry join. I would like to thank Prof. Bernd Schürmann, head of the Department for Neural Computations, for giving me the opportunity to be part of this excellent research environment and for initiating and promoting research activities in the field of computational biology.

I am very grateful to Prof. Wilfried Brauer at the Department of Computer Science, Technical University of Munich, for supervising and governing my thesis work, for his fruitful comments and for the interest he manifested for bioinformatics and computational biology.

I would like to thank Siemens AG, especially Dr. Hans Höller and Christine Kiehnlein, for the motivation and the financial support I got through the Ernst von Siemens Foundation scholarship.

I especially would like to thank Dr. Martin Stetter for supervising my work at Siemens Corporate Technology. Most of what I've learned in the past few years derives from his comments, ideas and discussions and without his input and his ”big pictures” my thesis would not be what it is.

In the last years I had the opportunity to collaborate with Dr. Anton Schwaighofer, Dr. Volker Tresp, Jan Scholz and Dr. Martin Greiner, for which I am very grateful; I would like to thank all the people I had the opportunity to work with in common projects; I would like to thank Andreas Nägele and Markus Obendorf, students with whom working was always a great pleasure; I would like to thank Holger Arndt for patiently answering all my, sometimes annoying, questions and all other group members for the great and inspiring working environment.

Last but not least many thanks to my family, my friends and especially to my girlfriend Claudia for supporting me all the time, for distracting me from work and for giving me the strength to keep up.

Contents

Chapter 1

Introduction

During their life span, living cells accomplish a vast multitude of different tasks which are all directly or indirectly controlled by their genome, encoded in the deoxyribonucleic acid (DNA) molecule. For example, cells have to maintain their organization, ensure their nutrition and synthesize and exchange new biomolecules – which they consist of – by metabolic processes. In response to external chemical signals, they eventually grow and undergo cell division, differentiate into specialized cell types, start or stop secretion of one or several substances, start proliferating, or migrating, leave their place within the cellular matrix, or initiate their own death (Lodish et al. 2000). But also without any direct external control, cells can vary their behavior: For example, specialized neurons underneath the optic nerve possess a machinery by which they produce oscillating concentrations of several proteins and thereby participate in encoding our wake-sleep or circadian rhythm (Hedges & Kumar 2003). According to a popular view, all these cellular life processes are coordinated and guided by the states of the underlying network of mutual interactions among the multitude of cell's aggregates, genes, RNAs, proteins and small molecules. Thus, cellular characteristics can be represented in terms of a complex *molecular network* composed of the interplay of millions of molecules. One prominent part of this network is constituted by the regulatory interaction network among genes, the *gene regulatory network* or *genetic network*.

When cells are assembled to form a higher organism, these and other cellular processes are being carried out in a concerted fashion, subtly coordinated by a mutual exchange of chemical signals. The resulting collective activities form the basis of growth and shaping of organs and body parts by self-organizational processes. They also guide the maintenance of the structure of the organism: for example the vast majority of the molecules of a human body are exchanged against other ones during a period of only two years, but despite our body more or less remains in its shape. In other words, our structure is maintained independently of the identity we are made of. Furthermore, the cellular machinery is involved in processes of adaptation to the environment such as learning or muscular training, and guides repair mechanisms involving wound healing and DNA repair (Bohr 2002) at the macroscopic and microscopic levels, respectively.

In light of the central role of cellular genetic network for the life of higher organisms, it seems not too surprising that many severe diseases including different kinds of cancer or Alzheimer's disease show a clear relationship to genetic disorders (Hanahan & Weinberg 2000, Strittmatter & Roses 1995). Further, fundamental restrictions in wound healing and repair exist for higher organisms (imputed limbs and removed organs do not replace themselves in humans), and seem to have their roots in a disability of differentiated cells to re-use the embryonic genetic machinery of morphogenesis. Finally, changes in the cellular machinery related to impairment in repair mechanisms also seem to be related to aging and death (Strehler 1995, Vijg & Dolle 2002).

1.1 Motivation

These and other insights claiming the importance of the genetic network, form an enormous drive towards understanding both, the principles and details of the machinery which underlies the operation of living cells and the pathogenesis of systemic diseases. There is a huge diversity of different molecular processes which have to be characterized and stored in data bases. Scientists are working since many years on this challenging tasks to collect data about genes, proteins, cellular pathways, disease markers and other molecular phenomena (Galperin 2004) to provide a comprehensive knowledge base, helpful for the elucidation of still unknown cellular procedures. However, biological diversity is enormous, as is the complexity of the data and knowledge to be stored, and existing data and knowledge bases are at the edge to turn into an inextricable jungle. Therefore, collecting the data is not enough: instead scientists have to additionally reveal unifying principles of operation of life processes.

Many researchers propagate the view, that networking is one such principle. All processes in a living cell are directly or indirectly related to and guided by complex, recurrent and mutually interacting signalling chains. Proteins are synthesized from genes, interact with each other and with smaller molecules but also act back onto the RNA and DNA where they regulate the production of other proteins. It is not sufficient to regard individual components of the cellular machinery anymore. Instead its full understanding is unresolvable coupled to the understanding of the concerted action of all molecules in the cell at a systems level. In light of this view, systems level modeling of the genetic regulatory network (by methods of artificial intelligence) represents one of the most powerful combinations on the way towards a complete understanding of the underlying cellular machinery and related inter- and intracellular processes. Computational models of genetic networks can be divided into mainly two classes: bottom-up approaches which use explicit molecular biological knowledge to reconstruct the genetic network and top-down approaches, which make use of massive biological data to learn genetic network principles without focusing explicitly on biochemical parameters. For a review see (Stetter et al. 2003).

The first class covers biochemically inspired models based on the reaction kinetics between the different components of the genetic regulatory network and can be associated with the level of detail of Figure 2.2a. Biochemically inspired models have the advantage that they can be more directly related to biological processes, but they also suffer from a number of difficulties. Firstly, many of the biochemically relevant reactions under participation of proteins do not follow linear reaction kinetics. For example, many proteins undergo conformational changes after reactions, which change their chemical behavior. In particular, in many regulatory DNA regions transcription factor binding can show cooperative or competitive effects, which are nonlinear and mostly unknown. Secondly, the full network of metabolic, enzymatic and regulatory reactions is very complex and hard to disentangle in a single step. To do so, the kinetic equations of all the different interactions (e.g., those in Figure 2.2a) would have to be written down, but the type of reactions and their parameters are often unknown. At present, the data basis seems not sufficient to globally understand regulatory networks at this level of detail. However, very well-examined regulatory sub-networks (Yuh et al. 1998, Yuh et al. 2001, de Jong et al. 2001) which are sufficiently well-characterized to be modelled in some detail exist. Other approaches use approximations to reaction-kinetics formulations to arrive at systems of coupled differential equations for describing the time course of gene expression levels (Chen et al. 1999, Kato et al. 2000, Sakamoto & Iba 2001). Other models which fall into this class are boolean networks, which have been among the earliest approaches towards genetic network modeling (Kauffman 1969).

The second class of models yet has gone through another step of abstraction, and treats the task of modeling microarray data as a data mining problem (Somogyi et al. 1997, D'haeseleer et al. 1997, Wang et al. 1999, Baldi & Brunak 1998, D'haeseleer et al. 2000, Slonim 2002). The goal of data mining is to explore a data set and to discover regularities and structures from it. As opposed to hypothesis-driven approaches, which search for a particular and pre-defined pattern in the data, data mining approaches specify autonomously which patterns are present in the data – they are exploratory and data driven. As gene expression data sets are high dimensional and noisy by nature, statistical methods play an important role in their interpretation by finding trends and patterns in the experimental results. In case of genetic networks, the patterns to be inferred may be for example clusters of genes which are coexpressed under a given mode of the cell, or the structure of regulatory relationships between genes. Clustering studies have revealed many extended clusters of genes, which collectively change their expression levels when a cell or tissue changes from one mode of life to another (Eisen et al. 1998, Spellman et al. 1998, Ben-Dor et al. 1999, Golub et al. 1999, Yeoh et al. 2002). In fact it has been shown that many coexpressed genes are known to share common molecular pathways which indicates that these global gene expression patterns reflect the execution of specific genetic programs.

However, clustering analysis can not provide an answer to the question of what is it that either stabilizes a global gene expression pattern or evokes a change to a new one. In other words, are there dominant genes or gene groups which are the underlying cause of a specific global pattern? It is also important to know, which global control functions might have failed when a pathological global gene expression pattern is observed in case of a disease. Knowing the basic origin of a disease might help identifying more potent and more selective drugs and to do so in shorter time. Due to these considerations, recent approaches have concentrated on inferring the structure of underlying genetic regulatory networks from microarray data. Finding a structure from high-dimensional data is a common problem in the machine learning community known as structure learning of graphical models (Jordan 1998). Friedman and colleagues (Friedman et al. 2000) introduced this framework to infer the structure of the underlying genetic regulatory networks from microarray data and initialized a series of related works (Pe'er et al. 2001, Hartemink et al. 2001, Imoto et al. 2002, Segal et al. 2003). In this approach, the set of measured gene expression vectors is considered to be drawn from a high-dimensional multivariate probability density function which is modeled by a Bayesian network with adaptive network structure. The basic idea is to display the associations among the variables, namely the conditional dependencies and independencies, by means of a directed acyclic graph. In the context of genetic pathway inference, each node of a Bayesian network is assigned to a gene, and can assume the different expression levels of this gene throughout the set of measurements. Each edge between genes hints towards a regulatory relationship between them. If this edge is directed, it can, under certain assumptions, be interpreted as a causal relationship: it can be inferred which gene controls another gene. Moreover, Bayesian networks can be used to pose probabilistic inference questions and, with an algorithm proposed in Chapter 6, also to conduct *what-if* scenarios, such as: ”what happens to the global genetic network if a certain gene turns into over-expressed?”. Even despite these advantages, Bayesian network learning is a data-driven method and therefore suffers from the sparseness of available data. Since most microarray data sets consist only of a few samples, one has to pose the question what can Bayesian networks, learned from sparse data, represent and what they can not represent (see Section 4.3.2).

Another representation of genetic networks by a member of the graphical models class puts particular emphasis on the modular way in which molecules act together to accomplish certain tasks (Schwaighofer et al. 2004). Decomposable models try to explain the statistics in a data set by the action of mutually linked functional modules, so-called cliques, represented by a structure of undirected links. By this, decomposable models should be able, with particularly high sensitivity, to detect the signature of a concerted action of gene modules in the data. In light of this rationale, cliques are likely to contain functionally high correlated genes, as opposed to gene clusters (see above), where genes are grouped together by mere coexpression. Hence, in contrast to clustering, the learned structure also reveals some information about possible statistical relationship of genes within a cluster and relationships between clusters (see Chapter 8).

1.2 Overview

The work presented in this thesis is concerned with revealing genetic network principles by means of graphical models focusing on Bayesian networks and decomposable models whose structural and parametric characteristics are estimated from molecular data.

Chapter 2 provides a brief summary of fundamental issues of molecular biology, the common view of genetic networks and their relation to disease mechanisms. Furthermore common screening and data retrieval methods which provide a basis for various data driven methods are presented (see (Stetter et al. 2003)).

In Chapter 3, I give a brief introduction to those two classes of graphical models which are used in the following chapters, namely Bayesian networks and decomposable models and motivate their descriptive power regarding molecular network principles (see (Stetter et al. 2004)).

In Chapter 4, I study the robustness of Bayesian networks under real-life conditions with the goal to find measures which can be accurately interpreted in Bayesian networks learned from microarray data. The main novel contributions are the definition of various measures to evaluate the robustness and correctness of learned networks and robustness tests to analyze the effect of search strategies and the effect of small sample sizes for structure learning (see (Dejori & Stetter 2003a)).

The approach presented in Chapter 5.1 concentrates solely on the structure of learned Bayesian networks aiming at drawing conclusions from characteristic topological patterns. The main contribution is the introduction of topological features which might provide additional information towards a deeper understanding of genetic network principles (see (Dejori & Stetter 2003c, Dejori, Schürmann & Stetter 2004, Scholz et al. 2005)).

In the approach of Chapter 6 I try the other way around and focus only on the parametric part with the intention to learn about genetic network mechanisms from the learned probabilistic densities. The main contribution here is a new approach to conduct *what-if* scenarios, by generating *in-silico* expression data, on the basis of probabilistic inference (see (Dejori & Stetter 2003b, Dejori & Stetter 2004)).

In Chapter 7 I propose two ways of using other sources of molecular data as a prior to guide Bayesian network learning. Novel contributions are the way in which appropriate biological prior knowledge is obtained and how it is incorporated as a probabilistic prior into the structure learning procedure (see (Dejori, Nägele & Stetter 2004)).

In Chapter 8 I propose an alternative way of modeling genetic networks, namely by means of decomposable models. The main contribution here is the application of this class of graphical models on gene expression data and the motivation for its use to reveal new genetic network principles (see (Dejori, Schwaighofer, Tresp & Stetter 2004)).

The appendices contain information about algorithms and data which have been used throughout this thesis.

Chapter 2

From Genes to Metabolites

In 1953 Watson and Crick discovered the double helix structure of the DNA and consequently the principle of DNA replication (Watson & Crick 1953). This was the breakthrough for molecular biology and provided evidence for the central role of the DNA sequence cellular processes. The era of molecular genetics was launched and with it production of huge amounts of genetic data, driven by a rapid development of new screening technologies. Almost 50 years later 90% of the human genome was announced to be sequenced (International Human Genome Sequencing Consortium 2001, Venter et al. 2001) and raised hope to hold in hand the Rosetta stone for solving the secret of life. But soon it became clear that discovering functional aspects of genes and their products, directly from the genomic sequence, is not feasible. Consequently, the era of functional genomics started and with it the effort to discover the function of individual genes and their related products by making use of different types of molecular data. Nowadays, attention is directed beyond the function of individual cellular components towards functional and systemic principles of the underlying complex system build by genes, proteins, RNA and other molecules. Only such an extensive view at a network level will put science forward to a complete insight into the cellular machinery.

This chapter provides a brief summary of molecular principles which underlie the idea of genetic networks and their relation to complex disease mechanisms. In favor of the grand picture we omit many important details and focus on eukaryotic cells. For further in-depth reading see (Lodish et al. 2000, Brown 1999).

2.1 The Cellular Machinery

Genetic information in living cells is encoded in the deoxyribonucleic acid (DNA) sequence, a linear polymer built up of 4 types of nucleotide bases: two purine bases, adenine (A) and guanine (G), and two pyrimidine bases, thymine (T) and cytosine (C). The major characteristic of DNA is the right-handed double helix structure composed of two antiparallel strands stabilized by hydrogen bonds between complementary purine and pyrimidine bases (A-T and G-C). The human genome consists of about 3 billion base pairs, the complete sequence of which has very recently been published (International Human Genome Sequencing Consortium 2004). The nucleus of an eukaryotic cell carries the entire genome in so called chromosomes, which consist of a set of DNA doublestrands and associated proteins. During each cell division, DNA is replicated with very high accuracy, and consequently almost every cell in an organism contains a virtually identical copy of the genome (see Figure 2.1).

Each DNA sequence can further be subdivided into functional information units called genes. The human genome is estimated to carry roughly 30000 genes, which all together cover only 2-3 % of the DNA double strand. A gene can be defined as a physical segment of varying length of a chromosome on one of the two DNA strands. This ”sense-making” strand codes for the amino acid sequence of one protein or a set of proteins. However, also noncoding DNA fragments are found within eukaryotic genes: a gene sequence can (functionally) be divided into exons, which explicitly encode the amino-acid sequence of corresponding proteins, and introns, which do not code protein sequence but can contain otherwise important information (Figure 2.2a top). Moreover a gene contains regulatory regions which do not code for corresponding proteins but which are necessary for their synthesis. Proteins are the working units of cells involved in nearly all cellular activities, e.g. as catalysts or signal carriers. Hence, as DNA stores all genetic information necessary for protein synthesis it also codes for the emerging cellular programs.

2.1.1 Gene Expression

As noticed before, genes are subunits of the DNA sequence, located in the nucleus and encode information for certain proteins. The way by which proteins is synthesized from the DNA sequence of corresponding genes follows a two-step procedure.

When a certain protein is needed, the DNA sequence of the corresponding gene is used to synthesize a complementary single stranded molecule of ribonucleic acid (RNA). RNA is very similar to single-stranded DNA, where thymidine is replaced by uracil (U). This transformation from DNA to RNA is called *transcription* and is mainly executed by the RNA-polymerase II (PolII) enzyme. PolII binds to a specific transcription start site (TSS), unwinds the double strand and adds nucleotides to the growing RNA strand. The primary RNA transcript is transcribed from the entire DNA sequence of the corresponding gene except regulatory regions. It therefore contains intron segments which do not code for aminoacids of the corresponding protein. Consequently, these non-coding intron regions are removed in a process called RNA splicing and the remaining exons in the primary RNA transcript are joined together to form the messenger RNA (mRNA).

In a second step, after mRNA has been transported out of the nucleus, it is used as a template for the synthesis of amino acids, which are chained together to form the desired polypeptide. This step, known as *translation*, is mediated by the cooperation of a ribosome, composed of numerous proteins, two ribosomal RNA molecules (rRNA) and transfer RNA (tRNA). The two-step mechanism of protein synthesis is referred as *gene expression* and forms the central dogma of molecular biology:

$$\text{DNA} \underbrace{\longrightarrow}_{\text{transcription}} \text{RNA} \underbrace{\longrightarrow}_{\text{translation}} \text{Protein}$$

RNA molecules are synthesized from a DNA template during transcription and proteins are synthesized from RNA molecules in the process of translation. Thus, when proteins are produced, the corresponding gene is said to be expressed. It is important to note that the stronger a gene is expressed, the more mRNA of this gene is present in the cytosol. Figure 2.1 provides a general sketch of gene expression and regulation.

2.1.2 Regulatory Mechanisms

The human genome is believed to encode about 1 million different proteins. Most interestingly, each cell of an organism contains only a subset of these proteins, the so-called proteome, at any time. In other words, whereas the genome is (almost) identical for each cell, the proteome varies strongly over different cell types, depends on the distinct cell-states (i.e., the phase of the cell cycle), and on external signals imposed on the cell (see Figure 2.1). This implies, that cells must have control over the use of their genome providing the possibility to change their proteome in a very flexible way. The regulation of the proteome is achieved by affecting the gene expression machinery at their different stages and different molecular products (see Figure 2.2a).

The first and most effective control mechanism is transcriptional regulation. DNA binding proteins, so called transcription factors (TF), bind to specific sites within the regulatory region of a gene and affect the transcription initiation of DNA into RNA. Usually up to a few tens of transcription factors can act on the same regulatory regions of a gene (Latchman 1998). Transcription factors can enable, disable, enhance or repress gene expression, and they can do so in a highly nonlinear collective way (Yuh et al. 2001). In turn, any given transcription factor could act on a few thousand of different genes (Brown 1999). TF-binding sites can be located immediately upstream of the transcription initiation site as well as up to many kilobases (kb) away from the start site. *Enhancers* for example can be located thousands of base pairs upstream, downstream or even within the gene they control. Another mechanism of transcriptional control is accomplished by DNA methylation of cytosine residues, specifically those cytosines which precede guanine nucleotides (CpG islands), in mammalian genome: methylated DNA-fragments in vicinity of promoters correlates with a reduced transcriptional activity of the corresponding genes. DNA methylation therefore is suspected to determine severe cellular changes (Widschwendter & Jones 2002).

Regulation also happens at the RNA level, for example when the primary RNA transcript is spliced to form mRNA. RNA splicing can occur in various different ways, so called alternative splicing, which is one way to produce a whole family of proteins from one gene. Alternative splicing is controlled by RNA-binding proteins that bind near regulated splice sites, such as splicing inhibitors or splicing activators. They can control which parts of mRNA are being translated, affecting the behavior of the resulting proteins. Moreover, mRNA concentration in the cytosol is controlled by regulating the rate of degradation or translation. Interaction of specific mRNA binding protein protects mRNA from degradation or blocks the translation process. Also translation can be inhibited by RNA interference (RNAi), where mRNA is marked for being degraded by short double-stranded RNAs (Fire et al. 1998).

Finally, the translated amino acid chains are subject of regulatory effects, so-called post-translational modifications. This includes attachments of peptides to the polypeptide chain and covalent modifications, e.g. phosphorylation. These chemical modifications alter the biological activity of the affected proteins and are argued to be crucial for the global cellular mode (Banin et al. 1998, Cortez et al. 1999). Moreover, by protein-protein interactions many proteins form various sets of aggregates and are only operative in these complexes.

Furthermore, the whole genomic/proteomic machinery is controlled by extracellular signals which form the interface between a cell and its environment. External signals can regulate gene expression either by directly acting as transcription factors, or by modifying transcription factors (Brown 1999). Due to a wealth of regulatory mechanisms, such as alternative splicing, post-translational modifications and aggregation, each gene produces a whole family of operative protein structures.

2.1.3 The Genetic Network

As noted in the previous section, gene expression is controlled by various mechanisms mainly conducted by special proteins which bind to DNA sequences, RNA molecules or other proteins. These regulatory mechanisms add a feedback step to the feed-forward process of protein synthesis since gene expression controls protein concentrations, and proteins in turn – either directly or indirectly – regulate gene expression levels. Hence, the genome and the proteome are linked together by a complex recurrent (and nonlinear) regulatory network, the genetic network.

Figure 2.2b summarizes, at a more abstract level, the pathways, by which protein A, synthesized from gene A, can regulate the expression level of gene B. Depending on the function of protein A, it controls the transcriptional, posttranscriptional or posttranslational level of gene B. In terms of a graph structure, this regulatory mechanism can be represented by a directed link from gene A to protein A back to gene B, or simply by a directed link from gene A directly to gene B. Roughly speaking, gene B is said to be regulated by gene

A. Indeed, this simplified view neglects diverse products emerging from gene A, but helps to gain a clear sketch of the basic regulatory relationship. Figure 2.2c, finally shows a highly abstracted graphical representation of a genetic network. Each gene regulates many (up to thousands) of other genes, and its own expression level may be regulated by up to a few tens of other genes. Some genes form a source of transcriptional control, they have a high fan out or divergence of edges (e.g. gene A). Other genes have no regulatory action at all but may be regulated by others. They are characterized by a high fan in or convergence of edges (e.g. gene C). This complex regulatory network between genes, RNA, proteins and other molecules forms a central part of the cellular machinery. Its different modes of operation control the multitude of biochemical processes in a living cell.

2.1.4 Disease Mechanisms

During the last decades, the hypothesis that different cellular states arise from execution of diverse genetic programs has been consolidated. Such genetic programs are accompanied and guided by certain functional states of the gene regulatory network of the cell. Thus, understanding the principles of normal cellular function, pathogenic mechanisms and the effect of drug treatment represents perhaps the most challenging issue of modern life sciences. Revealing the mechanisms which lead to fundamental aberration in the underlying genetic network will help to understand not only the cellular machinery but also the principles of genetically caused diseases. Many diseases are suspected and some known to be caused and influenced either directly or indirectly by genetic alterations (Dulbecco 1986, Hanahan & Weinberg 2000): as an effect of a local change in the function of one or a small collection of genes, the whole genetic program and the operational mode of the cell turns into a pathological one. This assumption is confirmed by the fact that pathological transformations are often paralleled by a change in the global gene expression profile (Ross et al. 2000, Yeoh et al. 2002). Malfunctions of regulatory mechanisms can have serious consequences especially in the context of tumor-specific abnormalities.

Cancer is a genetic disease triggered by the aberration of mainly two types of genes: proto-oncogenes and tumor-suppressor genes. Tumor-suppressor genes inherit special regulatory and repair functions to inhibit cell proliferation, and are crucial for the maintenance of a controlled cell state. They code either for proteins which are involved in DNA-repair, for proteins which promote the programmed cell death, the so-called apoptosis, or for proteins which arrest an ongoing cell-cycle if a previous step has carried out incorrectly. Proto-oncogenes are normal genes which, due to a genetic mutation, turn into oncogenes. In this transformed state, they cause a healthy cell to become a malignant one. Mutations can change their function, due to point-mutations, their expression level, due to re-duplications or deletions, or the environment in which they are usually expressed due to chromosomal translocations. Mutations in both gene types, caused by external factors (e.g. tobacco smoke or other carcinogenic agents), inherited by a genetic predisposition or by both, are the major molecular determinant of tumorigenesis. In the last decades, many tumor-suppressor genes have been identified, such as gene p53, coding for a protein which is involved in a variety of pivotal molecular processes such as DNA-repair or cell-cycle arrest (Symonds et al. 1994). Moreover, for some cancer types related genetic markers or typical expression profiles are known and provide new possibilities for diagnosis and prognosis (Golub et al. 1999, Yeoh et al. 2002). However, in most of the cases it is still unknown which genetic changes cause the disease, and how these changes are processed by the genetic network. Thus, a more complete description of tumorigenesis and pathogenesis in general will undoubtedly emerge only with the detailed elucidation of pathogenic mechanisms at a network level.

2.2 Screening the Cellular Machinery

Computational models of genetic networks depend on the availability of data that reflect the state of the cellular system. These data would ideally cover a wide range of genetic information such as genomic data (e.g. regulatory regions), transcriptomic data (e.g. expression rates of all genes), proteomic data (e.g. types, concentrations and states of all proteins in the cell) as well as metabolomic data (e.g. known biochemical pathways). This section summarizes *in vitro* as well as *in silico* methods which provide an important data basis for the genetic network inference step.

2.2.1 Genomic Data

As the human genome was announced to be sequenced, scientists realized that the genetic key for the complexity of life does not lie in the number of genes, but in the interplay of them with proteins, RNA and small molecules. Thus, from the DNA sequence only, functional aspects can not be retrieved. Although the genomic sequence is not enough to gain a full understanding of the cellular machinery, it stores almost all the information necessary for the execution of global cellular processes. The techniques presented in the following show that DNA sequence is a good starting point for deciphering the underlying genetic network.

Comparative Genomics

The great diversity of living organisms provides evidence, that mutations in the genetic material are crucial for the evolution of life, ranging from single point mutations to major structural modifications in chromosomes. Studying the evolution of life is hence strongly linked with studying the evolution of the genome. With the availability of genomes from different species it became possible to study not only individual sequences, but also evolutionary differences and homologies among them and even functional aspects by comparative studies.

Comparative genomics is based on the hypothesis that important biological sequences in the genome are conserved between species, due to functional constraints. The ideal pairwise comparison is between two organisms that share a common physiology or biology. Humans and mice, for example, have roughly the same number of protein-coding genes ($\sim$ 30000) and less than 1% of these have no ortholog in the other species (Mouse Genome Sequencing Consortium 2002). The mouse is hence a popular organism for the identification of functionally conserved sequences shared with the human genome. Comparative studies with the mouse genome led to the identification of new genes, the annotation of previously unknown genes, the identification of gene-regulatory elements and the detailed characterization of transcription factor binding sites (Nobrega & Pennacchio 2003, Pennacchio & Rubin 2003) (see Section 2.2.1 below).

Moreover, comparative studies are not limited to closely related genomes. Distant species comparisons revealed astonishing results: comparative studies between human and the pufferfish *Fugu rubripes* revealed more than 1000 genes that had previously been unidentified in the human genome (Aparicio et al. 2002).

Single-nucleotide polymorphisms

Single-nucleotide polymorphisms (SNPs) are individual point mutations in the DNA sequence and occur every 100 to 300 bases along the human genome (Collina et al. 1997). Besides the vast amount of point mutations in noncoding DNA regions, researchers focus on inter-individual variations in coding regions, since many of them are suspected to cause predisposition for diseases. The correlation of variants in the APOE4 gene with late-onset Alzheimers's disease (Strittmatter et al. 1993) or the protective effect of variants in the APOE2 gene against Alzheimer's are prominent examples for SNP-related disease predispositions.

Besides their importance as disease markers, identifying SNPs associated with significant biological effects in response to chemical drugs represents another major aim of SNP analysis. Mutations in drug-metabolizing enzymes such as cytochrome P450 are known to have a significant effect on drug response and effect. SNP profiling is therefore a promising way towards personalized medicine, because it helps to determine appropriate ways of drug-treatment (McCarthy & Hilfiker 2000). Another big advantage of SNPs for diagnosis is the fact that they can be measured with highthroughput DNA microarrays (see Section 2.2.2) thus the human genome can be screened for various SNP markers in one single experiment (Wang & et al. 1998).

Identification of regulatory regions

As noted in Section 2.1.3, proteins regulate gene expression at various levels, by binding to DNA, mRNA or to other proteins. Protein-DNA interactions form an important part in the global genetic network, the transcriptional regulatory network: proteins bind to specific binding sites within a regulatory region and regulate the expression of a certain gene. Thus, identifying DNA regions which hold regulatory functions, will provide new insights into transcriptional regulatory mechanisms. Regulatory regions mainly contain two classes of protein binding sites: binding sites which are involved in transcription initiation, so-called *transcription start sites* (TSS), and binding sites involved in transcriptional regulation *transcription factor binding sites* (TFBS).

Finding TSSs is already well established since the search space is limited to a relatively small region upstream of the gene. Proteins involved in the initiation process, such as the RNA polymerase (see Section 2.1.1), bind to short DNA sequences immediately upstream ($\sim$ 30 bp) of the gene on which they act. TATA-boxes or regions of C-G enrichment (CpG islands) are well known markers for TSSs and are consequently also helpful for detecting the starting point of a gene. Finding TFBSs is much more difficult since they are not limited to regions proximal to the transcription initiation site. *Enhancers* for example can be located thousands of base pairs upstream, downstream or even within the gene they control.

A general problem of finding regulatory elements in the DNA sequence is the fact that protein binding sites are specific to certain proteins but not identical in their sequences. In other words, there is no single sequence which is recognized, but a *sequence motif* where for some positions several choices of nucleotides are possible. Therefore, protein binding sites are not described by a unique sequence but by a collection of sites characterized by a *position weight matrix* (PWM) which provides a quantitative probabilistic description of a protein binding site (Stormo 2000).

Additionally, in higher eukaryotes the transcriptional regulatory network becomes very complex such that instead of independently binding to a single target, proteins form regulatory complexes by binding to multiple DNA regions and to each other. Transcription factors are known to compete or cooperate with each other. The identification of single binding sequences tends to consider proteins as binding independently and therefore neglects such collective effects. Nevertheless, the simplistic view of transcription initiation and regulation regions is a first step towards understanding transcriptional regulatory principles. Furthermore, the analysis of TFBSs can be improved through comparative studies, known as *phylogenetic footprinting*, and through a combination of sequence-analysis and transcriptomic data. Especially the combination with microarray data can improve our understanding of regulatory mechanisms.

2.2.2 Transcriptomic Data

The following discussed techniques make use of the fact that mRNA concentration in the cytosol reflects the level of expression of this gene. Since the measured mRNAs are further translated into proteins, screening the transcriptome also reflects, even if indirectly, the cellular behavior at the proteomic level. However, it needs to be noted that these techniques cannot capture the post-transcriptional processes downstream of mRNA processing and their effect on the proteome. This includes many protein protein interactions which can also form part of the regulatory mechanism within the genetic network (dashed arrows in Figure 2.2b). But direct regulatory mechanisms triggered by transcription factors, as well as indirect regulatory processes involving the modification of transcription factors by protein protein interactions, will in general be reflected in altered gene expression levels.

The popularity of transcriptomic data is primarily triggered by highthroughput methods which can monitor the expression of thousands of genes, and hence almost the entire transcriptome of a cell, in parallel. In the following, we will concentrate on this type of analysis since it provides a vast amount of data, necessary for the following statistical approaches. However, it should be noted, that various serial techniques are still used (e.g. RT-PCR (Heid et al. 1996) or SAGE (Velculescu et al. 1995)) not only to validate results of highthroughput experiments but also to provide highly accurate experimental results.

High-Throughput Gene Expression Profiling

During the last decade, techniques for large scale measurement of gene expression levels, based on DNA microarrays, have been developed (Fodor et al. 1993, Schena et al. 1995). For reviews see (Brown & Botstein 1999, Schena 2000, Baldi & Hatfield 2002).

DNA microarrays measure in parallel the cellular mRNA concentrations for many thousands of genes and make use of the following facts: (i) Complementary strands of RNA and DNA (a pair of sequences where each T combines with A/U and G with C on the two strands) bind to each other, a process which is called hybridization. (ii) Hybridization is highly selective for the sequence and is most stable when the two sequences are complementary to each other. For longer sequences of a few hundred base pairs, virtually only complementary strands bind at an adequate temperature. This effect can be used to very selectively filter out RNA strands of a certain sequence, by means of a complementary DNA (cDNA) strand. These cDNA strands are obtained by reverse transcription of the corresponding mRNA strands. The strength of hybridization is temperature-dependent which is used in many molecular biological techniques. (iii) The mRNA sequence for each gene is unique, in other words for every mRNA to a given gene there can be found a sequence that appears only in this mRNA. (iv) Due to whole genome projects, the sequences of entire genomes, but also the loci and sequences of an increasing number of genes for a collection of different organisms become available.

A gene expression measurement with microarrays is based on the selective hybridization of dissolved mRNA with probes of complementary sequences that are fixed on a substrate. The surface of a DNA microarray is divided into many spots. At each spot a number of nucleotide strands with a sequence complementary to the mRNA sequence of one gene are fixed. Hence, a DNA microarray with N spots can measure N mRNA concentrations at the same time. Microarrays use either short oligonucleotide probes with 15-25 base pairs, as manufactured by the company Affymetrix (Fodor et al. 1993), or longer strands of a few hundred bases. The latter microarrays are more selective and can be used for differential gene expression measurements (cf. (Brown & Botstein 1999)), whereas the former microarrays are mostly used to measure expression levels from a single cell.

Figure 2.3 illustrates schematically the procedure of a differential gene expression measurement. mRNA from a control cell and the cell to be measured are extracted, purified and labelled with two different dyes (often Cy3 and Cy5). They are brought together with the DNA probe spots on the microarray where they compete for hybridization. At the end of hybridization, the ratio of bound mRNA from both cells on each spot reflects the ratio of concentrations of this mRNA in both cells. After optical readout – usually with a confocal laser scanning microscope – an image with colored spot patterns results, where colors reflect the expression level of each gene in the measured cell, relative to the control cell. Image processing algorithms can semi-automatically or automatically analyze the spot pattern and transform it into numerical gene expression values (Hegde et al. 2000, Angulo & Serra 2003) which, after normalization procedures (Hegde et al. 2000, Bilban et al. 2002, Baldi & Hatfield 2002), are stored in a N dimensional gene expression vector x. There are many sources of noise in microarray experiments which include biological noise (variation of cellular states in a homogeneous strain, variations caused by RNA extraction procedures), finite sample effects such as fluctuations in the number of hybridizing molecules or optical readout noise. Consequently, the gene expression vector is of probabilistic nature, and often modelled as an instance of a random vector **X**.

2.2.3 Proteomic Data

Proteins are considered as the working units of the cell involved in almost all inter- and intra-cellular processes. In contrast to the genome, the proteome differs highly among cell types and states (cf. Figure 2.1). Measuring the proteome would therefore be an important step towards a better understanding of the cellular machinery. The widespread high-throughput microarray techniques capture the proteome only indirectly by measuring the transcriptome. Although the mRNA concentration might be related with the final protein quantity, microarray measurements miss important effects, such as post-translational mechanisms or varying protein degradation and hence ignore important influences. This makes such techniques uncapable of capturing the complete proteome. Measuring proteomic alteration directly is therefore one of the major issues of todays research. Present techniques are centered around two-dimensional gel electrophoresis (2DPAGE) which separates proteins by two physical characteristics. Their charge and their mass (Brown 1999, Yates 1998). Hanash and coworkers showed the power of protein analysis by successfully classifying different leukemia subtypes (Hanash et al. 1986, Hanash et al. 1989). However, these lowthroughput techniques only allow a limited view analyzing a small proportion of the proteome. Delineating the proteome as a whole, such as in DNA microarrays, is still difficult because proteins are much more difficult to handle than mRNA probes. Robust technologies for a parallel characterization of the proteome are not yet available, although high-throughput protein expression profiling technologies for large-scale proteome measurements are being put forward (Phizicky et al. 2003, Hanash 2003).

Yeast-two-Hybrid

As mentioned in section 2.1.3, many proteins interact with each other and thereby participate in molecular signalling or in reaction networks of the cell. Protein protein interactions control and regulate a large number of cellular processes, e.g. post translational modifications and are therefore crucial for the cellular machinery. The Yeast-two-hybrid method (Fields & Song 1989) is an *in vitro* assay to discover the ability of two proteins to interact with each other. It makes use of a mechanism of transcriptional regulation. A transcription factor regulates the expression of a gene by binding with its DNA-binding domain (BD) to an activation domain (AD) specific for the relevant gene. In a two-hybrid assay, protein X is fused to a DNA-binding domain of a transcription factor such as GAL4 and its potential binding partner Y is fused to the corresponding activation domain. If protein X interacts with protein Y, they form a transcriptional activator which activates the expression of a specific reporter gene, whose expression is measured.

Various studies (Uetz et al. 2000, Ito et al. 2001) use two-hybrid assays to systematically analyze the entire proteome to identify interactions which place functionally unclassified proteins into a biological context. Besides identifying novel protein-protein interactions, that were previously uncharacterized, two-hybrid assays are also used to understand the nature of protein protein interactions, by manipulating or inactivating certain proteins. Many diseases arise due to mutations causing the protein to be non-functional, or to have an altered function. The significance of such mutations can be studied by analyzing how much they affect following protein protein interactions.

2.2.4 Metabolomic Data

A fundamental task of proteins is to act as enzymes - catalysts which accelerate specific chemical reactions by lowering the activation energy. Enzymes accomplish the catalysis by binding to a substrate S forming a enzyme-substrate complex ES. While bound to the enzyme, the substrate is converted into the product P of the reaction which is then released from the enzyme.

$$E + S \rightleftharpoons ES \rightleftharpoons E + P \tag{2.1}$$

Metabolic reactions are a fundamental part of the cellular machinery: proteins, amino acids or other molecules (necessary for life processes) are generated but also degenerated through a sequence of different enzymatic reactions forming biochemical pathways which altogether create a complex entity, the *metabolic network*. Individual pathways are already well characterized and modeled at a varying level of detail. From the reactions kinetics of enzymatic processes modeled with Michaelis-Menten equations, to a static network level visualized as huge interaction maps, such as the Boehringer Mannheim wallchart, or stored in databases (Ogata et al. 1999). Together they form a huge knowledge basis for a better understanding of the metabolic network as well as the genetic network.

2.3 Summary

Molecular mechanisms, such as cell cycle control, DNA repair or apoptosis are of major interest for molecular biologists as well as for physicians or pharmacologists since their decipherment is expected to provoke a big step towards a better understanding of cellular modes, disease mechanisms or drug response. These underlying mechanisms are closely related to the perception that the diversity of cellular modes is guided by the execution of distinct genetic programs, assembled by the interaction of a huge number of molecules. In contrast to previous assumptions that only proteins regulate the cellular behavior, genetic programs emerge from the interplay of various types of molecules. Proteins are synthesized from DNA, bind to other proteins or act back to regulate the production of other proteins. Other regulatory systems emerge from the interaction of mRNA molecules with the DNA or with proteins or from enzymatic reactions.

To uncover the underlying molecular pathways and the emerging cellular phenomena, scientists are working on challenging tasks to collect molecular data. The first breakthrough was made with sequencing and annotating the genome from different species, providing information about the arrangement of genes and other functional units on the entire genome. But it became clear, that these data provide only one step and that, for a complete functional annotation of genes, all different molecular levels (genome, transcriptome, proteome, ..) need to be observed. Especially due to the invention of high-throughput techniques, such as DNA microarrays which allow the observation of thousands of mRNAs in parallel, it became possible to gain a global view of molecular activities in the cell.

However, the exploration and understanding of molecular interaction networks, their operational modes under different circumstances and their response to external signals, still remains one of the major challenging tasks of the post-genomic era. To solve this problem, data from different sources need to be analyzed and merged together in an integrated approach. Moreover, due to the complexity of the mapped system and the resulting data, information is not intuitively interpretable such that additional in depth statistical analyses are indispensable.

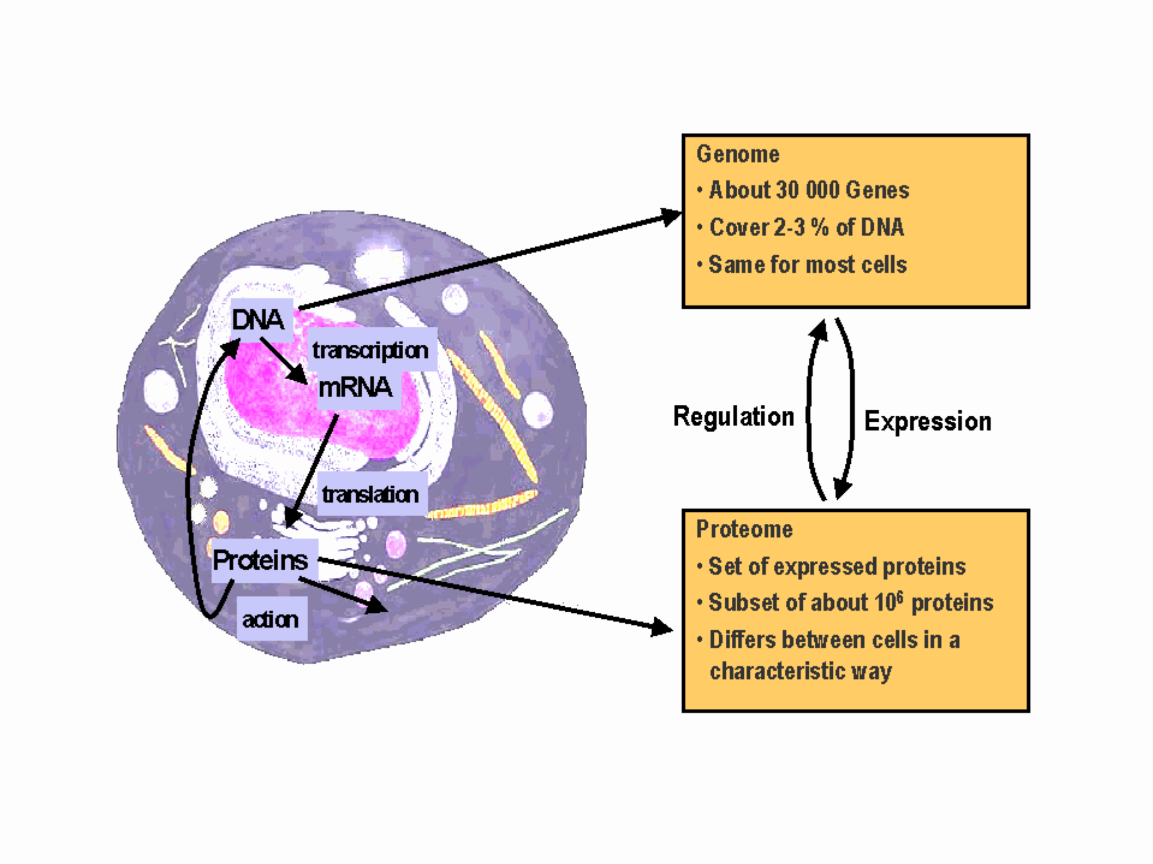


Figure 2.1: The genome and the proteome of a living cell. Transcription and translation are used to produce proteins from genes on the DNA. Proteins in turn regulate gene expression levels. For details see text.

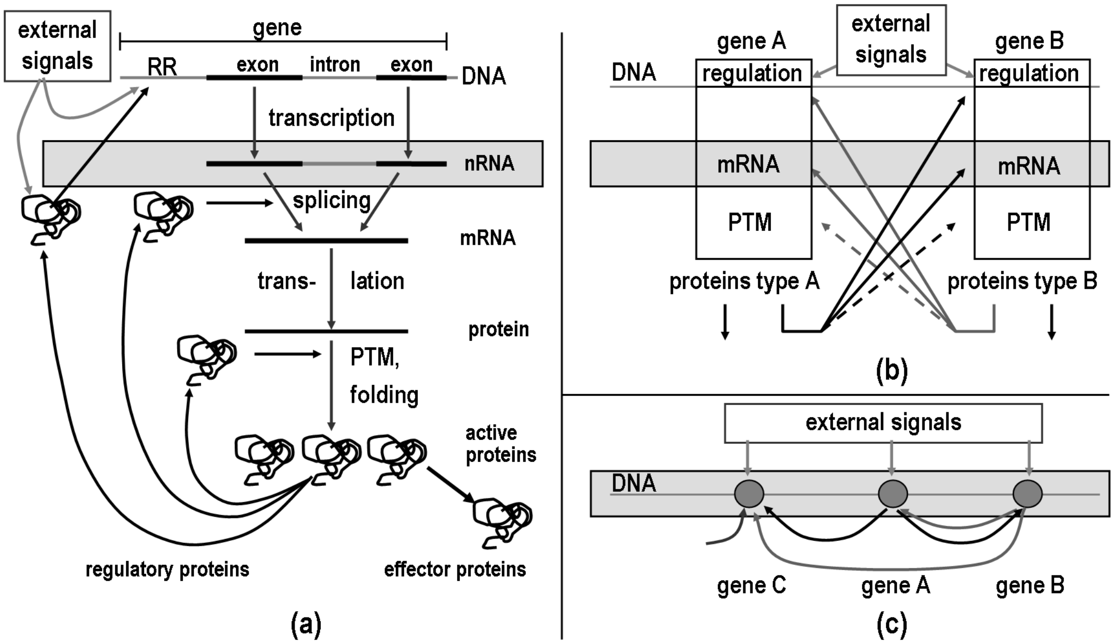


Figure 2.2: From cellular regulatory mechanisms to abstract genetic networks. **(a)** Transcription and translation mechanisms and control points for regulation. RR= regulatory region; PTM = post-translational modification (of proteins) **(b)** More schematic view of the interaction pathways between genes. **(c)** Abstract genetic network. Shaded boxes mark the part of the network measured by DNA microarray experiments.

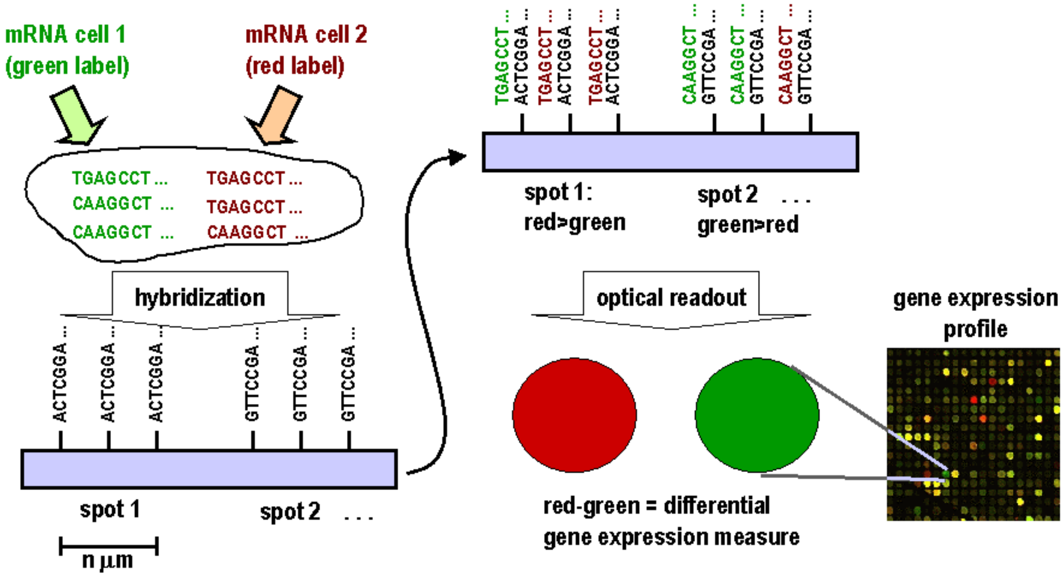


Figure 2.3: Schematic sketch of differential gene expression profiling with DNA microarrays. Left: each spot of a microarray contains single-stranded nucleotide sequences as probes which are complementary to a sequence of one gene. Center: differently labelled RNA molecules (light and dark gray) from two samples are brought in contact with the array, and are hybridized. Confocal fluorescence microscopy is used to optically determine the relative fraction of RNA from each cell and for each spot (gene). Right: a microarray usually contains many thousands of sample spots. The spot size n ranges from 25-500 μm, depending on the type of microarray.

Chapter 3

Genetic Network Inference with Graphical Models

The various modes of a cellular system are accomplished by the interplay of thousands of molecules, e.g. proteins which bind back to the DNA and regulate the synthesis of other proteins, or several proteins which form a functional complex by binding with each other. Altogether these molecular interactions form the molecular network. An intuitive way of modeling such a network might be to construct a graph structure G where nodes correspond to molecules and edges indicate relationships among them. Nodes can be associated with genes, proteins or other molecules and edges represent for example a transcriptional regulatory mechanism or a simple binding relation. This abstracted graphical representation might help to gain a clearer understanding of complex cellular mechanisms. Although graph structures are suitable for describing the qualitative nature of molecular relationships, they completely miss the quantitative nature of regulation, for example, if a transcription factor inhibits or promotes the expression of a certain gene and all other biochemical quantities.

Here the family of graphical models comes into play. Graphical models represent a family of probabilistic models which describe the relationships among a set of random variables $\mathbf{x}$ in terms of a directed or undirected graph structure G and a set of parameters Θ, given as probability distributions. This scheme fits the scenario mentioned above where variables correspond to molecules and edges denote general molecular interactions.

Thus, the intuitive representation as a graph structure and the probabilistic semantic fit the stochastic nature (Gillespie 1977, McAdams & Arkin 1997) and the network character of biological processes very well and make graphical models a good candidate for modeling genetic networks.

This chapter deals with the problem how to model the multi-dimensional probability distribution of microarray data sets in terms of a graphical model: how to learn the structure G and the set of parameters Θ from data D, a procedure known as structure learning. Since microarray data are argued to provide snapshots of the cellular system, the resulting model might represent some structural and functional aspects of the underlying genetic network. All members of the graphical model class consist of a network structure and a set of parameters. However, they differ in the way the joint probability distribution over variables is factorized into a set of conditional and marginal probability distributions.

In the following, two types of graphical models are used for genetic network inference: Bayesian networks and decomposable models. Due to their differences in factorizing the global probability distribution, each model might fit different aspects of the genetic network. Bayesian networks are useful for modeling molecular regulatory relationships, whereas decomposable models focus on the modular way in which different molecules act together to accomplish a certain task in the cellular machinery.

3.1 Learning Genetic Networks with Bayesian Networks

Molecular regulatory mechanisms form a fundamental part of the cellular machinery: proteins bind back to the DNA strand and bias the expression of certain genes, or they interact with other proteins, modify their three dimensional structure and – as a consequence – change their function. Thus, regulatory events appear at each molecular level, from the genome to the metabolome and, even more important, also between the different molecular scales of the cellular system (see Section 2.1.2 for a more detailed description). Figure 3.1a sketches several regulatory mechanisms by which different molecules can regulate each other. Protein A regulates the synthesis of protein B which in turn regulates together with protein D as a transcriptional activator the expression of gene C. Finally, on a proteomic scale, protein C alters protein E through a protein-protein interaction.

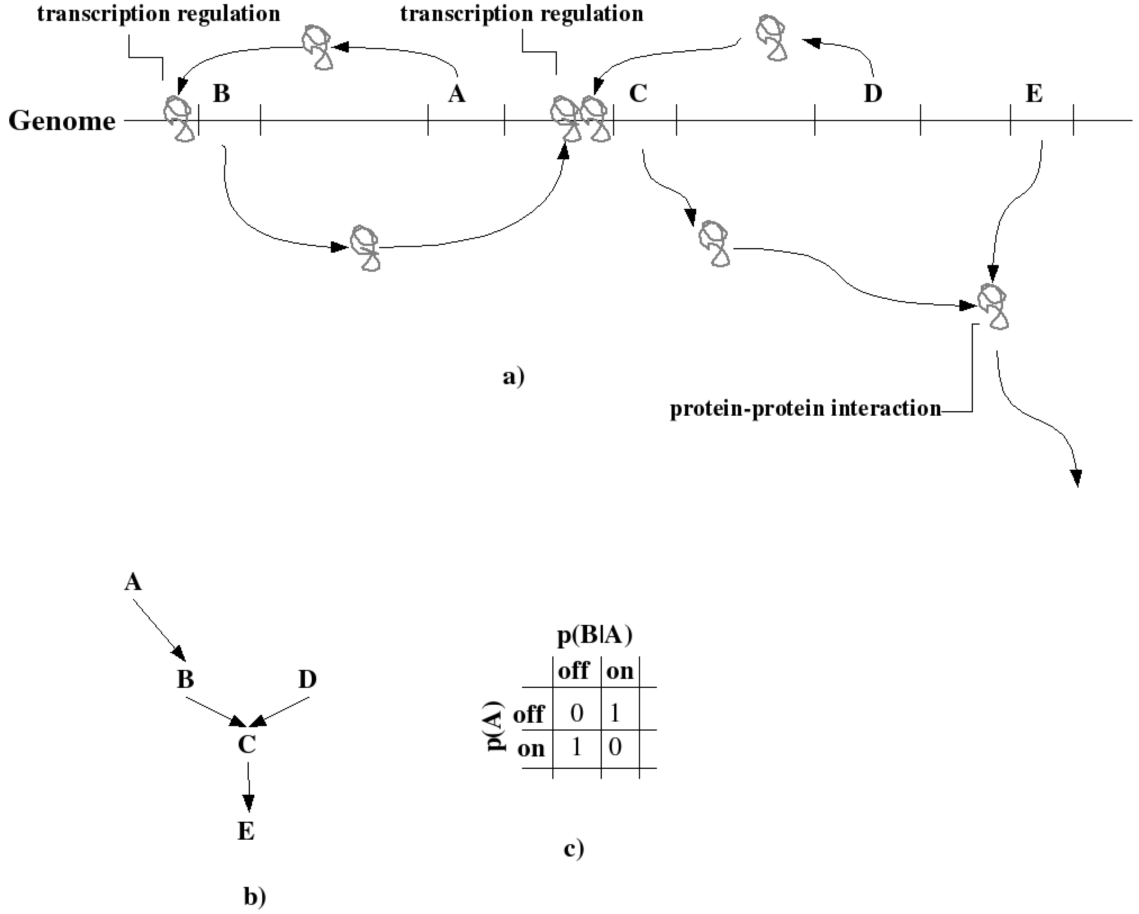


Figure 3.1: From cellular regulatory mechanisms to an abstracted probabilistic causal model. **a)** Regulatory mechanisms appear at various molecular levels. **b)** The set of regulatory mechanisms can be modeled by a directed graph whose edges code for causal relationship. **c)** Given a multinomial model, the regulatory mechanisms can quantitatively be described as a table of conditional probability entries.

Each of these events forms a causal relationship between two molecules and can be described by a directed edge pointing from the regulating molecule to the regulated one. Moreover, all processes in Figure 3.1a can be redrawn by means of a directed graph, where nodes correspond to genes, proteins or other molecules and edges describe regulatory mechanisms among them. This results in a simple graph structure, as shown in Figure 3.1b, of 6 variables interconnected by a set of directed edges. Edges symbolize causal relationships between the genes they connect. For example, the state of gene B is said to be caused by the state of gene A. This corresponds to the biological content of Figure 3.1a where the expression of gene B depends on the expression of gene A. As mentioned above the graph structure itself provides only a qualitative description of molecular interactions and lacks the quantitative aspect. Bayesian networks describe such regulatory relationships qualitatively, by a directed graph structure, as well as quantitatively, by a set of conditional probability distributions. Suppose each molecule can be in one of two states, on and off, the conditional probability distribution of gene B can be represented as a 2×2 table, such as in Figure 3.1b. The entries in the table state that A inactivates molecule B since the probability of B being active is 0, P(B=on|A=on)=0, when A is active, and 1 if A is inactive, P(B=on|A=off)=1. In addition, distributions are not limited to boolean functions, as sketched here, but entries are assigned with probability values. Thus, a Bayesian network model is well suitable to describe molecular regulatory mechanisms qualitatively as well as quantitatively, even though in a more abstract way as biochemical driven models. The following section deals with the problem of learning Bayesian networks out of microarray data and addresses the question of how well they can model complex molecular regulatory networks.

3.1.1 Bayesian Belief Network

A Bayesian network (Pearl 1998) is a probabilistic model which splits the joint probability distribution over a set of random variables $\mathbf{x} = \{x_1, ..., x_n\}$ into a set of local *conditional probabilities* in terms of a graph structure G and a set of parameters Θ such that

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i | x_{\{1,...,i-1\}}, \Theta, G). \tag{3.1}$$

Besides decomposing the probability distribution over **x** in a set of local probabilities, a Bayesian network also exploits independency relationships among these variables such that a set of variables $pa_i \in \mathbf{x}$ exists which renders variable x_i and $\{\mathbf{x} \setminus pa_i\}$ independent. Thus, Equation 3.1 can be rewritten as follows

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i | pa_i, \Theta, G). \tag{3.2}$$

The first part of a Bayesian network is a graphical structure $G = (V, E)$, with a set of nodes $V = \{1, ..., n\}$ and a set of edges E. Each node $i \in V$ corresponds to the random variable $x_i \in \mathbf{x}$, e.g. a molecule, edges represent the conditional dependencies and independecies among them. The structure of a Bayesian network is defined by a directed acyclic graph (DAG) which means that all edges are marked with a unique direction and no cycles appear (i.e. starting from any given node and following the direction of the edges, there is no way to cycle back to the original node). The conditional independencies among the variables defined in Equation 3.2 are encoded in the graph structure and can be explained by the *Markov independence* relation. It states that the state of each variable x_i depends only on the states taken by its parents pa_i, where $x_j \in pa_i$ is called a *parent* of x_i which is symbolized by an edge pointing from x_j to x_i. Consequently $x_i \in ch_j$ is called a *child* of x_j.

D-Separation Criterion

The complete relationship between probabilistic independence and the graph structure of a Bayesian network is given by the concept of d-separation (Pearl 1998) which states that: two variables a and b in a network G are d-separated given an intermediate variable c $(a \perp b | c)$ if for all paths between a and b

- c is a node of a serial or diverging connection and its state is known or
- c is a node of converging connection, called collider, and neither c nor any of its descendants is known.

It can be shown that the d-separation criterion results in the same set of conditional distributions as defined in Equation 3.2 (Verma & Pearl 1990).

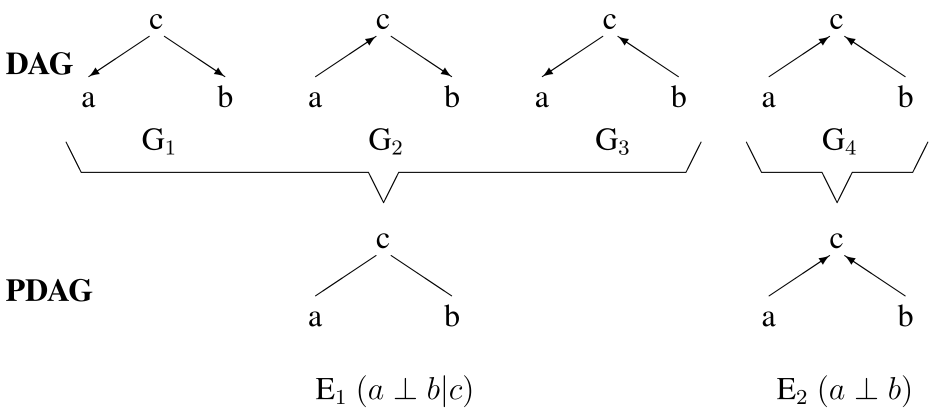


Figure 3.2: Instead of one DAG a set of DAGs, an equivalence class, describes a given joint density in an indistinguishable way. G_1, G_2 and G_3 code for the same joint density distribution and build the equivalence class E_1. G_3 on the other hand, represents a unique d-separation scheme and cannot be represented by another DAG.

Structure Equivalence

According to the d-separation criterion, the DAG of a Bayesian network graphically describes the conditional dependence and independence relationships encoded in the probability distribution among the set of variables. However, it is not guaranteed that the conditional dependencies and independencies lead to a unique DAG but instead to many DAGs which altogethers describe the same probability distribution equally. This problem is known as *structure equivalence* and can be formulated as follows: Two DAGs are equivalent if and only if they have the same set of edges and the same set of colliders.

This implies that two equivalent DAGs represent the same set of d-separations and therefore also the same probability distribution even though they differ in the direction of some edges.

G_1, G_2 and G_3 in Figure 3.2 for example state the same d-separation namely that variable a and b are d-separated given variable c. Thus, even though the structures differ in the direction of some edges they are all structure equivalent, denoted by $G_1 \sim G_2 \sim G_3$, and belong to the same equivalence class, namely E_1. Equivalent structures can be redrawn as a *partial* directed acyclic graph (PDAG) which can contain directed as well as undirected edges. Undirected edges have no direction whereas directed ones are labeled with an irreversible unique direction. The resulting PDAG for the equivalence class E_1 only contains undirected edges, since each edge varies in its direction across the class members. The assumption of structure equivalence can be easily proven by analyzing the probability distribution. Using Bayes' rule, the probability distribution of a DAG can be transformed into the distribution of any other member of the same equivalence class, e.g. for the example in Figure 3.2

$$\begin{aligned} E_1 :\ p(a,b,c) &= \underbrace{p(a|c)p(b|c)p(c)}_{p(\mathbf{x})_{G_1}} = \underbrace{p(c|a)p(b|c)p(a)}_{p(\mathbf{x})_{G_2}} = \underbrace{p(a|c)p(c|b)p(b)}_{p(\mathbf{x})_{G_3}} \\ E_2 :\ p(a,b,c) &= \underbrace{p(c|a,b)p(a)p(b)}_{p(\mathbf{x})_{G_4}} . \end{aligned}$$

Consequently a Bayesian network model can not necessarily be interpreted as a *causal* model since putative undirected edges of the corresponding PDAG do not represent causal relationships anymore. For example, part of the scenario of molecular regulatory mechanisms described in Figure 3.1a can be modeled by a Bayesian network structure in terms of causal relationships (see Figure 3.1b). However, the edge between A and B can be reversed in the graph (but not necessarily in the biochemical network) without changing the set of colliders. Hence for this relationship no unique graphical representation exists and no statement about the causal relationship among these two molecules can be made. In this case, additional prior knowledge is required to assign a direction to the undirected link. Given that the real scenario in Figure 3.1a is known, the edge pointing from A to B would be preferred to the reversed one, since protein A is known to regulate the transcription of gene B. Thus, the problem of structure equivalence can be best addressed by using additional prior domain knowledge. Chapter 7 presents two approaches which use biological prior knowledge to solve the problem of equivalent structures.

3.1.2 Parameter and Structure Learning

A Bayesian network models the conditional dependencies and independencies among a set of random variables **x** in terms of a network structure G and a set of parameters Θ. In the domain of structure learning the underlying probability distribution which encodes these conditional relationships is inferred from a finite database D of N cases, where each case includes observations of one or more variables in **x** (Buntine 1996).This unsupervised learning method can be divided into two problems. In the first case the structure is already known and only the parameters have to be learned from the data set (e.g. naïve Bayes approach). The second task is the more difficult, since, besides the parameters, also the structure has to be learned from the data set. This work focuses on the second problem, namely learning the parameters as well as the structure of genetic network systems from microarray data.

The procedure of structural learning can be summarized as follows: Let $D = \{d^1, ..., d^N\}$ be a data set of N independent observations, where each data point is an n-dimensional vector with components $d^l = (d^l_1, ..., d^l_n)$. Find a graph structure G and a parameter set Θ that best match D.

In the present context the data set is given by N independent microarray experiments, each observing the expression states of n probes or genes. Each node in the learned Bayesian network symbolizes a specific probe or gene and the structure represents the conditional dependency relationships among these molecules regarding the cellular conditions from which microarray samples where taken. In the same way as microarray data are considered to provide a snapshot of the cell at the observed state, the learned Bayesian network might provide patterns of molecular interplay, visible from the observed transcriptomic data.

Fitness Function

To evaluate the goodness of fit of a network graph structure G with respect to the data, a statistically motivated scoring function S assigns a score $S(G|D)$ to the graph G. In the following, data are supposed to be multinomial such that the Bayesian model is of multinomial nature too. In a multinomial model each gene can have several discrete states, e.g. $\{$underexpressed, normal, overexpressed$\}$. When each variable x_i can assume r_i different values k and the set of parents pa_i can assume q_i different values j, the local multinomial conditional probability distribution can be represented as a $r_i \times q_i$ table. Each parameter entry in the table is given by

$$\theta_{ijk} = p(x_i = k|pa_i = j, G), \tag{3.3}$$

where the parameters satisfy the constraints $0 \leq \theta_{ijk} \leq 1$ and $\sum_k \theta_{ijk} = 1$. Equation 3.3 can be approximated by taking the relative frequencies N_{ijk} as estimates for θ_{ijk} such that

$$\hat{\theta}_{ijk} = \frac{N_{ijk}}{N_{ij}}. \tag{3.4}$$

N_{ijk} denotes the number of cases in data set D in which $d_i^l = k$ and $pa_i(d^l) = j$, $N_{ij} = \sum_k N_{ijk}$.

Before discussing different scoring metrics, it is necessary to outline two fundamental properties of a scoring function.

Decomposability Since the conditional probability distribution of a Bayesian network decomposes into a product of local probabilities, see Equation 3.2, one might suspect that the same holds for the scoring function. In fact, if the data set contains neither missing nor hidden values, the score can be decomposed into a set of local scores such that

$$S(G|D) = \prod_{i=1}^{n} S_i(G|D). \tag{3.5}$$

$S_i(G|D)$ denotes the local score of variable i which only depends on the states of x_i and pa_i.

Score Equivalence Two structure equivalent DAGs belong to the same equivalence class since they represent the same conditional probability distribution. Only in the case of prior knowledge it might be possible to distinguish between equivalent DAGs. Otherwise there is no reason for favoring a particular equivalence class member. Due to this, Markov equivalent DAGs should also be assigned with the same score value to be score equivalent

$$G_1 \sim G_2 \Rightarrow S(G_1|D) = S(G_2|D). \tag{3.6}$$

Frequentist Score

According to the frequentist approach, where the data set D represents the ”true” joint probability distribution, the structure G which best fits the data is assigned with the highest score. The fitness of a network structure G according to a data set D can therefore be estimated in terms of the maximum likelihood function (cf. Equation A.11) which, taking the logarithm, is given by

$$\log p(D|\Theta, G) = \sum_{i=1}^{n} \sum_{j,k} N_{ijk} \cdot \log \theta_{ijk}. \tag{3.7}$$

One general problem of likelihood scoring metrics is the effect of overfitting which becomes even more problematic when the data contains only a small amount of samples. Thus, Equation 3.7 is extended by a penalty term p which penalizes the complexity of a network structure

$$S(G, \Theta, D) = \sum_{i=1}^{n} \sum_{j,k} N_{ijk} \cdot \log \theta_{ijk} - p. \tag{3.8}$$

A wide range of model complexity penalty functions has been suggested, among them the *Akaike Information Criterion* (AIC) (Bozdogan 1987) and the *Bayesian Information*

Criterion (BIC) (Schwarz 1978). The AIC complexity penalty is $|\Theta|$, the effective number of parameters in the model. The BIC term is $\frac{1}{2}|\Theta| \log N$, where N is the number of samples. Another prominent frequentist scoring function is the *Minimum Description Length* (MDL) which is identical to the negative BIC score but with a completely different origin, namely from coding theory (Wai & Fahiem 1994) (for a comparison of these model selection criteria see (Allen & Greiner 2000)).

Although the maximum likelihood score is quite useful for structure learning it has to be noted that only in the limit $N \rightarrow \infty$ the maximum likelihood estimate converges to the true value, whereas for small sample size the maximum likelihood produces biased results (Bishop 1995). This has to be taken into consideration especially in the domain of learning from microarray data where sample number is very low.

Bayesian Score

According to the Bayesian approach the data set does not represent the ”true” joint probability distribution and is therefore just used to revise our degree of belief including our already gained *a priori* knowledge. The Bayesian score is proportional to the posterior probability of the graph G given the data D,

$$S(G|D) = \frac{p(D|G)p(G)}{p(D)}, \tag{3.9}$$

where $p(G)$ is the prior probability of the structure, $p(D)$ is a normalization constant, and $p(D|G)$ is the marginal likelihood of the data given the structure G. In contrast to the frequentist scoring functions which relies on the maximum likelihood parameters $\hat{\theta}_G$ the Bayesian score treats the parameters of a model as random variables characterized by a distribution. This uncertainty over the parameters is expressed by marginalizing out the parameters. Thus, the marginal likelihood equals the integral:

$$p(D|G) = \int p(D|\Theta, G)p(\Theta|G)d\Theta, \tag{3.10}$$

where $p(\Theta|G)$ denotes the prior density of the model parameters Θ for a given structure, and $p(D|\Theta, G)$ is the likelihood of the data given a Bayesian network.

Given a *conjugate prior* which for the multinomial case implicates that $p(\Theta|G)$ follows a *Dirichlet distribution* and given other plausible assumptions such as data completeness, parameter independence and parameter modularity (see (Cooper & Herskovits 1991)), Equation 3.10 can be solved in closed form (see Appendix A for a more detailed description). The solution is a unique scoring function for the multinomial model namely the Bayesian Dirichlet (BD) score (Heckerman & Chickering 1995). Approximating Equation 3.3 by relative frequencies leads to a closed form solution for the BD score

$$p(D|G) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})}, \tag{3.11}$$

where N'_{ijk} express parameters of the Dirichlet prior distributions, $N'_{ij} = \sum_k N'_{ijk}$, and Γ denotes the gamma function.

Because member graphs of an equivalence class are indistinguishable, the Bayesian score in Equation 3.9 has to assume identical values for all members of a certain class to ensure score equivalence. To achieve this, Heckerman and co-workers proposed a non-informative parameter prior $N'_{ijk} = \frac{1}{q_i r_i}$. Together with the non-informative prior Equation 3.11 forms the so-called BD equivalent score (BDe).

As in the frequentist approach, also in this scoring metric the problem of overfitting needs to be reduced by punishing networks complexity. For this, the structure prior $p(G)$ can be used as a penalty term such that the prior probability of a structure G decreases the more complex it becomes. In (Heckerman & Chickering 1995), the number of edges δ is used as a measure for the complexity of structure G such that the structure prior is given as

$$p(G) = ck^{\delta}, \tag{3.12}$$

where k is a constant factor $0 < k \leq 1$ and c is a normalization constant. It has to be noted that the prior belief in Equation 3.12 does not reflect real expert knowledge regarding the observed (cellular) system, favoring the presence or absence of specific edges (cf. Chapter 7).

Search Strategies

Because the task of finding the optimal structure of a Bayesian network is NP-hard (Chickering et al. 1994), one needs to adopt heuristic search strategies which can efficiently determine a Bayesian network close to the optimum. As outlined above, the scoring function can be decomposed into a product of local scoring functions restricted to each family (a variable x_i and its parents pa_i). Each term can be defined as the local score of x_i which depends only on the state of x_i and pa_i. This decomposition property is crucial for learning structures, since a local search procedure that changes one edge at each move can efficiently evaluate the gains made by this change. This implies that at each step only the local scores of those variables whose set of parents has been changed have to be re-evaluated. For a structure G, the structures which differ only in the presence or absence of one edge and satisfy the acyclicity condition (using depth-first search), represent the so-called neighboring structures $nbg(G)$. The structure $G\prime \in nbg(G)$ which entails the best scoring function is selected as the next candidate structure $G\prime$. This technique is known as *local search* strategy and a commonly used approach in structure learning. If $G\prime$ is selected as the next structure G depends on the heuristic which used.

Greedy Search The simplest heuristic strategy is *greedy search*. In each iteration the space of neighboring structures is tested for an improvement in the score. The neighboring structure $G\prime$ which entails the largest improvement with respect to G becomes the next intermediate structure which is the starting point for the next iteration. Thus, the algorithm always moves across the model space in the direction of the greatest rate of decrease of the error which is quite similar to the *gradient descent* algorithm for training neural networks (see (Bishop 1995)). A big drawback of this search strategy is that instead of finding the global optimum it might get stuck in a local optimum.

Simulated Annealing Searching the model space by only accepting improvements in the model score might be a pitfall and leads to a suboptimal optimization result. Simulated annealing (SA) tries to overcome the problem of getting stuck in local optima by allowing up-hill as well as down-hill steps regarding the scoring function. Similar to greedy search, SA in each iteration selects the neighboring structure with the best score. If the change increases the score, the selected structure becomes the new intermediate structure. In case of a down-hill step, the neighboring structure is being accepted with a certain probability $p = e^{\frac{-\Delta}{T}}$, where $\Delta = S(G) - S(G\prime)$ and T is the current temperature. The algorithm is an extension of the Metropolis algorithm (Metropolis et al. 1958) and has its origin in cooling theory. It was originally proposed as a means of finding the equilibrium configuration of a collection of atoms at a given temperature (Pincus 1970) and further used as a global optimization technique (Kirkpatrick et al. 1983). Starting from a high value T_{start}, where almost every proposed transition is accepted, the pseudo-temperature T is cooled down to a low temperature where only up-hill steps are allowed.

If T is annealed logarithmically, SA is guaranteed to converge to the global optimum (Laarhoven & Aarts 1987). But since logarithmic cooling is of prohibitive computational expense, one has to find a compromise between search-accuracy and computational time. However, simulations showed that even with a fast cooling scheme, SA reaches better results than greedy-hillclimbing (Steck 2001, Dejori & Stetter 2003a).

3.2 Learning Genetic Networks with Decomposable Models

Besides regulatory relationships, the molecular interplay in the cellular system can be described in terms of functional modules, composed of different molecules which act together to accomplish a certain task. Functional modules are considered to be a critical aspect of biological organization (Hartwell et al. 1999) at each molecular scale. A functional module is defined as a discrete entity whose function is separable from those of other molecules. Most cellular processes are the result of a set of interacting molecules, rather than the result of the activity of an individual molecule. Transcriptional regulation for example is often not affected by a single transcription factor but by several ones which alter the transcription rate by forming a complex. Further examples of molecular modules are subunits of multimeric proteins, e.g. hemoglobin, where the subunits are coded by separate genes, or protein groups which associate into larger structures termed macromolecular assemblies or reaction chains.

Figure 3.3a shows two ways in which proteins form modular compositions. Protein A, B and D form a transcriptional regulatory complex which alters the expression rate of gene C. This complex might, for example, act like a logic AND-gate such that only when all 3 proteins bind the promoter region gene C is being expressed. Protein C in turn groups together with protein E and F to form a multimeric protein. This molecular system can be represented by an undirected graph, where molecular functional modules are characterized by a set of fully connected nodes, so-called cliques. The graph structure shown in Figure 3.3b consists of two cliques: one is composed of A, B and D and the other one of C, E and F. The decomposition of variables into cliques is provided by a decomposable model: variables are grouped into overlapping subsets of fully linked nodes (cliques) and nodes which participate in more than one clique form a separator which connects the cliques sharing this node. Figure 3.3c shows the corresponding decomposable structure which

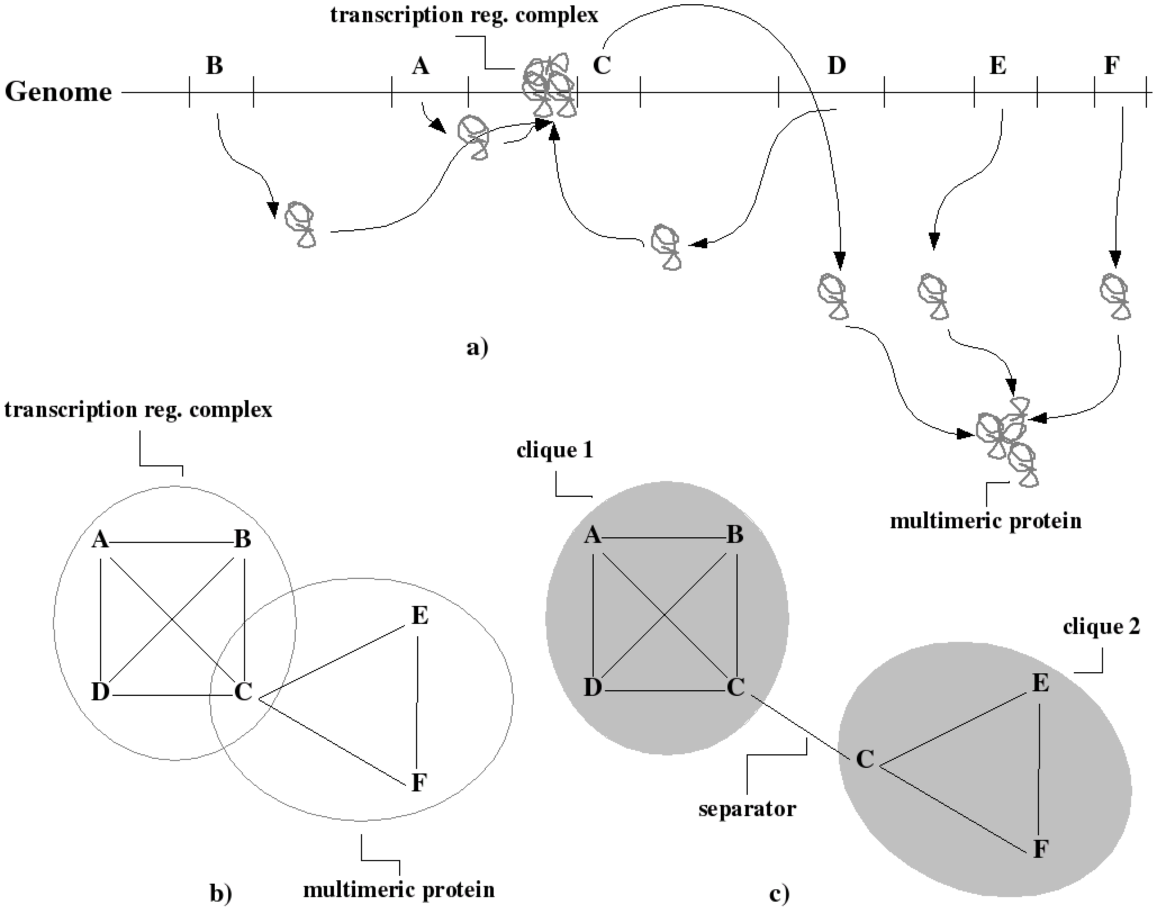


Figure 3.3: From molecular modules to an abstracted probabilistic decomposable model. **a)** Molecules group together and form functional modules to accomplish specific tasks in the cellular system. **b)** A functional module can be represented by a set of fully linked molecules, a clique. **c)** The whole set of modular relationships can be described by a decomposable model, in terms of cliques linked by separators.

models the relationships between the molecular functional modules Figure 3.3a.

The aim of learning decomposable models out of microarray data is to infer modular relationships among molecules and not, as with Bayesian networks, to reveal regulatory relationships. This modular view might contribute to a more profound knowledge of how molecules act together and form (previously unknown) functional groups.

3.2.1 Decomposable Model

Decomposable models belong to the family of graphical models. The probability distribution over a set of random variables **x** is described in terms of a graph structure G and a set of parameters Θ. The structure G is that of a general Markov model, namely fully undirected. Thus, in contrast to Bayesian networks, where each variable is characterized by its set of parents, variables are characterized by a set of adjacent variables sharing dependencies with it. Whereas the joint density in the Bayesian network model factorizes in a set of conditional densities, they decompose in a set of smaller joint densities in decomposable models. See Figure 3.4 for a schematic comparison of joint and conditional densities.

Therefore, the main characteristic of decomposable models is the way in which the joint density over **x** is decomposed into a set of local marginal densities. For this the notion of a *clique* is required. A clique C is defined as a maximal subgroup of nodes $\mathbf{x}_C \in \mathbf{x}$ which are mutually fully connected. $\mathbf{x}_S \in \mathbf{x}$ denotes the subset of variables that form separator S. The graph structure G can therefore be decomposed into a set of K cliques and Σ separators. Furthermore G can be transformed into a so-called *clique tree* a particular tree in which the cliques form the nodes. Each edge in the tree represents a *separator* which contains the nodes common to the cliques linked by the edge. The joint density over **x** of a decomposable model factorizes into the product of marginal densities over cliques, divided by the marginal densities over separators

$$p(\mathbf{x}) = \frac{\prod_{C \in K} p(\mathbf{x}_C)}{\prod_{S \in \Sigma} p(\mathbf{x}_S)}. \tag{3.13}$$

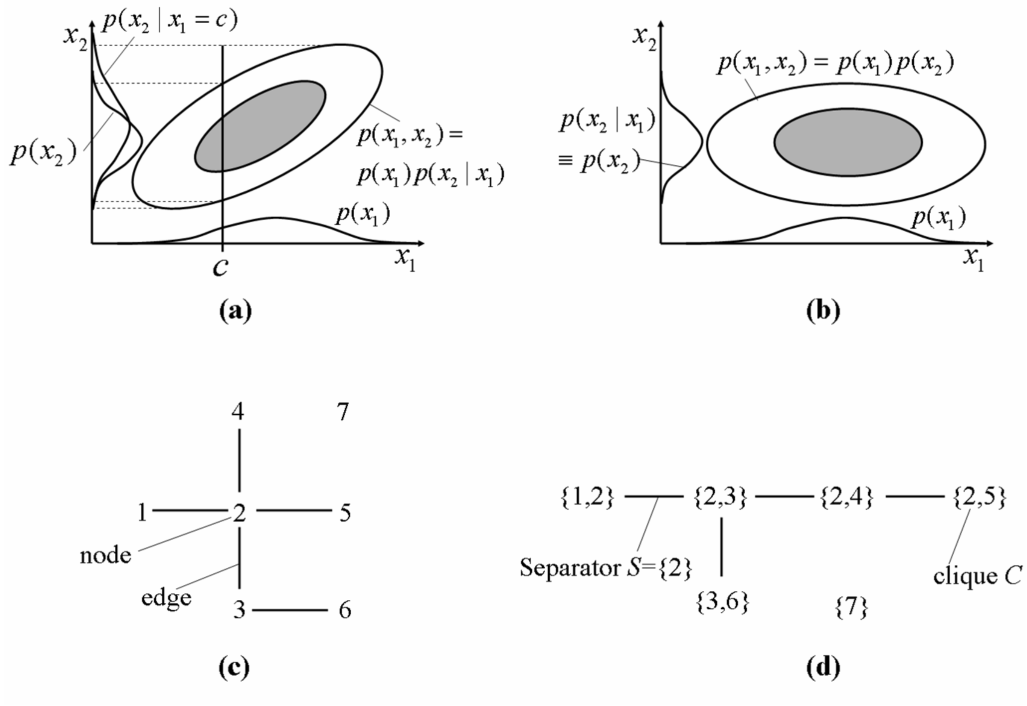


Figure 3.4: Schematic illustration of decomposable graphical models. **a)** Sketch of a two-dimensional joint probability density $p(x_1, x_2)$. The marginal densities $p(x_1)$ and $p(x_2)$ are obtained by integrating over the respective other variable. In this example, the conditional probability density $p(x_2|x_1)$ differs from the marginal density, reflecting a statistical dependency between x_1 and x_2. **b)** Same as **a)** but for statistically independent variables x_1 and x_2. Here conditional and marginal probability densities coincide and the joint probability density factorizes. **c)** Graph structure of a simple decomposable graphical model with 7 nodes. Each node i stands for a variable x_i, and each edge reflects a direct statistical dependency. **d)** A join tree equivalent to the graph structure of **c)**. Each node of the tree stands for a clique of fully connected variables, and each edge reflects a set of variables common to adjacent cliques: their separator.

Nonparametric Density Estimation

As previously shown, a decomposable model splits the joint density over **x** into a set of marginal densities. In the following a nonparametric probability density estimate, namely a kernel density estimator, is used for modeling these marginal clique and separator densities. A kernel density estimator consists of the superposition of N kernel functions g centered at each data point

$$p(\mathbf{x}|D, \Theta) = \frac{1}{N} \sum_{i=1}^{N} g(\mathbf{x}; \mathbf{x}^i, \Theta), \tag{3.14}$$

where N is the number of samples of data set D. In the following a Gaussian kernel with diagonal covariance matrix is used

$$g(\mathbf{x}; \mathbf{x}^i, \Theta) = \frac{1}{(2\pi)^{\frac{n}{2}} |\text{diag}\Theta|^{1/2}} \exp\left(\frac{1}{2} (\mathbf{x} - \mathbf{x}^i)^T (\text{diag}\Theta)^{-1} (\mathbf{x} - \mathbf{x}^i) \right). \tag{3.15}$$

In Equation 3.15, Θ denotes the vector of variances along the n dimensions of the data space. It is worth noting that although the kernel function is Gaussian, the nonparametric probability density estimate in Equation 3.14 is a superposition of Gaussians and can in principle describe any probability density function. This is an important extension to other approaches of structure learning from microarray data which are restricted to Gaussian densities or discrete data. Both assumptions are limited: the probability density of microarray data is not known but it is likely to be non-Gaussian and their discretization might mask statistical dependencies.

In (Schwaighofer et al. 2004), the variance parameters Θ are fitted once for a fully connected model, i.e. all nodes form one large clique, using a leave-one-out procedure. The densities for the low-dimensional marginal densities $p(\mathbf{x}_C)$ and $p(\mathbf{x}_S)$ can be calculated by marginalizing the joint density. This procedure has two important advantages: first the density estimate is consistent which means that calculating a marginal probability density with different sequences of marginalization always yields the same result. Second, computational complexity is reduced by avoiding a permanent re-fitting of the density model.

Thus, Equation 3.13 can be rewritten as follows

$$p(\mathbf{x}) = \frac{\prod_{C \in K} p(\mathbf{x}_C | D_C, \hat{\Theta}_c)}{\prod_{S \in S} p(\mathbf{x}_S | D_S, \hat{\Theta}_S)}, \tag{3.16}$$

where the vectors in D_C, $\hat{\Theta}_C$, D_S and $\hat{\Theta}_S$ contain only the dimensions (or nodes) which belong to clique C or separator S, respectively. The hat denotes the fitted values of a variable.

3.2.2 Structure Learning

Structure learning of decomposable models requires a criterion to score the quality of a given model structure for describing the data, such as a cross validation scheme. For this, the data set is divided into k disjoint sub sets D^k, then the model is learned using $k-1$ sets of the available data and is evaluated based on the log-likelihood of the remaining data points D^k which had not been used for training. The procedure is repeated k times with different sets and the final score is given by the average of the k test set log-likelihoods.

Since the joint probability density factorizes as given by Equation 3.13, the log-likelihood becomes

$$L(T) = \sum_{j=1}^{A} L(C_j) - \sum_{i=1}^{B} L(S_i) \tag{3.17}$$

with

$$L(C) = \sum_{k} \sum_{\mathbf{x}_c \in D_C^k} \log p(x_C | D_C \setminus D_C^k) \tag{3.18}$$

$$L(S) = \sum_{k} \sum_{\mathbf{x}_s \in D_s^k} \log p(x_S | D_S \setminus D_S^k), \tag{3.19}$$

and where A denotes the number of cliques and B is the number of separators. Cross validation enforces the model to explain previously unseen data. The log likelihood in Equation 3.17 is used to score the goodness of a model.

Besides the scoring scheme, an efficient search strategy to walk through the model space is required. While in Bayesian network learning the structure has to be acyclic, here in each search-step the condition of staying within the class of decomposable models has to be satisfied. This can be efficiently done with the *cordiality check procedure* introduced in (Ibarra 2000) and described in (Schwaighofer et al. 2004).

3.3 Summary

The various modes and states of the cellular machinery derive from the complex intercellular interplay of thousands of molecules. This wealth of molecular relationships can be described (however not completely) by means of a network structure where nodes represent molecules and edges stand for relationships among them. An appropriate statistical way for describing such molecular networks is given by the class of graphical models where the relationship among a set of random variables is represented in terms of a graph structure and a set of local probability distributions. Especially the combination of the intuitive graphical representation and their probabilistic nature makes this kind of models suitable for describing the complex interplay of molecules in the cellular system. This chapter presented two members of the graphical model class, each of which is appropriate to describe different properties of genetic network systems. Furthermore, it is shown how these models can be learned from microarray data sets to infer the genetic network of underlying cellular mechanisms encoded in the measured expression profiles. A Bayesian network model describes conditional dependencies among molecules in terms of a directed graph structure. This representation fits the way how molecules regulate and affect each other. Such molecular regulatory mechanisms denote causal relationships which can be modeled as directed edges. However, due to the problem of structure equivalence a learned Bayesian network might not be able to fill all relationships with causality. Another big problem of Bayesian networks is their demanded acyclicity which does not fit the picture of genetic regulatory feedback loops. One solution is to consider the time domain for the network design, as in dynamic Bayesian networks (Murphy & Mian 1999), such that acyclicity constraints are reduced. Moreover, experimentally and statistically results have shown that transcriptional networks are predominated by feedforward network units which fit Bayesian network architecture (Lee et al. 2002, Shen-Orr et al. 2002).

Decomposable models represent a probability distribution as a set of cliques connected by some separators. The graph structure is fully undirected and hence not able to model molecular regulatory mechanisms as Bayesian networks. However, the decomposition of a network structure into modules ties up with another fundamental principle of the molecular interplay, namely the union of several molecules into biological functional modules to accomplish a certain task. Moreover, this method permits working directly with continuous data instead of assuming a multinomial or Gaussian distribution.

Chapter 4

Robustness Analysis of Learned Bayesian Networks

Learning the Bayesian network structure from finite data suffers from different kinds of problems which altogether become even more evident in the approach described here. Especially in the field of bioinformatics, high throughput methods, such as DNA microarrays, produce data with a high number of observed variables but with a very small number of samples. Estimating the genetic network by learning a Bayesian network is a very hard task in this domain. This chapter focuses on the problems which arise from various limitations and studies to which extent such constraints influence the performance of network structure inference.

The first problem, that of finding the optimal structure, was introduced already in Section 3.1.2. With the number of variables the space of possible graph structures grows superexponentially and makes the learning problem NP-hard. Thus, one has to resort to heuristic search strategies which, however, do not guarantee to find the global optimum and thereby incompletely identify the underlying systems from which data derive. This poses the question, whether analyzing such a limited subnetwork allows meaningful statements about global cellular mechanisms. In particular, it needs to be assessed how the obtained results are affected by the high number of missing variables which are not considered for the inference step.

When learning Bayesian networks from microarray data, these two major problems, the high-dimensional space of variables on the one side and the limitation of available data points on the other side become even more evident: for thousands of genes only a few tens of samples are usually available. This shows the necessity for methods which quantitatively explore the robustness of Bayesian network structures learned under these circumstances. The following chapter addresses these problems with various approaches and tries to identify types of structural features which can be robustly identified when learning Bayesian networks from microarray data.

4.1 Data Bootstrap

Structure learning in graphical models is partially unstable which means that the learned structure can be sensitive to small modifications in the data. Variability in the structure can be assessed by retraining the network under a $Q-$fold non-parametric bootstrap procedure (Efron & Tibshirani 1993, Zoubir 1993). In this approach at each run q, a bootstrap data set D_q is generated by re-sampling, with replacement, N data points from the original data set D of size N. D_q is not a simple permutation of D. Because of the replacement, some data points may appear in multiple copies in D_q, while others may be missing. All bootstrap data sets have the same size as the original data set, and reflect the same underlying probability density, however in general contain different data points. Retraining Q networks based on Q bootstrap samples yields a sample from the distribution $P_Q(G)$ over graph structures G.

As a consequence, one can estimate the effect of finite sample size fluctuations on fluctuations in the learned structure which leads to an estimate of the confidence of network features. The confidence measure is reliable, when the data set D can be considered to really reflect the underlying probability distribution, e.g. when the data points ergodically spread over the distribution. Then, also the bootstrap replicas will resample the complete and true original distribution. If due to fluctuations or systematic errors, D does not uniformally cover the original distribution, the same will hold for the bootstrap replicas.

4.2 Feature Partial Directed Graph

The major difficulties for any data-driven approach regard the noise of microarray measurements and the lack of enough observations. Generally, given finite data, rather than a single best model a set of models exist which describe the underlying probability distribution equally well. As a result of this effect, called *model uncertainty* which comes along with the well known problem of over-fitting, one should not rely on the estimation of a single model but on several good models to gain a better understanding of the underlying distribution. In the case of bootstrap experiments, the fluctuation throughout the Q learned models reflects the uncertainty given by the observed domain.

The variability of the structure G over Q bootstrap steps can be described by a weighted graph, the so-called *feature* PDAG (fPDAG) which emerges from the superposition of Q learned PDAGs. An fPDAG addresses the uncertainty over structural features, such as the presence or absence of an edge or other topological features. Each feature F can be assigned with a probability, that is

$$p(F|D) = \sum_q p(G_q|D) f(G_q). \tag{4.1}$$

G_q is the PDAG learned at bootstrap step q and $f(G_q)$ denotes the value of feature F in G_q. It is 1 if feature F is contained in G and 0 otherwise. Given the Q graphs $G_q, q = 1, ..., Q$ trained by the bootstrap procedure, Equation 4.1 can be approximated by the empirical mean

$$p(F|D) = \frac{1}{Q} \sum_q f(G_q). \tag{4.2}$$

One example of a topological feature is the dependency relationship between a given pair of variables $\{x_i, x_j\}$. Since the PDAGs underlying a fPDAG, describe undirected as well as directed dependencies, such a feature can be described by a probability distribution $p_{i\leftrightarrow j}$ with 4 discrete states, that is

$$p_{i\leftrightarrow j} = \{p_{i\rightarrow j}, p_{i-j}, p_{i\leftarrow j}, p_{i\perp j}\} \equiv (p_{i\leftrightarrow j}(1), p_{i\leftrightarrow j}(2), p_{i\leftrightarrow j}(3), p_{i\leftrightarrow j}(4)). \tag{4.3}$$

$p_{i\rightarrow j}$ denotes the probability for a directed edge from i to j, p_{i-j} is the probability for an undirected edge between i and j, $p_{i\leftarrow j}$ denotes the probability for an edge from j to i and $p_{i\perp j}$ denotes the probability that i and j are independent from each other. Each term satisfy the constraints $0 \leq p_{i\leftrightarrow j}(k) \leq 1$ and $\sum_k p_{i\leftrightarrow j}(k) = 1$. Each of these terms is calculated according to Equation 4.2, where $f(G_q)$ is again 1 if G contains this dependency and 0 otherwise. Moreover, the probability for an edge between node i and j, regardless of its direction, is given by $c = 1 - p_{i\perp j}$. It reflects the probability of a Markov relation between x_i and x_j and is, in the following, named as *edge confidence*.

A fPDAG over a set of n random variables **x** can therefore be represented by a *feature* distribution p which decomposes into $\frac{n(n-1)}{2}$ local probability distributions each of which describes the uncertainty for the dependency relationship between two variables. The mean uncertainty per edge of a fPDAG's probability distribution p is given quantitatively by

$$H(p) = -\frac{1}{n(n-1)} \sum_{\substack{i,j\in\mathbf{x}\\ i\neq j}} \sum_{k=1}^{4} p_{i\leftrightarrow j}(k) \log_2 p_{i\leftrightarrow j}(k), \tag{4.4}$$

which is the entropy (Shannon 1948) or the degree of disorder per edge of the probability distribution described by the fPDAG. $H(p) = 0$ if for each pair of nodes $\{x_i, x_j\}$ there is an absolute certainty about $p_{i\leftrightarrow j}$, and $\in]0, 2]$ otherwise.

Although Equation 4.4 provides a measure for the uncertainty of the learned fPDAG and therefore for the robustness of the estimated features, it does not evaluate to which extent the learned model approximates the true one which underlies the observed data.

This requires a measure for the distance between the true and the learned fPDAG, or rather, between their feature distributions. For a quantitative comparison in terms of the divergence between the feature distribution q of the true fPDAG and the estimated feature distribution p of the learned fPDAG the Kullback-Leibler divergence (Kullback & Leibler 1951) per edge was used, such that

$$K(q,p) = \frac{1}{n(n-1)} \sum_{\substack{i,j \in \mathbf{x} \\ i \neq j}} \sum_{k=1}^{4} q_{i \leftrightarrow j}(k) \log \frac{q_{i \leftrightarrow j}(k)}{p_{i \leftrightarrow j}(k)}. \tag{4.5}$$

$K(q,p)$ is non symmetric, however it represents a quasi-distance that is always non-negative and zero if and only if for each pair of variables $\{x_i, x_j\}$ $q_{i \leftrightarrow j} = p_{i \leftrightarrow j}$. The readers should be reminded, that the KL-distance in Equation 4.5 only relates to the structural component of a fPDAG and not to the probability density represented by the full Bayesian networks the fPDAG is composed of. For this see (Whittaker 1990).

Another important issue is the question which kind of deviations from the learned distribution p to the underlying true distribution q contribute to the distance $K(q,p)$. Especially one might be interested to which extent the distance results from erroneously overestimated dependencies, $K(q,p)_+ = K(q,p) \mid p_{i \perp j} < q_{i \perp j}$, or from erroneously underestimated ones, $K(q,p)_- = K(q,p) \mid p_{i \perp j} > q_{i \perp j}$. Both measures are related to the estimates of false positive edges (fp) which appear in the learned but not in the true network, and false negative edges (fn), those which appear in the true but not in the learned network. $K(q,p)_+$ relates to the divergence caused by false positives and $K(q,p)_-$ to the divergence caused by false negatives, in the framework of a weighted graph.

4.3 Analysis of Robust Features in Benchmark Networks

In order to be able to interpret the structures learned from microarray data in a correct manner, robustness parameters have to be extracted which can then be used to discriminate between true and false features. Therefore, in this section, we conduct several robustness tests on 3 benchmark data sets addressing two problems of structure learning: the variability of common heuristic search strategies and the effect of sparse data. The first benchmark case is the well known ALARM network (Beinlich et al. 1989). This established benchmark network consists of 37 discrete variables, interconnected by 46 arcs. However this network might not be an ideal benchmark to address problems related to learning genetic networks from gene expression data since it does not reflect the ”statistics” of real microarray data. For this purpose, two Bayesian networks, ALL-SIM and E2APBX1-SIM, have been constructed based on a real microarray data set (cf. Appendix B.2.1). The first one consists of 271 discrete variables each of which can have 3 values (-1, 0, +1), connected by 300 directed edges. The second one consists of 39 discrete variables each of which can have 3 values (-1, 0, +1), connected by 43 directed edges. All three networks are used in the following as generative models to generate data sets of varying sample size by use of the algorithm 6.1.1. Based on these data, structure learning results can be evaluated according to Equation 4.5, because the corresponding ”true” model is known. In order to smoothen the effect of outliers, all benchmarks are performed 5 times, with constant learning parameters but different random seeds.

4.3.1 Variability of Heuristic Search Strategies

As noted in Section 3.1.2 finding a model which best fits the data is NP-hard. To solve this problem one has to use heuristic search strategies, for example a simple greedy hill-climbing search or a more costly simulated annealing scheme. This section addresses the question if the complexity of the search strategy positively correlates with the quality of the learned structure. For this a 20-fold data bootstrap learning procedure with different search strategies has been applied on three data sets of 1000 samples. On each data set the following different heuristics have been applied: greedy search, and simulated annealing with different start temperatures from a very low temperature, $T_{\text{start}} = 0.5$, up to a high

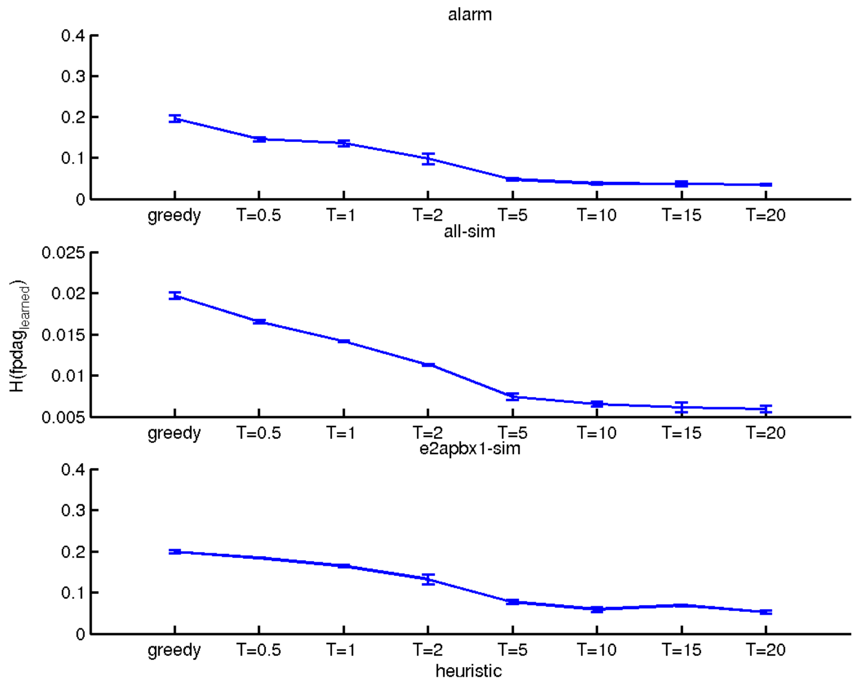


Figure 4.1: Entropy of learned fPDAGs as a function of the heuristic search strategy. The more costly the search strategy the lower the network entropy of the learned fPDAG.

value $T_{\text{start}} = 20$.

For each learned fPDAG the entropy $H(p)$ was calculated and plotted against the applied heuristic search strategy in Figure 4.1. For each network, the mean entropy (black solid line) is highest for greedy search and decays with increasing the start temperature of the SA scheme. As the computation cost increases with T_{start}, it can be concluded that the main entropy decreases the more costly the heuristic search strategy. The dashed lines represent the course of the entropy for each of the 5 runs. It can be seen that the entropy decreases consistently and almost independent of the random seed. The variability over the 5 runs remains relatively low. Thus, the more expensive the applied heuristic the more robust the learned fPDAG.

However, this does not explain if and to which extent the expense of the heuristic search leads to a closer approximation of the true graph. For this each fPDAG has been compared with the true fPDAG, from which the data set has been sampled, according to Equation 4.5. It should be noted, that since the true fPDAG is based on only one Bayesian network, the distribution over $p_{i\leftrightarrow j}$ is concentrated on a single value.

Figure 4.2 plots the KL-distance between true fPDAG and learned fPDAG (black solid line), according to Equation 4.5, as a function of the search strategy. The distance as well as the variability across training runs decay with the expensiveness of the applied heuristic. Already at a relatively low start temperature of $T_{\text{start}} = 0.5$, where only a few down-hill steps are allowed, simulated annealing performs much better than greedy search. Start temperatures higher than 5 do not result in a significant improvement. Consistently across networks it is observed that (i) the more expensive the heuristic the better the learned fPDAG approximates the true one and (ii) start temperatures $T_{\text{start}} \geq 5$ result in a good and constant learning performance.

4.3.2 The Effect of Small Sample Size

Another problem related to current microarray data is their small sample size N. From a statistical point of view, a data set of finite sample size is always incapable of exhaustively reflecting the entire density over the observed variables of the underlying system. Thus, especially for learning the genetic network of thousands of genes from a few microarray data vectors this limiting factor has to be examined. To address the impact of data sparseness on the learning result, a 20-fold bootstrap procedure has been applied 5 times on data sets of different sample sizes, ranging from 1000 down to 50 cases. All networks were learned by using a simulated annealing scheme with $T_{\text{start}} = 5$. As before first the entropy of learned networks was examined. Figure 4.3 plots the entropy as a function of increasing sample size. The result is quite comprehensible: The larger the data set the more clearly defined the underlying density which results in more robust network estimates. Figure 4.4 illus-

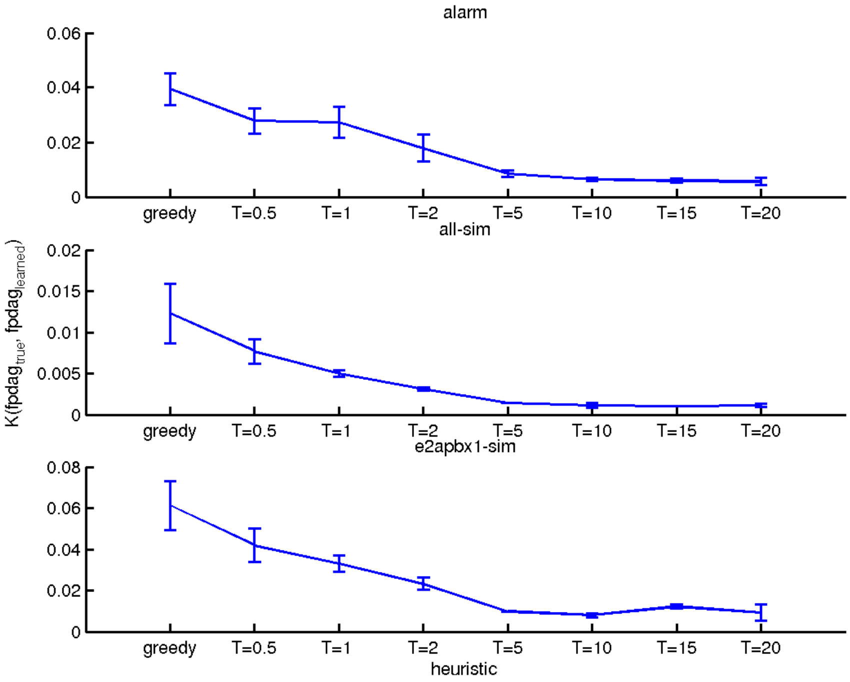


Figure 4.2: KL-distance between true and learned fPDAG as a function of the heuristic search strategy. The more expensive the heuristic the better the learned fPDAG approximates the true one.

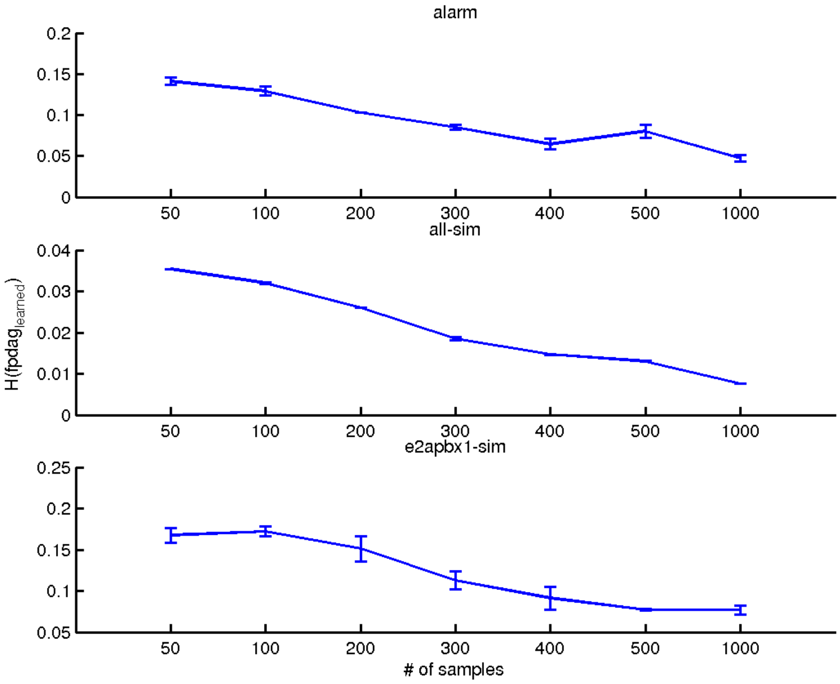


Figure 4.3: Entropy of learned fPDAGs as a function of the heuristic search strategy. With increasing the number of samples the learned fPDAG becomes more robust and the entropy decreases.

trates the evolution of the relative entropy between the true and learned fPDAG, $K(q,p)$ (black solid line) when increasing the sample size N. As expected, with an increasing number of samples, the learned fPDAGs better approximate the true one, and the KL-distance decreases. Thus, with a sufficiently large sample size the difference between the original underlying system and the learned one diminishes, but it increases when scaling down the data set size.

However, if one focuses on the distance which results from erroneously overestimated dependencies $K(q,p)_+$ (black dashed lines) one can see that it is relatively low and remains also relatively constant compared to the total distance. Thus, we can draw the important conclusion that deviations from the true network is primarily caused by erroneously underestimated dependencies $K(q,p)_-$ (black dashed-dotted line), and not by erroneously overestimated ones. This hypothesis becomes even more evident when looking at Figure 4.5 which shows how the confidence of false positives and true positives is distributed, for two sample runs with 1000 and 50 samples, respectively, for each of the three networks. The solid line indicates the edge confidence distribution of the true network and dots that of the learned network. Thus, the area to the left of the solid line shows the confidences of the true positives and the area to the right shows the confidences of the false positives. With a small sample size (N=50) the number of falsely learned dependencies is much higher compared to a relatively large size data set (N=1000). However, since they are mainly of low confidences $K(p,q)_+$ remains relatively constant and does not highly contribute to the total distance. Further, a small sample size results in a decrease of originally highly confident edges and in an increase of low confident ones. Thus, $K(p,q)_-$ states the biggest part of the total distance. A small sample size causes a decrease of truly estimated dependencies but does not strongly increase the belief in falsely estimated dependencies. Hence, structures learned from sparse data deviate from the underlying true system but those edges which have a high confidence in the learned fPDAG are likely to be correct. However, Figure 4.5 also shows that given a very small sample size of $N = 50$ no threshold c_τ for the edge con-

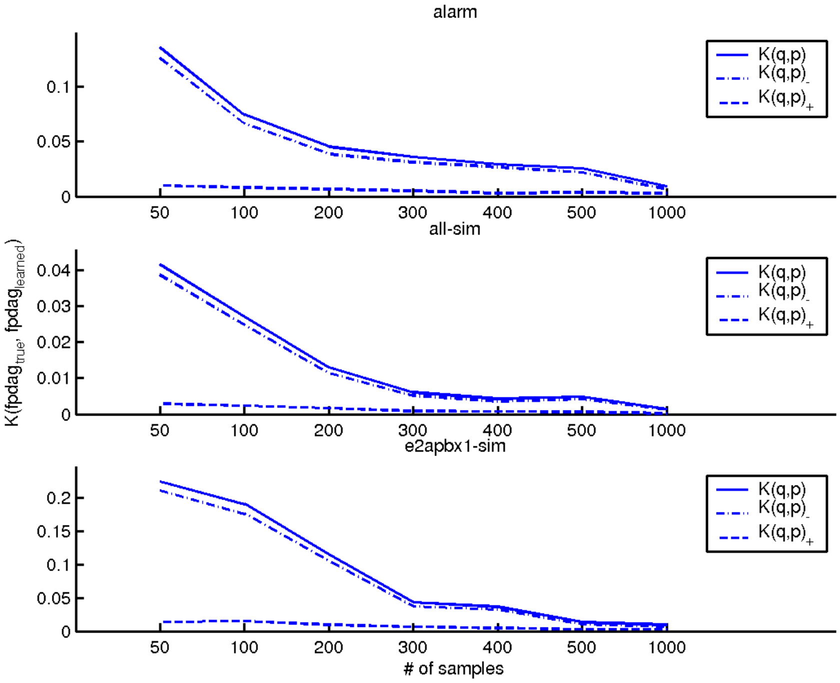
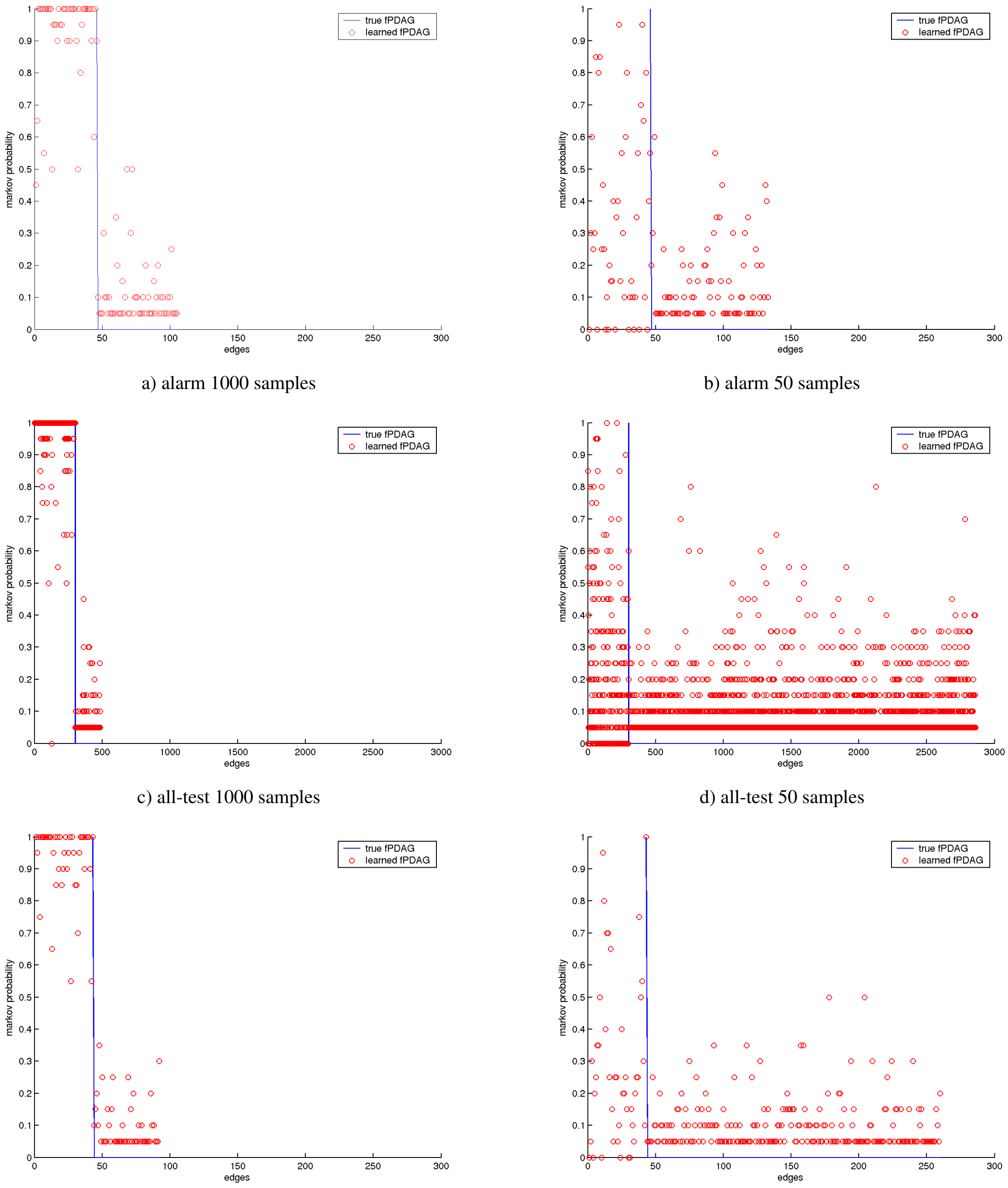


Figure 4.4: Comparison of true fPDAG and learned fPDAG from generated data sets with different sample sizes. When increasing the sample size, the distance between the true and the learned model decreases. The major part of this distance is caused by false negatives whereas the contribution of false positives is low and remains relatively constant.



a) alarm 1000 samples

b) alarm 50 samples

c) all-test 1000 samples

d) all-test 50 samples

e) e2apbx1-test 1000 samples

f) e2apbx1-test 50 samples

Figure 4.5: Confidence distribution (dots) of true and false positive edges learned from data sets of 1000 samples **(a**, **c**, **e)** and 50 samples **(b**, **b**, **f)**. The area to the left of the solid line shows the confidences of the true positives and the area to the right shows the confidences of the false positives.

	$N = 50$			$N = 300$			$N = 1000$		
benchmark	c_{τ}	s	p	c_{τ}	s	p	c_{τ}	s	p
ALARM	0.8	0.174	0	0.5	0.89	0.04	0.5	0.97	0.03
ALL-SIM	0.8	0.078	0.02	0.5	0.81	0.03	0.5	0.99	0.02
E2APBX1-SIM	0.8	0.093	0.05	0.5	0.95	0.03	0.4	1.0	0.03

Table 4.1: Confidence threshold and sensitivity for all three benchmark cases, with $N = 50$, $N = 300$ and $N = 1000$, given $p \leq 0.05$.

fidence exists which fully separates the true from the false positives, since both are of low confidences. Thus, a compromise between predicting most of the underlying dependencies and accepting false predictions has to be made. This condition can be best described by means of two parameters: the sensitivity ($s(c_{\tau}) = \frac{tp}{tp+fn}$) which is the probability of predicting the true dependencies, and the error probability measure or p-value ($p(c_{\tau}) = \frac{fp}{fp+tp}$) which denotes the probability that the estimated dependencies are wrong an therefore false positives. Given these two parameters, each confidence threshold c_{τ} can be assessed for its accuracy with the goal to find a cut-off which perfectly separates true from false positives such that $s(c_{\tau}) = 1$ and $p(c_{\tau}) = 0$.

Varying the confidence threshold c_{τ} and plotting the sensitivity versus the corresponding p-value results in a curve which illustrates how much error one has to tolerate to achieve a certain sensitivity. Figure 4.6 plots these curves for all three benchmarks with different sample sizes ($N = 1000 - 50$). In all three cases lowering the sample size results in a decrease of the area under the curve meaning that to achieve a certain sensitivity a bigger error has to be accepted. Thus, to stay below a certain p-value (e.g. $p \leq 0.05$), the achieved sensitivity respectively the number of true positives decreases with the number of available samples.

As listed in Table 4.1, for sample sizes of $N = 50$ and a p-value of $p \leq 0.05$, a sensitivity of around $0.078 \leq s \leq 0.174$ can be achieved. This might sound very little, but readers should be advised that revealing $\sim 10\%$ of the underlying genetic network would already

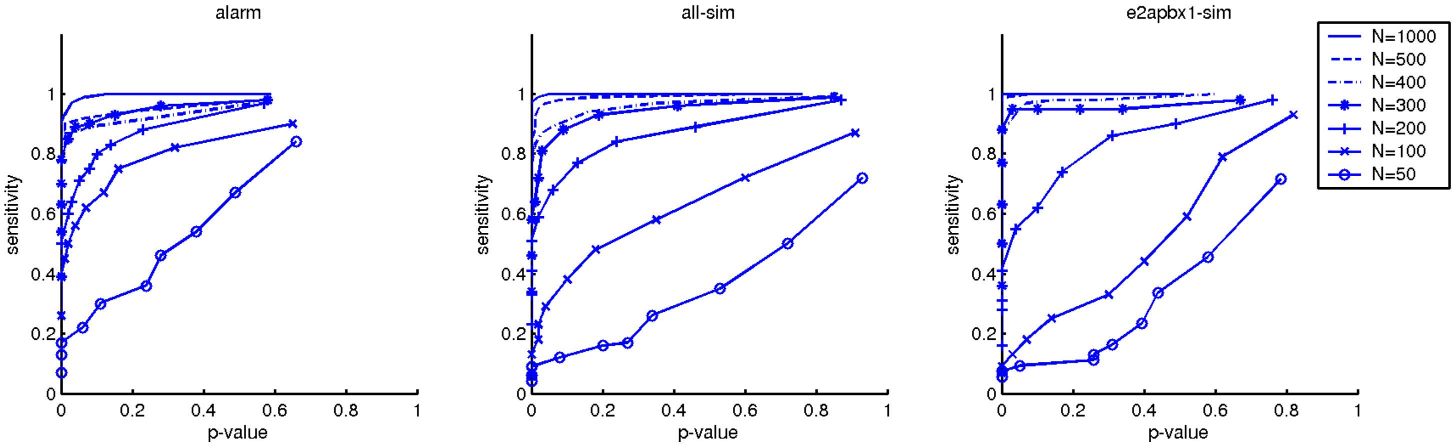


Figure 4.6: Sensitivity versus p-value plotted for different sample sizes. The fewer samples available the more error one has to accept to achieve a certain sensitivity.

be a promising result. In addition there are experiments with much more than 50 samples. So, for a data set of 300 samples we achieve a sensitivity of around $0.89 \leq s \leq 0.95$. Moreover, Table 4.1 shows a constant confidence cut-offs up to which the corresponding dependencies can be believed in with a high accuracy: for $N = 50$ the threshold is $c_\tau \geq 0.8$ whereas for 300 samples it is $c_\tau \geq 0.5$. Thus, for further analyses of Bayesian networks learned from sparse microarray data sets on should primarily focus on edges whose edge confidence exceeds the corresponding threshold since these dependencies are likely to part of the underlying genetic network.

4.4 Robustness of Bayesian Networks Learned from Microarray Data

The simulations of the previous section show that even from data with a low sample size robust structural features can be extracted. These benchmarks can now be used to estimate features, learned from microarray data which probably belong to the underlying genetic network. As a data basis for the following examinations we chose the leukemia microarray data set of (Yeoh et al. 2002) since it provides a relatively high number of samples, namely 327.

4.4.1 The ALL Data Set

The acute lymphoblastic leukemia (ALL) study (Yeoh et al. 2002) provides measurements of 12.000 probes in 327 samples collected from patients with different pediatric ALL subtypes. The goal of this study was to use expression profiling for identifying each of the known prognostically and therapeutically relevant subgroups and for the identification of patients who are at high risk for failing conventional therapeutic approaches.

Out of the 12.000 measured genes, those genes have been selected that best define the individual subtypes using χ^2 statistic according (Yeoh et al. 2002). The final data set (271 genes $\times$ 327 samples) is composed of the 40 most discriminative genes for each of the 7 subtypes, whereby 9 genes appear in more then one cluster but only once in our final dataset. Be precise from the beginning, the data corresponds to 271 gene probes. Some genes are represented by multiple probes (e.g. to overcome problems with alternative splicing or to test for variability within a microarray measurement). In accordance with the original experiments (Yeoh et al. 2002), we leave these duplicates in the data.

Next, gene expression levels were discretized to three levels, over-expressed, unchanged and under-expressed, thresholded by the standard deviation of the expression levels over the whole data set, to learn a multinomial Bayesian network. Since this model can describe any discrete conditional distribution, all algorithms also work for higher classes of ordinal data. However, given the low signal to noise ratio of current microarray data with a polynomial scaling of computational expense, finer discretization might result in noise-contaminated data.

Based on these data we trained 20 networks according to the non-parametric data bootstrap approach described in Section 4.1 with simulated annealing $(T_{\text{start}} = 10)$ as a heuristic search strategy. According to the previous benchmark simulation and the size of the used data set, namely $N = 327$, we only took those dependencies with a confidence ≥ 0.5. We found 81 edges exceeding this threshold whereby 29 dependencies, listed in Table 4.2 exceed the confidence threshold of $0.8.$, whereof 20 represent dependencies between probes of the same gene. Since a gene is by definition linked most strongly to itself, this result demonstrates the power of Bayesian network learning to capture statistical dependencies and thereby to hint towards functional gene gene relationships even from sparse data.

Furthermore also functional dependencies have been estimated. For example, CD3D and CD3E are both part of the T-cell surface glyoprotein CD3 and therefore both involved in the same cellular pathways. Another example is the estimated dependency between HLA-DRA and HLA-DPB1. Both genes are part of the major histocompatibility complex, class

$i_{\text{Affymetrix ID}}$	i_{Symbol}	$p_{i\rightarrow j}$	p_{i-j}	$p_{i\leftarrow j}$	$j_{\text{Affymetrix ID}}$	j_{Symbol}
39730_at	ABL1	0.55	0.175	0.275	1636_g_at	ABL1
37014_at	MX1	0.625	0.05	0.2	39061_at	BST2
330_s_at	TUBA1	0.55	0.1	0.35	36591_at	TUBA1
1287_at	PARP1	0.425	0.15	0.275	41146_at	PARP1
40648_at	MERTK	0.475	0.125	0.225	1786_at	MERTK
717_at	TRIB2	0.625	0.075	0.3	40113_at	TRIB2
38679_g_at	SNRPE	0.65	0.05	0.1	33859_at	SAP18
38459_g_at	CYB5	0.625	0.1	0.075	31492_at	eIF3k
34374_g_at	UREB1	0.45	0.1	0.25	40998_at	TNRC11
39402_at	IL1B	0.65	0.025	0.325	1520_s_at	IL1B
955_at	CALMI	0.45	0.1	0.425	41288_at	CALM1
40518_at	PTPRC	0.825	0.075	0.1	40519_at	PTPRC
1126_s_at	CD44	0.45	0.05	0.325	2036_s_at	CD44
38242_at	BLNK	0.55	0.05	0.225	34168_at	DNTT
31892_at	PTPRM	0.425	0.175	0.375	995_g_at	PTPRM
41266_at	ITGA6	0.45	0.2	0.225	33410_at	ITGA6
33410_at	ITGA6	0.475	0.15	0.175	33411_g_at	ITGA6
38747_at	CD34	0.475	0.2	0.325	538_at	CD34
38319_at	CD3D	0.225	0.375	0.2	36277_at	CD3E
38017_at	CD79A	0.6	0.125	0.275	38018_g_at	CD79A
37039_at	HLA-DRA	0.35	0.3	0.25	38095_i_at	HLA-DPB1
38095_i_at	HLA-DPB1	0.625	0.3	0.075	38096_f_at	HLA-DPB1
2059_s_at	LCK	0.3	0.425	0.25	33238_at	LCK
41165_g_at	IGHM	0.5	0.05	0.45	41166_at	IGHM
36239_at	POU2AF1	0.575	0.075	0.225	40570_at	FOXO1A
38652_at	C10orf26	0.475	0.125	0.25	32224_at	FCHSD2
41097_at	TERF2	0.5	0.2	0.3	1299_at	TERF2
1980_s_at	NME2	0.425	0.175	0.4	33415_at	NME2
1336_s_at	PRKCB1	0.45	0.1	0.45	160029_at	PRKCB1

Table 4.2: Statistical dependencies estimated from the ALL data set with a confidence greater or equal 0.8. Most of them are between probes of the same gene. Dependencies between distinct genes are marked gray.

II, and therefore functionally related. Further emphasis should be put on the analysis of dependencies which predict causal relationships and hence regulatory relationships among gene, such as the dependency between MX1 and BST2 or between SNRPE and SAP18. However, we do not conduct such a profound analysis since it would be beyond the scope of this section.

4.5 Summary

By learning Bayesian networks from microarray data we assume to be able to capture gene regulatory relationships which are part of the whole genetic network. But as a data driven approach this method suffers from the fact that common microarray data sets are of small sample size and therefore insufficient to reflect the entire ”statistic” of the underlying system. Generally given finite data, rather than a single best model, several models describe the underlying probability distribution equally well and should therefore be taken into account. This results in a weighted graph structure, where each dependency is labeled with a probability value, the so-called edge confidence, addressing the uncertainty given the observed domain.

Several benchmarks have been performed in this chapter to simulate the effect of small sample sizes with the goal to extract features which are robust against data sparseness. For this, a measure has been introduced which accounts for the deviation of the learned model from that model where the data are sampled from. The results show that most of the error is induced by falsely underestimated dependencies and only weakly by falsely overestimated ones. Thus, we can conclude that the sparser the data the less dependencies can be learned, but that dependencies of high confidence are likely to be true estimates. We further examined this results more precisely with the aim to find a confidence threshold which discriminates between false and true estimates such that the p value is still acceptable $p \leq 0.05$. Simulations showed that for a sample size of $N = 50$ and an error probability of $p \leq 0.05$ only edges with a confidence of 0.8 or higher should be considered as true estimates. Thus, for microarray data sets of around 50 samples, structure learning should be able to detect $\sim 10\%$ of the underlying genetic network and $\sim 80\%$ with data sets of 300 samples. However, a major point of criticism regarding these results is the fact that Bayesian network learning has been tested on data sampled again from Bayesian networks. We therefore need to investigate in further benchmark cases including other types of generative models, such as systems of coupled differential equations or boolean networks which is left to future work.

Chapter 5

Topological Analysis of Bayesian Networks

As shown in the previous chapter, structures learned from sparse data have to be interpreted carefully especially when analyzed edge by edge. However, further topological features exist whose exploration might be important in light of a deeper understanding of genetic network principles. The here presented approach puts emphasis on global topological features of the learned network structure, with the aim of finding basic principles of the learned network structure and hence of the underlying genetic regulatory system. This work is summarized in (Dejori, Schürmann & Stetter 2004).

5.1 Network Topology and Scale-Free Architecture

A Bayesian network is specified by a pair (G, Θ), where Θ describes the probability distribution of the variables as a set of conditional probability distributions and $G = (V, E)$ denotes the graph structure with a set of vertices V and edges E. The approach proposed in this chapter is based on the analysis of estimated topological features of the trained Bayesian networks. Thus, from a learned Bayesian network, only the graph structure G

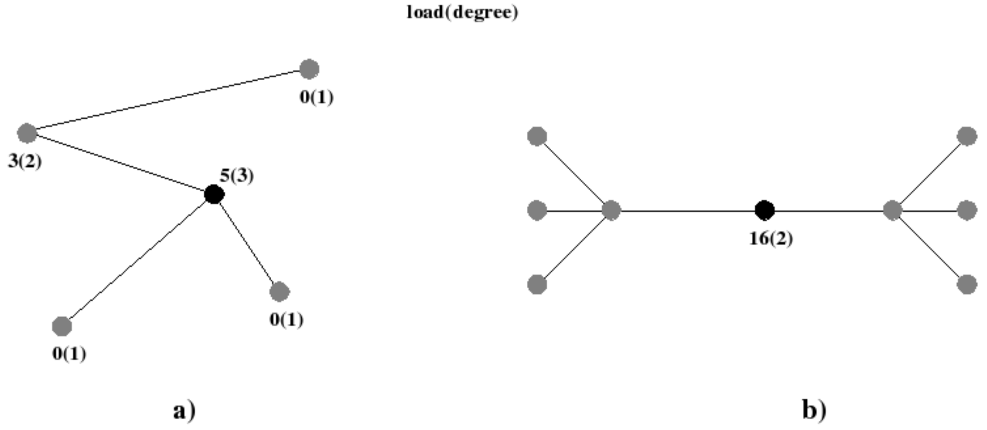


Figure 5.1: **(a)** Load and degree (in brackets) of a small example network. **(b)** A case in which the degree and the load differ strongly.

is utilized for a further analysis whereas the set of parameters Θ is fully neglected. Moreover, rather then studying the quality of a learned structure G by looking at each edge independently, the following approach focuses on global topological features of a single network G or of a set networks $\{G_1, ..., G_q\}$ resulting from a Q-fold bootstrap procedure and represented by an fPDAG.

To this end, the topology T of obtained graph structures is assumed to be a direct estimate of the topology of the underlying genetic regulatory network. More precisely, using the posterior distribution $p(G_q|D)$ of graph structures, it is possible to equip for example each feature F with its confidence $p(F|D)$ as shown in Section 4.2.

5.1.1 Scale-free Topology

As mentioned at the beginning of this thesis the cellular system is often characterized by a network structure, where nodes are joined together by links indicating an interaction or association among molecules, e.g. genes or proteins. Such complex regulatory systems appear on various molecular levels, from the genetic regulatory network up to better known metabolic pathways. Besides molecular network systems many other complex organizations such as social relationships among individuals or the spreading of diseases are often analyzed at a network level. Many studies in this field try to find fundamental procedures which let such networks emerge and common principles which might help gaining a better understanding of complex systems in general. Empirical studies of the past few years have shown that many large-scale networks, such as metabolic pathways or protein-protein networks, share a common topological feature, namely a *scale-free topology* (see (Barabasi & Bonabeau 2003) for a review).

To define a scale-free property of a network, a new feature has to be introduced namely that of the degree k of a node. The *degree* is defined as the number of connections (edges) k to or from it. In contrast to a random network, whose degree distribution is that of a Poisson function, in a scale-free network the degree k of a node is distributed according to a power law of the form

$$P(k) \sim k^{-\gamma}, \tag{5.1}$$

where γ denotes the scaling exponent. Such networks are characterized by a fairly small amount of nodes which show a much higher degree than the average. An interesting phenomenon found for the operation of scale-free networks is that they display a high degree of robustness against random failures of nodes but a high vulnerability for targeted attacks on the few highly connected nodes (Albert et al. 2000).

Another important topological feature of scale-free networks has been introduced by Motter et al. (Motter et al. 2002): the *load* c of a node is defined by the total number of shortest paths between all possible pairs of nodes that pass through it (also referred to as *node betweenness centrality* (Barthelemy 2004)). Depending on the scaling exponent γ, nodes with high degree or nodes with a high load represent spots of high vulnerability of the network. For exponents around 3, a high load has been shown to mark stabilizing nodes. Thus, scale-free topology of a network with exponents in this range implies that only a small number of nodes, characterized by a high load, control the global network behavior.

Both topological features suggest that in scale-free networks only a small fraction of nodes play a crucial role for the integrity of the system and stable network operation.

Assuming a scale-free topology of the molecular network such nodes might represent the Achilles heel of the cellular machinery and are hypothesized to play a crucial role in pathogenesis. This view matches perfectly the biological principles of tumorigenesis: tumor-suppressor genes are crucial for the maintenance of a controlled cell state and mutations on them can turn the cell into a cancer cell. Since a learned Bayesian network is argued to reflect the underlying genetic network, it might be interesting to test learned structures for their scale-free behavior and if this holds to raise the following question: does the topological importance of nodes correlate with what is known about the biological importance of the proteins they represent. For example, are such genes known to be involved in oncogenesis or related critical processes?

5.1.2 Calculation of Mean degree and Directed Load

The result which emerges from a structure learning procedure is not a single fully directed graph but a set of Q partial directed graphs, which together form a fPDAG. Thus, as mentioned in Section 5.1, both features, degree and load, can be formulated by a feature likelihood using Equation 4.2. The mean degree of a variable x_i is given by:

$$k_i = \frac{1}{Q} \sum_q K_i(G_q), \tag{5.2}$$

where $K_i(G_q)$ is the degree of variable x_i in graph G_q which can easily be estimated by counting all directed and undirected edges related to node i.

The mean directed load of a variable x_i is defined as:

$$c_i = \frac{1}{Q} \sum_q C_i(G_q). \tag{5.3}$$

The directed load $C_i(G_q)$ is calculated by searching in G_q for each of the $n(n-1)$ pairs of nodes for the shortest connecting path through the network, which is consistent with eventual edge directions, and incrementing the load C_i of each node i on this shortest path by 1. In case of a directed edge the path can only pass in the way of the direction whereas in case of an undirected edge both paths are possible. Figure 5.1a depicts a small example network and lists the values of the load and the degree (brackets). In general the load and the degree will be correlated, however they can also differ drastically, as for the case of a node connecting two large communities, as shown in Figure 5.1b.

5.2 Using Scale-free Topology to Estimate Critical Genes

The basis for the following examinations consists of a fPDAG (cf. Section 3.1) composed of Bayesian networks learned from 20 bootstrap replicas of the leukemia ALL-microarray data set B.2.1 as described in Section 4.1. ALL is a heterogeneous disease. It appears in various subtypes which differ markedly in their response to medical treatment. Apart from T-cell-related ALL, T-ALL, the pathogenesis of which is not yet well-understood, the subtypes related to B-cells can be retraced to specific genetic lesions, namely genetic translocations t(9;22) (BCR-ABL), t(1;19) (E2A-PBX1), t(12;21) (TEL-AML1), t(4;11) (MLL) and hyperdiploid karyotype with >50 chromosomes (hyperdip>50). The data set used to train our model B.2.1 partitions into markedly different gene-expression patterns, which are characterized by different over- or underexpressed gene clusters and which can be assigned either to the six known ALL subtypes or a seventh new subtype (see Figure Figure 5.2 shows the resulting ALL-fPDAG connecting the 271 investigated genes. The small red numbers (also reflected by the linewidth) denote the likelihood of the edge (confidence) as the result of a $Q = 20$ fold bootstrap procedure. The position of the 271 nodes, each representing a certain gene i, results from the projection of the corresponding expression vector across experiments, $d_i = (d_i^1, ..., d_i^N)$, onto the plane spanned by the first

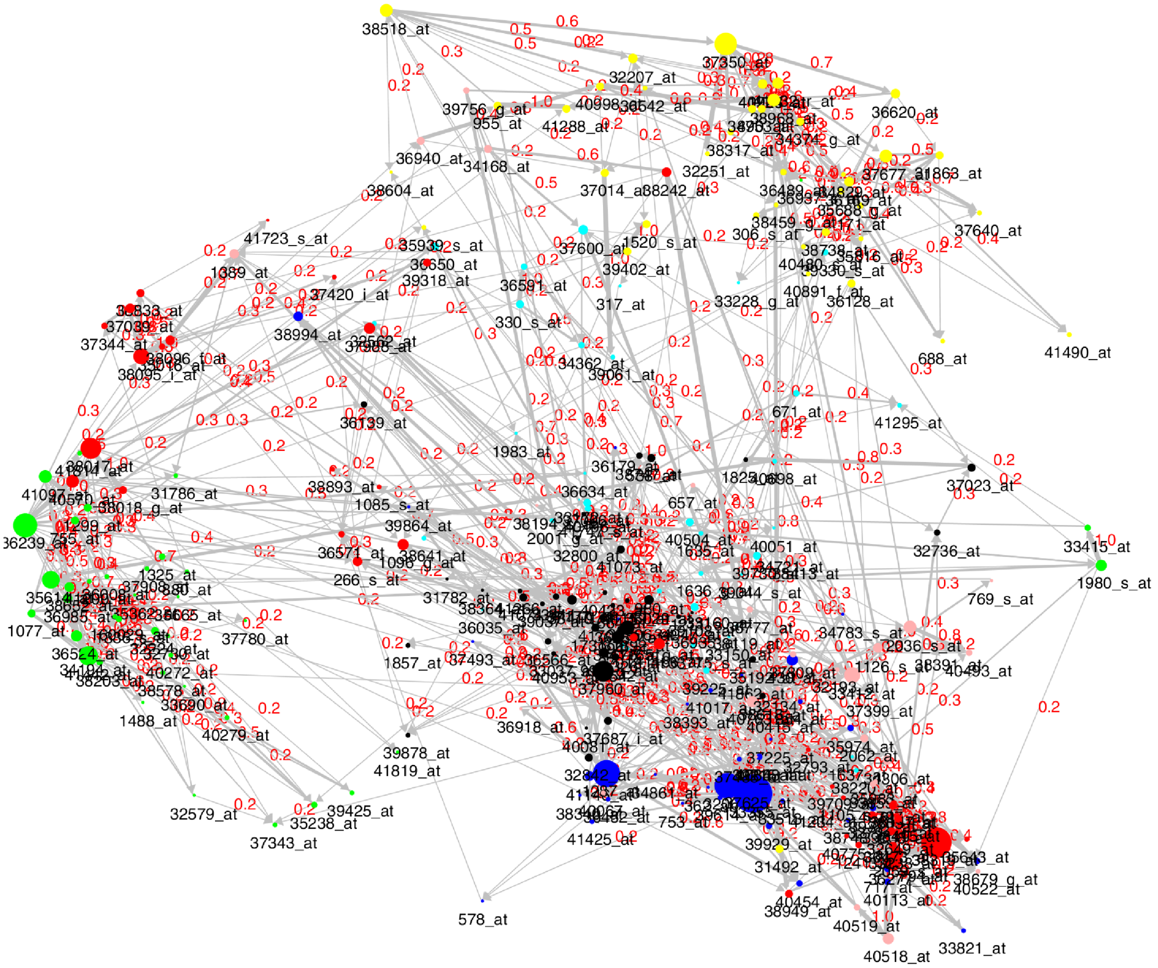


Figure 5.2: Genetic network structure obtained by a learning Bayesian network from the ALL gene expression data. Each node represents a gene and each edge an estimated regulatory relationship. The area of each gene node is proportional to its degree, the color reflects the gene cluster. Genes are denoted by their Affymetrix-IDs. Red numbers indicate the confidence of each edge according to the bootstrap procedure.

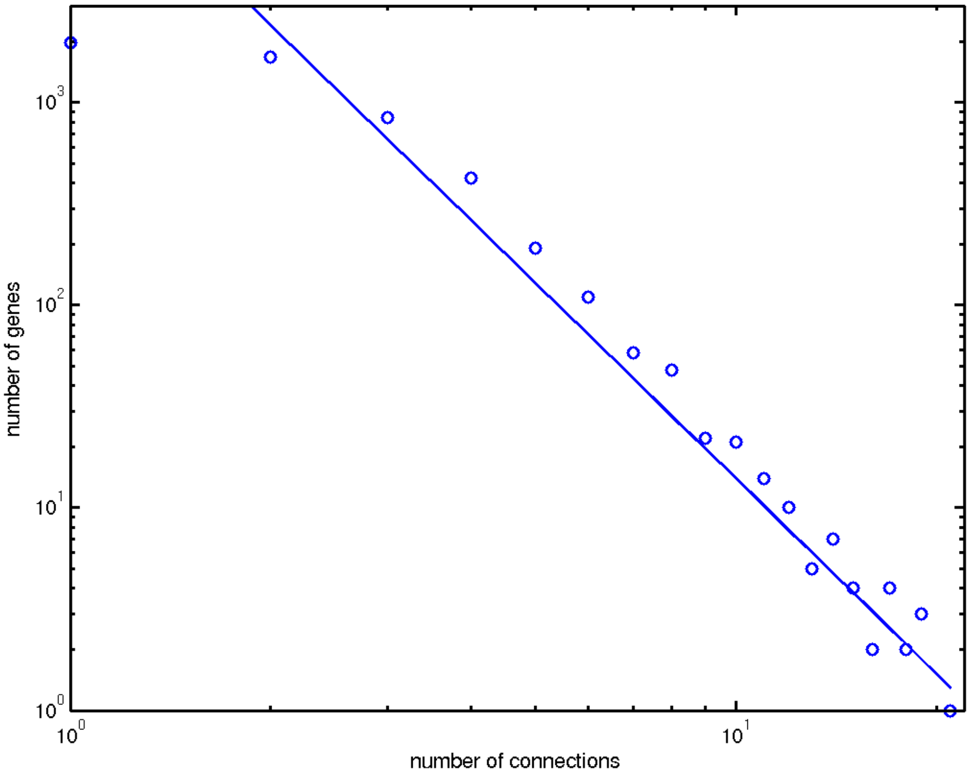


Figure 5.3: Degree distribution of the ALL network. Apart from genes with degree 1, the network follows closely a scale-free topology with $\gamma = 3.2$.

two principal components over these vectors $d_1, ..., d_n$. This representation already allows a first coarse classification of the high-dimensional gene space into several gene clusters. The area of each node is proportional to the mean degree, see Equation 5.2, of the corresponding gene. Already visual inspection reveals that there is only a small number of genes with a high degree whereas the majority of nodes have small degree, which is a qualitative indication for the scale-free characteristic of the fPDAG network. B.1) (Yeoh et al. 2002). We tested the scaling of the ALL network by calculating its degree distribution, i.e. the histogram of the number of genes with degree k. The degree distribution, summed up over

Affy.-ID	Gene	Putative function
36239_at	*POU2AF1*	transcription cofactor, anti-pathogene response
1287_at	*ADPRT*	DNA repair, apoptosis
38017_at	*CD79A*	Anti-pathogene response
38319_at	*CD3D*	T-cell receptor, cellular defense response
33355_at	*PBX1*	Proto-oncogene, transcription factor
41442_at	*CBFA2T3*	cell proliferation, transcription factor
2059_s_at	*LCK*	tyrosin kinase, cell cycle regulation
37988_at	*CD79B*	Anti-pathogene response
38095_i_at	*HLA-DPB1*	MHC, Anti-pathogene response
37350_at	*PSMD10*	proteosome subunit, protein degradation

Table 5.1: Genes with highest directed load

all $Q = 20$ bootstrap runs, is displayed in a log-log plot in Figure 5.3. It clearly shows a power-law decay, as given in Equation 5.1. This demonstrates the scale-free characteristic of the ALL network, for which we find a scaling exponent of $\gamma = 3.2$. The slightly too low number for genes with only one link might arise from the fact that we consider only a subset of genes. Next, we make use of this scale-free property to define a measure for the importance of a gene for the network operation. For this we consider the propagation of information through a gene within the network. In the context of biological regulatory networks, a direct link between two genes corresponds to a chemical regulatory relationship: one gene provides information about its state to another gene by a chemical link. Likewise, an arbitrary, indirect path between two genes can be interpreted as a chemical signalling chain, through which information from a source gene to a target gene is propagated as a chemical reaction cascade, for example a cascade of transcription factor bindings on regulatory upstream regions. The information load or traffic load of a gene can then be interpreted as the total chemical information, which flows through that gene while forming indirect multi-step regulatory relationships between pairs of genes in the network. Besides this intuitive motivation for the load of a gene as an important topological feature, we can now make use of the findings from the systematic analysis of scale-free networks, (Motter

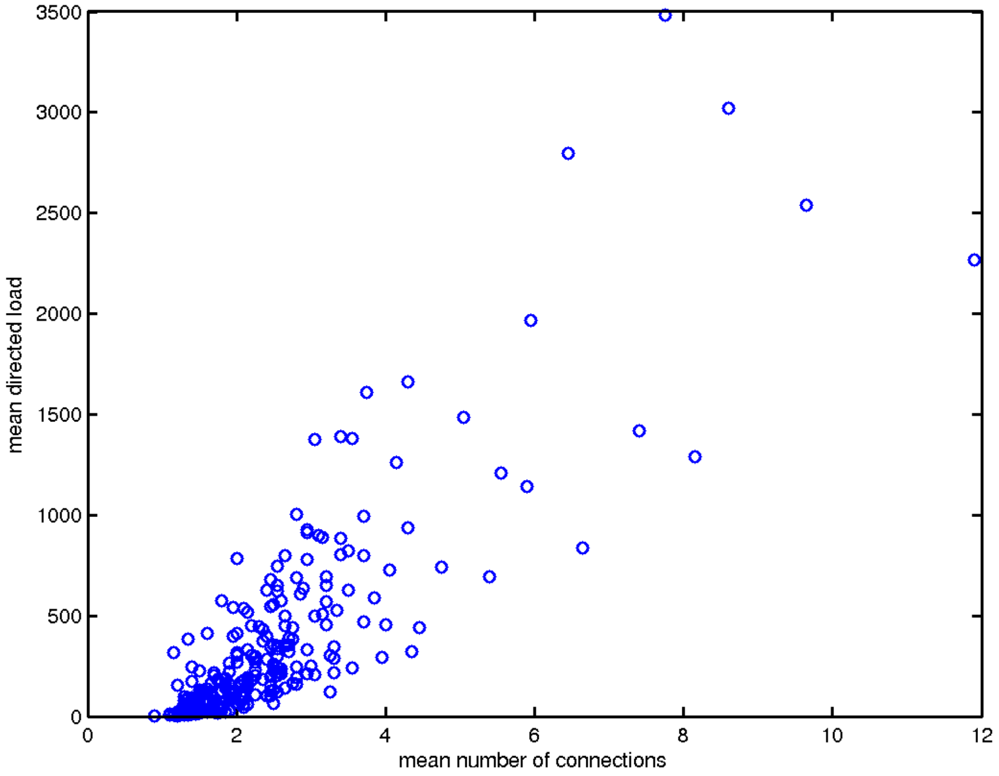


Figure 5.4: Scatter plot of load against degree for the 271 genes. Both features are correlated, but mostly for high-load and high-connected genes, load and degree differ from each other.

et al. 2002), namely that for scale-free networks with scaling coefficients around 3 a high load c_i of node i renders it important for a globally stable operation for the network. In other words: damage to this node will have a particularly large impact on the global network behavior. Now we identify the total chemical information as motivated above with the directed load as defined in Section 5.1.2. Then a high load c_i in the genetic regulatory network represents a good measure for the network's vulnerability to a local attack to the expression of this gene i. We predict that if genes with high load are damaged, the normal operation of the regulatory network will break down more likely than for damage of low-load genes. In particular, we predict damage on genes with a high load to be the cause for pathological cellular function. These genes should be responsible for oncogenesis, tumor development or other critical processes. Consequently, high-load critical genes are estimates for pharmaceutical drug targets.

Table 5.1 shows the annotation (from the Affymetrix database) of the 10 genes with highest mean directed load. Some of them are known as oncogenes or proto-oncogenes whereas others are involved in critical processes such as DNA repair, induced cell death or cell-cycle regulation. All genes with high load are involved in critical cellular processes. *POU2AF1*, the gene with the highest load, is annotated as a proto-oncogene, acting as a B-cell-specific transcriptional coactivator. The results seem to confirm that a high load is a good indicator for genes which occupy a central position in normal cellular function, and whose failure is likely to cause severe cellular malfunction such as cancer. As the 10 highest degree-genes seem to be related, by annotation, with oncogenesis , it might be useful to rate genes by topological features in order to identify disease-critical genes.

But one criticism still remains, namely, that the network consists of a small selection of genes where most of them can be related with oncogenic processes. In fact 36% of the 271 genes are related with oncogenesis [1]. However, as shown in Figure 5.5, for the first 5 highest-degree genes the relative frequency of oncogenes is 75%, taking the first 10 of

[1] listed in CancerGene database

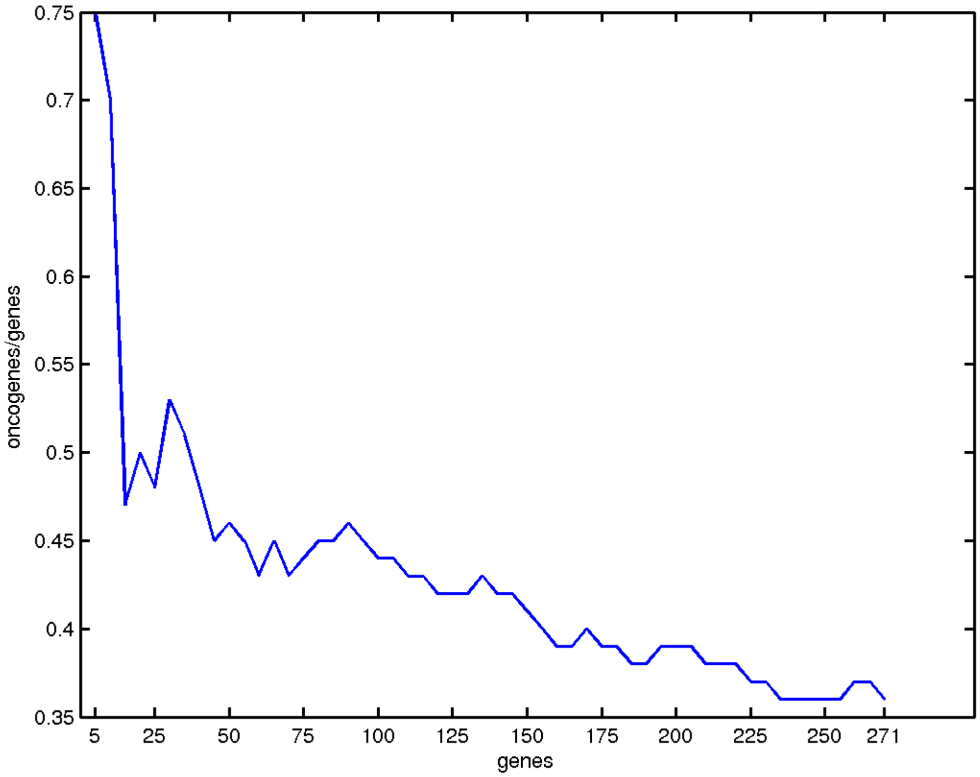


Figure 5.5: Relative frequency of annotated oncogenes as a function of the number of genes selected by their node's degree in descending order.

the list the percentage is still 70% but deacys the more genes from the sorted list are included. This underlines the predictive power of topological features for the characterization of critical genes.

5.3 Summary

In the here presented approach we focus on the topology of mutual dependencies between genes, and rank the importance of genes for the global operation of the network by topological measures including the degree of a node (the number of links to and from other nodes of the network) and a measure for the traffic load. For a set of genes, whose expression levels have high covariance with the individual subtypes of the disease, we find that the graph structure of the learned Bayesian network of statistical dependencies follows a scale-free network topology (Motter et al. 2002), a structure which has been observed also for other cellular molecular networks such as in metabolic networks (Jeong et al. 2000) or in the *S. cerevisiae* protein-protein interaction network (Jeong et al. 2001). According to the theory of scale-free networks, genes with high load or high degree are important for stable network operations, and are points of high vulnerability of the complex system. We provide the load and degree distribution of the ALL-relevant genes and find a small number of high-load genes, which are often annotated with central functions.

In order to judge the robustness of the present method in light of the data sparseness, preparatory benchmark tests in Chapter 4 found, that the edges of our learned network might reflect the subset of the strongest and therefore most significant ”backbone” regulatory relationships, which form a robust skeleton of the complete network. Analyzing the topology of a subsystem and making hypothesis for the underlying global one might be a point of criticism. But in (Scholz et al. 2005) it has been shown that random link removal conserves the scale-free organization as well as the respective exponent of the true network even up to a considerable level of noise, which underlines the relevance of our findings for the exploration of genetic network principles.

Chapter 6

Generative Inverse Modeling in Bayesian Networks

Bayesian networks belong to the class of graphical models where graph theory and probability theory are related to each other. In Section 5.1 we presented one way of interpretation, by considering only the structure G of a learned Bayesian network and neglecting its probabilistic density Θ. In the here presented approach which is summarized in (Dejori & Stetter 2004),the graph structure of the Bayesian network is not explicitly analyzed, and edges are not put in a direct relation to regulatory relationships between genes or molecules. Instead the whole Bayesian network is interpreted as a density estimator which consequently can be treated as a generative statistical model. In the case of learning from microarray data such a generative model can be used to generate artificial gene expression profiles which should reflect the statistics of the original data and therefore patterns of the underlying system. Much more interesting it might be to alter the learned model and to observe the consequences, in form of artificial gene expression profiles, which arise from this intervention.

In the following, we present a probabilistic graphical modeling technique which allows to relate local genetic interventions to alterations in the global gene expression profile as well as the other way around, namely, to infer the local genetic change responsible for a global gene expression alteration. The underlying part of our approach consists of a number of Bayesian belief networks, learned from gene expression measurements, which represent dependency relationships among a set of genes. The learned Bayesian networks are used as generative models for sampling simulated expression profiles, while at the same time genetic changes are imposed on these networks. Genetic changes are induced by fixing the expression level of one or a few genes to a constant value and calculating all other expression levels conditioned on this intervention. Fixing gene expression levels can be thought to model the effects of biological transcriptional signals, mutations, drug treatment, gene knock-out or other interventions into the biological regulatory network. Therefore, this technique might be useful to infer stabilizing or destabilizing genes or pathways from microarray data.

6.1 Generative Inverse Modeling

Generative inverse modeling is a three-step approach using gene expression profiles as data basis (see Figure6.1). In the first step, based on microarray data, a Bayesian network is learned which, as described in Section 3.1, can be used to extract the statistical dependencies and regulatory relationships hidden in the observed microarray data set. In the second step this statistical model is being used to simulate artificial gene expression profiles, that follow either the same statistical structure or a changed statistics as an effect of an intervention (e.g. by switching a certain gene on or off). When the generated patterns closely resemble measured patterns under a certain condition, such as a specific disease, the intervention is considered likely to be involved in the mechanism causing this disease. Thus, in a third step, artificial gene expression patterns are compared with gene characteristic expression patterns such that generative inverse modeling can provide a ranking of a certain intervention to be the cause of an observed global expression pattern.

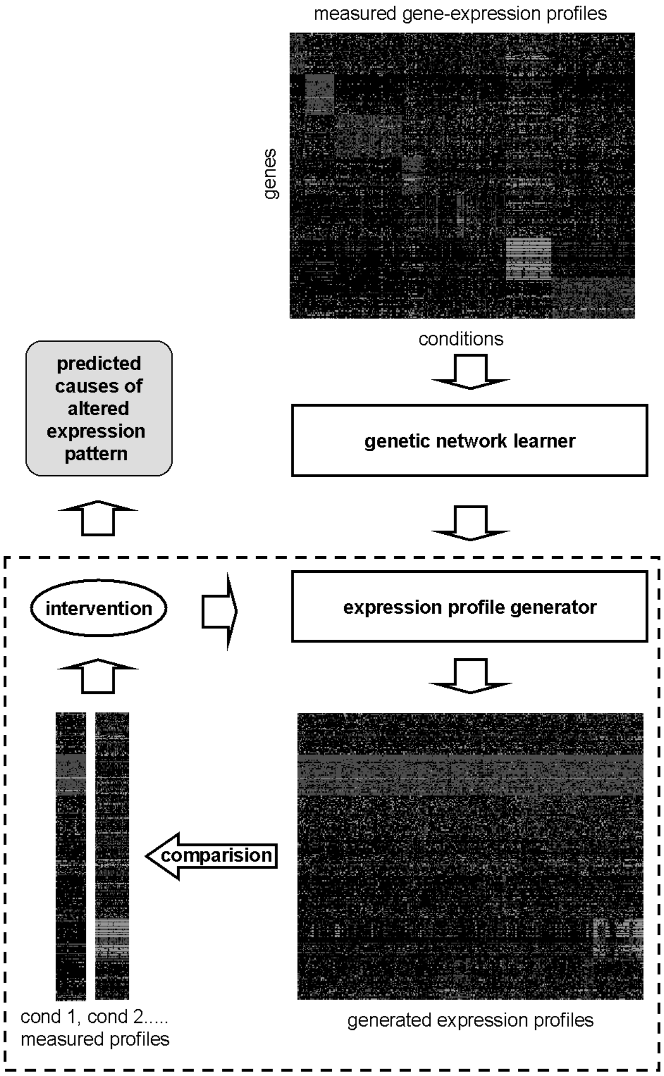

measured gene-expression profiles
genes
conditions
predicted causes of altered expression pattern
genetic network learner
intervention
expression profile generator
comparision
cond 1, cond 2.....
measured profiles
generated expression profiles

Figure 6.1: Illustration of generative inverse modeling. From a set of gene-expression patterns measured under different conditions, the genetic network learner extracts the statistical dependencies and regulatory relationships that best describe the input expression set. This statistical description is then used by the expression profile generator to produce artificial gene expression profiles, that follow either the same statistical structure or a changed statistics as an effect of an intervention. The dashed box contains the process of generative inverse modeling.

```
Algorithm 1 Sampling (B,N)
Input:
  $B$ - Bayes-net;
  $N$ - Number of independent samples.
Output:
  $D_g$ - Data set of $N$ independent samples.

1. Order variable-set x consistent with the condition
   that parents $pa_i$ are sorted before $x_i$
2. For $s = 1, ...., N$
3.   For $i = 1, ...., n$
4.     Let $x_i$ be the highest ordered node not instantiated in sample $s$
5.     Select state with probability $p(x_i = \text{state} | pa_{i,g})$
6.     Update $s$-th sample of $D_g$
7.     Instantiate $x_i$=state
```

Figure 6.2: Algorithm to sample from a Bayesian network without interventions

6.1.1 Data Generation without Intervention

Being a density estimator, the trained Bayesian network $B = (G, \Theta)$, Equation 3.2, can be used as a generative probabilistic model to produce a data set D_g that mirrors the probability distribution, that had been learned previously from the original data set D (cf. Section 3.1). Drawing gene expression profiles without an intervention works as follows (cf. Algorithm 1 in Figure 6.2): First, all variables are ordered such that the parents pa_i of each variable x_i are instantiated before x_i itself. Next, variables are selected according to this ordering and instantiated with a value, $x_i = x_{i,g}$. The value of each variable is selected with probability $p(x_i|pa_{i,g})$, where $pa_{i,g}$ denotes the already selected states for x_i's parents. This procedure is repeated until all variables are instantiated to form a generated global gene expression profile $\mathbf{x}_g$, and until N gene expression patterns are drawn to form an artificial data set D_g.

6.1.2 Interventional Modeling

The approach of interventional modeling estimates the effect of a certain intervention on the behavior of the learned Bayesian network using a combination of probabilistic inference and data sampling. The aim is to draw gene expression patterns and to form an artificial data set $D_{g|E}$ under a set of interventions, which are imposed as a set of evidences E. Possible interventions can be for example (i) clamping a subset $\mathbf{x}_E$ of genes to certain values and/or (ii) clamping parts of the graph structure G to certain values yielding a new posterior distribution $p'(G) \neq p_Q(G)$ (e.g., forcing an edge to be present or absent). In this work we focus solely on the first type of intervention, namely clamping genes to certain expression levels and measuring the expression behavior of all other genes given this intervention, by generating data.

Generating data under interventions (cf. Algorithm 2 in Figure 6.3) is done by propagation of evidence through the Bayes-net, that is, by obtaining the posterior distributions of the subset $\mathbf{x}_q = \mathbf{x} \setminus \mathbf{x}_E$ of free expression levels. The posterior distribution follows

$$p(\mathbf{x}_q|E) = \sum_G p(\mathbf{x}_q|E, G)p'(G), \tag{6.1}$$

were $p(\mathbf{x}_q|E, G)$ denotes the joint probability to measure gene expression levels $\mathbf{x}_q$ in a network with structure G, given certain genes have been fixed to expression levels by an intervention E. Before instantiation, the free variable set $\mathbf{x}_q$ is sorted as described in the previous section, such that for each variable $x_i \in \mathbf{x}_q$ its parents pa_i are ordered before the variable itself. In contrast to the sampling procedure without intervention, the distribution over values of x_i depends on its parents pa_i and on the set of gene expression levels $\mathbf{x}_E$ instantiated through the intervention E. Thus, the conditional probability has to be calculated by performing Bayesian inference

$$p(x_i|pa_{i,g}, E) = \frac{p(x_i, pa_{i,g}, E)}{p(pa_{i,g}, E)}, \tag{6.2}$$

Algorithm 2 Interventional Sampling (B,E,N)

```
Input:
  $B$ - Bayes-net;
  $E$ - Set of interventions;
  $N$ - Number of independent samples.
Output:
  $D_{g|E}$ - Data set of $N$ independent samples given $E$.

  $\mathbf{x}_E$ - Set of observed variables;
  $\mathbf{x}_q = \{\mathbf{x} \backslash \mathbf{x}_E\}$ - Set of query variables.
Order $\mathbf{x_q}$ consistent with the condition
  that parents $pa_i$ are sorted before $x_i$
For $s = 1, ...., N$
  For $i = 1, ...., n$
    Let $x_i$ be the highest ordered node not instantiated in sample $s$
    Select state with probability $p(x_i =$state$|pa_{i,g}, E)$
    Update $s$-th sample of $D_{g|E}$
    Instantiate $x_i$=state
```

Figure 6.3: Algorithm to sample from a Bayesian network with interventions

where the numerator is computed by marginalizing the joint distribution over all variables except x_i, pa_i and $\mathbf{x}_E$ and the denominator is obtained by a subsequent marginalization over x_i:

$$p(x_i|pa_{i,g}, E) = \frac{\sum_{\mathbf{x}\backslash x_i, pa_i, \mathbf{x}_E} p(\mathbf{x})}{\sum_{\mathbf{x}\backslash pa_i, \mathbf{x}_E} p(\mathbf{x})} \tag{6.3}$$

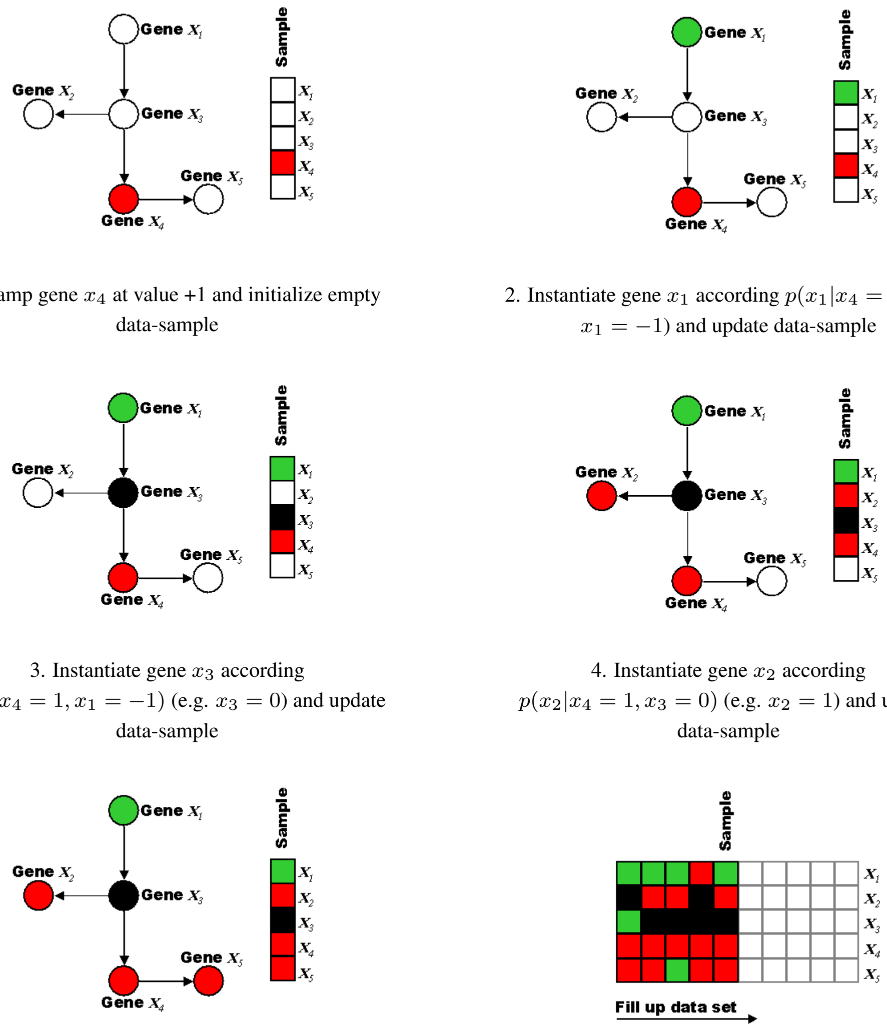
In order to efficiently solve Equation 6.3, we use bucket elimination (Dechter 1996), an exact inference algorithm in which variables are summed out one at a time. Each gene $x_i \in \mathbf{x}_q$ is then instantiated according to Equation 6.3 until the full vector $\mathbf{x} = (\mathbf{x}_q, \mathbf{x}_E)$ of gene-expression levels is instantiated.

Figure 6.4 illustrates the generation of one data-sample from a 5-gene network, where each gene can be either under-, normal- or overexpressed, given gene x_4 is observed as overexpressed ($x_4 \equiv x_E = 1$). After gene x_4 is fixed at its observed state and the empty data-sample is initialized (step 1), each non-observed gene, beginning with the highest ordered one, gene x_1, is instantiated with an expression value according Equation 6.3. The data-sample is updated with the selected instantiation until the full vector is drawn (step 2 - step 5). Note, that to solve Equation 6.3 only those variables are required, that are conditionally relevant to x_i (Shachter 1998). Thus, for gene x_1 Equation 6.3 can be written as:

$$p(x_1|x_4 = 1) = \frac{\sum_{x_3} p(x_4 = 1|x_3)p(x_3|x_1)p(x_1)}{\sum_{x_1}\sum_{x_3} p(x_4 = 1|x_3)p(x_3|x_1)p(x_1)} \tag{6.4}$$

Step 1 - step 5 are repeated until the full data set of N samples is filled up (step 6).

The marginalization over the graph structure in Equation 6.1 is approximated by drawing expression patterns from all Q graph structures obtained from the bootstrap procedure after application of structural expression interventions E, until $D_{g|E}$ is complete. For example, data of Table 6.1 and Table 6.2, were obtained by drawing 100 samples from each of the $Q = 20$ graph structure to form $N = 2000$ artificial patterns, which were compared to the measured patterns as described below.



1. Clamp gene x_4 at value +1 and initialize empty data-sample

2. Instantiate gene x_1 according $p(x_1|x_4 = 1)$ (e.g. $x_1 = -1$) and update data-sample

3. Instantiate gene x_3 according $p(x_3|x_4 = 1, x_1 = -1)$ (e.g. $x_3 = 0$) and update data-sample

4. Instantiate gene x_2 according $p(x_2|x_4 = 1, x_3 = 0)$ (e.g. $x_2 = 1$) and update data-sample

5. Instantiate gene x_5 according $p(x_5|x_4 = 1)$ (e.g. $x_5 = 1$) and update data-sample

6. Proceed with step 1 until data set is drawn

Figure 6.4: Interventional modeling scheme. Step 1 - step 5 illustrate the generation of one data-sample from a 5-gene network. As an intervention gene x_4 is kept fixed on its over-expressed state. Conditioned to its parents and the observed intervention, each unobserved gene is instantiated with an expression-level and the current data-sample is filled up with the selected value. This procedure is repeated until the full data set is generated (step 6). The drawn samples reflect the impact of the observed intervention on the global expression behavior.

6.1.3 Statistical Comparison of Data Sets

To estimate the effect of an intervention E on the global expression behavior, the artificial data set $D_{g|E}$ is compared with gene expression patterns of different conditions $D_t, t = 1, ..., T$. For the following approach we use a dissimilarity measure between two gene expression vectors, the generated **X** and measured **X'** one, $d(\mathbf{X}, \mathbf{X'})$, which was chosen as the Euclidean distance. Dissimilarity measures based on the city-block metric or mismatch count turned out to yield very similar results.

In order to estimate the correlation between an evidence E and a condition t we have to calculate the posterior probability $p(t|E)$ for the appearance of condition t given evidence E (note that, as our goal it is to examine disease related mechanism, the terms condition and disease will be used likewise). We therefore estimate the number of samples $\mathbf{x}_{g|E}$ of $D_{g|E}$ that are closest to D_t, $N_{t|E}$, by calculating the distance between each generated sample and each condition D_t and assigning the sample to the closest condition t. In the limit as N becomes large, the maximum likelihood estimate for the conditional probability is given as

$$p(t|E) = \frac{N_{t|E}}{N}, \tag{6.5}$$

were the denominator $N = \sum_{t=1}^{T} N_{t|E}$ is given as the total number of samples of $D_{g|E}$.

Finally, we propose a criterion for the probability of an intervention E being the cause for a condition t, whose effect is given by its specific gene expression profile. The pathogenic score of an intervention E for a condition t is defined as

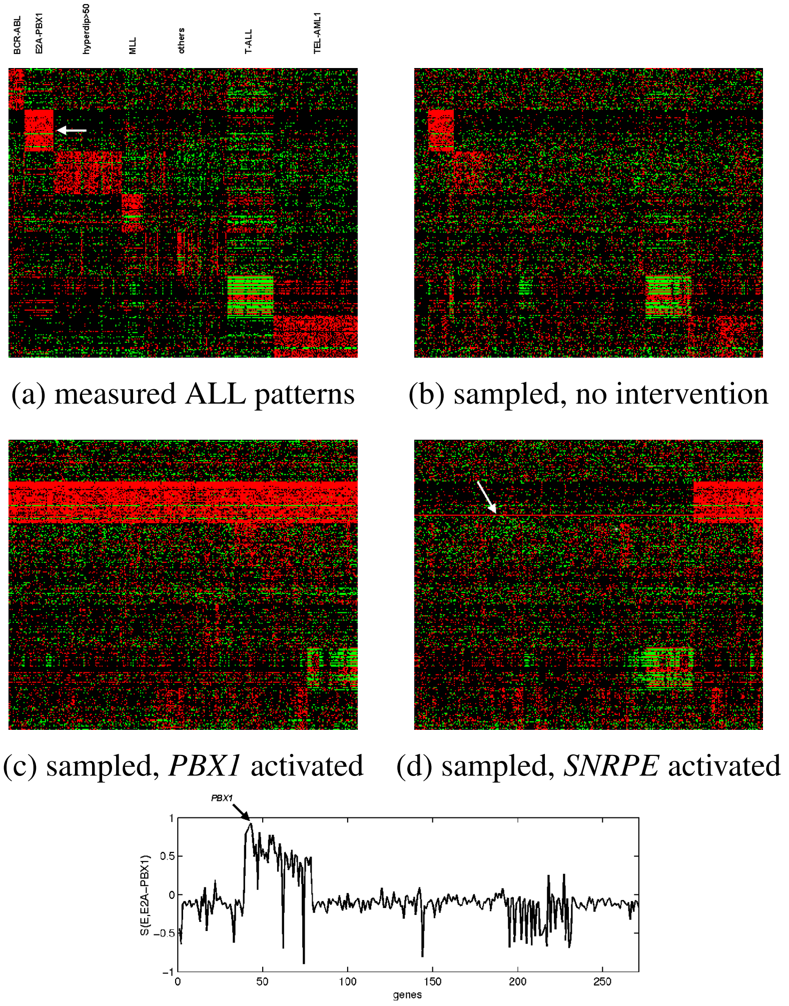
$$S(E,t) = \begin{cases} \frac{p(t|E)-p(t)}{1-p(t)} & p(t|E) > p(t) \\ \frac{p(t|E)-p(t)}{p(t)} & \text{else} \end{cases} \tag{6.6}$$

where $p(t)$ is the relative frequency of the condition t pattern without intervention. The pathogenic score corresponds to the normalized increase or decrease in relative frequency of a generated, condition specific pattern caused by an intervention E. It ranges within $[-1, 1]$ and represents a probabilistic measure of how much the intervention stabilizes or destabilizes a disease specific global expression profile and therefore the corresponding disease t.

6.2 Estimation of Oncogenes by Generative Inverse Modeling

This chapter presents an application of the generative inverse modeling procedure, based on the ALL microarray data set (see B.2.1), for the case of imposing interventions by means of clamping genes on a fixed expression value. The underlying model consists of 20 Bayesian networks learned from the leukemia data set with a simulated annealing scheme ($T_{start} = 5$). In this study we wanted to estimate those gene expression alterations which are most likely to be the cause of one of the 7 ALL subtypes, the original data is composed of. Thus, for the estimation of the pathogenic score, t runs over the different ALL subtypes where each of them is represented by the mean vector D_t over samples of the same subtype. As already mentioned in the previous section, results were obtained by drawing 100 samples from each of the $Q = 20$ graph structure to form $N = 2000$ artificial patterns.

For the following computer experiment, we generated artificial data sets while fixing the state of gene *PBX1* to be permanently overexpressed (activating *PBX1*). This was done by using Algorithm 2 in Figure 6.3. While enforcing $x_E \equiv x_{\text{PBX1}} = 1$, instances for the remaining gene expression levels $\mathbf{x}_q = \mathbf{x} \setminus x_{\text{PBX1}}$ were drawn from the resulting conditional probability $P(\mathbf{x}_q | x_{\text{PBX1}} = 1)$. For a better visual comparison of training and artificial data, Figures 6.5b-d consist of 327 randomly selected samples from the corresponding drawn data set (N=2000), whereas for all calculations the entire data sets were used. Figure 6.5b shows 327 samples drawn from the learned network without any intervention using Algorithm 6.2, after average linkage hierarchical clustering has been applied to the columns (samples). A comparison of Figure 3b with 3a shows, that the artificial data resembles



(a) measured ALL patterns

(b) sampled, no intervention

(c) sampled, *PBX1* activated

(d) sampled, *SNRPE* activated

(e) pathogenic score as a function of monogenic activations

Figure 6.5: **(a)** Training data consisting of 327 gene-expression profiles from patients with seven different subtypes of childhood acute lymphoblastic leukemia (ALL). Each column contains the expression profile for one patient, each row the expression levels of one of 271 genes selected to be most discriminating between the subtypes. The different ALL subtypes are accompanied by globally different expression profiles. **(b)** 327 artificial expression profiles produced by the generative model without intervention, in the same format as **(a)**. **(c)** 327 generated expression profiles, when proto-oncogene *PBX1* is clamped to the over-expressed state. **(d)** 327 generated patterns, when gene *SNRPE* is clamped to the over-expressed state (red line marked with arrow). Activating *SNRPE* has not such a strong effect on the global expression pattern as activating *PBX1*. **(e)** Pathogenic score for ALL-subtype E2A-PBX1, plotted against the 271 interventions, where each of the genes is individually clamped to the over-expressed state. The global maximum of the score is reached for *PBX1*: E(PBX1=1,E2A-PBX1)=0.923.

the measured expression patterns both in appearance and frequency of occurrence. Only two of the subtype patterns which are rare in the data (BCR-ABL and MLL) are less well reproduced. However, the general similarity reflects the capability of the model to correctly capture the statistical structure from the size-limited data set.

When artificial data are generated under this intervention, the relative frequency of those artificial patterns that are most similar to the E2A-PBX1 subtype patterns, dramatically increases from 0.07 to 0.93. This vast over-representation of the E2A-PBX1 typical expression pattern as a consequence of the intervention is reflected in Figure 6.5c where a random selection of 327 drawn samples with activated *PBX1* is shown. In contrast, clamping gene *SNRPE* (small nuclear ribonucleoprotein) to the active state (cf. arrow in Figure 3d) does not strongly increase the relative frequency of these patterns. This finding is remarkable, because both *PBX1* and *SNRPE* belong to the same gene cluster (cf. arrow in Figure 6.5a) and are both over-expressed in E2A-PBX1 ALL patients. In other words, whereas standard cluster analysis cannot distinguish between these two genes, generative inverse modeling clearly ranks both genes differently in light of their role for pathogenesis: *PBX1* is predicted to cause stabilization of the pathologically altered expression pattern in E2A-PBX1 ALL, whereas *SNRPE* is predicted to be over-expressed as an effect.

In fact, *PBX1* is hypothesized to be a proto-oncogene: Due to the chromosomal translocation t(1:19), *PBX1* fuses with gene *E2A* and converts to a potent transcriptional activator (van Duk et al. 1993) for which accumulating evidence supports its importance for causing ALL subtype E2A-PBX1 (Aspland et al. 2001). Hence, the model correctly predicted both the identity of the oncogene and the type of pathogenic transition (the gene is being activated).

Generative inverse modeling was then performed by searching through the space of all monogenic interventions (i.e., clamping all genes one by one) and calculating the pathogenic score for ALL-subtype E2A-PBX1. This pathogenic score is shown in Figure 6.5e, where the abscissa runs over all 271 genes being individually set to the active, over-expressed state. Many of the genes that belong to the most significant active gene cluster in E2A-PBX1 ALL (arrow in Figure 3a) show increased scores, however the absolute maximum is given by *PBX1*.

Next, the pathogenic scores for all known subtypes were calculated as a function of monogenic activation and suppression, respectively, resulting in 542 scores for each subtype. Table 6.1 summarizes the five genes that reached the highest scores for each subtype. All of the highest scoring interventions were activations. In all but one case, the highest scoring genes are substantially involved in oncogenic processes by annotation. For the remaining case – the hyperdipd>50 subtype – the model predicts an over-expressed proteosome 26S subunit gene *PSMD10* as pathogenic. This seems reasonable, as 26S is involved in general protein degradation and could be hyperactive in response of excess protein production by the hyperdiploidy. The side effects of its hyperactivity might then be driving the pathogenesis of this subtype.

For the subtypes where the putative disease causing proto-oncogenes were part of the training data set (*PBX1* (Aspland et al. 2001) for E2A-PBX1 and *ABL* (Fainstein et al. 1987) for BCR-ABL), they reached the highest scores. However, for BCR-ABL all scores are relatively small and of limited robustness, probably caused by the small sample size of these profiles in the training data. For the TEL-AML and MLL subtypes, the corresponding pathogenic genes *AML* and *MLL* were not contained in the training set used. This probably deteriorates the predictive power of the model for the highest scoring genes. For T-ALL, the examination reached high but unspecific scores, which reflects the extremely tight dependencies of the genes in the active gene cluster.

Finally we examined higher order effects by setting interventions to selected pairs of genes. The top part of Table 6.2 exemplifies a synergistic pathogenic effect predicted by the model for the hyperdip>50 subtype. *PSMD10* and the metallo-endopeptidase *MME* are both involved in protein degradation, but activating *MME* alone has no detectable impact on the score. However, when both genes are coactivated, the score substantially exceeds

Score	Affy-name	Gene	Putative function
hyperdip>50			
0.535	37350_at	*PSMD10*	proteosome subunit, protein degradation
0.513	41132_r_at	*HNRPH2*	RNA-binding
0.432	38518_at	*SCML2*	embryogenesis, transcription factor
0.382	38317_at	*TCEAL1*	repression of Pol II transcription
0.378	36169_a	*NDUFA1*	energy generation
E2A-PBX1			
0.923	33355_at	*PBX1*	transcription factor, proto-oncogene
0.899	32063_at	*PBX1*	transcription factor, proto-oncogene
0.834	1287_at	*ADPRT*	DNA repair, apoptosis
0.809	39614_at	*KIAA0802*	unknown
0.790	41146_at	*ADPRT*	DNA repair, apoptosis
BCR-ABL			
0.144	39730_at	*ABL1*	apoptosis related, proto-oncogene
0.140	36591_at	*TUBA1*	tubulin cytoskeleton associated
0.130	34362_at	*SLC2A5*	glucose transport
0.129	1636_g_at	*BCR*	signal transduction, oncogenesis
0.118	330_s_at	–	tubulin cytoskeleton associated
TEL-AML1			
0.518	41442_at	*CBFA2T3*	cell proliferation, transcription factor
0.501	36524_at	*ARHGEF4*	cytoskeleton
0.495	35614_at	*TCFL5*	cell proliferation, differentiation
0.484	1299_at	*TERF2*	telomerase-dependent telomere maintenance
0.473	36985_at	*IDI1*	isoprenoid biosynthesis
MLL			
0.139	33412_at	*LGALS1*	control of proliferation, apoptosis
0.128	2036_s_at	*CD44*	matrix adhesion
0.114	1126_s_at	*CD44*	matrix adhesion
0.099	40763_at	*MEIS1*	Pol II transcription, oncogenesis
0.087	38413_at	*DAD1*	anti-apoptosis
T-ALL			
0.833	38319_at	*CD3D*	T-cell receptor, cellular defense response
0.832	39709_at	*SEPW1*	unknown
0.812	36277_at	*CD3E*	cellular defense response
0.811	38415_at	*PTP4A2*	hydrolase
0.811	33238_at	*LCK*	regulation of cell cycle

Table 6.1: The five highest-scoring interventions (clamping all genes one by one at its over-expressed state) for each ALL-subtype

Score	Gene 1	Gene 2	Putative function
			hyperdip> 50
0.535	*PSMD10*		proteosome, protein degradation
-0.030	*MME*		protease, protein degradation
0.764	*PSMD10*	*MME*	
0.201	*TGFB1*		anti-apoptosis
0.608	*PSMD10*	*TGFB1*	

Table 6.2: Score changes for selected bi-genic interventions (clamping genes at its over-expressed state)

the sum of both individual scores. When, in contrast, a second gene from the *PSMD10* gene cluster is coactivated, both scores sum up roughly linearly or less (Table 6.2, bottom). This observation hints towards a synergy between the *PSMD10* and *MME* gene products in agreement with the annotation. This synergy appears particularly interesting because *MME* is not in the same gene cluster as *PSMD10*. This finding thus underlines the analytical power of generative inverse modeling by Bayesian networks for detecting functional relations between genes from expression data.

6.3 Summary

Revealing the underlying cause of pathogenic mechanisms and the related alteration of genetic programs represent one of the major challenging tasks of the post-genomic era (de Jong 2002, Stetter et al. 2003). Standard clustering methods were among the earliest tools for grouping genes by their functions from expression measurements (Eisen et al. 1998, Yeoh et al. 2002) and although experiments showed, that groups of genes are often coexpressed in a characteristic manner while a certain cellular function is carried out, no causal relationships and functional roles can be assigned to the different genes within a cluster or between the clusters itself.

One major contribution of the here presented approach is to suggest an alternative way of using the dependency structure found by a Bayesian belief network. We bypass structural considerations of the graph structures and do not directly interpret edges of a graph as causal regulatory relationships. Instead, the Bayesian network is treated as a density estimator and as a generative model with the ability to produce artificial expression data sets. With this approach, the impact of interventions on the global expression behavior is shown in an intuitively fashion, namely as artificial expression profiles. Thus, one can perform what-if scenarios on genetic regulatory systems *in silico*, rather than in lab. Moreover, with the so-called pathogenic score it is then possible to label each intervention with a probability of how likely it is triggering a certain gene expression profile and therefore the disease the expression profile is characteristic for.

Chapter 7

Biological Priors for Bayesian Network Learning

The results of previous sections have shown that learning Bayesian networks from microarray data can be useful to gain a better understanding of genetic network principles. The decomposition into conditional dependencies can describe the statistics of regulatory relationships, such as transcriptional mechanisms, and the probabilistic nature renders Bayesian networks and graphical models in general ideal candidates for modelling cellular systems.

However, structure learning on the basis of microarray data suffers from a number of difficulties such as small sample size and noisy measurements.

Another problem is related to the type of data which DNA microarrays provide. As noted in Section 2.2.2 transcriptomic data are not sufficient to capture the complexity of the whole cellular machinery. Post-translational effects, methylation, protein-protein interactions or metabolic changes are not measurable with DNA microarray techniques and might not enter the learned structure. On the other hand, as shown in Chapter 2.2, a huge amount of additional information in form of genomic, proteomic and metabolomic data is available and should be used for the reverse engineering procedure to gain a complete view of an observed cellular state. This again underlines the advantage of the Bayesian frame-

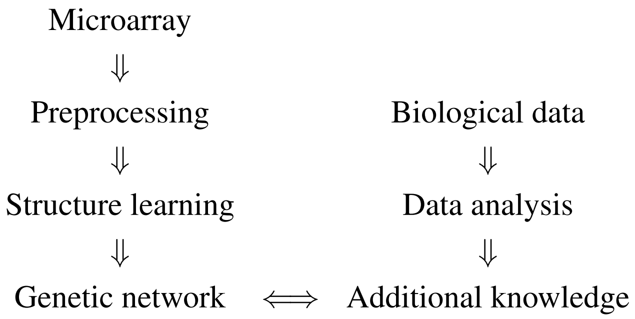


Figure 7.1: Schematic illustration of how prior knowledge is incorporated so far. Only microarray data are used to learn a genetic network whereas additional biological information from other data-sources, e.g. gene annotation, is only used after the learning step, for example to explore the relevance of learned structural features.

work which can easily use such additional information by integrating it as a probabilistic prior into the structural learning process.

7.1 Using Biological Prior Knowledge to Guide Network Inference

In the Bayesian scoring approach the fitness function of a structure G given a data set D is

$$\log p(G|D) = \underbrace{\log p(D|G)}_{\text{likelihood}} + \underbrace{\log p(G)}_{\text{prior}} \tag{7.1}$$

where $p(D|G)$ represents the data likelihood, our belief in the probability that G is correct given the data set D, and $p(G)$ represents the prior probability, our belief in G before we see the data. Together, both estimates form our updated belief about G after we saw the data. In contrast to the frequentist approach which derives the posterior from the data solely, the

Bayesian approach makes use of a prior distribution to guide the posterior belief.

Given a set of n variables, the prior belief of the Bayesian network structure $p(G)$ can be written as a $n \times n$ matrix p, where p_{ij} is our prior belief for variable x_i being a parent of variable x_j. p_{ij} implicates that x_j conditionally depends on x_i which is symbolized by an edge i→j. Biologically speaking, p_{ij} quantifies our prior belief that gene i regulates, or influences in some other way, the expression behavior of gene j. $p_{i \perp j}$ denotes our prior estimate for an independency of x_i and x_j and is given by

$$p_{i \perp j} = 1 - (p_{ij} + p_{ji}) \tag{7.2}$$

By this means any prior knowledge about gene-gene, protein-gene or protein-protein relationships can be included into the prior probability of the network structure $p(G)$.

If no prior knowledge regarding the structure is available, because for example nothing is known about the relationships between the corresponding set of genes, the biological prior belief is reflected by an non-informative biological prior, by giving equal weights to all possible structures. The only restriction comes from the acyclicity-condition of Bayesian networks: a self-cycle, where variable x_i points to itself, is excluded $p_{ij} = 0 \mid x_i = x_j$. Without prior knowledge, a conditional independency between x_i and x_j ($p_{i \perp j}$) is as likely as a conditional dependency of j from i (p_{ij}) and vice versa (p_{ji}) such that

$$p_{i \perp j} = p_{ij} = p_{ji} = \frac{1}{3} \tag{7.3}$$

and

$$p = \begin{pmatrix} 0 & \frac{1}{3} & . & . & \frac{1}{3} \\ \frac{1}{3} & 0 & . & . & \frac{1}{3} \\ . & . & 0 & . & . \\ . & . & . & 0 & . \\ \frac{1}{3} & \frac{1}{3} & . & . & 0 \end{pmatrix}$$

The probabilistic prior of network G can then be written in decomposed form such that

$$p(G) = \prod_{i=1}^{n} \prod_{j \in pa_i} p_{ji} \prod_{k \in ch_i} p_{ik} \prod_{l \in \mathbf{x} \backslash pa_i, ch_i} (1 - (p_{li} + p_{il})) \tag{7.4}$$

where pa_i denotes the parents of x_i and ch_i the children of x_i.

So far, we used additional biological knowledge only after the learning procedure to examine the biological relevance of learned network structures (see Figure 7.1). For example, consider the case of learning a Bayesian network of two genes, gene a and gene b. Assume that the data likelihood favors a dependency between a and b, however, as shown in Section 3.1.1, due to structure equivalence and the resulting likelihood equivalence, one can not favor $a \rightarrow b$ over $a \leftarrow b$. Thus, the result is a PDAG with an undirected edge. According to Figure 7.1, one could now use additional information about these two genes to prove the data-driven result. Let, for example, gene a be a known transcription factor which is known to regulate gene b. This fact confirms the dependency between a and b learned from the data. But additionally one can mark this (undirected) edge with an well-defined direction, namely from a to b. In this case, additional information was used as a prior to bias the evaluation of the learned structure.

The following sections present two approaches which use additional biological knowledge already in early stages of the genetic network inference framework. The goal of both approaches is to improve the robustness of learned structures by guiding structure learning. Moreover, by explicitly incorporating supplementary knowledge about the cellular system the learned genetic network is drawn closer to the true genetic network.

7.1.1 Biasing Gene Selection and Structure Learning

Due to the sparseness of given datasets and computational limitations of the learning procedure only a relatively small set of genes is usually considered: One focuses on a significant gene subset **x** of presumably important genes from the entire gene set **X**. The selection

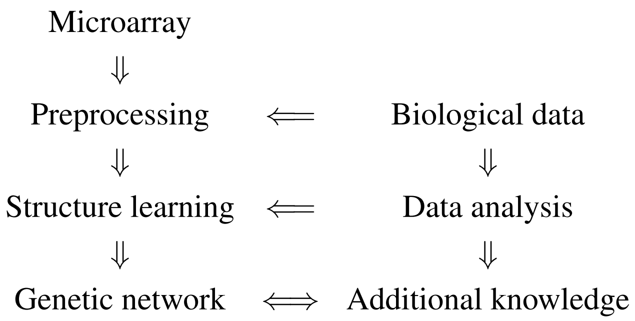


Figure 7.2: The preprocessing step which is crucial for the further structure learning procedure is biased by additional biological knowledge.

is usually done by choosing highly differentially expressed genes using various statistical tests such as ANOVA or t-test. Scoring genes according their expression profile seems to be a reasonable and sufficient way to select genes important for the learning process. Genes showing an altered expression in a certain cell state might play a crucial role in the underlying cellular machinery (Yeoh et al. 2002) and might therefore also be important for genetic network inference.

However, by focusing only on their expression profile, genes with weak differential expression, e.g. some transcription factors, will be neglected. Transcription factors play a central part in several regulatory mechanisms and are crucial for capturing the gene regulatory network of the cellular system (see Section 2.1.3). The lack of such important factors renders learned genetic networks incomplete and even potentially wrong. Such missing variables might bias structure learning especially if they are crucial for the exact understanding of the underlying system rendering learned genetic networks incomplete and even potentially wrong.

Thus, the preselected gene set has to be completed with those genes which are not in the initial preselection but are highly relevant for modelling the underlying genetic regulatory network.

Transcription Factor Binding Site Matches

Transcription factor binding sites (TFBSs) are commonly described by a W-length position weight matrix (PWM), w, where w_{bj} gives the probability for finding base b at position j of the binding site. The PWM of a transcription factor can be used to search for putative binding sites in the upstream DNA sequence of a gene suggesting a putative protein-DNA interaction. The estimation of putative binding sites for a certain TF is done by scoring the quality of an alignment of w with a nucleotide sequence s, usually located in the upstream region of a certain gene. For this, w is slid along s and at each position i a similarity score between the matrix and the W-length subsequence of s is calculated. Quandt et al. (Quandt et al. 1995), for example, proposed the following score: first, a consensus index is calculated meaning to what extent each position of a motif is variable. The conservation of the individual positions j in the matrix M is given by

$$\begin{aligned} &\mathbf{Ci}(j) = (100/\ln 5)(\sum_{b \in A,C,G,T} w_{bj} \ln w_{bj} + \ln 5) \\ &0 \leq \mathbf{Ci}(j) \leq 100 \end{aligned} \tag{7.5}$$

The resulting vector **Ci** provides a measure for the entropy at each position where $\mathbf{Ci}(j) = 100$ indicates a total conservation of one nucleotide at position j and $\mathbf{Ci}(j) = 0$ a flat distribution over all nucleotides. Using the consensus index, each motif is scored according to

$$\begin{aligned} &\text{mat\_sim} = (\sum_{j=1}^{W} \mathbf{Ci}(j) w_{bj})/(\sum_{j=1}^{W} \mathbf{Ci}(j) \arg\max_b w_{bj}), \\ &0 \leq \text{mat\_sim} \leq 1 \end{aligned} \tag{7.6}$$

where w_{bj} is the value for base b at position j in the PSSM. Multiplying each score by the **Ci** vector emphasizes the fact that mismatches at less conserved positions are more easily tolerated than mismatches at highly conserved positions.

To decide whether or not a score is high enough such that a transcription factor is able to bind to specific position, a threshold τ is introduced which should discriminate between matches and non-matches. A transcription factor is considered to bind only to those genes which have at least one position in their upstream region with a score higher than τ for the binding motif of this TF. Finding a *match* of a binding site for transcription factor a in the promoter region of gene b might therefore indicate a directed protein-DNA relationship, namely the regulation of gene b through transcription factor a.

Estimation of Prior-relevant Transcription Factors

Our task is to find TFs which relate to the observed gene expression alterations and which therefore should be enter structure learning. Consequently, we search for binding sites which are found more frequently in the gene preselection **x** than in the entire gene set **X**, as their overrepresented TFBSs might be an indicator for a relationship between gene altered expression and the corresponding TF. For example, the accumulation of a certain TFBS in a preselected list of strongly co-expressed genes might indicate that this co-expression is caused by means of a common TF.

To evaluate this, the statistical significance of each TF is characterized by a p-value, derived from the hypergeometric distribution. The hypergeometric distribution models the number of elements k with a certain property in a sample of size $|\mathbf{x}|$, selected (without replacement) from a total of $|\mathbf{X}|$, K of which have that property. The probability for this observation, arising by chance, is

$$p(k) = \frac{\binom{K}{|\mathbf{x}|}\binom{|\mathbf{X}|-K}{|\mathbf{x}|-k}}{\binom{|\mathbf{X}|}{|\mathbf{x}|}} \tag{7.7}$$

and the probability of observing k or more elements (with a certain property) in a randomly selected subset of size $|\mathbf{x}|$ is given by

$$p = \sum_{j=k}^{|\mathbf{x}|} \frac{\binom{K}{j}\binom{|\mathbf{X}|-K}{|\mathbf{x}|-j}}{\binom{|\mathbf{X}|}{|\mathbf{x}|}} \tag{7.8}$$

In our case the problem can be stated as follows: given a total number of $|\mathbf{X}|$ genes, of which K show a potential TFBS in their upstream region, and given a preselected gene set of size $|\mathbf{x}|$ of which k contain a potential match, is the relative frequency $\frac{k}{|\mathbf{x}|}$ significantly different from $\frac{K}{|\mathbf{X}|}$? Equation 7.8 quantifies this significance by means of a p-value. Thus, the lower $p_{\text{TF}\,i}$, the higher our belief belief in the significance of TF i for the preselected gene set. TFs with a low p-value and their potential regulated genes are presumable causing, directly or indirectly, the observed altered expression patterns. These TFs are therefore of major importance for inferring the underlying regulatory network structure and should be included to the initial gene set.

Protein-DNA Interaction Prior

Besides using significant TFBSs to refine the step of gene selection, we make use of the emerging knowledge about putative regulatory mechanisms to build a probabilistic prior which guides the structure learning process. If transcription factor TF i is suspected to act on gene j by binding at a specific upstream located binding site, our prior belief for an edge pointing from i to j increases in favor of our prior belief for the reversed variant or for the absence of this link, namely by

$$p_{ij} = 1 - p_{\text{TF}\,i} \tag{7.9}$$

$$p_{ji} = p_{i\perp j} = 0.5 p_{\text{TF}\,i}, \tag{7.10}$$

where $p_{\text{TF}\,i}$ denotes the p-value for TF i. Thus, prior $p(G)$ which previously was a non-informative, now turns into a (partially) informative prior. The smaller the p-value of a significant TF, the higher the prior belief for links pointing from this TF to those genes

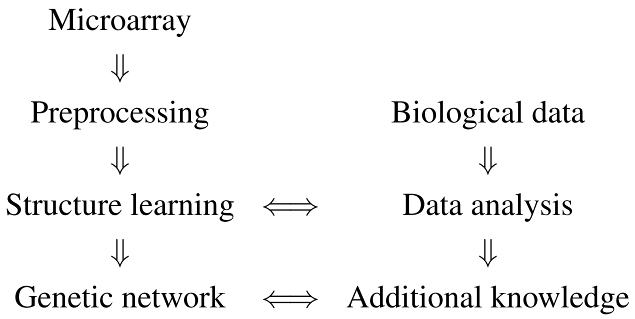


Figure 7.3: Structure learning is on the one hand guided by additional biological data but on the other hand structure learning itself guides the knowledge discovery of the additional data

owning a match for the corresponding TFBS. It should be noted that for the prior construction only highly significant TFs are considered ($p_{\mathrm{TF}\,i} \leq 0.01$) whereas for low significant TFs the prior remains flat.

7.1.2 Elimination of Equivalence Classes via Online Motif Search

A major problem of Bayesian network learning is the fact, that the conditional probability distribution learned from a data set D can not always be mapped to one single DAG. Instead, as shown in Section 3.1.1, a set of Bayesian networks can represent the same probability distribution such that some edges lack a unique direction and therefore a causal meaning. Thus, instead of obtaining a single Bayesian network where each edge represents a regulatory relationship among genes, structure learning from microarray data results in a PDAG where some dependencies are not uniquely defined.

This section presents an approach to resolve ambiguities in equivalence classes by making use of additional biological knowledge to discriminate between previously equivalent structures. Consequently, ambiguous relationships can be resolved, edges can be equipped with a unique direction and causal meaning is added. The probabilistic prior derives from motif discovery results, where a set of upstream sequences is scanned for a common pattern indicating a common TFBS and hence a putative regulation by the same TF. As shown in Figure 7.3, motif discovery, biological prior estimation and structure learning are done simultaneously and cooperatively. The result is a Bayesian network structure, learned from microarray data and guided by an online constructed structure prior, and a set of putative TFBSs.

Resolving Equivalence Classes

Bayesian network structures which share the same skeleton and the same set of colliders fall into the same equivalence class, whose member DAGs are assigned with the same score if no additional prior knowledge is available. However, another biologically relevant information lies in the network structure and is probably helpful to differentiate between equivalent structures. To explain this, let the complete non-singleton set of variables which share a common parent i be a ***regulon*** ϱ_i such that $\varrho_i = \{j | i \in pa_j\}$. From a statistical point of view, all members of ϱ_i causally depend on a common variable x_i which in terms of transcriptional regulatory mechanisms might refer to a common transcription factor TF i. As noted in Section 2.2.1, transcription factors regulate the genetic activity by binding to a specific domain which, in a simplified view, is located upstream of the considered gene. Genes which are regulated by the same transcription factor might therefore share a common TFBS in their promoter region. Consequently, since all genes of regulon ϱ_i are hypothesized to be regulated by a common gene i they probably share a common DNA sequence pattern or motif w_i in their upstream region, given as a position weight matrix (PWM). Even more important, finding a common DNA sequence pattern in the upstream region of a set of genes which share a common parent, confirms their affiliation to a common regulon structure.

Motif discovery in DNA sequences is a well known problem for which a variety of methods have been proposed, such as Gibbs sampling (Lawrence et al. 1993) or expectation maximization (Bailey & Elkan 1994). In our case the problem can be stated as follows: given a regulon ϱ_i find an unknown motif w_i that is shared by the upstream sequences s of ϱ_i's members and score regulon ϱ_i according its motif. In this work we use the EM algorithm of Bailey an co-workers (Bailey & Elkan 1994), whose simplest variant, the so-called OOPS algorithm, estimates one PWM by searching in each sequence exactly one substring. Given the found motif w_i, each regulon ϱ_i can be assigned with an additional score, namely, by evaluating the quality of the corresponding PWM. A common score is given by the information content of the weight matrix

$$I_{w_i} = \sum_{j=1}^{W} \sum_{b \in A,C,G,T} w_{bj} \log_2 \frac{w_{bj}}{w_b} \tag{7.11}$$

where w_b is the background frequency of base b which we assumed to be uniform. Given Equation 7.11 we define a probabilistic score for regulon ϱ_i as follows

$$S(\varrho_i) = \frac{1}{2W} I_{w_i}. \tag{7.12}$$

where W is the number of positions w_i consists of.

Given a structure G, a set of regulons $\{\varrho_1, ..., \varrho_n\}$ and the corresponding set of motifs $\{w_1, ..., w_n\}$ the estimates $\{p(w_1|\varrho_1), .., p(w_n|\varrho_n)\}$ can in the following be used to construct a probabilistic structure prior $p(G)$. A high value of $p(w_i|\varrho_i)$ which indicates a highly conserved TFBS shared within the members of regulon ϱ_i confirms the occurrence of this regulon as it is reflected in the structure G. Consequently, a directed link from gene i to gene j in regulon ϱ_i should be higher scored than a link from gene j to gene i. This altered prior belief for a specific link can be manifested in the structure prior matrix p by setting

$$p_{ij} = S(\varrho_i) \;\; \forall j \in \varrho_i \tag{7.13}$$

$$p_{ji} = p_{i \perp j} = 0.5(1 - S(\varrho_i)) \;\; \forall j \in \varrho_i \tag{7.14}$$

TF i	$p_{\text{TF } i}$	description	subtype
HFH-3	0.06	forkhead box I1	E2A-PBX1
HFH-3	0.06	forkhead box I1	BCR-ABL
GATA-x*	0.03	T-cell specific transcription factor	BCR-ABL
GATA-x*	0.007	T-cell specific transcription factor	TEL-AML1
AML-1*	0.01	runt-related transcription factor 1	TEL-AML1
Thing1-E47*	0.07	transcription factor E2-alpha	TALL

Table 7.1: List of TFs estimated as significant ($p_{\text{TF } i} < 0.1$) for a subtype specific expression pattern. $*$-labeled TFs are associated with leukemia oncogenesis by annotation.

As the structure prior $p(G)$ turns into a (partially) informative one, equivalence classes with at least one regulon can be dissolved completely into a set of unique Bayesian networks or at least partitioned into several equivalence classes where the number of undirected edges is reduced.

In contrast to the static prior estimate in Section 7.1.1, in this approach the probabilistic prior is calculated dynamically during the structure learning procedure. At each learning step the next candidate structure G' is estimated according to Equation 3.9 where the structure prior $p(G)$, given by Equation 7.4, is constructed with respect to the found motifs.

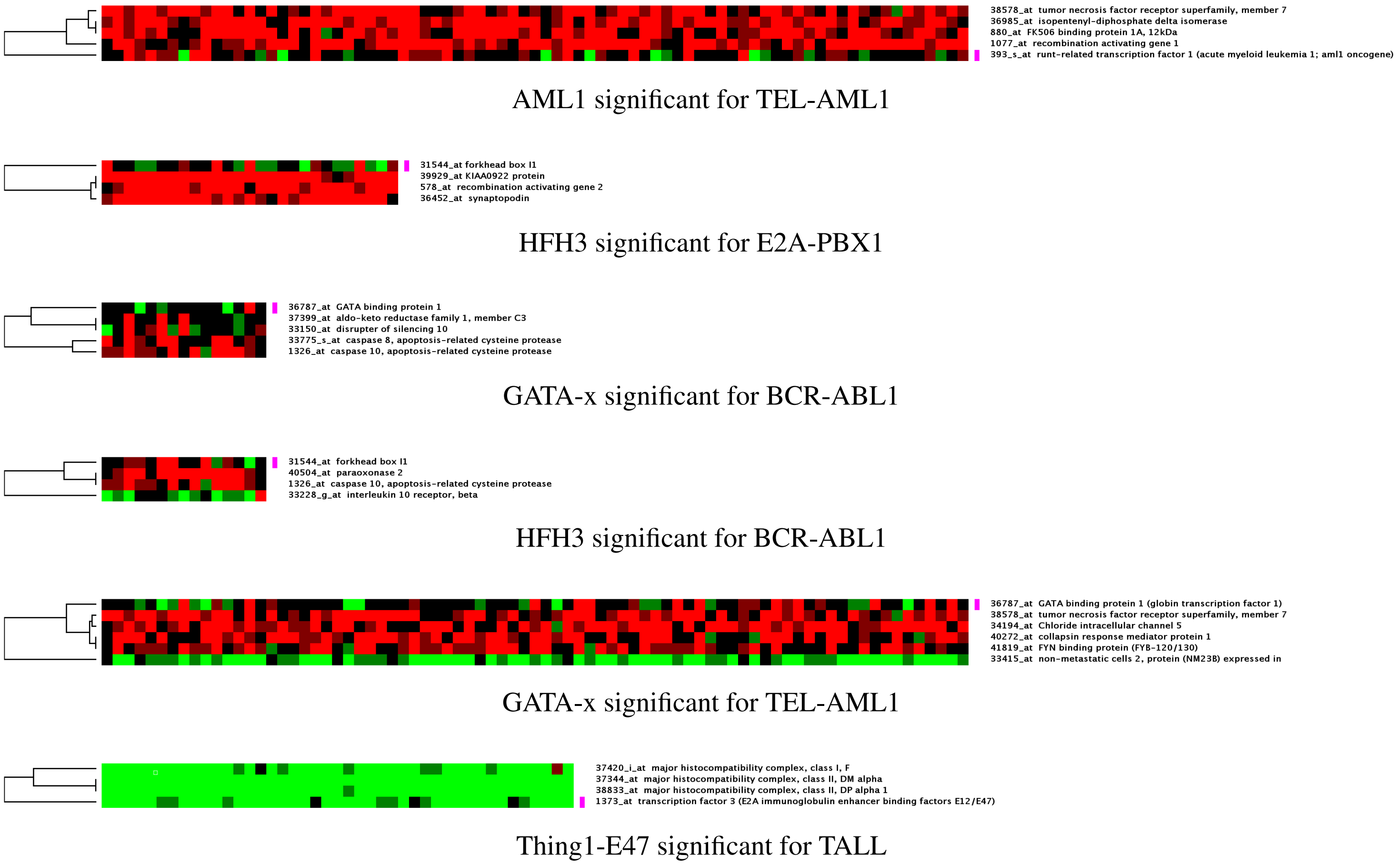
7.2 Using TFBS Information to Bias Genetic Network Estimation

As shown in Figure B.1 the prefiltered ALL data set decomposes into 7 clusters of co-expressed genes each of which corresponds to a specific ALL subtype. Studies of Yeoh and co-workers (Yeoh et al. 2002) as well as results from previous sections of this thesis have underlined the significance of the preselected genes for the particular subtypes.

However, as TFs are known to be weakly expressed, they might be missed in this preselection in spite of their putative importance. In the following we combine expression and sequence data analysis to extract significant TFs. Moreover, we refine the selected gene set and construct a probabilistic prior which is then used to guide structure learning. For this, a collection of 109 position weight matrices of different TFs from human has been derived from TRANSFAC and JASPAR database. Each of the weight matrices was then tested for an alignment in the upstream region (500 bp) of each gene present in the entire data set (12.000) using Equation 7.6. A common threshold of $\tau \geq 0.96$ was used to discriminate between matches and non-matches. Given these estimates each of the 109 TFs was assigned with a p-value for its significance regarding the 7 subtype-specific gene clusters. As listed in Table 7.1 only 4 TFs achieved a sufficiently low p-value $p_{\text{TF}_i} < 0.1$ to be labeled as significant. Their annotations underline our findings as most of the TFs (labeled with $*$) are associated with leukemia oncogenesis by annotation [1]. For example, we found the transcription factor AML-1 to be highly significant for the co-expression pattern in the ALL-subtype TEL-AML1. In fact, this subtype correlates with a chromosomal rearrangement resulting in the fusion of gene TEL and transcription factor AML-1 (Golub et al. 1995). Generally, as shown in Figure 7.4, the TFs show only a moderate differential expression which does not strongly correlate with the expression profile of the regulated gene, except the transcription factor Thing1-E47 which strongly correlates with the corresponding gene cluster. Moreover, all 3 genes related to Thing1-E47 are part of the Major Histocompatibility complex (MHC) which suggests a dependency between transcription factor Thing1-E47 and the Major Histocompatibility complex (MHC).

After integrating those genes which corresponds to the 4 TFs in our initial gene list the knowledge about putative genetic regulatory mechanisms is next used to alter the structure prior according to Equation 7.10. As in Section 4.4 we trained 20 networks according to the non-parametric data bootstrap and simulated annealing $(T_{\text{start}} = 10)$ with the difference that the current network is composed of 275 variables and the structure prior is partially informative based on these 4 TFs. As shown in Figure 7.5 the edge confidences strongly

[1] According to the CancerGene database.



AML1 significant for TEL-AML1

HFH3 significant for E2A-PBX1

GATA-x significant for BCR-ABL1

HFH3 significant for BCR-ABL1

GATA-x significant for TEL-AML1

Thing1-E47 significant for TALL

Figure 7.4: Gene expression profiles of significant TFs (marked magenta) and of their putative regulated genes. Shown are microarray experiments for patients suffering the respective subtype.

correlates with those of the network learned with an uniform prior (see Section 4.4). Especially high confident edges ($\tau \geq 0.5$) seem to be robust against changes in the gene selection or alteration of the structure prior. This gives us some trust in the accuracy of the confidence thresholds estimated in Section 4.3.2.

7.3 Online Motif Discovery on an Artificial Network

We designed an artificial network structure, shown in Figure 7.6 with 23 genes connected by 22 edges without any collider structure. From this network an artificial microarray data set with 1000 samples has been generated. The network consists of 5 regulons, each of which was assigned to a PWM taken from the Jaspar database. For each gene we randomly generated artificial DNA sequences and inserted a pseudo motif according to the corresponding PWM. First, we learned 20 networks only on the sampled microarray data with a simulated annealing scheme ($T_{\text{start}} = 5$). All 20 learned structures fall into the equivalence class of the original network. But since no prior has been used we can not discriminate between equivalent structures and that the resulting fPDAG is fully undirected. On the one hand, the results show the power of structure learning to train the original network from the data. But on the other side, they reveal the weakness of Bayesian networks regarding equivalent structures.

We next repeated the learning procedure with the extension that at each step candidate structures are weighted not only according to their expression profile but also according to the motifs estimated from the corresponding set of regulons respectively the constructed structure prior. The resulting structure falls into the same equivalence class as in the approach without prior. However, as one can see from Figure 7.7, we are now able to discrimante between equivalent structures with the effect that the learning procedure returns a fully directed structure. Except the reversion of one edge, namely from gene 36959_at to gene 39864_at, the highest scoring structure corresponds to the original network. This deviation is caused by the fact that the motif discovery procedure was unable to find the specific motif in gene 36959_at with the consequence that the regulon was higher scored without gene 36959_at than with it.

7.4 Summary

Besides gene expression data a variety of other data exist that are useful for genetic network inference. As Bayesian network learning requires infinitely many data points to infer the complete underlying density distribution, data sparseness is a big problem. However, the big advantage of the Bayesian approach in general is the ability of using a-priori information to refine the posterior estimate.

In this chapter we proposed two different approaches which estimate a structure prior from DNA motif discovery estimates and then guide the structure learning procedure. Furthermore, we addressed the problem of restricting the genetic network space to relatively small networks. Usually gene selection is done by focusing on gene expression data only, e.g. by selecting highly differentially expressed genes. However, as it is known that many weakly expressed genes are crucial for the underlying genetic network, such missing variables bias the resulting learned network. By scanning the entire set of genes for putative binding sites of known TFs and by extracting those TFs whose TFBS appears more often than by chance our approach is able to enlarge the initial gene selection by weakly expressed but still highly significant genes. The fact that none of the found TFs was present in the preselection, although their putative important role for the underlying genetic network, emphasizes the need of integrating and analyzing various types of data. Furthermore, the structure prior proposed in this chapter is not restricted solely to data used here, but can be adapted to integrate a variety of other biological knowledge ranging from quantitative data, e.g. protein-protein interaction networks, to qualitative statements which altogether assist structure learning.

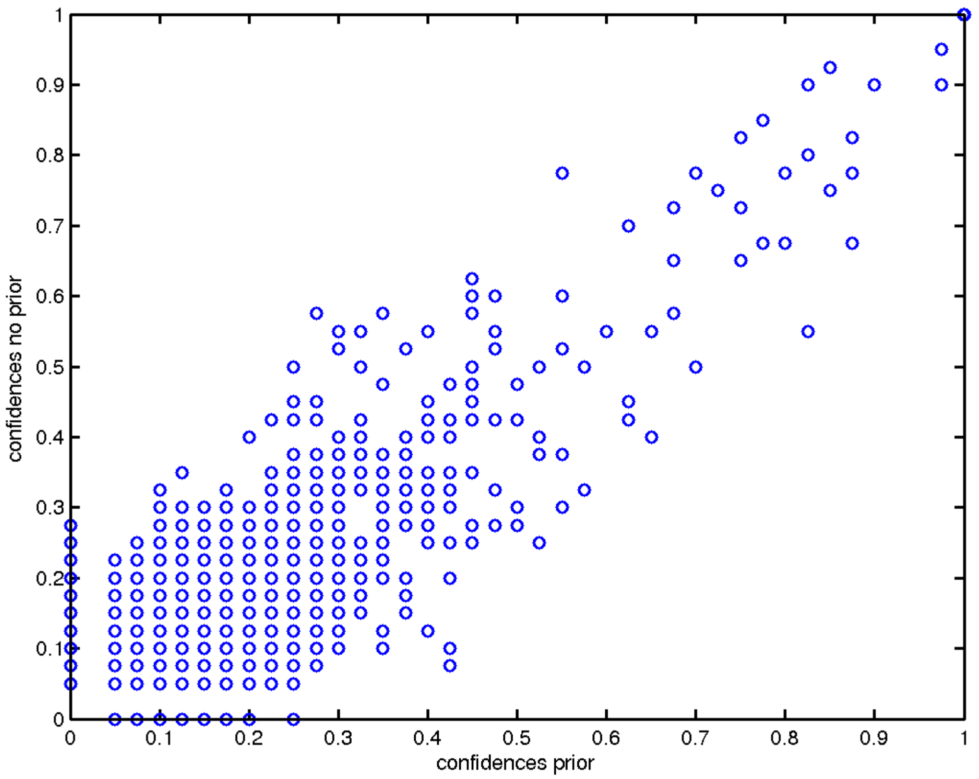


Figure 7.5: Edge confidences of the ALL network learned with uniform prior versus ALL network extended by significant TFs and learned with a (partially) informative prior. Especially high confident edges ($\tau \geq 0.5$) are not strongly affected by the extensions through additional genes and prior knowledge.

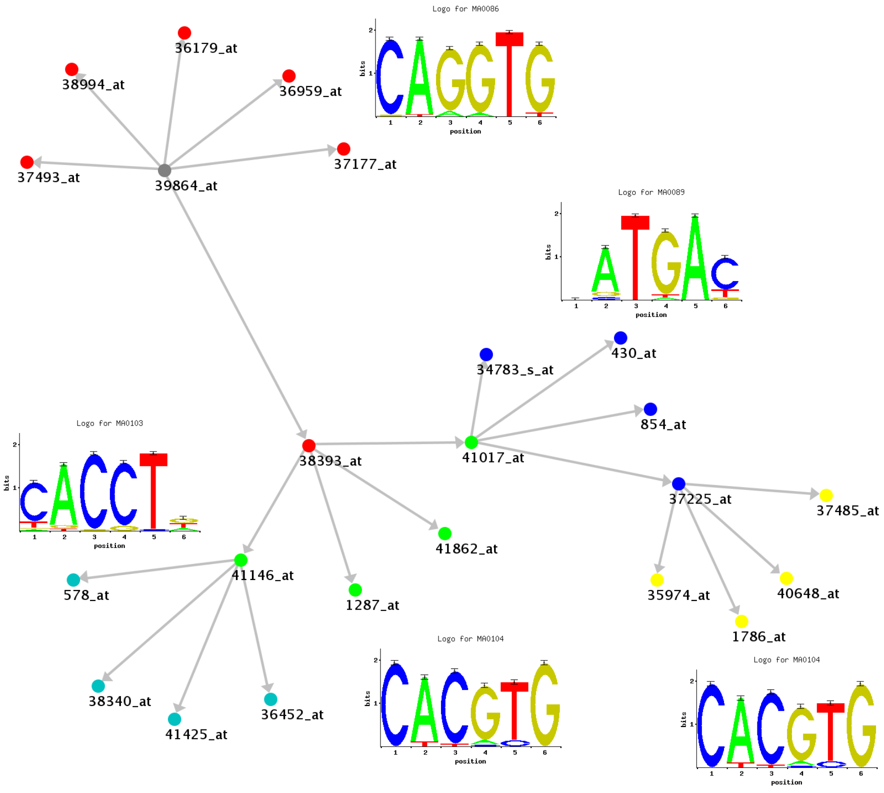


Figure 7.6: A designed network consisting of 5 regulons each of which is characterized by a specific TFBS. As no collider exists in the network no causal regulatory relationship can be learned.

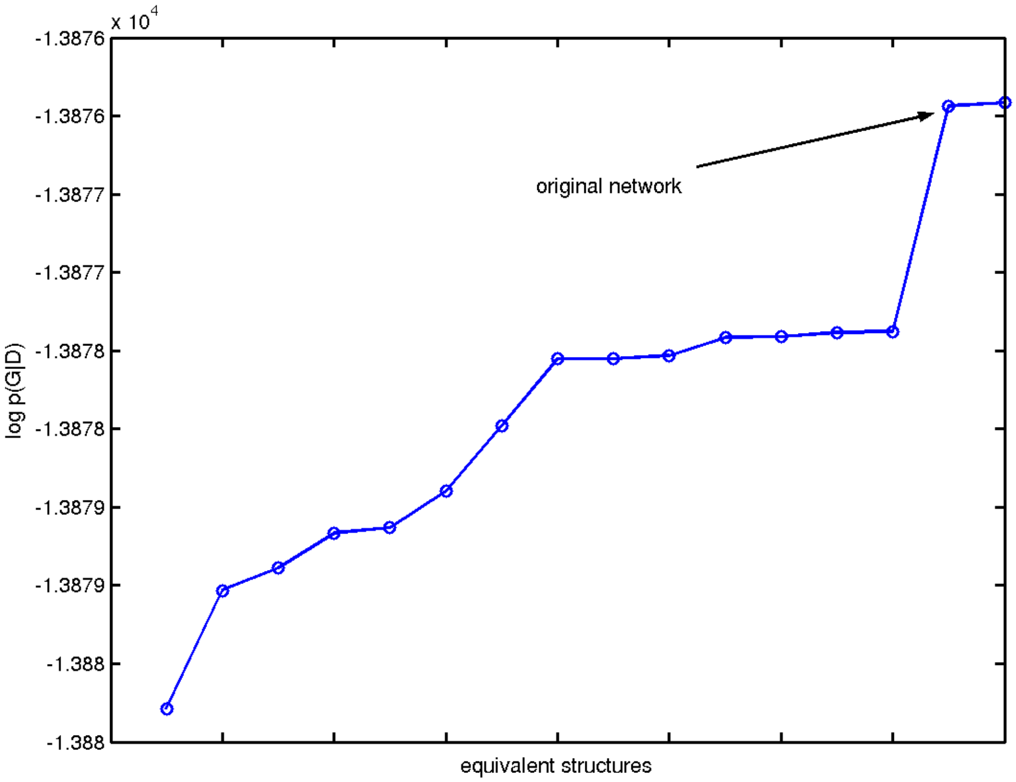


Figure 7.7: Given the informative structure prior estimate, equivalent structures are scored differentially according to their decomposition into high scoring regulons. Except the reversion of one edge the highest scoring structure corresponds to the original network.

Chapter 8

Mining Functional Modules with Decomposable Models

The descriptive power of decomposable models in light of the modular decomposition of molecular networks has already been motivated in Section 3.2. We therefore want to apply this approach on real microarray data to test its capacity of revealing molecular functional modules. This chapter is summarized in (Dejori, Schwaighofer, Tresp & Stetter 2004)

8.1 Mining Functional Modules in Acute Lymphoblastic Leukemia

The approach presented in Section 3.2 is capable to efficiently learn the graphical structure of a decomposable model from continuous data. By its structure, the decomposable model explains the statistical structure of the data in terms of functionally linked genes (the nodes). The first section is concerned with a topological analysis of the dependence graph with the goal to identify central genes. Moreover, the decomposable model can be represented as a clique tree, where each clique represents a group of genes (the nodes of the clique) which are fully connected, and therefore explains interacting functional modules (represented by cliques). This dense statistical dependency structure is thought to reflect the dense biological link of a functional module causing the gene expression patterns which will be examined in the second section.

To ensure a robust learning result, a bootstrap scheme was used. Moreover, only edges that have a confidence of 90 % or above, i.e., edges that were found in at least 18 out of the 20 replications, were considered for the following analysis. This thresholding by edge confidence may lead to a non- decomposable resulting model. To again obtain a decomposable model, some of the lowest confidence edges were pruned until decomposability was reached.

8.1.1 Analysis of the Dependence Graph

The network topology output by the structure learning algorithm (with restriction to high confidence edges, as described in the previous section) shows a few highly connected genes, with most edges connecting genes belonging to the same ALL subtype. Thus, most genes are conditionally independent from each other, given one of these highly connected genes. For example, in Figure 8.1, gene 37350_at renders most of its adjacent genes pairwise conditionally independent. Biologically speaking, the expression behavior of many genes only depends on a set of few genes which therefore are supposed to play a key role in the underlying genetic network. Since the structure is learned from leukemia data, these few highly connected genes are predicted to be important for leukemogenesis or for tumor development in general. In fact, as shown in Table 8.1, highly connected genes are either known to be genes with an oncogenic characteristic or known to be involved in critical biological processes, such as DNA repair or proteolysis.

Figure 8.1 shows part of the decomposable model learned on the ALL data set. Gene PSMD10 (Affymetrix-ID 37350_at, at the top of the figure) is found to be linked to a high number of other genes, and is therefore predicted by the model as important for the stability of cellular function. In fact, PSMD10 is a regulatory subunit of the 26S proteasome, a

Gene	Affymetrix ID	# of connections	Putative function
PSMD10	37350_at	17	proteosome, protein degradation
HLA-DRA	37039_at	13	immune response, antigen presentation
SCML2	38518_at	9	embryogenesis, transcription factor
POU2AF1	36239_at	7	transcription cofactor, anti-pathogene response

Table 8.1: Genes in the ALL data set, ranked by the number of connections. Each row lists the gene's name and Affymetrix ID, the number of connections and the putative function of this gene

Gene	Affymetrix ID	# of cliques	Putative function
PSMD10	37350_at	12	proteosome, protein degradation
HLA-DRA	37039_at	8	immune response, antigen presentation
SCML2	38518_at	6	embryogenesis, transcription factor
POU2AF1	36239_at	6	transcription cofactor, anti-pathogene response

Table 8.2: Genes in the ALL data set, ranked by the number of cliques they are contained in. Each row lists the gene's name and Affymetrix ID, the number of cliques and the putative function of this gene

protein complex which—in agreement with the model topology—degrades a large family of proteins that are marked to be destroyed, and thus helps regulating the protein turnover in eukaryotic cells. Hence, it is known to be crucial for normal cellular function. In particular, a malfunction of PSMD10 is known to result in a defective regulation of a large number of intracellular proteins that govern cell division, tumor growth, and tumor survival, and which are functionally altered in cancer cells. Indeed, recent work has shown that the PSMD10 pathway is often the target of cancer-related deregulation and can underlie processes, such as oncogenic transformation or tumor progression.

8.1.2 Analysis of Functional Modules

In a second step of analysis, we consider individual cliques of the learned decomposable model as functional modules.

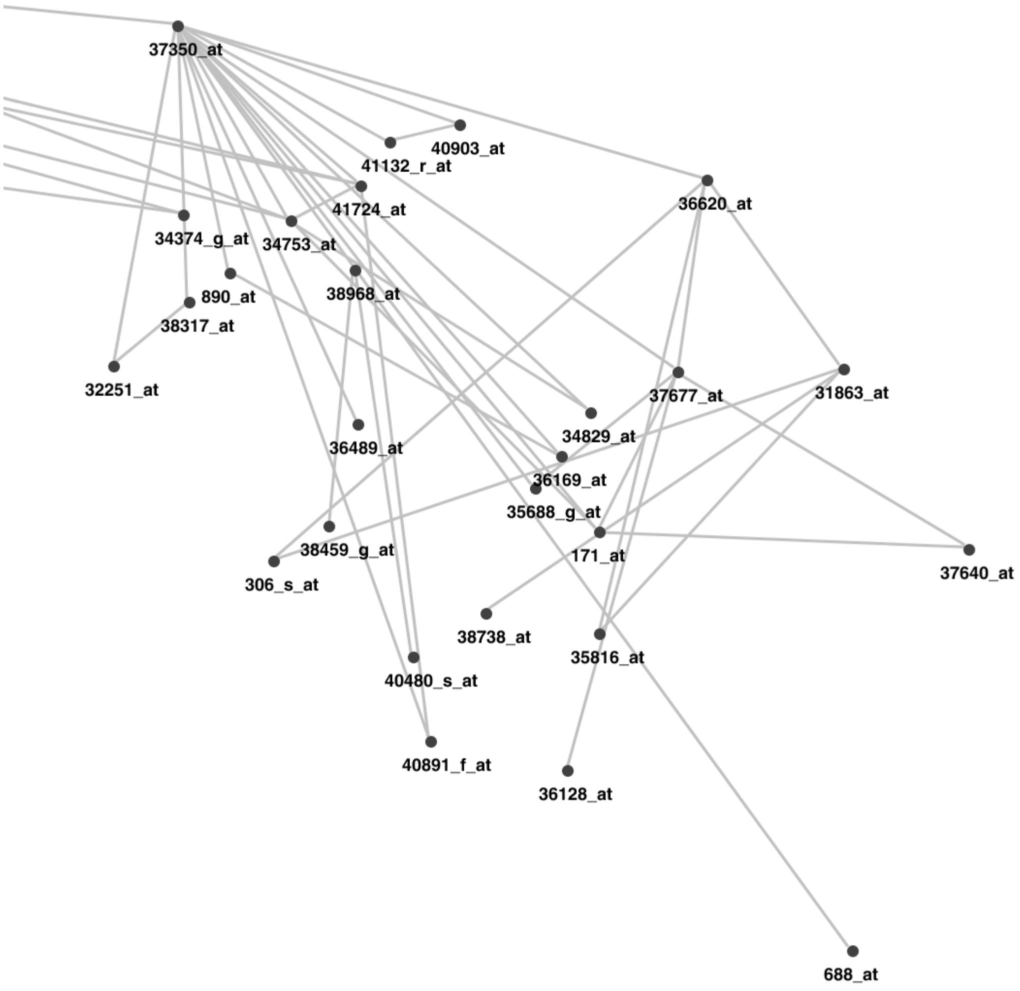


Figure 8.1: Part of the decomposable model structure learned on the ALL data set. The highly connected gene PSMD10 (Affymetrix-ID 37350_at, at the top of the figure) is known to be involved in cellular deregulations that potentially leads to oncogenesis

Gene	Affymetrix ID	# of cliques	Gene title
HLA-DRA	37039_at	8	MHC II, DR alpha
HLA-DMA	37344_at	4	MHC II, DM alpha
HLA-DPB1	38095_i_at	3	MHC II, DP alpha 1
HLA-DPA1	38833_at	2	MHC II, DP alpha 1
HLA-DPB1	38096_f_at	1	MHC II, DP beta 1
HLA-DRB1	41723_s_at	1	MHC II, DR beta 1

Table 8.3: Genes of the MHC II complex ranked by the number of cliques they are contained in. Each row lists the gene's name and Affymetrix ID, the number of cliques the particular gene is contained in, and the title of this gene

A first observation is concerned with multiply occurring genes in the data. In the given data set, 13 genes are represented twice and 2 thrice on the chip. We noticed that these multiply occurring genes always formed unique cliques which affirms our assumption that cliques contain (functionally) highly correlated genes.

We next focused on genes known to be subunits of a common functional module. From annotation data we found three functional modules, the *major histocompatibility complex class II* (MCH II), the *proteosome 26s* (p26S), and the *T-cell antigen receptor complex* (T3) with more than one member present in the data set. Interpreting cliques as functional modules, these genes should also belong to the same clique. And in fact it turned out that they were always put into one clique, or into adjacent[1] cliques. Furthermore, we ranked these genes by the number of cliques they are contained in. We suggest that the higher the number of cliques a gene is contained in the more important is its role in a module.

Genes listed in Table 8.3 are part of the MHC II complex. Class II molecules are composed of two polypeptide chains, α and β chains. The MHC II molecules themselves are highly polymorphic (meaning that there are many different variants of these genes within the population), forming different MHC II variants for different antigenes. Yet, HLA-DRA

[1] By adjacent we mean cliques that are adjacent in the clique graph (Galinier et al. 1995), that is, cliques that have some common nodes.

Gene	Affymetrix ID	# of cliques	Gene title
PSMD10	37350_at	12	26S proteosome, non-ATPase regulatory subunit 10
PSMC1	688_at	1	26S proteosome, ATPase regulatory subunit 1

Table 8.4: Genes of the p26 proteosome complex ranked by the number of cliques they are contained in. Each row lists the gene's name and Affymetrix ID, the number of cliques the particular gene is contained in, and the title of this gene

Gene	Affymetrix ID	# of cliques	Gene title
CD3D	38319_at	12	T3 complex, delta polypeptide subunit
CD3E	36277_at	1	T3 complex, epsilon polypeptide subunit

Table 8.5: Genes of the T3 complex ranked by the number of cliques they are contained in. Each row lists the gene's name and Affymetrix ID, the number of cliques the particular gene is contained in, and the title of this gene

itself is monomorph, thus it is present in almost each of the MHC modules. This agrees with the fact that each clique containing a MHC II member also contains HLA-DRA itself.

Genes listed in Table 8.4 are subunits of the 26S proteosome complex. Whereas PSMD10 is present in many cliques the other subunit,PSMC1, is present only in one clique, namely with PSMD10. This can be evidence for a more dominant role of PSMD10 in protein degradation than PSMC1.

8.2 Summary

A decomposable model tries to explain the statistics in a data set by the action of mutually linked functional modules, so-called cliques. Decomposable models with continuous variables have significant advantages for this application domain:

(i) Previous approaches (Friedman et al. 2000, Pe'er et al. 2001) have mostly concentrated on learning discrete valued models from such data. Hence, one needs to first discretize the continuous-valued expression level. This is a crucial and quite delicate pre-

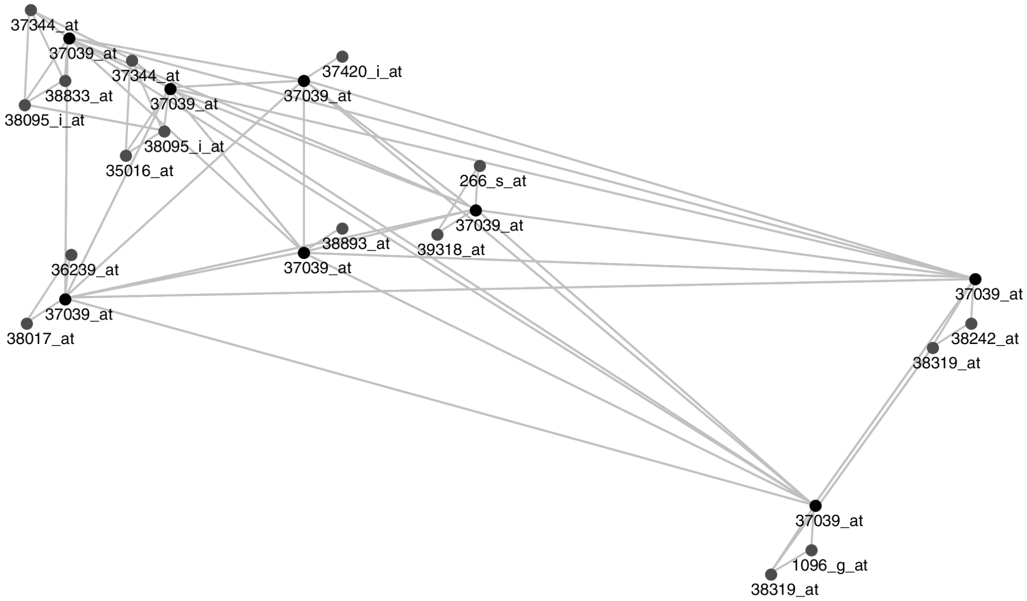


Figure 8.2: A part of the clique tree learned on the ALL data set. The highly connected gene HLA-DRA (Affymetrix-ID 37039_at) is present in each clique

processing step that needs to be conducted carefully (Friedman et al. 2000). In contrast, the approach adopted here accounts in a natural way for the continuous nature of the measurements, and for their unknown and probably non-Gaussian joint probability distribution.

(ii) Molecular networks often show a small-world topology (Jeong et al. 2000), in which the network is decomposable into smaller groups of densely connected clusters (Watts and Strogatz, 1998). This finding might render decomposable models with their intrinsic modular or clique-like structure particularly suitable for describing genetic networks.

(iii) Functional modules are considered to be a critical level of biological organization (Hartwell et al. 1999). One example are modules in transcriptional regulation. Transcription factors work by binding to DNA-motifs and affecting the rate of transcription. Many binding sites occur in spatial and functional clusters called enhancers, promoter elements, or regulatory modules. Thus, the promoter regions suggest a hierarchical or modular style of the transcription complex. Two further examples of molecular modules are subunits of multimeric proteins, where the subunits are coded by separate genes, or protein groups which associate into larger structures termed macromolecular assemblies. In the latter two cases, the genes for the different subunits or the genes that code for proteins of the same macromolecular assembly are functionally grouped to a module. Finally, gene products can also form a functional module by carrying out a certain cellular function in a concerted way, but without being physically grouped to a molecular assembly. The inherent modular structure of a decomposable model imposes a strong drive for it towards explaining the data in terms of densely linked gene groups. By this, decomposable models should be able to detect with particularly high sensitivity the signature of a concerted action of gene modules in the data. In light of this rationale, cliques are likely to contain functionally highly correlated genes, as opposed to gene clusters (Eisen et al. 1998, Yeoh et al. 2002), where genes are grouped together by mere coexpression. Hence, as opposed to clustering, the learned structure also reveals some information about the possible statistical relationship of genes within a cluster.

A decomposable model approach preserves some of advantages of the related systems-level modelling techniques by Bayesian networks (Friedman et al. 2000, Dejori & Stetter 2003a, Dejori, Schürmann & Stetter 2004), namely (i) it takes into account the systemic nature of many biological processes which arise from the interactions of many genes rather than from actions of an individual gene, and (ii) it accounts for the statistical and noisy nature of the data by adopting a probabilistic approach. A difference is that Bayesian networks allow a causal interpretation whereas decomposable models are restricted to identifying strongly coupling sets of genes. The two main advantages of our approach are that (i) it directly works with the continuous expression data and does not depend on preprocessing by discretization or assumptions like Gaussianity of the data. (ii) The approach is tailored to account for the modular nature of biological molecular life processes which frequently involve the collective action of protein subunits, protein assemblies, and other functional modules.

An apparent restriction of a decomposable graphical model, namely its special structure as a set of linked cliques, turns out to be its strength: Decomposable models are particularly sensitive to the signature of functional modules in the data, because they are designed to explain all the statistics in terms of interacting cliques of genes. In applying the model to ALL data, genes known to encode subunits of known complexes were correctly inked into individual or closely linked adjacent cliques, demonstrating a high sensitivity for detecting functional modules. If subunits were not grouped into a single clique but in adjacent cliques, the link connecting them was very strong. This means that for such adjacent cliques only one or few edges in the graph were missing to render them a single clique which might be a consequence of statistical fluctuations due the limited number and the noisiness of the data set. Based on the clique and the link structure learned from the data, it became possible to formulate two new scores for ranking genes according to their putative importance in cellular processes: the number of links to a gene and the number of cliques it participates in.

One important ingredient of cellular processes is their dynamical nature which is linked with molecular reaction constants (Stetter et al. 2004). Unfortunately, due to the vast complexity of these dynamics and the small amount of data available the dynamics of molecular networks at a large scale have rarely been investigated so far. At present, dynamical considerations can be at most applied to small and experimentally very well-characterized subsystems. Decomposable models might be able to simplify a dynamic analysis by suggesting small (i.e., low-dimensional) tightly coupled functional modules. These modules might be natural breakpoints for the separation of scales, for example by assuming an adiabatic approximation within a functional module and by explicitly modelling only the time constants for interactions between modules.

Decomposable models are not restricted to analyzing the transcriptome of a cell and its changes under various pathological conditions. As cellular life processes are strongly affected and even dominated by a enormous multitude of protein-protein interactions, the technique presented here will be suitable to analyze whole proteome measurements from cells in laboratories of the near future, putting emphasis in the level of interaction between proteins. Of similar importance is the extraction of the modularity of these interactions, where small functional modules will be continuously grouped together to accomplish more and more complex tasks, up to whole cellular genetic programs. In light of this view, genome wide and proteome wide modular analysis and related techniques might form a key ingredient of modern functional genomics and proteomics.

Chapter 9

Conclusions

The perhaps most important signaling network in living cells is constituted by the interactions of proteins and other molecules with the genome – the gene regulatory network of the cell. In various stages of the cell cycle, genetic regulatory mechanisms are of fundamental importance for a controlled action, starting with the initial cell differentiation and ending up with the final programed cell death. From a system level point of view, the various interactions and control loops which form a genetic network, represent the basis upon which the vast complexity and flexibility of life processes emerges. Especially for pharmacology and healthcare industry knowledge about genetic network principles will help opening the gate towards a deeper understanding of morphogenesis and pathogenesis and towards the development of new tissue engineering techniques and drug discovery methods, just to mention a few.

A quantitative understanding of the regulatory genetic network represents therefore one of the major challenges of the post-genomic era not only for biologists but also for computer scientists as their methods and algorithms might contribute to decipher regulatory mechanisms and their related outcomes.With the increased availability of genomic data, mainly caused by high-throughput techniques, an enormous data basis has become available which manifests the need of bioinformatics for analyzing, storing and managing the produced data. Moreover, these data provide an ideal basis for data-driven genetic network modeling approaches to learn genetic network principles from generated data. The learned models might then assist the research of biologists, pharmacologists and physicians to decipher for example disease mechanisms or fundamental regulatory rules.

The way how genetic networks are tried to be modelled with computational approaches ranges from biochemically inspired models based on the reaction kinetics between the different components to more abstracted data driven approaches with the aim to explore a data set and to discover regularities and structures from it. With the invent of high-throughput methods such as gene expression profiling and the resulting massive amount of data, data driven approaches have become of major interest for the genetic network modelling community and their descriptive power has been proven in many studies even though the biochemical part is fully neglected. Classification approaches, e.g. clustering or SVM, are in the mean time standard tools to analyze and mine gene expression data. However, these approaches are limited in their contribution to the genetic network inference problem as they do not provide an answer to the regulatory mechanisms or relationships among genes.

Graphical models partly overcome these problems as they estimate the dependency structure between entities which in case of gene expression data can be interpreted as dependencies among genes and therefore might hint towards gene regulatory relationships. Learning graphical models on the basis of genomic data has been first applied by Friedman and colleagues when learning Bayesian networks from DNA microarray data (Friedman et al. 2000). Especially their probabilistic nature and the intuitive graphical representation makes graphical models an ideal approach towards genetic network modeling from noisy omics data. In many studies, learning graphical models, e.g. Bayesian belief networks, dynamic Bayesian networks, module networks and decomposable models, from microarray data has been discovered as a potentially useful tool for estimating principles of genetic regulatory networks (Friedman et al. 2000, Pe'er et al. 2001, Hartemink et al. 2001, Imoto et al. 2002).

9.1 Contribution and Future Work

The major problem of data driven methods is their need of large data sets to correctly estimate the underlying probability density. Unfortunately, this conflicts with the sparseness of todays genomic data, a fact that challenges the robustness of structure learning from microarray data. We therefore studied structure learning of Bayesian networks under real-life conditions with the goal to estimate parameters which lead to a more accurate interpretation of Bayesian networks learned from microarray data. With various measures we evaluate the robustness and correctness of learned networks, analyze the effect of search strategies and, most importantly, the effect of small sample sizes for structure learning. As a result, we showed that even with a sparse data structure learning provides robust results and we have been able to define parameters which discriminate between false and true estimates. Another way to increase robustness is the use of additional data sources as a prior to guide structure learning. We presented two approaches of constructing a probabilistic prior, namely from motif discovery and motif binding site estimates. Moreover, we addressed the problem of neglecting potentially important genes when selecting genes on the basis of their expression profile only. It is known that transcription factors are weakly expressed even when they play crucial roles in the underlying genetic network. Consequently, our approach used data of putative binding sites to extract previously missed genes, to add them to the initial gene selection and to further use this knowledge to build a probabilistic prior for the learning process.

A major strength of graphical models in general is the intuitive representation of the statistics which underlies the data. Variables are drawn as nodes and edges, directed or undirected, represent dependencies among variables. Both together results in an ideal approach in a complex network structure. Based on this, we presented a method for estimating genes that play a key role in controlling the state of regulatory genetic networks by analyzing the network topology. For this, each gene in the network is equipped with several features, that are all derivable from the network structure. Two well-known features are the connectivity of a gene, given by the number of input and output connections, and the type of connectivity determined by the fraction of fan-in and fan-out. In addition to these two features, we propose a new one that characterizes the importance of a gene for the stability and operational mode of a genetic network. We annotate each gene by its load, that is the number of shortest paths passing through it. In conjunction with a scale-free structure, this property is hypothesized to have a big biological impact since it was shown that such nodes are spots of high vulnerability of a scale-free network and that their failure can cause the network to collapse. By introducing a new topological feature we are able to estimate the effect of genes on the stability of scale-free genetic networks finding those ones that represent the Achilles Heel of a molecular interaction network, for example the putative oncogene PBX1.

In another approach, we bypass structural considerations of the graph structures and do not directly interpret edges of a graph as causal regulatory relationships. Instead, we treat the Bayesian network as a density estimator and as a generative model to produce artificial expression datasets. This data-driven method called generative inverse modeling, simulates the effect of local genetic changes on the global cellular state, as reflected by an altered genome-wide expression profile. For each genetic change we define a pathogenic score by calculating to what extent it transforms the simulated expression patterns into patterns measured for pathologically altered tissues. The method can be used to estimate the relevance of genes for disease-specific genetic mechanisms, e.g. as presented here for pathogenesis. With this approach, the impact of interventions on the global expression behavior is shown in an intuitive way, namely as artificial expression profiles. Thus, one can perform what-if scenarios on genetic regulatory systems *in silico*, rather than in the wet lab.

We finally presented a novel approach towards a systems level analysis of concerted cellular mechanisms, and applied the model to a set of genomewide expression profiles from

ALL patients. The approach is based on a graphical modelling technique called decomposable model which puts particular emphasis on the modular way, in which biomolecules act together to accomplish a certain task, and on the continuous yet noisy nature of the data to be analyzed. A decomposable model tries to explain the statistics in a data set by the action of mutually linked functional modules, so-called cliques.

Computational biology is driven by the need of new algorithms and methods but also by the necessity of new tools and applications. Thus, all algorithms presented in this thesis have been implemented to build the core modules of a microarray analysis platform, called `GeneSim` [1]. Besides standard microarray data analysis tools, such as hierarchical clustering or principal component analysis, `GeneSim` provides the ability to learn graphical models from microarray data, to incorporate prior knowledge from other data sources, to visualize these models graphically and to perform in-silico what-if scenarios. The fact that `GeneSim` is used by several companies to assist their drug-discovery workflow gives evidence for the strength of our methods.

[1] All figures showing microarray data or graphical model structures are generated with `GeneSim`

Appendix A

Bayesian Score

A.1 Bayesian Dirichlet Equivalent Score

To evaluate the goodness of fit of a network G with respect to the data set D, a score $S(G)$ is assigned to the graph G. Using Bayesian statistics $S(G)$ is given by

$$S(G) = \frac{p(D|G)p(G)}{p(D)} \tag{A.1}$$

were $p(D|G)$ is the marginal likelihood, $p(G)$ is the prior probability of structure G and $p(D)$ a normalization constant. Given a uniform structure prior $p(G)$, Equation A.1 reduces to the marginal likelihood

$$S(G) \simeq p(D|G) = \int p(D|\Theta, G, \xi)p(\Theta|G, \xi)d\Theta \tag{A.2}$$

where Θ is the set of parameters and ξ denotes our entire background knowledge. Given that data set D consists of N independent samples, such that $p(D|G)$ is

$$p(D|\Theta, G, \xi) = \prod_{l=1}^{N} p(d^l|\Theta, G, \xi) \tag{A.3}$$

, where d^l represents the lth case in the data set. Hence, Equation A.2 can be written as

$$p(D|G) = \prod_{l=1}^{N} \int p(d^l|\Theta, G, \xi) p(\Theta|, G, \xi) d\Theta \tag{A.4}$$

To solve Equation A.2 in closed the following 5 assumptions have to be made (Cooper & Herskovits 1991):

Assumption 1 Multinomial Distribution A.1 *Let d_i^l and $d_{pa_i}^l$ denote the variable x_i and the parent set pa_i in the lth case of data set D, respectively. Then,*

$$p(d_i^l = k|d_{pa_i}^l = j, \Theta, G, \xi) = \theta_{ijk} \quad \in [0,1] \quad \forall x_i, pa_i. \tag{A.5}$$

Assumption 2 Parameter Independence A.2 *Given network structure G, the parameters associated with each variable are independent from each other such that $p(\Theta|G,\xi)$ decomposes into*

$$p(\Theta|G,\xi) = \prod_{i=1}^{n} p(\Theta_i|G,\xi). \tag{A.6}$$

Due to the local independence of each instance of parents of a variable x_i, $p(\Theta_i|G,\xi)$ decomposes into

$$p(\Theta_i|G,\xi) = \prod_{j=1}^{q_i} p(\Theta_{ij}|G,\xi) \quad \forall i = 1, ..., n, \tag{A.7}$$

where q_i is the number of values the set of parents, pa_i, can assume.

Assumption 3 Parameter Modularity A.3 *Given two network structures G_1 and G_2, if x_i has the same parents in G_1 and G_2, then*

$$p(\Theta_{ij}|G_1,\xi) = p(\Theta_{ij}|G_2,\xi) \quad \forall j = 1, ..., q_i \tag{A.8}$$

Assumption 4 Dirichlet Prior A.4 *Given a network structure G,* $p(\Theta_{ij}|G,\xi)$ *is a priori Dirichlet distributed,* $\theta_{ij} \sim D(N_{ij1},....,N_{ijr_i})$*, byexists exponents* $N^{'}_{ijk}$*, which depend on*

$$p(\Theta_{ij}|G,\xi) = \frac{\Gamma(\sum_{k=1}^{r_i} N^{'}_{ijk})}{\prod_{k=1}^{r_i}\Gamma(N^{'}_{ijk})}\prod_{k}\theta_{ijk}^{N^{'}_{ijk}-1}, \tag{A.9}$$

where $\Gamma(.)$ denotes the Gamma function and r_i is the number of values of variable x_i. The hyperparameters $N^{'}_{ijk}$ are given as

$$N^{'}_{ijk} = N^{'}p(x_i = k, pa_i = j|\xi), \tag{A.10}$$

where $N^{'}$ denotes the equivalent sample size and can be seen as the dimension of the imaginary data set from which the a priori knowledge is extracted.

Assumption 5 Complete Data A.5 *The data set is complete. That is, D contains no missing values.*

From the multinomial sample assumption (Assumption A.1) and the assumption of complete data (Assumption A.2) $p(D|\Theta,G)$ factorizes into

$$p(D|\Theta,G,\xi) = \prod_{l=1}^{N}\prod_{i=1}^{n} p(d_i^l = k|d_{pa_i}^l = j,\Theta,G,\xi) = \prod_{i=1}^{n}\prod_{j,k}\theta_{ijk}^{N_{ijk}}, \tag{A.11}$$

where N_{ijk} is the number of cases in the data set D in which $x_i = k$ and $pa_i = j$.

With Equation A.9 and Equation A.11 the marginal likelihood in Equation A.2 can be re-written as

$$\begin{aligned} p(D|G,\xi) &= \int p(D|\Theta,G,\xi)p(\Theta|G,\xi)d\Theta & \text{(A.12)}\\ &= \int \prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\theta_{ijk}^{N_{ijk}}\frac{\Gamma(\sum_{k=1}^{r_i} N^{'}_{ijk})}{\prod_{k=1}^{r_i}\Gamma(N^{'}_{ijk})}\prod_{k}\theta_{ijk}^{N^{'}_{ijk}-1}d\Theta & \text{(A.13)}\\ &= \prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\sum_{k=1}^{r_i} N^{'}_{ijk})}{\prod_{k=1}^{r_i}\Gamma(N^{'}_{ijk})}\int\prod_{k=1}^{r_i}\theta_{ijk}^{N_{ijk}+N^{'}_{ijk}-1}d\Theta & \text{(A.14)} \end{aligned}$$

Since the Dirichlet distribution is conjugate for this domain the posterior of each parameter remains in the conjugate family and the integral results in

$$\int \prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}+N'_{ijk}-1} d\Theta = \frac{\prod_{k=1}^{r_i} \Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ij} + N_{ij})} \tag{A.15}$$

where $N_{ij} = \sum_k N_{ijk}$ and $N'_{ij} = \sum_k N'_{ijk}$.

Thus, finally the marginal likelihood results in

$$p(D|G,\xi) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})}. \tag{A.16}$$

Appendix B

Data Sets

B.1 Benchmark Data Sets

B.1.1 Alarm Network

This Bayesian network was constructed from expert knowledge as a medical diagnostic alarm message system for patient monitoring and has become the most popular benchmark-network for assessing structural learning algorithms. The domain has 37 discrete variables taking between 2 and 4 values, connected by 46 directed edges, which can be interpreted in a causal manner. The set of conditional probability distributions and the structure, used for generating the data sets are described on the *Netica* homepage (Corp. 2003).

B.1.2 ALL-SIM Network

This Bayesian network was learned from the ALL microarray data set (Yeoh et al. 2002) (cf. Section B.2.1) with a fast simulated annealing scheme. The network consists of 271 discrete variables each of which can have 3 values (-1, 0, +1), connected by 300 directed edges. Data sets of different sample size were generated from the underlying probability distribution (cf. Algorithm 6.1.1).

B.1.3 E2APBX1-SIM Network

This network was learned from the same data set as with a fast simulated annealing scheme. The network consists of 39 discrete variables each of which can have 3 values (-1, 0, +1), connected by 41 directed edges. Data sets of different sample size were generated from the underlying probability distributions (cf. Figure 6.1.1).

B.2 Microarray Data Sets

B.2.1 St. Jude ALL Data

The acute lymphoblastic leukemia (ALL) study provides measurements of 12.000 probes in 327 samples collected from patients with different pediatric ALL subtypes. The goal of this study was to use expression profiling for identifying each of the known prognostically and therapeutically relevant subgroups and for the identification of patients who are at high risk for failing conventional therapeutic approaches.

Out of the 12.000 measured genes, we selected those genes that best define the individual subtypes using the χ^2 statistic according (Yeoh et al. 2002). The final data set (271 genes $\times$ 327 samples) is composed of the 40 most discriminative genes for each of the 7 subtypes, whereby 9 genes appear in more then one cluster but only once in our final data set.

Finally, gene expression levels were discretized to three levels, over-expressed, unchanged and under-expressed, thresholded by the standard deviation of the expression levels across samples for each gene separately, to learn a multinomial Bayesian network. Since this model can describe any discrete conditional distribution, all algorithms also work for higher classes of ordinal data. However, given the low signal to noise ratio of current microarray data with a polynomial scaling of computational expense, finer discretization might result in noise-contaminated data.

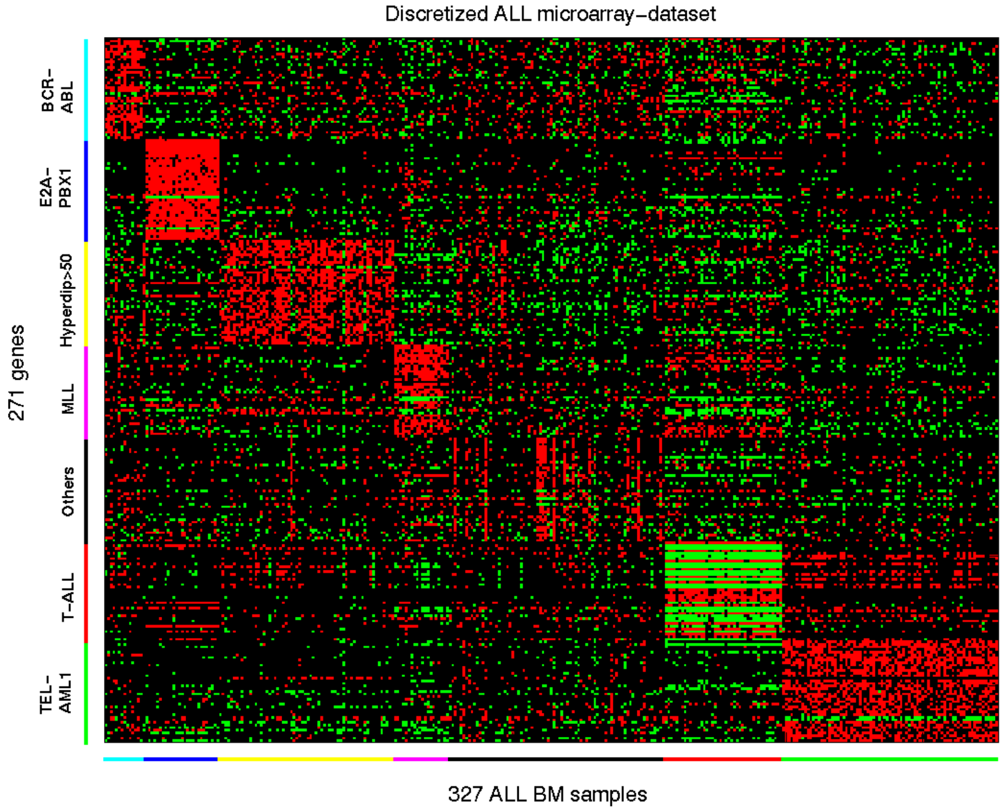


Figure B.1: Acute Lymphoblastic Leukemia is a heterogeneous disease consisting of various subtypes that go back to distinct genetic lesions. The different subtypes which appear in the data set are manifested by distinct gene expression patterns (marked at the left hand side).

Bibliography

Albert, R., Jeong, H. & Barabasi, A.-L. (2000). Error and attack tolerance of complex networks, *Nature* **406**: 378–381.

Allen, T. V. & Greiner, R. (2000). Model selection criteria for learning belief nets: An empirical comparison, *Proceedings of the Seventeenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., pp. 1047–1054.

Angulo, J. & Serra, J. (2003). Automatic analysis of DNA microarray images using mathematical morphology, *Bioinformatics* **19**(5): 553–562.

Aparicio, S., Chapman, J., Stupka, E., Putman, N., Chia, J. & Dehal, P. (2002). Whole-genome shotgun assembly and analysis of the genome of fugu rubipres, *Science* **297**: 1301–1310.

Aspland, S. E., Bendall, H. H. & Murre, C. (2001). The role of E2A-PBX1 in leukemogenesis, *Oncogene* **20**: 5708–5717.

Bailey, T. L. & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, California, pp. 28–36.

Baldi, P. & Brunak, S. (1998). *Bioinformatics, the Machine Learning Approach.*, MIT Press, Cambridge, MA.

Baldi, P. & Hatfield, G. W. (2002). *DNA microarrays and gene expression*, Cambridge university press, Cambridge MA.

Banin, S., Moyal, L., Shieh, S. Y., Taya, Y., Anderson, C. W., Chessa, L., Smorodinsky, N. I., Prives, C., Reiss, Y., Shiloh, Y. & Ziv, Y. (1998). Enhanced phosphorylation of p53 by ATM in response to DNA damage, *Science* **281**(5383): 1674–1677.

Barabasi, A. L. & Bonabeau, E. (2003). Scale-free networks, *Scientific American* **05**: 50–59.

Barthelemy, M. (2004). Betweenness centrality in large complex networks, *Europ. Phys. Jour. B* **38**(163).

Beinlich, I. A., Suermondt, H. J., Chavez, R. M. & Cooper, G. F. (1989). The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks, *Proc. of the Second European Conf. on Artificial Intelligence in Medicine*, Vol. 38, pp. 247–256.

Ben-Dor, A., Shamir, R. & Yakhini, Z. (1999). Clustering gene expression patterns, *J. Comput. Biol.* **6**(3-4): 281–297.

Bilban, M., Buehler, L. K., Head, S., Desoye, G. & Quaranta, V. (2002). Normalizing DNA microarrays, *Curr. Issues Mol. Biol.* **4**: 57–64.

Bishop, C. M. (1995). *Neural Networks for pattern recognition*, Clarendon press Oxford.

Bohr, V. A. (2002). DNA damage and its processing, relation to human disease, *J. Inherit. Metab. Dis* **25**(3): 215–222.

Bozdogan, H. (1987). Model selection and Akaikes's information criterion (AIC): The general theory and its analytical extensions, *Psychometrika* **52**(3): 345–370.

Brown, P. O. & Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays, *Nature Genetics* **21**: 33–37.

Brown, T. A. (1999). *Genomes*, Bios Scientific Publishers, Oxford.

Buntine, W. (1996). A guide to the literature on learning probabilistic networks from data, *Ieee Trans. On Knowledge And Data Engineering* **8**: 195–210.

Chen, T., Le, H. L. & Church, G. M. (1999). Modeling gene expression with differential equations, *Proc. Pacific Symp. Biocomputing* **4**: 29–40.

Chickering, D. M., Geiger, D. & Heckerman, D. (1994). Learning Bayesian Networks is NP-Hard, *Technical Report MSR-TR-94-17*, Microsoft Research.

Collina, F. S., Guyer, M. S. & Chakravarti, A. (1997). Variations on a theme: cataloging human DNA sequence variation, *Science* **278**: 1580–1581.

Cooper, G. & Herskovits, E. (1991). A Bayesian method for constructing Bayesian belief networks from databases, *Uncertainty in Artificial intelligence*, pp. 86–94.

Corp., N. S. (2003). Alarm network, http://www.norsys.com/netlib/.

Cortez, D., Wang, Y., Qin, J. & Elledge, S. J. (1999). Requirement of ATM-dependent phosphorylation of brca1 in the DNA damage response to double-strand breaks, *Science* **286**(5442): 1162–1166.

de Jong, H. (2002). Modeling and simulation of genetic regulatory systems: A literature review, *Journal of Computational Biology* **9**(1): 67–103.

de Jong, H., Page, M., Hernandez, C. & Geiselmann, J. (2001). Qualitative simulation of genetic regulatory networks: method and application, *in* B. Nebel (ed.), *Proceedings of the 17. International Joint Conference on Artificial Intelligence IJCAI-01*, Morgan Kauffman, San Mateo CA, pp. 67–73.

Dechter, R. (1996). Bucket elimination: A unifying framework for probabilistic inference, *Uncertainty in Artificial intelligence*, pp. 211–219.

Dejori, M., Nägele, A. & Stetter, M. (2004). Using TFBS information to bias genetic network estimation via Bayesian network learning, *Proceedings of the 12th Conference on Intelligent Systems for Molecular Biology, Poster Abstracts*.

Dejori, M., Schürmann, B. & Stetter, M. (2004). Hunting drug targets by systems-level modeling of gene expression profiles, *IEEE Transaction NanoBioscience* **3**(3): 180–191.

Dejori, M., Schwaighofer, A., Tresp, V. & Stetter, M. (2004). Mining functional modules in genetic networks with decomposable models, *OMICS: A Journal of Integrative Biology* **8**(2): 176–188.

Dejori, M. & Stetter, M. (2003a). Bayesian inference of genetic networks from gene-expression data: convergence and reliability, *Proceedings of the 2003 International Conference on Artificial Intelligence (IC-AI'03)*, pp. 323–327.

Dejori, M. & Stetter, M. (2003b). Estimation of oncogenes by Bayesian inverse modeling of gene-expression patterns, *Proceedings of the 11th Conference on Intelligent Systems for Molecular Biology, Poster Abstracts*.

Dejori, M. & Stetter, M. (2003c). Using scale-free topology to estimate critical genes from regulatory networks, *Proceedings of the 11th Conference on Intelligent Systems for Molecular Biology, Poster Abstracts*.

Dejori, M. & Stetter, M. (2004). Identifying interventional and pathogenic mechanisms by generative inverse modeling of gene expression profiles, *Journal of Computational Biology* **11**(6): 1135–1148.

D'haeseleer, P., Liang, S. & Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering, *Bioinformatics* **16**: 707–726.

D'haeseleer, P., Wen, X., Fuhrman, S. & Somogyi, R. (1997). Mining the gene expression matrix: inferring gene relationships from large-scale expression data, *in* M. Holcombe & R. Paton (eds), *Information processing in cells and tissues*, Plenum Press, pp. 203–212.

Dulbecco, R. (1986). A turning point in cancer research: Sequencing the human genome, *Science* **231**: 1055–1056.

Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstrap*, Chapman and Hall, New York.

Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA* **95**(25): 14863–14868.

Fainstein, E., Marcelle, C., Rosner, A., Canaani, E., Gale, R. P., Dreazen, O., Smith, S. D. & Croce, C. M. (1987). A new fused transcript in Philadelphia chromosome positive acute lymphoblastic leukemia, *Nature* **330**: 386–388.

Fields, S. & Song, O. (1989). A novel genetic system to detect protein-protein interactions, *Nature* **340**: 245–246.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E. & Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans, *Nature* **391**(6669): 744–745.

Fodor, S. P., Rava, R. P., Huang, X. C., Pease, A. C., Holmes, C. P. & Adams, C. L. (1993). Multiplexed biochemical assays with biological chips, *Nature* **364**: 555–556.

Friedman, N., Linial, M., Nachman, I. & Pe'er, D. (2000). Using Bayesian network to analyze expression data., *J. Comput. Biology* **7**: 601–620.

Galinier, P., Habib, M. & Paul, C. (1995). Chordal graphs and their clique graph, *in* M. Nagl (ed.), *Graph-Theoretic Concepts in Computer Science, WG'95*, Vol. 1017 of *Lecture Notes in Computer Science*, Springer Verlag, pp. 358–371.

Galperin, M. Y. (2004). The Molecular Biology Database Collection: 2004 update, *Nucl. Acids Res.* **32**(90001): D3–22.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions, *Journal of Physical Chemistry* **81**(25): 2340–2361.

Golub, T., Barker, G., Bohlander, S., Hiebert, S., Warda, D., Bray-Ward, P., Morgan, E., Raimondi, S., Rowley, J. & Gilliland, D. (1995). Fusion of the TEL gene on 12p13 to the AML1 gene on 21q22 in Acute Lymphoblastic Leukemia, *PNAS* **92**(11): 4917–4921.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. & Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* **286**(5439): 531–537.

Hanahan, D. & Weinberg, R. A. (2000). The hallmarks of cancer, *Cell* **100**: 57–70.

Hanash, S. (2003). Disease proteomics, *Nature* **422**(6928): 226–232.

Hanash, S. M., Baier, L. J., McCurry, L. & Schwartz, S. A. (1986). Lineage-related polypeptide markers in acute lymphoblastic leukemia detected by two-dimensional gel electrophoresis, *PNAS* **83**(3): 807–811.

Hanash, S. M., Strahler, J., Richardson, B., Reaman, G., Stoolman, L., Hanson, C., Nichols, D. & Tueche, H. J. (1989). Identification of a cellular polypeptide that distinguishes between acute lymphoblastic leukemia in infants and in older children, *Blood* **73**(2).

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S. & Young, R. A. (2001). Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks, *Proceedings of the 6th Pacific Symposium on Biocomputing*, pp. 422–433.

Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. (1999). From molecular to modular cell biology, *Nature* **402**: C47.

Heckerman, D. & Chickering, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data, *Machine Learning* **20**: 197–243.

Hedges, S. B. & Kumar, S. (2003). Genomic clocks and evolutionary timescales, *Trends Genet* **19**: 200–206.

Hegde, P., Qi, R., Abernathy, K., Gay, C. & et al., S. D. (2000). A concise guide to cDNA microarray analysis, *Biotechniques* **29**(3): 548–550,552–554,556.

Heid, C. A., Stevens, J., Livak, K. J. & Williams, P. M. (1996). Real time quantitative PCR, *Genome Research* **6**(10): 986–994.

Ibarra, L. (2000). Fully dynamic algorithms for chordal graphs and split graphs, *Technical Report DCS-262-IR*, IVictoriaCS.

Imoto, S., Goto, T. & Miyano, S. (2002). Estimation of genetic networks and functional structures between genes by using Bayesian network and non-parametric regression, *Pacific Symposium on Biocomputing*, pp. 175–186.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome, *Nature* **409**: 860–921.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome, *Nature* **431**: 931 – 945.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome, *PNAS* **98**(8): 4569–4574.

Jeong, H., Mason, S. P., Barabasi, A. L. & Oltvai, Z. N. (2001). Lethality and centrality in protein networks, *Nature* **411**: 41–42.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. & Barabasi, A. (2000). The large-scale organization of metabolic networks, *Nature* **407**: 651 – 654.

Jordan, M. I. (ed.) (1998). *Learning in Graphical Models*, MIT Press.

Kato, M., Tsunoda, T. & Takagi, T. (2000). Inferring genetic networks from DNA microarray data by multiple regression analysis, *Genome informatics* **11**: 118–128.

Kauffman, S. A. (1969). Homeostasis and differentiation in random genetic control networks, *Nature* **224**: 177–178.

Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983). Optimization by simulated annealing, *Science* **220, 4598**: 671–680.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency, *Annals of Mathematical Statistics* **22**: 79–86.

Laarhoven, P. J. M. V. & Aarts, E. H. L. (1987). Simulated annealing: theory and applications.

Latchman, D. S. (1998). *Eukaryotic transcription factors. 3rd edition*, Academic Press, San Diego, CA.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Lui, J. S., Neuwald, A. F. & Wootton, J. C. (1993). Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment, *Science* **262**(5131): 208–214.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J., Volkert, T. L., Fraenkel, E., Gifford, D. K. & Young, R. A. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae, *Science* **298**(5594): 799–804.

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D. & Darnell, J. (2000). *Molecular Cell Biology*, W. H. Freeman and Company.

McAdams, H. H. & Arkin, A. (1997). Stochastic mechanisms in gene expression, *Proc. Natl. Acad. Sci. USA* **94**: 814–819.

McCarthy, J. & Hilfiker, R. (2000). The use of single-nucleotide polymorphism maps in pharmacogenomics, *Nature Biotechnology* **18**: 505–508.

Metropolis, N., Rosenbluth, A., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1958). Equations of state calculations by fast computing machines, *Journal of Chemical Physics* **21**: 1087–1092.

Motter, A. E., Nishikawa, T. & Lai, Y. C. (2002). Range-based attacks on links in scale-free networks: are long-range links responsible for the small-world phenomenon?, *Phys. Rev. E* **66**: 065103.

Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome, *Nature* **420**: 520–562.

Murphy, K. & Mian, S. (1999). Modelling gene expression data using dynamic Bayesian networks, *Technical report*, Computer Science Division, University of California, Berkeley, CA.

Nobrega, M. A. & Pennacchio, L. A. (2003). Comparative genomic analysis as a tool for biological discovery, *J. Physiol* **554**: 31–39.

Ogata, H., Goto, S., Fujibuchi, W., Bono, H. & Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Res.* **27**(1): 29–34.

Pearl, J. (1998). *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Francisco, CA.

Pe'er, D., Regev, A., Elidan, G. & Friedman, N. (2001). Inferring subnetworks from perturbed expression profiles, *Bioinformatics* **338**(1): 1–9.

Pennacchio, L. A. & Rubin, E. M. (2003). Comparative genomic tools and databases: providing insights into the human genome, *J. Clin. Invest.* **111**: 1099–1106.

Phizicky, E., Bastiaens, P. I., Zhu, H., Snyder, M. & Fields, S. (2003). Protein analysis on a proteomic scale, *Nature* **422**(6928): 208–215.

Pincus, M. (1970). A monte carlo method for the approximate solution of certain types of constrained optimization problems, *Oper. Res.* **18**: 1225–1228.

Quandt, K., Fech, K., Karas, H., Wingender, E. & Werner, T. (1995). MatInd and MatInspector: new fast and versatile matches in nucleotide sequence data, *Nucleic Acids Research* **23**(23): 4878–4884.

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., de Rijn, M. V., Waltham, M., Pergamenschikov, A., Lee, J. C. & Lashkari, D. (2000). Systematic variation in gene expression patterns in human cancer cell lines, *Nature Genetics* **24**(3): 208–209.

Sakamoto, E. & Iba, H. (2001). Inferring a system of differential equations for a gene regulatory network by using genetic programming, *Congress on evolutionary computation* pp. 720–726.

Schena, M. (2000). *Microarray biochip technology*, Eaton Publishing Co., Natick.

Schena, M., Shalon, D., Davis, R. & Brown, P. O. (1995). Quantitative monitoring of gene-expression patterns with a cDNA microarray, *Science* **270**: 467–470.

Scholz, J., Dejori, M., Stetter, M. & Greiner, M. (2005). Noisy scale-free networks, *Physica A* **350**(2005): 622–642.

Schwaighofer, A., Tresp, V., Dejori, M. & Stetter, M. (2004). Structure learning for non-paramteric decomposable models, *Journal of Machine Learning Research* p. submitted.

Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics* **7**(2): 461–464.

Segal, E., Shapira, M., Regev, A., Pe’er, D., Botstein, D., Koller, D. & Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene-expression data, *Nat. Genet.* **34**(2): 166–176.

Shachter, R. (1998). Bayes-ball: The rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams), *Proceedings of the 14th Conference on Uncertainty and Artificial Intelligence (UAI)*, Vol. 14, pp. 480–487.

Shannon, C. E. (1948). A mathematical theory of communication, *Bell System Technical Journal* **27**: 379–423, 623–656.

Shen-Orr, S., Milo, R., Mangan, S. & Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli, *Nature Genetics* **31**(1): 64–68.

Slonim, D. K. (2002). From patterns to pathways: gene expression data analysis comes of age, *Nature Genetics* **32 Suppl**: 502–508.

Somogyi, R., Fuhrman, S., Askenazi, M. & Wuensche, A. (1997). The gene expression matrix: towards the extraction of genetic network architectures, *Nonlinear Analysis, Proc. of the Second World Congress of Nonlinear Analysis (WCNA96)*, Vol. 30, pp. 1815–1824.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, *Molecular Biology of the Cell* **9**: 3273–3297.

Steck, H. (2001). *Constraint-based structural learning in Bayesian networks using finite data sets*, PhD thesis, Doctoral thesis, Munich, Germany.

Stetter, M., Deco, G. & Dejori, M. (2003). Large-scale computational modeling of genetic regulatory networks, *Artificial Intelligence Review* **20**: 75–93.

Stetter, M., Schürmann, B. & Dejori, M. (2004). Systems level modeling of gene regulatory networks, *in* W. Dubitzky & F. Azuaje (eds), *Artificial Intelligence Methods and Tools for Systems Biology*, Vol. 5 of *Computational Biology*, Springer Verlag, p. 221.

Stormo, G. D. (2000). DNA binding sites: representation and discovery, *Bioinformatics* **16**(1): 16–23.

Strehler, B. L. (1995). Deletional mutations are the basic cause of ageing: historical perspectives, *Mutat. Res* **338**(1-6): 3–17.

Strittmatter, W. J. & Roses, A. D. (1995). Apolipoprotein E and Alzheimer disease, *Proc. Natl. Acad. Sci. USA* **92**: 4725–4727.

Strittmatter, W. J., Saunders, A. M., Schmechel, D., Pericak-Vance, M., Enghild, J., Salvesen, G. S. & Roses, A. D. (1993). Apolipoprotein E: High-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease, *PNAS* **90**(5): 1977–1981.

Symonds, H., Krall, L., Remington, L., Saenz-Robles, M., Lowe, S., Jacks, T. & Dyke, T. V. (1994). p53-dependent apoptosis suppresses tumor growth and progression in vivo, *Cell* **78**(4): 703–711.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A. & Judson, R. S. (2000). A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae, *Nature* **403**(6770): 623–627.

van Duk, M. A., Voorhoeve, P. M. & Murre, C. (1993). PBX1 is converted into a transcriptional activator upon acquiring the N-terminal region of E2A in pre-b-cell acute lymphoblastic leukemia, *Proc. Natl. Acad. Sci. USA* **90**: 6061–6065.

Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. (1995). Serial analysis of gene expression, *Science* **270**(5235): 484–487.

Venter, J. C., Adams, M. D. & Myers, E. W. (2001). The Sequence of the Human Genome, *Science* **291**(5507): 1304–1351.

Verma, T. & Pearl, J. (1990). Equivalence and synthesis of causal models, *Proceedings of the 6th Conference on Uncertainty and Artificial Intelligence (UAI)*, Vol. 6, pp. 220–227.

Vijg, J. & Dolle, M. E. (2002). Large genome rearrangements as primary cause of ageing, *Mech Ageing Dev* **123**(8): 907–915.

Wai, L. & Fahiem, B. (1994). Learning Bayesian belief networks: An approach based on the MDL principle, *Computational Intelligence* **10**(3): 269–293.

Wang, D. G. & et al., C. J. S. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome, *Science* **280**: 1077–1082.

Wang, T. J. L., Shapiro, B. A. & Shasha, D. (1999). *Pattern Discovery in Biomolecular Data: Tools, Techniques and Applications.*, Oxford University Press, Oxford.

Watson, J. D. & Crick, F. H. C. (1953). Molecular structure of Nucleic Acids, *Nature* **171**: 737–738.

Whittaker, J. (1990). *Graphical models in applied multivariate statistic*, John Wiley and Sons, New York.

Widschwendter, M. & Jones, A. (2002). DNA methylation and breast carcinogenesis, *Oncogene* **21**(35): 5462–5482.

Yates, J. R. (1998). Mass spectrometry and the age of the proteome, *J. Mass Specrom.* **33**: 1–19.

Yeoh, E. J., Ross, M. E., Shurtleff, S. A., Williams, W. K. & et al., D. P. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling, *Cancer cell* **1**: 133–143.

Yuh, C. H., Bolouri, H. & Davidson, E. H. (1998). Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene, *Science* **279**: 1896–1902.

Yuh, C. H., Bolouri, H. & Davidson, E. H. (2001). *Cis*-regulatory logic in the *endo* 16 gene: switching from a specification to a differentiation mode of control, *Development* **128**: 617–629.

Zoubir, A. M. (1993). Bootstrap: Theory and applications, *in* F. T. Luk (ed.), *Advanced signal processing algorithms, architectures and algorithms*, Vol. 2027 of *Proceedings of SPIE*, SPIE.