Bayesian network models for inferring cancer pathogenetic and gene regulatory pathways

Zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften
von der Fakultät für Informatik
der Universität Karlsruhe

genehmigte Dissertation
von
Svetlana Bulashevska
aus Karlsruhe

Tag der mündlichen Prüfung: 14 Juli 2004
Erster Gutachter: Prof. Dr. Roland Vollmar
Zweiter Gutachter: Dr. Roland Eils

Contents

Abbreviations

BUGS - Bayesian Updating with Gibbs Sampling
CAP - catabolite activator protein
CDK - cyclin-dependent kinase
cDNA - complementary DNA
CGH - comparative genomic hybridization
CPT - conditional probability table
DAG - directed acyclic graph
DKFZ - Deutsches Krebsforschungszentrum
DNA - deoxyribonucleic acid
FISH - fluorescence in situ hybridization
GIST - gastrointestinal stromal tumor
GVS - Gibbs Variable Selection
ISCN - International System for Human Cytogenetic Nomenclature
LOH - loss of heterozygosity
Mb - million base pairs
MCMC - Markov Chain Monte Carlo
MFISH - multicolor fluorescence in situ hybridization
mRNA - messenger RNA
NIPT - number of imbalances per tumor
PCA - principal component analysis
PCR - polymerase chain reaction
PDAG - partially directed acyclic graph
REVEAL - reverse engineering algorithm
RFLP - Restriction Fragment Length Polymorphism
RNA - ribonucleic acid
SSVS - Stochastic Search Variable Selection
UC - urothelial carcinoma

List of Figures

List of Tables

Abstract

The present thesis is a result of an interdisciplinary work conducted in the German Cancer Research Center (DKFZ). The major goal of this research institution is the development of molecular genetics methods allowing to gain insight into mechanisms underlying the tumor pathogenesis. Functional genome research and understanding the process of genetic regulation play an important role in this goal. The experimental molecular genetics techniques produce a huge amount of data which must be analysed by computational methods. Novel model-based approaches are required capable to capture biological processes, to extract new patterns between biological entities and to provide new hypotheses and predictions.

The biological processes are stochastic in their nature, and the experimental measurements are noisy. Hence, modelling approaches and learning models from data must be based on statistics.

The present thesis focuses on probabilistic graphical models. These models represent probabilistic dependencies between variables. Learning the structure and parameters of the models from data is facilitated by the Bayesian methodology which is a modern Bayesian statistics approach.

The first part of this thesis concerns the analysis of data about chromosomal abnormalities in tumor cells (in particular, allelic losses). The challenge was to reconstruct from the data the possible flow of progression of genetic abnormalities during the development of tumor. I employed the probabilistic graphical model Bayesian network and used Bayesian network learning. This approach allowed to discover patterns of allelic losses in urothelial cancer and to suggest primary and secondary genetic events in the tumor pathogenesis.

The second part of this thesis deals with the gene expression data obtained with microarray experiments. The challenge was to infer the gene regulatory interactions from the data that enable to get insight into the mechanisms of genetic regulation. I proposed a model for the gene regulatory interactions which is a probabilistic graphical model, hence being able to confront noisy biological process and data. I have developed an approach for learning the model from data based on Bayesian approach. The method utilizes Markov Chain Monte Carlo simulation techniques, in particular Gibbs sampling. I tested the method with previously published data of the *S.cerevisiae* cell cy-

cle and inferred relations between genes consistent with biological knowledge. Both methods presented in this thesis contribute to further development of the field bioinformatics.

Chapter 1

Introduction

The present thesis is a result of an interdisciplinary work conducted in the German Cancer Research Center in the field of bioinformatics.

Bioinformatics, originally used for the analysis and storage of the genome sequence data, after the completion of the Human Genome Project, has shifted its focus on the data of another quality. The recent advent of new efficient molecular genetics technologies allow to gain insight into the function of the genome. These techniques produce a huge amount of experimental data, which must be analysed by new computational methods. Hence, bioinformatics experiences a shift from the "genomic era", where the emphasis was on database construction and the analysis of DNA sequence data, to the "post-genomic era", where the focus is on knowledge discovery or data mining.

The recently developed high-throughput microarray technology (cDNA chip) allows to measure expression levels of thousands of genes simultaneously, as they change over time and react to external stimuli. A great challenge for bioinformatics is to develop computational methods for inferring gene regulatory pathways from gene expression data and to reconstruct the genetic regulatory network. The microarray technology makes it possible to profile the gene expression in tumor genomes and to gain insight into the molecular mechanisms underlying such complex diseases like cancer.

The recent advancements of cytogenetics has made it possible to screen the whole genome for chromosomal abnormalities by means of one experiment. These methods include comparative genomic hybridization (CGH), array-based comparative genomic hybridization (matrix-CGH), methods for detection of allelic instabilities like loss of heterozygosity (LOH), and various in situ hybridization techniques, such as fluorescence in situ hybridization (FISH) and multicolor fluorescence in situ hybridization (MFISH). It was shown that chromosomal abnormalities are related to the initiation and progression of tumor. The cytogenetic methods provide the researchers with experimental data to infer hypotheses on the tumor pathogenetic pathways.

Novel exploratory data analysis methods (also called *data mining* methods) are required to meet the new biological data. The techniques being in use include unsupervised methods like clustering, supervised classification methods, techniques for modelling and simulation. The data analysis attempts to extract new patterns, new relations between the biological entities, to construct models capturing the biological processes and thus capable to provide useful biological insights and predictions.

The characteristical feature of the bioinformatics problem space is that the biological processes are stochastic, and experimental measurements are noisy. The modelling systems must be robust against noise and possess high inferential power. For these reasons bioinformatics approaches have to speak the language of probability theory and statistics.

The models, required in bioinformatics, are often much more complex than "classical" statistical models. They have a great number of parameters which often cannot be derived with classical statistical methods like maximum-likelihood estimation.

The present thesis focuses on probabilistic graphical models that represent probabilistic dependencies (independencies) between variables. The models have a graphical component, which is an important property for knowledge representation, especially in an interdisciplinary field like bioinformatics. Learning such models from data enables to uncover multivariate probabilistic dependencies between variables. For learning the structure and parameters of the model from data, I employed Bayesian learning, which is a modern Bayesian statistics approach. Bayesian approach treats the uncertainty on model structure and parameters in a unified fashion, defining the priors on these quantities, and performs the probabilistic inference based on Bayes' theorem. Bayesian approach allows for flexibility by dealing with complex models with many parameters due to the possibility of hierarchical formulation of the model: the prior on model parameters can be defined with the help of further parameters. Unlike classical model estimation methods such as maximum likelihood, Bayesian methods are free from the assumptions of asymptotic normality, and therefore are more appropriate for learning from sparse datasets, as the biological datasets are.

In Chapter 2 I employ the probabilistic graphical model Bayesian Network. Bayesian Network represents the multivariate probabilistic dependencies between variables. I provide an introduction into the Bayesian network model and principles of probabilistic reasoning in Section 2.2. Bayesian network model can be learned from empirical data as described in Section 2.4. I introduce my approach for reconstructing the possible flow of progression of genetic abnormalities with Bayesian network in Section 2.5. I applied this approach to the allelotyping data (LOH). This allowed to discover patterns of allelic losses in urothelial cancer and to suggest primary and secondary genetic events in tumor pathogenesis (Section 2.6).

In Chapter 3 I deal with inferring genetic regulatory interactions from microarray data. I propose a model for the genetic regulatory interactions which is also a probabilistic graphical model, hence being able to confront noisy biological process and data. In Section 3.5 I introduce the model which originates from the field of probabilistic graphical models. I have developed an approach for learning the structure and parameters of the model from gene expression data. I employed the methodology of Bayesian learning (Bayesian model selection). The general introduction into this approach is given in Section 2.3. In Section 3.6 I present the application of this approach for learning my model. I tested my approach on a previously published dataset of budding yeast cell cycle and present the results in Section 3.7.

In Chapter 4 I present an outlook of my work for biological and bioinformatics research.

Chapter 2

Reconstructing cancer pathogenetic pathways with Bayesian networks

2.1 Biological motivation

During the last few decades of cancer research it was shown that initiation and progression of tumor is related to *chromosomal abnormalities.* A number of recurrent *translocations* were found to be characteristically associated with different forms of leukemia, lymphoma and soft tissue tumors. It was shown that *deletions* (losses of chromosomal material) can lead to inactivation or loss of the functionality of the so called "tumor supressor genes", which are involved in the maintenance of normal cell growth and differentiation. *Amplifications* (gains, i.e. multiple copies of the same chromosomal region) can lead to the abnormal activation or overexpression of "oncogenes" responsible for the abnormal cell growth. The genes responsible for regulation in the cell (transcription factors, growth factors, receptors of growth factors, etc.) become deregulated in consequence of abnormality and contribute to the pathological phenotype. Hence, specific chromosomal aberrations have been shown to be significant markers of tumor progression.

Solid tumors appear to have a much larger and heterogeneous set of chromosomal aberrations. Genetic changes accumulate during tumor progression. Many of the genetic changes might be random due to the general genomic instability in tumor, but there must be certain significant genetic events and certain significant dependencies among the events relevant to the formation of tumor. The interplay of genetic changes disrupt the normal cell cycle, promoting other genetic changes, - the cell "goes out of control". While relating the accumulation of genetic alterations to tumor progression, it is important to indicate which events tend to occur early in tumor progression and which events tend to occur together, possibly identifying a tumor subclass.

One of the first attempts to describe the tumor pathogenetic process as the accumulation of multiple genetic alterations was the work of Fearon and Vogelstein (1990), who had shown that at least four different genetic changes are needed to transform a normal cell into a malignant cell corresponding to different stages in progression of colorectal cancer. This model represented the genetic changes as a single *path* going from normal cell to advanced tumor.

Many types of cancer are too heterogeneous in their causes to be described with such simple path models.

The advances of molecular genetics enable to screen the whole genome for chromosomal abnormalities. The methods like LOH (loss of heterozygosity), CGH (comparative genomic hybridization), matrix-CGH provide researchers with experimental data to infer hypothesis about tumor pathogenesis. The challenge of this work was to introduce a computational method enabling to reconstruct the possible flow of progression of genetic abnormalities (cancer pathogenetic pathways) from the single "snapshot" of abnormalities provided by the cytogenetic experiments.

The first attempt to employ mathematical models to infer the order of genetic changes from cytogenetic data was made by R. Desper (see Desper *et al.* 1999, 2000). They described the progression of alterations as a tree with a root representing the normal cell and used two different tree models. In a distance-based tree leaf nodes represent genetic aberrations. The distance function between the nodes in the tree was defined based on probabilities of the co-occurrence of genetic aberrations. A distance-based phylogenetic tree-building algorithm was used to infer the tree structure that fits the pairwise distances at best. Alternatively, a maximum weight branching algorithm was used to reconstruct a tree, in which both internal nodes and leaf nodes correspond to aberrations. A major limitation of these models is that they describe the progression of genetic events as trees, whereas the biological intuition suggests that this is a more complex process and does not necessarily evolve as a tree structure. Pathways of genetic events might occur in parallel and converge in one or more common events. Tree models postulate that there is only one path from the root of the tree passing through a particular genetic event. Accordingly, a tree model would not detect alternative pathways of genetic events that converge in a certain aberration pattern characterizing particular tumor subtypes.

In the present work I employ a more general mathematical model, i.e. a Bayesian network model. Bayesian network model can be learned from data due to its statistical foundations. This enables to uncover the probabilistic dependencies between genetic abnormalities. In the following I will give an introduction into the Bayesian networks formalism and Bayesian networks learning. Then I will present my approach for reconstructing the flow of progression of genetic abnormalities with the help of Bayesian network model.

2.2 Bayesian Networks. Foundations

Bayesian networks (Cowell 1999, Heckerman 1998, Jensen 1996, Pearl 1998) are *probabilistic models* that use probability as a measure of *uncertainty.* In their original concept Bayesian networks were designed to represent the qualitative knowledge in expert systems and to provide the mechanism for reasoning under uncertainty. There are also other techniques to represent uncertainty like e.g. *certainty factors* used in the medical expert system MYCIN (Shortliffe 1976), *fuzzy logic* (Zadeh 1983) and *belief functions* (Dempster 1967, Shafer 1976). However, probability theory has a long history in representing uncertainty and provide a good theoretical basis for the uncertain inference. Heckerman (1986) established a connection of certainty factors to the probability theory by redefining the interpretation of certainty factors as monotone functions of likelihood ratios.

In contrast to the *rule-based* expert systems, the Bayesian networks describe the relationships among objects (variables) with the help of a *joint probability distribution.*

Let $\{X_1, \ldots, X_n\}$ be a set of random variables and $\{x_1, \ldots, x_n\}$ be a set of their possible instantiations. The joint probability distribution of the variables in X is

$$P(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n).$$

Unfortunately, the direct specification of the joint probability distribution involves a huge number of parameters. The joint probability distribution over n binary variables has 2^n parameters (the probabilities $P(x_1, \ldots, x_n)$ for every possible realization $x_1, \ldots, x_n$ of the variables). This was one of the early criticisms of using probability in expert systems. In most domains, however, many subsets of variables can be independent or conditionally independent. Simplification of the joint probability distribution can be obtained by exploiting the independence structure among the variables.

Let X and Y be two random variables with joint probability distribution P(X,Y). We say that variables X and Y are *independent* if for all possible values x, y of X and Y

$$P(x, y) = P(x)P(y).$$

Otherwise, the variables are *dependent.* Since

$$P(x|y) = \frac{P(x,y)}{P(y)}, \qquad P(y|x) = \frac{P(x,y)}{P(x)},$$

the condition for the independence between two variables can be equivalently formulated as: $P(x|y) = P(x)$ or $P(y|x) = P(y)$. When X and Y are independent, learning the value of Y gives us no information about X, and vice

versa. Note that the relation is symmetric. We can represent the dependencies between variables using a graph, in which each variable is denoted by a node. When two variables are dependent, there is an undirected edge between them. Missing an edge between two variables in a directed graph implies the independence of these variables.
Consider a relation involving three random variables. We say that X and Y are *conditionally independent given Z*, when for all possible values x, y, z of X, Y and Z

$$P(x|y,z) = P(x|z). \qquad (*)$$

We denote this as $I(X;Y|Z)$ and refer to as *conditional independence statement.* The definition of conditional independence conveys the idea that once Z is known, knowing Y can no longer influence the probability of X. The dependency between X and Y is *mediated* through Z.
An alternative but equivalent definition of conditional independence is given by:

$$P(x,y|z) = P(x|z)P(y|z).$$

Proof:

$$P(x,y|z) = \frac{P(x,y,z)}{P(z)} = \frac{P(x|y,z)P(y,z)}{P(z)}.$$

Since

$$P(y,z) = P(y|z)P(z)$$

and using $(*)$:

$$P(x,y|z) = P(x|z)P(y|z).$$

The conditional independence between variables X and Y given Z can be represented with the help of a directed graph as in Figure 2.1 by missing an edge between X and Y, although there is a path between X and Y going through Z. The conditional (in)dependency cannot be represented in an undirected graph; two independent variables will be connected if there exists some other variable that depends on both. The undirected graphs are not able to represent nontransitive dependencies. The directed graph in Figure 2.2 represents that variables X and Y are independent (they are not connected with an edge), but not conditionally independent.

The definition of conditional independence can be generalized to sets of variables. Denoting by boldface capital letters the sets of variables **X**, **Y** and **Z**, I(**X**;**Y**|**Z**) means that the set **X** is independent on **Y** conditioned on **Z**. Two nodes in the graph have no common edge if and only if they are conditionally independent given the set of all other variables in the graph.
Probabilistic graphical models based on undirected graphs are called *Markov networks* (Lauritzen 1996).
Bayesian networks are probabilistic models based on directed graphs which represent *conditional independencies* between variables. A Bayesian Network

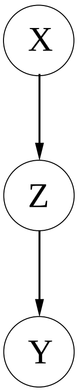

Figure 2.1: A directed graph representing the conditional independence of the variables X and Y given Z.

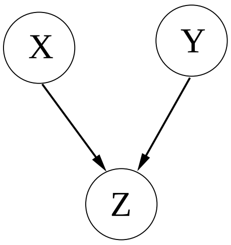

Figure 2.2: A directed graph representing that variables X and Y are independent but not conditionally independent given Z.

model consists of two components. The first is a directed acyclic graph G, whose vertices correspond to variables $X_1, \ldots, X_n$. The graph G encodes conditional independencies between the variables: each variable X_i is independent of its non-descendants, given its parents in G. Bayesian Network is a representation of a joint probability distribution over a set of random variables $X_1, \ldots, X_n$. Due to conditional independencies, the joint probability distribution of the variables can be decomposed in the product form:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Pa_i),$$

where Pa_i is the set of parents of X_i in G. Conditional probabilities appearing in the product form describe the conditional probability distribution for each variable given its parents in the graph G. This is the second component of the Bayesian network. The conditional probability distributions are normally stored in *conditional probability tables* (CPTs). Figure 2.3 shows a simple example of a Bayesian network and the set of independencies it encodes. Table 2.1 presents the conditional probability table for the variable C.

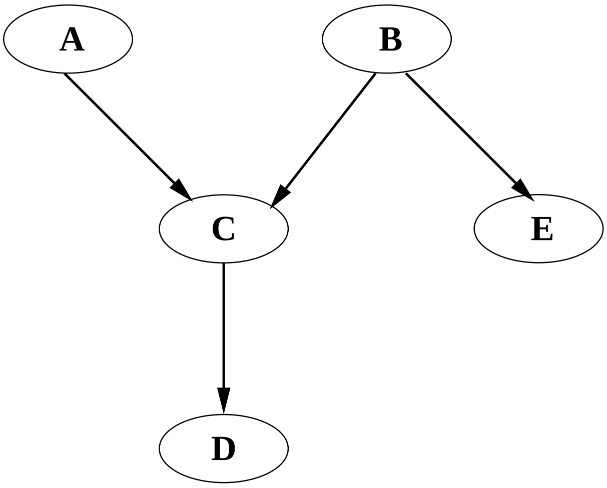

Figure 2.3: An example of a simple Bayesian network. The network structure implies that the joint probability distribution has the product form P(A, B, C, D, E) = P(A)P(B)P(C | A, B)P(D | C)P(E | B). The network structure also implies the following conditional independence statements: I(A; B,E), I(B;A), I(C;E | B,A), I(D;A,B,E | C), I(E;A,C,D | B).

An interesting feature of the variables in Bayesian networks is their Markov relations, i.e. whether the variable Y is in the *Markov blanket* of X. The *Markov blanket* of a node is the set of its parents, children and parents of

		C	
A	B	0	1
0	0	0.8	0.2
0	1	0.6	0.4
1	0	0.7	0.3
1	1	0.1	0.9

Table 2.1: The conditional probability table defines the conditional probability of the variable C given the combinations of its parents' values.

the children. It is the minimal set of variables that ”shields” X from the rest of the variables in the network. Given its Markov blanket, X is independent from the remaining variables. Nodes from one Markov blanket probably indicate some common process modelled by the variables.

Bayesian networks provide the mechanism for *probabilistic inference* (*probabilistic reasoning*). One can solve problems such as ”What is the probability of $X_i = x_i$ given observation of some of the other variables?”. For each node in a Bayesian network a *marginal probability distribution* is computed from the joint probability distribution $P(x_1, \ldots, x_n)$:

$$P(X_i = x_i) = \sum_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n} P(x_1, \ldots, x_n).$$

This is an initial (*prior*) probability distribution computed before any evidence is available. (Note that for continuous variables the summation is replaced with integration.) Knowledge about the value of a variable in the network can modify the probabilities of other variables. When a particular variable in the network is observed to have a certain value, this *evidence* will be *propagated* through the network causing updating the probability distributions of other variables in the network (*belief updating*). This will lead to the *marginal posterior distributions* of the nodes that incorporate the evidence:

$$P(X_i = x_i | X_j = e), \qquad i \neq j,$$

which is the conditional distribution. Propagation of evidence in Bayesian networks exploit the convenient factorized representation of the joint probability distribution. J. Pearl showed that evidence propagation in Bayesian networks can be transformed into the local computations on the graphical structure of the Bayesian network (see Pearl 1988, also Castillo 1997, Neapolitan 1990).

A Bayesian Network structure implies a set of independence statements. More than one graph can imply exactly the same set of independencies. These graphs are called *independence equivalent.* Equivalent graphs have the same underlying undirected graph but might disagree on the direction of some of the edges. Pearl and Verma (1991) showed that DAGs are independence equivalent iff they have:

- the same underlying undirected graphs and
- the same *v-structures*, i.e. a DAG on triplet of nodes X, Y, Z with edges $X \to Y$ and $Z \to Y$, where X and Z are not adjacent.

Chickering (1996 b) showed that an equivalence class of network structures can be uniquely represented by a partially directed acyclic graph (PDAG), which contains all the edges from the underlying directed graph G. The edges are directed in the PDAG, if in graph G they are part of a v-structure or could participate in a v-structure by reversal. The other edges are undirected in the PDAG and might be reversible in equivalent directed graphs, so that some members of the class contain the edge $X \to Y$, while others contain the edge $X \leftarrow Y$. For example, a PDAG corresponding to the graph in Figure 2.3 will contain the directed edges $A \to C$ and $B \to C$, because they comprize the v-structure. The edge $C \to D$ will also keep its directionality, since it could participate in the v-structure otherwise. The edge $B - E$ will be undirected in the PDAG.

Given a directed graph G the PDAG representation of its equivalence class can be constructed efficiently (see Chickering 1996 b). Learning procedures based on scoring metrics as those described later in Section 2.4 can find exact network structure up to the correct equivalence class. The presentation of the joint probability distribution will be the same for all members of the equivalence class, unless we use a subjective prior for the network structure enabling to distinguish among them (see Heckerman, 1995). The reversibility of edges in equivalent graphs does not effect belief updating in a Bayesian Network.

The advances in research on Bayesian modelling led to the development of methods to infer Bayesian networks from databases of cases rather than from the insight of an expert. *Bayesian Network learning* consists of two major tasks:

- the induction of the graphical structure G, specifying the conditional independence assumptions among the variables in X;
- the estimation of conditional probabilities defining the dependencies in the given graphical model G.

The first of these tasks involves searching in the space of possible graphical models for one or more structures consistent with the conditional independence relationships suggested by the data. A scoring function that evaluates each network with respect to the training data was derived by Cooper and Herskovits (1992). They applied the methodology of Bayesian statistics. In

the next section I provide a general introduction into this methodology. Section 2.4 presents the derivation of the Bayesian scoring metric using Bayesian statistics.
In the last years Bayesian network models have gained increasing popularity for the possibility they provide to represent uncertain knowledge in complex domains and to make predictions. Due to the development of Bayesian network learning from data, Bayesian network models have become an important data mining tool. Here I mention only several applications. Bayesian network is the central part of the expert system HUGIN (Andersen et al. 1989), which was used e.g. in the commercial tool BayesCredit that predicts the risk of a credit (see http://www.hugin.com). Bayesian networks were broadly used in medical applications (see Lucas et al. 2001). The Bayesian expert system DIAVAL was built for diagnosis of heart diseases (Diez et al. 2001). Bayesian networks can be used for classification purposes (Friedman and Goldszmidt 1996). In molecular genetics domain Bayesian network learning was applied to reveal the genetic interactions from gene expression data (Friedman et al. 2000), this is presented in Section 3.4. The present thesis demonstrates the application of the Bayesian network model in cancer genetics domain.

2.3 Bayesian modelling

Probability and statistics provide a basis for addressing two crucial problems in artificial intelligence - how to reason in the presence of uncertainty, and how to learn from experience. The central quantity in statistics is *likelihood*, that is the probability that a model with particular parameter values assigns to the observed data. Assume, cases $X_1, \ldots, X_C$ have been observed and θ is the vector of the model parameters, then the likelihood is:

$$L(\theta|x_1, \ldots, x_C) = P(x_1, \ldots, x_C|\theta) = \prod_{i=1}^{C} P(x_i|\theta).$$

The likelihood is regarded as a function of the model parameters given data. It encapsulates the relative abilities of the various parameter values to "explain" the observed data, which may be considered as a measure of how plausible the parameter values are in light of the data.
Determining the parameters of the model from empirical data is a task of *statistical inference* and corresponds to the concept of *learning* in artificial intelligence. In modern statistics there are two approaches to learning. The conventional *frequentist* approach addresses this task by attempting to find *estimators* for unknown quantities. The widely used *maximum likelihood* procedure estimates the parameters of the model to be those that maximize the likelihood given the observed data. For a large class of models, the maximum

likelihood procedure has the frequentist justification that it converges to the true parameter values in the limit as the number of observed cases goes to infinity. This is not always the case, however, and even when it is, the quality of such estimates when based on small amounts of data may be poor.

In contrast, the Bayesian approach reduces statistical inference to probabilistic inference by defining a joint distribution for both the parameters and the observable data. Conditional on the data actually observed, posterior probability distributions for the parameters and for future observations can then be obtained.

This is a crucial aspect of Bayesian methods, in contrast to frequentist procedures, - to regard θ as a random quantity. Uncertainty concerning the parameters of the model is expressed by means of a probability distribution over the possible parameter values. A *prior* probability distribution for the parameters, $P(\theta_1, \ldots, \theta_p)$, is required, which embodies our judgement, before seeing any data, of how plausible it is that the parameters could have values in the various regions of parameter space. The introduction of a prior is the crucial element that converts statistical inference into an application of probabilistic inference.

When we combine a prior distribution for the parameters with the conditional distribution for the observed data, by Bayes' rule, we get a joint probability distribution for all quantities related to the problem, the *full probability model*:

$$P(\theta_1, \ldots, \theta_p, x_1, \ldots, x_C) = P(\theta_1, \ldots, \theta_p) P(x_1, \ldots, x_C | \theta_1, \ldots, \theta_p) =$$

$$= P(\theta) \prod_{i=1}^{C} P(x_i|\theta).$$

Using the Bayes' rule we can derive the *posterior distribution of the parameters*, given observed values for $X_1, \ldots, X_C$:

$$P(\theta|x_1, \ldots, x_C) = \frac{P(\theta, x_1, \ldots, x_C)}{P(x_1, \ldots, x_C)} = \frac{P(\theta) \prod_{i=1}^{C} P(x_i|\theta)}{\int P(\theta) \prod_{i=1}^{C} P(x_i|\theta) d\theta}$$

Since the denominator does not depend on θ, the posterior can be expressed as a proportionality in terms of the likelihood:

$$P(\theta|x_1, \ldots, x_C) \propto P(\theta) L(\theta|x_1, \ldots, x_C)$$

(the operator $\propto$ means "is proportional to"). In words:

posterior distribution $\propto$ likelihood $*$ prior distribution.

This shows how the introduction of a prior converts the expression of relative plausibility contained in the likelihood into an actual probability distribution over parameter space.

In general, Bayesian methodology use probability to express all forms of uncertainty: uncertainty about parameters of the model and uncertainty about the model itself. The results of Bayesian learning are expressed in terms of probability distributions over all unknown quantities.

Learning the model from data is often called in the literature *model selection* or *model choice.* A Bayesian approach to model selection is a problem of calculating the posterior probability of a model given data for a collection of candidate models. Then the model with the maximum posterior probability will be selected.

Suppose that the data D have been generated by a model m, one of a set M of candidate models $(m \in M)$. If p(m) is the *prior probability* of model m, then the *posterior model probability* is given by Bayes rule:

$$p(m|D) = \frac{p(m)p(D|m)}{\sum\limits_{m \in M} p(m)p(D|m)},$$

where $p(D|m)$ is the *marginal likelihood* calculated by Total Probability Theorem:

$$p(D|m) = \int p(D|m, \theta_m)p(\theta_m|m)d\theta_m, \qquad (*)$$

and $p(\theta_m|m)$ is the conditional prior distribution of model parameters θ_m for model m. Since the denominator is constant for different models, it is sufficient to compute the marginal likelihood $p(D|m)$ for the model m, in order to select the most probable model. Again, the definition of the prior on model parameters is required.

Bayesians divide into two schools on the point of prior definition. Some try to produce "objective" (*noninformative*) priors that represent complete ignorance about the parameters. Others, while finding "subjective" priors useful on occasion, regard the requirement for complete objectivity as unnecessary.

A common approach for prior elicitation is to choose a prior distribution with density function similar to the likelihood function (Bernardo and Smith, 1984). In doing so, the posterior distribution of θ will be in the same class of distributions as the prior. The prior is said to be *conjugate* to the likelihood. Conjugate priors play an important role in Bayesian methods, since they can simplify the integration procedure in $(*)$. A list of important conjugate distributions can be found in (Bernardo and Smith, 1984).

Learning Bayesian network model from data was facilitated based on the general Bayesian approach.

2.4 Bayesian Network learning

Here I shortly present how the Bayesian learning approach was used by Cooper and Herskovits (1992) to facilitate the Bayesian network learning from data, i.e. to derive the Bayesian scoring metrics.
Let $X = (X_1, \dots, X_n)$ be a set of n discrete variables, i.e. a multinomial random variable, where each variable X_i can take one of r_i distinct values $1, \dots, r_i$. We will denote by x_{ik} a state k of a variable X_i, $k \in \{1, \dots, r_i\}$.
A conditional dependency links a variable X_i to a set of parent variables, and it is defined by the conditional distributions of X_i given each *configuration* $\pi_{i1}, \dots, \pi_{iq_i}$ of the parent variables, q_i is the number of all possible configurations of the parents of X_i. We denote by $\theta_{ij} = (\theta_{ij1}, \dots, \theta_{ijr_i})$ the parameter vector associated to the conditional distribution $X_i|\pi_{ij}, j \in \{1, \dots, q_i\}$.
Let D be a database of N independent cases, where each case l is a random sample $(x_{1l}, \dots, x_{nl}), l \in \{1, \dots, N\}$ from a multinomial distribution.
The following assumptions were made to facilitate the derivation of the scoring measure:

- Assumption 1: the data set D is *complete*, there are no missing values;
- Assumption 2: *parameter independence*;
  The parameter vectors θ_i and $\theta_{i'}$ associated to different variables X_i and $X_{i'}$ are independent for $i \neq i'$ (*global independence*), so:

  $$p(\theta|D) = \prod_{i=1}^{n} p(\theta_i|D)$$

  The parameters θ_{ij} and $\theta_{ij'}$ associated to the distributions of X_i given different parent configurations, are also assumed to be independent (*local independence*), then:

  $$p(\theta_i|D) = \prod_{j=1}^{q_i} p(\theta_{ij}|D)$$

- Assumption 3: *parameter modularity*;
  If a node has the same parents in two distinct networks, the probability distribution functions of the parameters, associated with this node are identical in both networks;

  $$p(\theta_i|m_j) = p(\theta_i|m_{j'})$$

- Assumption 4: the prior probability distribution of the parameters θ_{ij} is the Dirichlet distribution with hyperparameters $\{\alpha_{ij1}, \dots, \alpha_{ijr_i}\}$:

  $$D(\alpha_{ij1}, \dots, \alpha_{ijr_i}),$$

$$p(\theta_{ij}) \propto \prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}-1},$$

then:

$$p(\theta|D) \propto \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}-1}.$$

Let $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$. Let N_{ijk} be the number of cases in D in which variable X_i has the value k and parents π_i are instantiated with the values of configuration j. Let $N = \sum_{k} N_{ijk}$. Since the cases are independent, the likelihood of the data D can be written as:

$$L(\theta) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}}$$

(multinomial distribution). The reason for using Dirichlet distribution for the prior of the parameters is that it is the natural conjugate for the multinomial distribution. The posterior distribution of the parameters given D, then, is

$$p(\theta|D) \propto \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}+\alpha_{ijk}-1},$$

that is the Dirichlet distribution:

$$D(\alpha_{ij1} + N_{ij1}, \ldots, \alpha_{ijr_i} + N_{ijr_i}).$$

The information conveyed by the sample can be therefore incorporated by simply updating the hyperparameters of the distribution of θ_{ij} by increasing them of the frequency of cases with particular parent-child configurations observed in the sample.

Under the previously introduced assumptions, the marginal likelihood

$$p(D|m) = \int p(D|m, \theta_m) p(\theta_m|m) d\theta_m$$

is a Dirichlet integral and has a closed form solution:

$$p(D|m) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})},$$

where $\Gamma()$ is the Gamma function satisfying $\Gamma(1) = 1$ and $\Gamma(n+1) = n\Gamma(n) = n!$. Cooper and Herskovits suggested the noninformative assignment for the prior distribution of the parameters: $D(1,\ldots,1)$, i.e. hyperparameters $\alpha_{ijk} = 1$. Then $\alpha_{ij} = \sum\limits_{k=1}^{r_i} 1 = r_i$. Thus, $p(D|m)$ can be expressed as:

$$p(D|m) = \prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\prod_{k=1}^{r_i}N_{ijk}!$$

and

$$P(m,D) = P(m)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\prod_{k=1}^{r_i}N_{ijk}!.$$

This result is referred to as the *Bayesian scoring* of the quality of a network structure or *K2 metric*, because it is used in the algorithm referred to as K2-algorithm (Cooper and Herskovits 1992).

In practical implementations, to simplify the computations often the logarithm of this equation is used:

$$logP(m,D) = logP(m) + \sum_{i=1}^{n}\sum_{j=1}^{q_i} log\frac{(r_i-1)!}{(N_{ij}+r_i-1)!} + \sum_{k=1}^{r_i} logN_{ijk}!.$$

An important characteristic of the Bayesian scoring is its decomposability. Denote the local contribution of a node X_i and its parents Pa_i to the overall score of the model by

$$S_{local} = \sum_{j=1}^{q_i} log\frac{(r_i-1)!}{(N_{ij}+r_i-1)!} + \sum_{k=1}^{r_i} logN_{ijk}!.$$

Then the total score is:

$$S = logP(m) + \sum_{i=1}^{n} S_{local}(X_i, Pa_i).$$

Hence, the contribution of each variable X_i to the total network score depends only on the values of X_i and the values of its parents Pa_i in the training instances.

The optimization problem of learning the structure G that maximizes the score is known to be NP-hard (the number of candidate structures grows exponentially with the number of nodes), see Chickering (1996 a). Heuristic methods have to be used to reduce the search space to the subset of models. Cooper and Herskovits proposed a *greedy (hill climbing)* search algorithm that takes advantage of the decomposability of the score and works locally (K2-algorithm). There is also an *arc inversion* algorithm. The algorithms assume that there is an ordering among the variables. Both algorithms start from an initial network structure containing no edges. For each variable they examine the set of its possible parents by adding the edges incrementally until the score of the structure does not further increase. During the greedy search, each node in the order is tested as child only of the lower nodes. During the arc inversion procedure, higher nodes will be tested first and, if not set as children, they will be tested as parents of lower nodes in the order. In this work I used the software Bayesware Discoverer (http://www.bayesware.com), which implements both of these two heuristic procedures: the greedy search and the arc inversion.

2.5 The method of reconstructing the flow of progression of genetic abnormalities using Bayesian network

In the present work I introduce the method of reconstructing the flow of progression of genetic abnormalities in tumor samples. The main idea is to apply Bayesian network learning to the data of genetic abnormalities. The probabilistic semantics of Bayesian network allows to model the stochastic nature of occurrence of genetic abnormalities and to handle noisy experimental data. Once a Bayesian network model is learned, one can exploit it for discovering important aspects of the variables representing genetic abnormalities.

The induced Bayesian network model can be used to study dependencies and independencies between the genetic abnormalities. These are multivariate dependencies describing the probability of one event dependent on configuration of one or more predecessor-events. The dependencies are much more complex than pairwise co-occurrences of events. "Negative" dependencies describing the occurrence of an event, conditioned that other events do not occur, can also be captured.

One can investigate the graphical component of the Bayesian network. The lack of edges in the graph going from one variable to another suggests an independence between these variables. One can investigate which variables are directly connected with each other, and which are connected via other variables. Also, by considering Markov blankets of the nodes one can reveal interesting groups of directly related abnormalities.

Further, the quantitative component of the Bayesian network can be examined, namely the conditional probability distributions representing the dependence of one variable on its parent variables in the network.

Exploiting the mechanism of probability propagation and belief updating in the Bayesian network one can assess the probability of one abnormality given the observation of some other abnormality. To gain an insight into the effect of one abnormality on the other connected with it by an edge, for each edge $X_i \to X_j$ in the network, one can calculate the ratio of the posterior marginal probability and the prior marginal probability:

$$\frac{P(X_j = 1|X_i = 1)}{P(X_j = 1)}$$

This ratio reflects the change in the probability distribution of X_j after observing X_i. The ratio is larger than 1 if the occurrence of X_i increases the probability of X_j ("positive" connection) and is less than 1 if the occurrence of X_i decreases the probability of X_j("negative" connection).

The induced Bayesian network model can be used to infer hypotheses about early events that initiate the accumulation of abnormalities and about events associated with late stages of tumor progression. If a particular alteration is a late event, the probability of other alterations to be observed simultaneously will be high. Abnormalities which lead to malignant transformation of the cell will give rise to many other abnormalities and, therefore, in the Bayesian network they will have high degree of connectivity and high impact on other abnormalities according to their conditional dependency strength.

As already outlined, the mechanisms of probability propagation in the Bayesian network allow to assess the probability of one genetic event conditional on the observation of some other events. This leads to the idea to investigate in a quantitative way the progression of abnormalities along the hypothetic pathways as they are being accumulated during oncogenesis. One can insert in the Bayesian network evidence on occurrence (or no-occurrence) of abnormalities from some interesting patterns of genetic events and investigate the probability of particular abnormalities conditional on this evidence. For each genetic event of interest one can inspect what other events make this event mostly probable. The predictions on genetic abnormalities under the induced Bayesian network model allows to infer hypotheses about the progression of these events.

An important issue with this approach to analysis of molecular genetics data is to induce the Bayesian network model with high degree of confidence.

One problem while applying the K2-algorithm for learning the Bayesian network from data is that this heuristic search algorithm depends on the ordering of the variables. A normal practice in the Bayesian networks community is to generate as many Bayesian networks as possible while presenting to the algorithm different random variable orderings. Then, the Bayesian network with the highest scoring will be chosen.

When working with sparse dataset and relatively high number of variables, the induction of the highest scoring Bayesian network model might not be sufficient (Friedman *et al.* 1999 a, b). The amount of data might not be enough to induce a high scoring network. Network structures with almost equal scores might have very different structures. The score of the network

reflects how well does the network fits the data. The data might contain a lot of noise, and some relations found might be spurious (random). Friedman *et al.* studied the problem of assessing a confidence measure on features of the learned Bayesian network structure. The authors proposed to use such confidence measures to induce better Bayesian network structure from the data. Following Friedman *et al.*, a "feature" of the model could be considered as reliable, if this feature appears in the majority of the models learned from data. Also, features of the Bayesian network model could be considered as more reliable, if they were induced from "perturbed" data, i.e the model does not only fit the training data, but also generalizes from the data. The perturbed data can be generated from the original dataset by *re-sampling (with or without replacement).* This is, generally, the idea of the *non-parametric bootstrap* method, originally developed by Efron and Tibshirani (1998). Friedman *et al.* apply the method for the estimation of confidence in the features of learned Bayesian networks.
Consider a set of discrete random variables $X = \{X_1, \ldots, X_n\}$ where each variable X_i may take on values from a finite set. We denote by x assignments of values to the variables in the set X. Suppose we are given N observations $\{x[1], \ldots, x[N]\}$,- the dataset D of size N. The non-parametric bootstrap begins by re-sampling N times (with or without replacement) from the dataset D. This results in a sequence of instances:

$$D_1^* = \{x_1^*[1], \ldots, x_1^*[N]\}.$$

The procedure will be repeated m times, the i^{th} replicate is then:

$$D_i^* = \{x_i^*[1], \ldots, x_i^*[N]\}, \qquad i = 1, \ldots, m.$$

From each replicate dataset D_i^* the Bayesian network structure will be learned $G(D_i^*)$. Let $f(G_i^*)$ be a feature of the Bayesian network structure $G(D_i^*)$. Define the following quantity

$$p_N(f) = \frac{1}{m}\sum_{i=1}^{m} f(G_i^*),$$

where

$$f(G_i^*) = \begin{cases} 1, f \in G(D_i^*) \\ 0, otherwise \end{cases}$$

In words, $f(G_i^*)$ is equal to 1, if this feature is present in the network structure. The quantity $p_N(f)$ is called the confidence in the feature of the Bayesian network learned based on the dataset D. In the following, we are interested in the structural feature of the model, i.e. whether two nodes are directly connected.

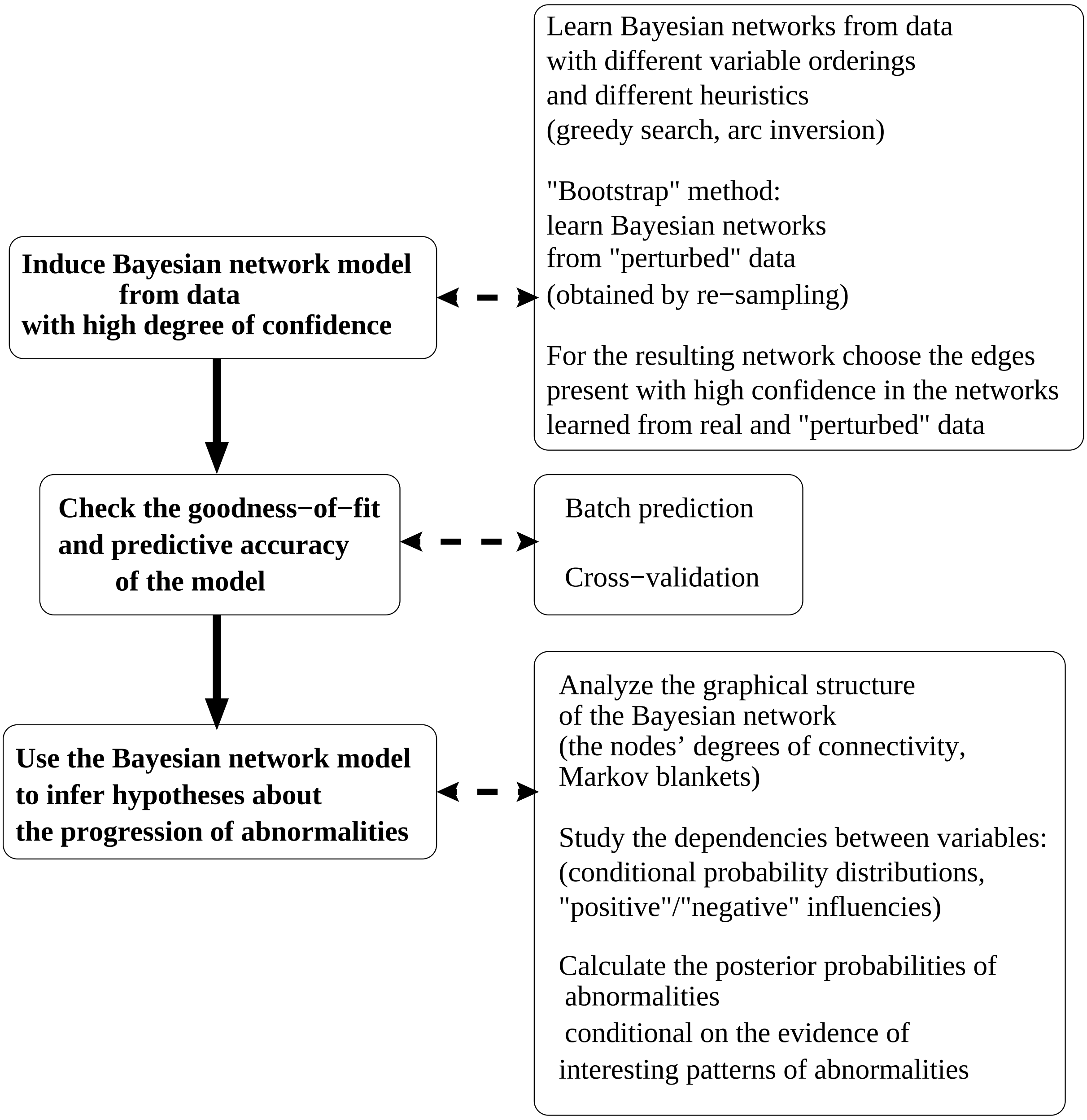
Learn Bayesian networks from data
with different variable orderings
and different heuristics
(greedy search, arc inversion)
"Bootstrap" method:
learn Bayesian networks
from "perturbed" data
(obtained by re−sampling)
For the resulting network choose the edges
present with high confidence in the networks
learned from real and "perturbed" data
Induce Bayesian network model
from data
with high degree of confidence
Check the goodness−of−fit
and predictive accuracy
of the model
Batch prediction
Cross−validation
Use the Bayesian network model
to infer hypotheses about
the progression of abnormalities
Analyze the graphical structure
of the Bayesian network
(the nodes' degrees of connectivity,
Markov blankets)
Study the dependencies between variables:
(conditional probability distributions,
"positive"/"negative" influencies)
Calculate the posterior probabilities of
abnormalities
conditional on the evidence of
interesting patterns of abnormalities

Figure 2.4: Analysis of genetic abnormalities data with Bayesian network model.

My approach for the analysis of the data of genetic abnormalities is presented in Figure 2.4. For the induction of the Bayesian network model with high degree of confidence the following procedure is proposed. Generate Bayesian networks with different random variable orderings using heuristic algorithms (e.g. greedy search, arc inversion). For each undirected edge from the generated networks calculate the percentage of its occurrence in the networks. Further, generate Bayesian networks from "perturbed" data (bootstrap method) and calculate the degree of confidence of this edge. For the resulting Bayesian network, select the edges that were most frequently induced by the network generation procedure, and still have a high degree of confidence as obtained by bootstrap method.

If for some edges the directions in different Bayesian networks do not coincide, for the resulting Bayesian network choose the direction of an edge going from the more frequent abnormality to the less frequent. This is a biologically plausible assumption, since more common abnormalities are likely to appear earlier than less frequent ones.

Once the Bayesian network model is induced, the goodness-of-fit of the model must be checked. The idea of the *Bayesian model checking* is to assess the posterior predictions for the variables made under the model and compare them to the observed values. If the model fits, then the model predictions should be similar to the observed data. Two procedures can be performed: *batch prediction* and *cross-validation*. The batch prediction predicts the value of a single variable in each case given the evidence on other variables contained in the case. For the cross-validation method the dataset is divided in parts. The cross-validation method predicts the value of the variable in one part of the dataset with the conditional probabilities of the model estimated based on the other part of the dataset.

After induction of the reliable Bayesian network model, it can be used to infer hypotheses about the progression of genetic abnormalities.

I applied the described method for the analysis of data of losses of heterozygosity in urothelial cancer samples. This will be presented in the next section.

2.6 Applying Bayesian network analysis to the allelotyping data (LOH)

2.6.1 The data of losses of heterozygosity. Previous studies of urothelial cancer.

In tumor samples the condition called "loss of heterozygosity" (loss of alleles on one chromosome) can be detected for particular markers for which an individual is heterozygous. Losses of heterozygosity indicate the presence of tumor suppressor genes inactivated due to the allelic loss and are integral parts of tumor progression. Screening for LOH is of great significance for understanding the tumorigenesis. In LOH experiments, several chromosomal regions (loci) are being mapped with microsatellite markers labeled with fluorescent substrates. Then PCR (polymerase chain reaction) is performed. The PCR product is then analysed to measure the intensity of fluorescent signal with an automated fluorescent DNA sequencer. Loss of heterozygosity is identified at particular microsatellite loci when either no peak or a very weak peak is detected in the tumor DNA as compared to the normal DNA (for example, from blood). For the introduction into the molecular genetics methods see for example Strachan and Read (1996).

Urothelial carcinoma of the bladder (UC) comprise biologically and morphologically heterogenous groups of neoplasms. During the last decade a broad spectrum of genetic alterations has been described in UCs. Cytogenetic, CGH and microsatellite analyses revealed loss, gain and amplification of DNA sequences at several chromosomal regions (Knowles, 2001). Hemi- and homozygous deletion at and methylation/mutation of the *CDKN2A* gene at chromosome 9p21 is considered to be an early genetic event (Berggren *et al.* 2003). The vast majority of urothelial carcinomas acquire several additional genetic alterations during progression, including deletion of chromosome 2q, 5q, and 8p, deletion/mutation of the *p*53 and *Rb* genes or amplification and overexpression of the *ERBB-2* gene (Langbein *et al.* 2002, von Knobloch *et al.* 2000, Muschek *et al.* 2000, Habuchi *et al.* 1993, Logothetis *et al.* 1992, Mellon *et al.* 1996). Although some of the genetic alterations occur at random, the recurrent changes may refer to a network of genes that are specifically involved in tumor development and progression.

Several models indicating a step-by-step order of genetic changes as a single pathway from normal urothelial cell to malignant tumor have been proposed (Dalbagni *et al.* 1993, Spruck *et al.* 1994, Reznikoff *et al.* 1996).

Desper *et al.* (1999) and Schäffer *et al.* (2001) applied tree models to infer the order of genetic changes during progression of urothelial carcinomas. As already pointed out in Section 2.1, tree models are too restrictive to model the heterogeneous process of tumor progression.

I use Bayesian network model to describe the complex dependencies between the genetic abnormalities. I apply the framework presented in section 2.5.

2.6.2 Induction of the Bayesian network from LOH data of urothelial cancer

The dataset I was working on contained the data of allelic losses from 123 cases of papillary urothelial carcinomas of the bladder.

I considered 17 random variables representing the stochastic genetic events "complete loss of heterozygosity at a given chromosomal region", which are encoded for example such as 9p for loss of the specific region at the short arm of chromosome 9. These random variables have binary values 0 or 1 representing the occurrence or non-occurrence of LOH in a case, respectively.

I have generated 100 Bayesian networks with different random variable orderings with greedy search and with arc inversion strategies. For each undirected edge from the generated networks, I have calculated the percentage of its occurrence in the networks (see Table 2.2).

Further, I performed the non-parametric bootstrap method. I generated 30 instances from the dataset by re-sampling without replacement. From each of the resulting datasets I learned 3 Bayesian networks with different random ordering of the variables and different search strategies, thus obtaining 90 Bayesian network structures. Again, for each undirected edge, I estimated the degree of confidence by calculating in how many network structures this edge was induced. For the resulting Bayesian network, I have selected edges that were most frequently induced by the network generation procedure, and still have a high degree of confidence as obtained by bootstrap method (see Tab. 2.2). For the edges, for which the directions in different Bayesian networks did not coincide, I chose the direction of an edge going from the more frequent abnormality to the less frequent. The resulting induced Bayesian Network is presented in Figure 2.5.

Once I have induced the Bayesian network model, I checked the goodness-of-fit of the model to the data. I have performed batch prediction and cross-validation as described in the section 2.5. For the cross-validation I have divided the dataset in two parts and repeated the cross-validation procedure 100 times. The results are presented in Table 2.3.

2.6.3 Results of the application of Bayesian network analysis to the LOH data of urothelial cancer

The resulting induced Bayesian Network is presented in Figure 2.5. It is a representation of the most significant features of the underlying probability distribution. Some arc connections (see Table 2.2) like 13q – 3p, 11p – 11q, 18q – 6q have poor level of confidence, but excluding them from the model decreased the overall predictive accuracy of the model. Analysis of the graph revealed only few edges that might be reversible in equivalent graphs: 9p – 9q, 8p – 1q, 8p – 18q, 8p – 2q. Example of the conditional probability tables quantifying the graph structure and the respective distributions variances are displayed in Tables 2.4 and 2.5. Analysis of the distribution variances showed that they were small indicating a high degree of confidence for the network model. The predictive accuracy of the model obtained by batch

Edge		**Number of BNs, %**	**Confidence**
17p	16q	100	1
2q	14q	100	1
18q	11p	100	1
9p	9q	100	1
10q	16q	100	0.99
5p	3p	99	1
17p	3p	99	1
17p	13q	99	1
17p	5q	99	1
8p	17p	97	1
17p	2q	97	0.93
8p	1q	96	1
10q	3p	96	0.98
8p	13q	95	1
17p	11p	95	0.98
8p	10q	88	0.97
8p	6q	87	0.99
14q	3p	85	0.88
11q	3p	85	0.87
5p	13q	78	0.77
17p	11q	77	0.96
9q	17p	75	0.91
17p	5p	74	0.83
11p	10q	61	0.72
5p	14q	61	0.46
18q	11q	56	0.71
2q	16q	53	0.38
8p	14q	47	0.24
8p	2q	45	0.59
8p	18q	35	0.83
5q	14q	25	0.28
5q	16q	12	0.11
5q	6q	10	0.14
13q	3p	10	0.07
11p	11q	10	0.07
18q	6q	10	0.07
10q	6q	10	0.01

Table 2.2: Percentage of generated Bayesian networks containing the undirected edge defined by two leftmost columns, and the degree of confidence of the edge as obtained by bootstrap method.

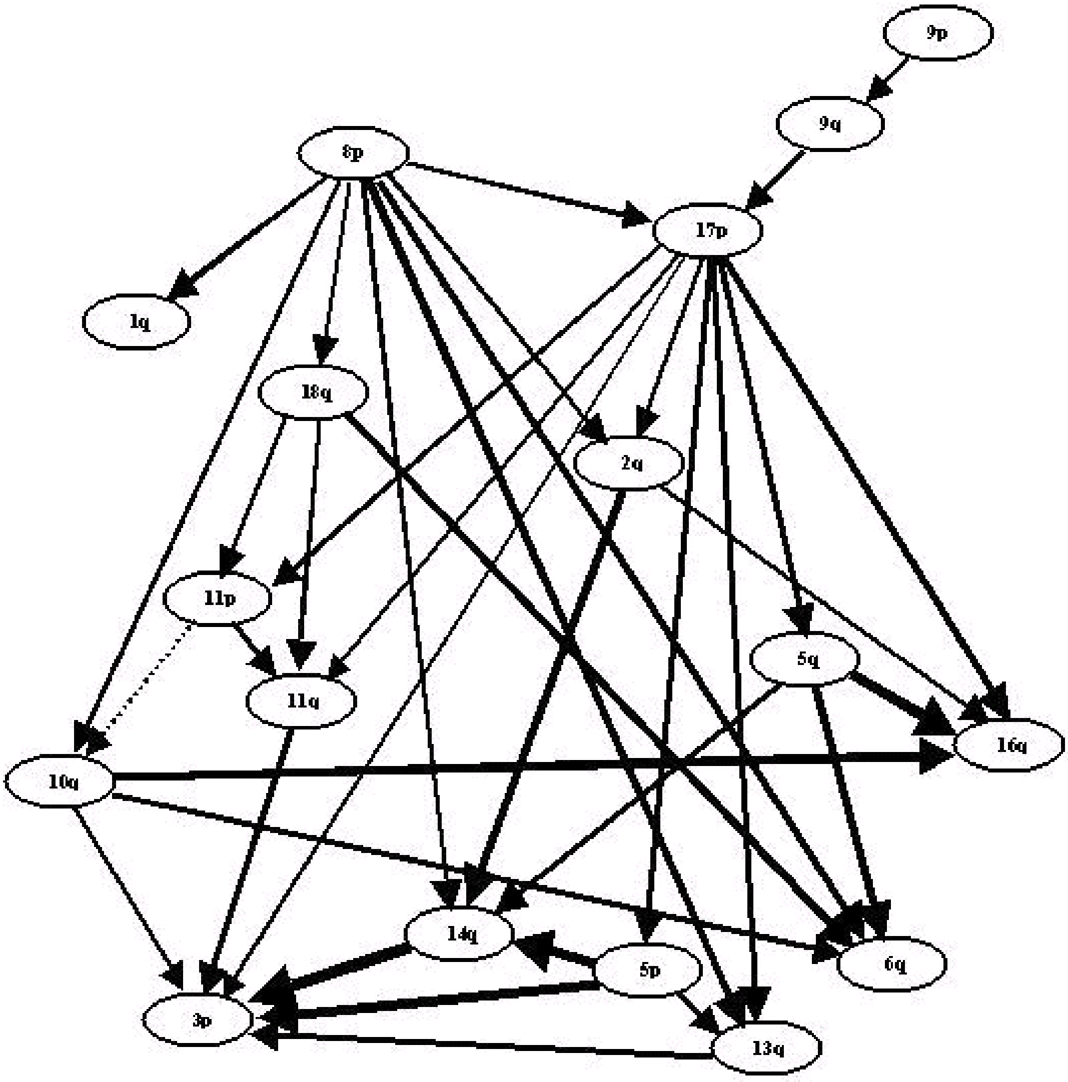

Figure 2.5: Bayesian network induced from LOH data of urothelial carcinomas. The width of the arc represents the strength of the probabilistic dependencies between genetic losses. Full arcs represent "positive" connections, dashed arcs represent "negative" connections.

	9p	**9q**	**8p**	**17p**	**1q**	**18q**	**2q**	**10q**	**11p**	**11q**	**5p**	**5q**	**14q**	**3p**	**13q**	**6q**	**16q**	
Batch prediction	74.8%	74.0%	83.9%	91.5%	90.1%	79.7%	85.1%	88.4%	78.0%	83.9%	91.6%	89.1%	94.2%	92.8%	85.4%	89.3%	93.6%	
Cross-validation	74.8%	71.5%	74.7%	81.8%	90.1%	73.3%	78.2%	82.2%	71.7%	76.3%	87.1%	83.1%	86.4%	86.8%	83%	85.5%	89.0%	
St. deviation of cross-validation	3.9	4.1	3.9	3.5	2.7	3.4	3.7	3.5	4.1	3.8	3.0	3.4	3.1	3.1	3.4	3.2	2.8	

Table 2.3: Predictive accuracy of the model for each of the variables obtained by batch prediction and cross-validation methods.

prediction and cross-validation is summarized in Table 2.3. The results show that the model is capable of making accurate predictions.

		2q	
8p	17p	0	1
0	0	0.824	0.176
0	1	0.745	0.255
1	0	0.602	0.398
1	1	0.270	0.730

Table 2.4: Example of the conditonal probability table for the variable 2q.

After induction of the Bayesian network and checking the goodness-of-fit of the model, I used its graphical and quantitative components to study (in)dependencies between variables representing allelic losses.

As outlined in the section 2.5, abnormalities which lead to malignant transformation of the cell will give rise to many other abnormalities and therefore in the Bayesian network they will have high degree of connectivity and high impact on other abnormalities according to their conditional dependency strength. The graphical structure of the Bayesian network shows that the nodes 8p and 17p are highly connected with other nodes, whereas the nodes 9p and 9q are rather ”isolated”. The Markov blanket of the node 9q consists only of the nodes 9p, 17p, 8p. This “isolation” of the nodes 9p and 9q in view of the fact that these are frequent losses suggests that they are early events

2q			
0	0	0.034	0.034
0	1	0.045	0.045
1	0	0.056	0.056
1	1	0.046	0.046

Table 2.5: Example of the distribution variancies for the conditional probability distribution of the variable 2q.

which may constitute the primary condition for tumor development. The Markov blanket of the node 8p contains nodes: 9q, 17p, 1q, 18q, 14q, 13q, 6q, 2q, 10q, 11p, 5q, 5p. Only nodes 11q, 3p, 16q and 9p do not belong to the Markov blanket of 8p. The nodes 11q, 3p, 16q are not connected directly with 8p, but via other nodes. The Markov blanket of the node 17p consists of nodes 9q, 8p, 11p, 11q, 3p, 2q, 5p, 5q, 13q, 16q, 10q, 14q. The nodes 1q, 6q, 18q and 9p do not participate in Markov blanket of 17p. The nodes 1q, 6q, 18q are not directly connected with 17p. The dependence of 6q on 17p is mediated via the node 5q. The nodes 1q and 18q are independent of 17p given 8p. The large Markov blankets of the nodes 8p and 17p suggest that these abnormalities give rise to many other abnormalities. The conditional independence of the nodes 1q and 18q from 17p given 8p might indicate a pathway of progression distinct from those going through the LOH of 17p region. The abnormality 6q is likely to be more related to 8p. The loss of 17p is obviously dependent on the losses of chromosome 9p/9q regions and the loss of 8p. The conditional probability tables determine the strength of the dependency of an event dependent on the states of its parents. The conditional probability distribution of the node 17p (Table 2.6) shows that the probability of loss in 17p given both losses of 8p and 9q equals 0.8, whereas the probabilities of loss of 17p given occurrence of only one of the described parent events are both 0.27. The propagation of the evidence on observing

17p			
9q	**8p**	0	1
0	0	0.895	0.105
0	1	0.732	0.268
1	0	0.733	0.267
1	1	0.200	0.800

Table 2.6: Conditional probability table representing the probability distribution of the loss of heterozygosity of 17p conditional on the losses of 9q and 8p. The value 0 (1) indicates loss (no loss).

losses 9p, 9q and 8p increased the marginal probability of the loss of 17p from its prior value 0.312 to the posterior value 0.803. This suggests that loss of 17p occurs mostly after the losses of chromosome 9 and 8p.

The type of the effect ("positive" or "negative") of one abnormality on another connected by an edge and the strength of this connection was calculated as described in the section 2.5. All arc connections in the Bayesian network except 11p → 10q were found to be "positive" indicating that the occurrence of one loss makes the occurrence of another loss more probable. The only negative connection of weak strength was found between nodes 11p and 10q. The loss of chromosome 11p might not be associated with the loss of chromosome 10q alone. Instead, in combination with the loss 8p, 11p influences the occurrence of loss 10q positively.

The mechanisms of probability propagation and belief updating in the Bayesian network allow to assess the probability of one genetic event conditional on the observation of some others. I have inserted in the Bayesian network evidence on occurrence (or no-occurrence) of losses from some interesting patterns and investigated the probability of genetic losses conditional on this evidence (Table 2.7). The patterns of evidence are presented in the most left column of the table. In the first line of the Table 2.7 the prior marginal probabilities of single losses (before any evidence is observed) are presented. The observation of loss of chromosome 9 changes the probability of 17p loss from its marginal probability 0.312 to 0.451. The conditional probability of 17p given the evidence that the losses 9p/9q do not occur is estimated to be 0.163.

Of special interest was to determine which genetic losses are associated with the loss of 8p and the loss of 17p. Consider the patterns of evidence {not (9p, 9q, 17p) and 8p}, {9p, 9q, 17p and not 8p} and {9p, 9q, 8p, 17p} (see Tab. 2.7). The conditional probability of the loss of 18q to occur given the evidence of losses 8p and 17p was 0.467, almost the same (0.464) when only 8p is present, and is estimated to be 0.219 when 17p but not 8p is present. This suggests that the loss of 18q might be more associated with the loss of 8p. Similary, the loss of 10q is likely to be improbable if 8p is not present (0.079) and has the estimated probability 0.263 to occur when only 8p has occurred. Further, it has almost the same probability (0.266) when also losses 9p/9q and 17p have occured. Thus, the loss 17p is likely to have no influence on the occurrence of 10q. I suggest that the losses of 1q, 18q, and 10q are more likely to be associated with 8p. With similar examinations the losses of 5p, 5q, 16q were revealed to be likely associated with the loss of 17p. The losses 2q, 14q, 13q, 3p were found to be associated with both losses 8p and 17p. For example, the probability of 2q to occur if only 8p is present is estimated to be 0.396 and if 17p (but not 8p) occurs it is 0.255, while the evidence of both 8p and 17p increases the estimated probability of 2q to 0.731.

For each loss (each column) in Tab. 2.7 I have highlighted in bold the highest probabilities to occur conditional on the evidence. Inspection of these numbers revealed that the loss of chromosome 2q has the estimated probability 0.926 to occur when observing the group of losses 9p, 9q, 8p, 17p, 14q. It indicates a strong connection between losses 2q and 14q. The loss 14q appears to have high probability (0.837) to occur when losses of chromosome 5p/q occur. The loss 14q is also highly connected with the loss 3p. The losses 18q, 11p, and 11q were revealed to comprize a group of closely related events. The loss 6q appears to be almost improbable (0.03) when 8p is absent. The loss 6q is likely to be highly probable in connection with the losses of 18q and of chromosome 11 (0.773). This suggests the pattern of related losses: 8p, 18q, 6q. The losses 16q and 10q were found to be strongly associated, since the loss of 16q is most probable under the model when the loss of 10q is evident. The loss of 3p is obviously a late event, because many other losses (18q, 2q, 11p, 11q, 5p, 14q, 13q) appeared to be highly probable while considering the evidence of losses 9p/q, 8p, 17p and 3p. One can say that different pathways "converge" in this event. The loss of 13q is probable (with the estimated probability 0.359) already with the evidence of abnormalities 9p/9q, 8p and 17p.

	9p	**9q**	**8p**	**17p**	**1q**	**18q**	**2q**	**10q**	**11p**	**11q**	**5p**	**5q**	**14q**	**3p**	**13q**	**6q**	**16q**
Marginal distribution	0.472	0.513	0.338	0.312	0.092	0.301	0.322	0.165	0.304	0.207	0.067	0.153	0.101	0.129	0.094	0.126	0.079
not 9p	0	0.279	0.34	0.243	0.091	0.302	0.307	0.165	0.285	0.207	0.058	0.13	0.093	0.13	0.084	0.116	0.067
not (9p, 9q)	0	0	0.34	0.163	0.091	0.302	0.288	0.166	0.265	0.207	0.048	0.109	0.083	0.131	0.071	0.108	0.053
9p	1	0.777	**0.339**	0.387	0.092	0.302	0.34	0.164	0.323	0.207	0.076	0.174	0.111	0.129	0.108	0.135	0.092
9p, 9q	1	1	**0.34**	0.451	0.093	0.302	0.356	0.164	0.339	0.207	0.083	0.191	0.119	0.129	0.12	0.142	0.102
9p, 9q, 8p	1	1	1	**0.803**	**0.19**	0.466	0.665	0.265	0.475	0.243	0.125	0.288	0.255	0.192	0.311	0.356	0.175
9p, 9q and not 8p	1	1	0	0.268	0.04	0.217	0.197	0.108	0.267	0.189	0.061	0.137	0.049	0.099	0.020	0.030	0.064
not (9p, 9q) and 8p	0	0	1	0.274	0.19	0.465	0.487	0.263	0.369	0.267	0.062	0.141	0.153	0.174	0.181	0.265	0.091
9p, 9q, 17p and not 8p	1	1	0	1	0.042	0.219	0.255	0.079	0.523	0.25	0.149	0.343	0.057	0.043	0.072	0.032	0.201
not (9p, 9q, 17p) and 8p	0	0	1	0	0.19	0.464	0.396	0.263	0.314	0.28	0.029	0.065	0.1	0.165	0.114	0.218	0.048
9p, 9q, 8p, 17p	1	1	1	1	0.19	0.467	0.731	0.266	0.514	0.234	0.149	0.342	0.293	0.199	**0.359**	0.39	0.206
not (9p, 9q, 8p, 17p)	0	0	0	0	0.040	0.217	0.176	0.121	0.174	0.168	0.028	0.064	0.047	0.117	0.002	0.030	0.014
9p, 9q, 8p, 17p, 18q	1	1	1	1	0.19	1	0.731	0.274	0.498	0.203	0.149	0.343	0.293	0.196	0.359	0.408	0.211
9p, 9q, 8p, 17p, 2q	1	1	1	1	0.19	0.467	1	0.266	0.514	0.233	0.15	0.343	0.371	0.217	0.358	0.391	0.164
9p, 9q, 8p, 17p, 11p	1	1	1	1	0.19	0.452	0.73	0.275	1	0.337	0.149	0.342	0.292	0.217	0.358	0.391	0.209
9p, 9q, 8p, 17p, 11p, 11q	1	1	1	1	0.19	**0.531**	0.73	0.279	1	1	0.148	0.34	0.286	0.323	0.358	0.394	0.211
9p, 9q, 8p, 17p, 10q	1	1	1	1	0.19	0.48	0.729	1	**0.532**	0.239	0.147	0.326	0.281	0.327	0.358	0.524	**0.414**
9p, 9q, 8p, 17p, 5q	1	1	1	1	0.19	0.468	0.733	0.254	0.514	0.232	0.154	1	0.471	0.247	0.357	**0.797**	0.292
9p, 9q, 8p, 17p, 5q, 5p	1	1	1	1	0.19	0.468	0.748	0.243	0.511	0.223	1	1	**0.837**	**0.69**	0.032	**0.794**	0.278
9p,9q,8p,17p,18q,11p	1	1	1	1	0.19	1	0.73	0.284	1	**0.397**	0.149	0.343	0.291	0.228	0.358	0.412	0.214
9p,9q,8p,17p,18q,11p,11q	1	1	1	1	0.19	1	0.73	0.286	1	1	0.148	0.342	0.287	0.327	0.358	0.413	0.215
9p,9q,8p,17p,18q,11p,11q,10q	1	1	1	1	0.19	1	0.731	1	1	1	0.158	0.353	0.298	**0.686**	0.354	**0.773**	**0.426**
9p, 9q, 8p, 17p, 2q	1	1	1	1	0.19	0.465	1	0.267	0.515	0.234	0.152	0.338	0.369	0.220	0.357	0.377	0.169
9p, 9q, 8p, 17p, 14q	1	1	1	1	0.19	0.466	**0.926**	0.261	0.513	0.23	0.327	0.552	1	0.45	0.29	0.516	0.203
9p, 9q, 8p, 17p, 16q	1	1	1	1	0.19	0.465	0.592	**0.541**	0.522	0.235	0.152	0.491	0.292	0.248	0.357	0.506	1
9p, 9q, 8p, 17p, 6q	1	1	1	1	0.19	0.496	0.730	0.350	0.516	0.233	0.152	**0.691**	0.387	0.238	0.357	1	0.279
9p, 9q, 8p, 17p, 3p	1	1	1	1	0.19	**0.547**	**0.793**	0.418	**0.563**	**0.386**	**0.439**	0.415	**0.639**	1	**0.366**	0.442	0.255
9p, 9q, 8p, 17p, 13q	1	1	1	1	0.19	0.465	0.730	0.267	0.515	0.234	0.011	0.338	0.236	0.208	1	0.377	0.208

Table 2.7: Posterior probability distributions of the LOH of chromosome regions conditional on the evidence presented in the leftmost column. The first line presents the marginal distributions of losses. For each loss the highest probabilities to occur conditional on the evidence are highlighted in bold.

2.6.4 Discussion of the application of Bayesian network analysis to the LOH data of urothelial cancer

In this work I induced the Bayesian network model from the LOH data of urothelial carcinomas. The model contained the most reliable dependencies between the losses of heterozygosity. This enabled to extract patterns of DNA losses in bladder cancer and to reveal primary and secondary abnormalities associated with the tumor development. I suggested a possible flow of progression of allelic losses in bladder cancer as heterogeneous, distinct and converging genetic pathways.

Earlier models suggested that the development of papillary UCs is associated with LOH at chromosome 9p followed by mutation of the p53 gene in invasive tumors, whereas flat tumors, e.g. carcinoma in situ (Tis) are initiated by mutation of the p53 gene (Spruck *et al.*, 1994). Based on RFLP analysis, Dalbagni *et al.* 1993 have proposed that the transition of noninvasive UCs into invasive ones is marked by alteration of chromosomes 5q, 3p, 17p, 11p, 6q, 13q and 18q. Reznikoff *et al.* (1996) suggested three different pathways. In the first, alteration of chromosome 9q/9p (*CDKN2A*) is followed by inactivation of the p53 gene, LOH at chromosomes 3p, or 6q. The second pathway is characterised by inactivation of the p53 gene and subsequently by LOH at 3p and 6q/or 9p. The third proposed pathway is initiated by inactivation of the *RB* gene at chromosome 13q, followed by the duplication of chromosomes 5 and 20q, LOH at 10p and inactivation of the p53 gene at 17p. The three pathways converge into one common pathway showing LOH at chromosomes 4p, 8p, 11p and amplification of the *ERBB2* gene at 17q. Summing up the CGH data, Schäffer *et al.* (2001) applied tree models for dependent copy number changes in UCs. In both kinds of trees the aberrations -9p, -9q were close to the root suggesting that they are early aberrations. The trees contained two groups of closely related aberrations +17q, +20q, +20p and +5p, -8p, -17p, +10p, which were suggested to be late aberrations. The aberrations -18q, -11p, -11q were found to be strongly related by the branching tree model.
Höglund *et al.* (2001) have evaluated the genetic data obtained by karyotyping 200 UCs. They described two distinct pathways for UCs. One of them is characterized by deletion of chromosome 9p region, which was followed by -11p, +1q, -17p, -10, -15, and -16. The other pathway was marked by trisomy 7, followed by -8p, +8q, -17p and -3p. During late progression tumors of both pathways acquire -2p, -4p, -18, -22p, -6q, -5q, and +5p. Höglund *et al.* define early and late imbalances as those predominantly present in tumors with few and many imbalances, respectively. They used the modes of the distributions of number of imbalances per tumor (NIPT) as a measure for

time of occurrence of an imbalance. They applied principal component analysis (PCA) and assumed that well separating groups of imbalancies belong to distinct pathways of progression of bladder cancer. In contrast to this discriminative approach, my method provides an explicit and quantitative description of the relations of genetic events. Also, it allows assessing the probabilities of abnormalities along the oncogenic pathways.

In agreement with several other publications our data suggest that most UCs arise after alterations of the chromosome 9p (*CDKN2A*) region. This change occurs frequently together with LOH of three tumor gene regions at chromosome 9q (*PTCH, DBRCC1, TSC1*) due to loss of the entire chromosome 9. The loss of chromosome 8p and 17p significantly contributes to the formation of tumor phenotype and is followed by other genetic abnormalities. The loss of 17p relies on the genetic changes of chromosome 9 and 8p. There are two genetic pathways going through 8p and 17p. The losses of 18q, 10q are likely to be related only to the 8p pathway. The loss of 1q region is also not involved in the 17p pathway. Losses of 2q, 5p/5q, 14q, 3p, 13q, 6q, 16q regions are late events in the progression of cancer that occur in both pathways going through 8p and 17p. The group of strongly related losses 18q, 11p, and 11q probably indicates a certain genetic subgroup of bladder cancer. Another group of closely associated genetic events comprise losses of 5p, 14q, 13q and 3p.

Schäffer *et al.* (2001) have found the strong dependency between loss in 8p and loss in 17p but they were not able to distinguish between two pathways going through 17p or not. Notably, this is due to the nature of their mathematical tree model and pairwise dependencies they consider. In contrast, my model captures more complex dependencies between events based on parent-child configurations and thus is able to reveal heterogeneous ways of tumor progression.

The techniques applied to detect genetic alterations in tumor cells may also have influenced these results. Karyotyping and CGH give an overview of gross genomic alterations, but the former may lead to several in vitro artefacts and both methods can detect only genetic alterations larger than 10 Mb (million base pairs of DNA) (Schäffer *et al.* 2001, Höglund *et al.* 2001). In the dataset I have worked on, specific chromosomal sites have been analysed by microsatellite deletion mapping, and narrowed down alterations at chromosome 2q, 5q, 8p, 9p, 9q and 11p by saturation of these regions with several microsatellites. Specific genetic changes in tumors could be detected at higher percentage than others by CGH or karyotyping. For example, cytogenetic as well as CGH studies found loss of chromosome 9 or 9p in approx. 50-55% of the UCs whereas in our dataset allelic changes at the *CDKN2A, PTCH, DBRC1* and *TSC1* genes on the short and long arms of chromosome 9 including small hemi- or homozygous deletion at the *CDKN2A* were detected in 80% of the tumors.

I have presented a new approach for analysing allelotyping data. The method is based on learning Bayesian network model from data. My approach provides an explicit and quantitative description of the relations of genetic events. I were not only able to demonstrate the dependencies between allelic losses based on their pairwise correlations, but rather to uncover multivariate dependencies reflecting the heterogeneous pathways of tumor progression. The Bayesian network analysis suggested primary events that initiate the accumulation of abnormalities and late events, accumulated at the late stages of cancer. It revealed interesting patterns of allelic losses. The analysis enabled to suggest the possible order of allelic losses during oncogenesis. The discovered "high-level" genetic information can give an insight into the underlying mechanisms of unnormal gene regulation resulting in cancer formation.
The presented approach is the further step on the way of mathematical modelling of tumorigenesis.

2.7 Future work

2.7.1 Applying Bayesian network analysis to the comparative genome hybridization data (CGH)

The approach I have developed and tested on the allelotyping data can be readily applied to the data obtained with another kind of molecular cytogenetics experiments, e.g. comparative genomic hybridization (CGH).
CGH is a method for screening a tumor genome for chromosomal imbalances like gains and losses of chromosomal regions by means of a single experiment. The genomic DNA from tumor cells is hybridized with the chromosomes from a normal cell and is detected, for example, through a green fluorescent dye. Simultaneously, the DNA from normal cells is co-hybridized and is detected by means of a red fluorescent dye. After an overlapping of the signals, mixed colors are formed. Chromosome regions that are overrepresented in tumor appear in green, while underrepresented regions in red. Balanced chromosome regions appear in yellow. Image processing techniques permit to obtain the description of the whole tumor genome profile. CGH does not detect balanced chromosomal translocations or inversion, and it is particularly difficult to analyse small interstitial deletions by CGH.
We are currently working on the analysis of the CGH data of gastrointestinal stromal tumors (GISTs).
The data from the CGH experiments is conventionally notated by the biologists in a special notation ISCN (The International System for Human Cytogenetic Nomenclature 1995). One of my projects was to develop a software tool for parsing the cytogenetic data (see the list of my publications).

The concept of the ISCN-Parser, briefly, is that it can interpret the ISCN as a formal language. I have developed the grammar rules for the ISCN in Extended Backus-Naur Form (EBNF).

Chapter 3

Inferring genetic regulatory pathways from expression data

3.1 Biological motivation

As already underlined in Introduction, we are now entering the post-genomic era. The main focus in genomic research switches from sequencing to using the genome sequences, in order to understand how genomes are functioning. The advent of microarray technology is a revolutionary milestone in biological research. The microarray technology allows to measure the expression levels of thousands of genes simultaneously. This provide a "snapshot" of the cell's transcriptions as the genes change over time and react to external stimuli. The term "transcriptome" has been introduced to refer to this new type of data, comprising the expression levels of all genes of a genome in a given regulatory state of a cell. The gene expression differs in various cell types, tissues, in various cell-cycle or developmental stages and under different conditions like compound treatment or disease. These different expression patterns are achieved as the result of the complex process of genetic regulation. Knowing the gene transcript abundance gives a hope to answer the questions arising:

- how do genes and gene products interact;
- how is the gene expression regulated and controlled;
- in what pathways or cellular processes the genes participate;
- what is the functional role of different genes.

To decipher the mechanisms of transcriptional regulatory machinery is the great challenge of functional genomics.

The transcription of the genes is controlled by multiple transcription factors

or signalling molecules that are the gene products themselves. Genes can be thought of as information processing units "wired" into the regulatory network. Reconstructing the gene regulatory network from gene expression data (reverse engineering of genetic networks) is an area of an active research. I introduce this area in the section 3.4.
In the next section I will make a short introduction into the microarray technology, and present briefly the typical method of microarray data analysis.

3.2 Microarray technology and expression data analysis

In the process of gene transcription, messenger RNA (*mRNA*) is being constructed based on the gene-coding sequence. The mRNA leaves the cell nucleus and, in the surrounding cytoplasm, each mRNA molecule conducts the synthesis of the particular protein encoded. Presence and amount of a particular mRNA regulates the presence and the amount of the encoded protein. The level of mRNA can be measured in parallel for thousands of genes by microarray technology.
RNA is prepaired from the cells and is reversely transcribed into more stable DNA (called *cDNA*). The samples are labeled with distinct fluorescent (or radioactive) dyes. The labeled DNA is then applied to a microarray. The microarray consists of a solid support material (nylon, polypropylene or glass), onto which DNA fragments of different sequences, representing genes, have been spotted. The roboter facilities make it possible to spot tens thousands of DNA fragments onto microarray. For DNA of the same kind, complementary single strands will bind (*hybridize*), resulting in double stranded DNA. A laser-scanning microscope reads the microarray, and image analysis programs are used to determine spot intensities, that is, to measure the amount of label for each spot. The level of labeled DNA bound to a particular spot will correspond to the level of the particular kind of mRNA in the cells. In differential experiments, researchers label two samples with fluorescent dyes of different colors (usually red for a sample, and green for the reference or control populations). This allows one to determine the relative amount of transcript present in the pool by the relative intensities of the fluorescent signals generated at each spot. After scanning, if the amount of RNA expressed by a gene being studied exceeds that of the reference sample, the spot turns red. If it is less than the reference sample, the spot turns green.
The data produced by image analysis is usually a matrix, with rows corresponding to genes, and columns corresponding to different experimental conditions or different time points. Each row is the gene expression pattern of a particular gene across all conditions characterizing the dynamic func-

tioning of each gene in the genome (*gene expression profile*). Each column is called the *profile of the condition.*

Microarray technology provides insight into the transcriptional status of the cell, measuring RNA levels for thousands of genes at once. The benefit of this technology lies in its broad spectrum of applications. The applications range from the study of organisms with a particular gene inactivated (*knockout* mutants) to the investigation of the adaptation of cells to different environmental conditions (*time-series* experiments). In the time-course analysis, snapshots of the entire transcriptome are taken at successive time points after inducing a change in the regulatory state of a cell culture or a tissue. Microarrays can be used to study the differential gene expression in tumor samples, in the cells after the treatment with different chemical substances or drugs. They can be used in drug discovery, in pharmaceutical and medical studies, for diagnostic purposes.

Large data sets of gene expression have been collected for model organisms such as yeast *Saccharomyces cerevisiae.* To mention also is the "compendium" of expression profiles corresponding to 300 diverse mutations and chemical treatments of the yeast (Hughes *et al.* 2000, Gasch *et al.* 2000). Chu *et al.* (1998) studied the genes involved in completion of the sporulation program of the yeast. J.deRisi *et al.* (1997) produced a 7-point time-series dataset for each gene of the yeast envolved in the diauxic shift from sugar metabolism to ethanol metabolism. Spellman *et al.* (1998) was the first to provide the biological community with the comprehensive gene expression data of cell cycle-regulated genes. Cho *et al.* (1998) produced 17-point time series data characterizing the cell cycle of the yeast monitoring transcripts of almost all of the 6000 genes.

While studying the temporal responses of gene expression patterns during the development or due to perturbations, one of the difficulties is to determine, what is the proper time step across which experiments need to be acquired and interpreted. While producing "perturbed" expression profiles, it is unknown how many perturbations will be necessary to capture sufficient diversity of gene control mechanism.

An initial and broadly used approach for analyzing gene expression data obtained with microarray experiments is *clustering* (see Eisen *et al.* 1998), that is the detection of groups of genes that exhibit similar expression patterns. The goal is to partition the elements into clusters, so that elements in the same cluster are highly similar to each other, and elements from different clusters have low similarity to each other. A measure of similarity is defined between pairs of vectors (expression profiles). This approach relies on the biological hypotheses that genes with correlated expression changes are likely to be regulated by common transcription factors, and might be involved in similar functions or cellular processes. Gene expression clustering allows also not to focuse on individual gene, but to handle data in a global fashion. In many experimental settings, this approach was demonstrated to be useful for summarizing data and identifying common data patterns. Figure 3.1 displays hierarchically clustered gene expression profiles of human foreskin fibroblasts infected with toxoplasma (see Blader *et al.* 2001). In Figure 3.2 the hierarchically clustered gene expression profiles of the yeast cell cycle-regulated genes are displayed (see Spellman *et al.* 1998).

Clustering is a rather crude method, as it is based on pairwise comparisons.

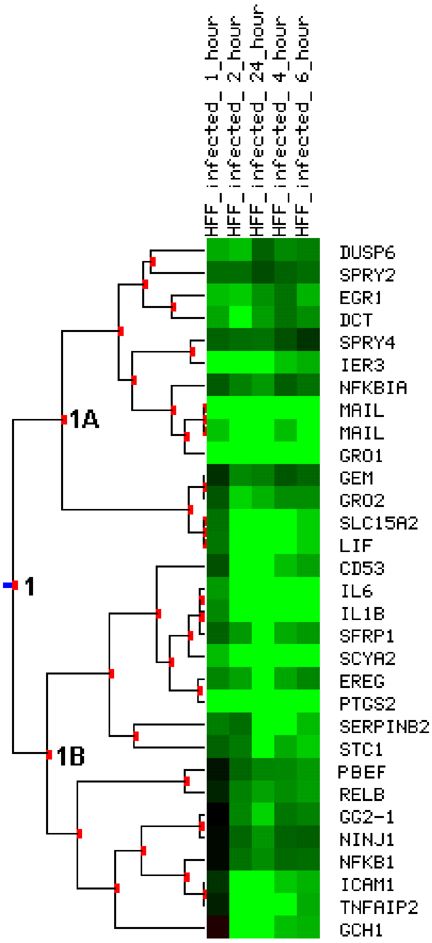

Figure 3.1: Hierarchical clusters of the gene expression profiles (Blader *et al.* 2001).

In general, co-expression does not imply co-regulation. The genes assigned to one cluster by clustering analysis might not share common regulatory region, they might in fact be secondary response genes, and often could belong to different regulatory or signalling pathways. Deeper inference of gene relations at a higher level of complexity is required. Beyond cluster analysis lies the more ambitious challenge: the reconstruction of the underlying gene

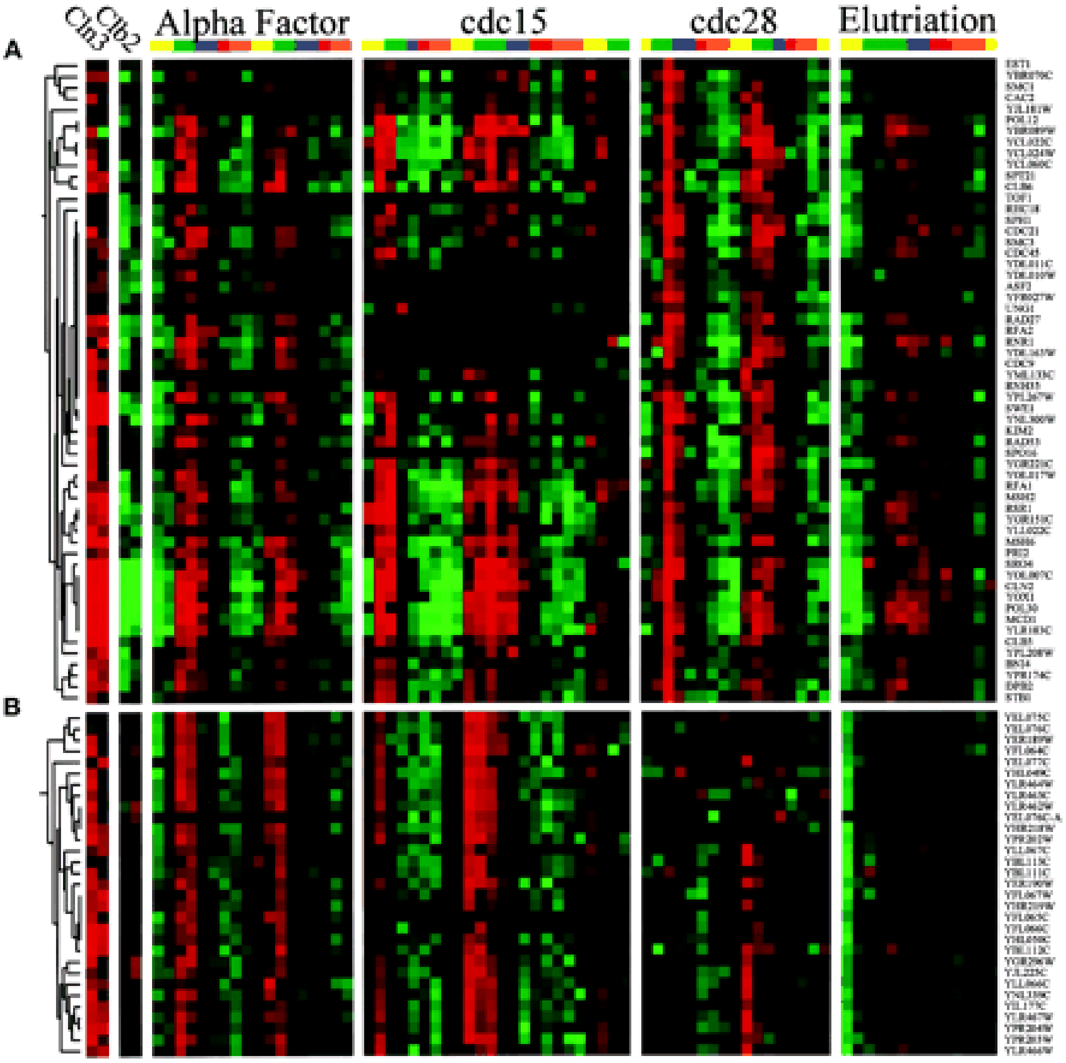

Figure 3.2: Hierarchical clusters of the gene expression profiles (Spellman *et al.* 1998).

regulatory interactions from expression data.

3.3 Transcriptional regulation

In this section I will give a short introduction into the general mechanisms of transcriptional regulation in the cell.
There are several levels at which a cell can control which proteins are expressed at a given time. The most important is transcriptional control, that is, the control of when and how often a gene is transcribed. Other controls exist at the level of RNA processing (such as alternate splicing), RNA transport, RNA translation, RNA degradation, and at the post-translational level such as protein phosphorylation, inactivation and compartmentalization.
In order the transcription of a gene can be initiated, the RNA polymerase must recognize the *promoter*, a sequence of nucleotides preceding the actual coding sequence of a gene. There is also a region of DNA upstream of the gene comprizing one or more *binding sites*, to which distinct regulatory proteins, known as *transcription factors* or *signalling molecules*, are capable to bind. This sequence of DNA, called *operon*, may be subject to negative or positive control through *repressor* or *activator* proteins, respectively. Positive control occurs when an activator protein binds to a DNA site near the promoter. The activator recruits the RNA polymerase to promoter, increasing the transcription of the gene from its low "basal" level to the tens-fold higher level (*regulated recruitment*). In negative control, a repressor protein binds to a DNA site preventing RNA polymerase to assess the promoter and to start transcription. Binding sites of genes to which regulatory proteins bind are called *cis-regulatory elements.* Complexes of regulatory proteins are called *regulatory modules.*
An example of a transcriptional regulatory module is a *"lac operon"* that regulates the lacZ gene in *E.coli.* It was first investigated by Jacob and Monod, 1961. The product of the gene lacZ cleaves lactose, therefore the gene is transcribed if, and only if, lactose is present in the medium. The activator CAP (*catabolite activator protein*) recruits the polymerase to promoter increasing the transcription of the gene from its "basal" level to the 40-fold higher level. In the absence of lactose Lac-repressor binds to the operon excluding the polymerase from binding, whether or not the activator is present. There are also other activators capable to influence the transcription of the lacZ gene: cAMP, CORE, σ (see Figure 3.3). Another example is the transcriptional regulation of the Gal4-gene in yeast. The Gal4-activator is inhibited by Gal80. In the presence of galactose, a protein Gal3 binds Gal80, an interaction that frees Gal4's activating region and the gene is transcribed. In higher organisms, the process of regulating transcription is more complicated. Activators and repressors are capable of influencing transcription from thou-

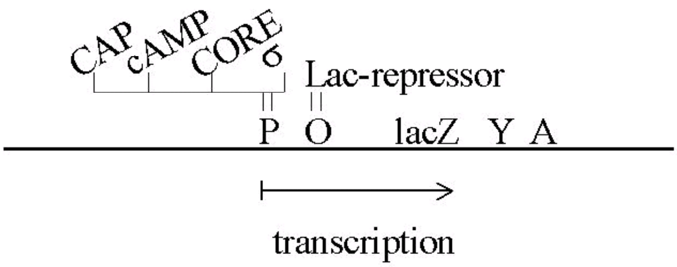

Figure 3.3: Transcriptional regulation of the genes lacZ, Y and A in *E.coli* by the set of activators and the repressor.

sands of nucleotide base pairs away from the start site of the transcription. Further examples of the principles of genetic regulation and control can be found in Ptashne and Gann (1998).

The regulatory control is provided by cooperative binding of multiple proteins.

The regulatory proteins usually act as complexes: the simultaneous interaction of two or more factors result in a high level of transcriptional activation. A gene can have one or more activators which are necessary to start the transcription. A gene can have one or more inhibitors which prevent the expression of a particular gene, even in the presence of an apropriate activator. In some cases two or more regulatory proteins, independently binding to DNA, synergistically activate transcription. In some cases competition for overlapping sites leads to a mutually exclusive binding. In other cases, regulatory proteins can bind to DNA simultaneously, but binding of a repressor "masks" an activation domain of an activator.

Combinatorial nature of transcriptional regulation is one of the fundamental principles of genome functioning. The genomic regulatory code in this way enables each gene to be expressed specifically only in those cell types in which the appropriate combinations of transcription factors is present. This enables the gene to respond to all of the ambient situations to be encountered throughout the life cycle. The transcription factors composing a regulatory element can have different functional features: they can be tissue-, cell type- and stage-specific, or cell-cycle dependent. In this way the specific spatio-temporal gene expression profiles are achieved.

Because of the combinatorial nature of gene interaction, in order to correctly infer the regulation of a single gene, one need to observe the expression of that gene under many different combinations of expression levels of its regulatory inputs. This implies a wide variety of different environmental conditions and perturbations, and thus an enormous experimental effort. This is where the computer science comes to help the biological research.

The principles of the transcriptional regulation may be captured by computational models. The cis-regulatory regions can be seen as information processing elements. Each element acts as a conditional logic gate, the output of which depends on its various inputs. Each logic gate realizes a logical (boolean) function prescribing the output of the gate. There are genes with regulatory elements demanding that two or more (according the biological evidence, up to 8) transcription factors must all be bound to activate the gene ("AND" logic). There are genes that can be activated by one of a few different possible transcription factors ("OR" logic). The transcriptional activation of some genes may be inhibited by one of a few possible repressor proteins ("NOT OR" logic, we denote this by "NOR"). There are genes that can be inhibited by a synergistic effect of some repressor proteins ("NOT AND" logic, - "NAND"). In case of "OR-NOR" logic, a gene is regulated by a set of possible activators, combined with "OR"-function, and a set of possible inhibitors, also combined with "OR"-function. The gene is transcribed if and only if one of its possible activators is active, and it is not repressed by one of its possible repressors. The "OR-NAND" logic implies that the gene is regulated by a set of activators, combined with "OR"-function, and a set of inhibitors combined with "AND"-function. The genes regulatory interactions can be presented by means of logic gates as shown in Figure 3.4.

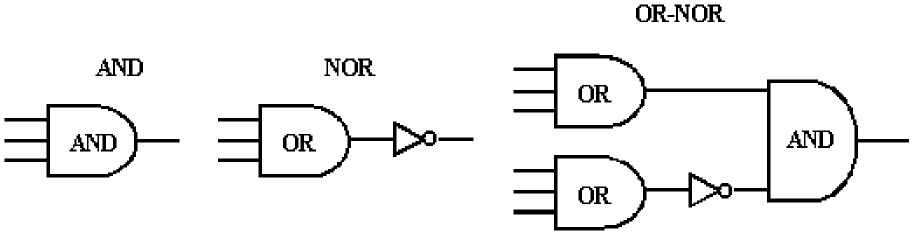

Figure 3.4: Genetic regulatory functions represented as logic gates.

To decipher the mechanisms of transcriptional regulation is the great challenge of functional genomics. In the past, the investigation of transcriptional control was based on the statistical analysis of gene regulatory regions. Structurally similar regulatory elements and modules are present in several different genes, which probably implies that they are functionally significant. The problem is that such an analysis brings too much "false positives", since the regulatory elements have short sequences relative to the promoters (regulatory regions) of the genes.

The advent of high-throughput mRNA quantitation technologies like microarrays enable to obtain "snapshots" of gene expression under different conditions and over time. This data can be analysed to reveal the possible regulatory interactions of the genes. The task can be defined as follows: for each gene identify a set of candidate regulatory genes. I address this problem as the task of inferring the genes interaction governed by a particular logical function.

3.4 Previous approaches for modelling genetic regulatory interactions

The system of gene interactions has the architecture of the network. A gene regulatory network defines the complicated structure of gene products which activate/inhibit other genes.

There have been a number of attempts to model the mechanism of gene regulation and to reconstruct the genetic regulatory network from data. The mathematical and computational approaches include linear differential equations models (D'Haeseleer, 1999, Chen *et al.*, 1999, van Someren *et al.*, 2000), Boolean networks (Somogyi, Sniegosky, 1996; Liang *et al.*, 1998), recurrent neural networks (Wahde and Hertz, 2000), Bayesian networks (Friedman *et al.* 2000), qualitative modelling (Akutsu *et al.* 2000, deJong 2002) and biochemical models (Arkin *et al.*, 1998).

Boolean network model was originally proposed and theoretically investigated by Kauffman (Kauffman, 1993). A gene is modelled as a binary variable which has value 1, if the gene is active, i.e. transcribed (sometimes the gene is referred to be "ON"). The variable has value 0, if the gene is silent, i.e. not transcribed (the gene is "OFF"). In some context, for example, when the amount of transcription is being measured relative to control sample, it is appropriate to say that the gene is overexpressed or underexpressed. In the microarray experiments the continuous measurements of the gene transcription are obtained. Working with discrete models, the continuous domain need to be appropriately discretized into discrete values. Usually, two states of gene expression 0 and 1 depend on whether the expression level is significantly lower or significantly greater than the respective control. These two states reflect the situations of underexpression or overexpression of a gene, respectively.

In Boolean network model, the expression state of each gene is functionally related to the expression states of some other genes using logical rules. Boolean networks is a biologically plausible conceptual framework for representing genetic regulation. In the previous section it was demonstrated that many simple Boolean operators are present in real transcriptional regulatory modules.

The problem of inferring the logical rules in Boolean networks from experimental data (Reverse Engineering of Genetic Networks) was considered by Somogyi, Sniegosky (1996), Liang (1998). They have developed an algorithm

REVEAL which, shortly, proceeds as follows. For a particular gene Y, a minimal set of genes $\{X_1, \ldots, X_n\}$ will be identified that has the same Shannon's entropy as the set $\{Y, X_1, \ldots, X_n\}$:

$$H(X_1, \ldots, X_n) = H(Y, X_1, \ldots, X_n).$$

Once the input elements have been identified, it remains to find the logical rules that specify the state of Y given the combination of input states for the set. This is done by examining the tables of the input values and the values of Y. The problem of the REVEAL approach was, that the computational expense of determining Shannon's entropy as well as of rules governing the table relationships increased with the *in-degree* of a gene Y (the number of genes in the input set). It was needed to limit the in-degrees of genes, i.e. the connectivity of the genetic network. Although this modelling assumption is biologically justified, since the genes are believed to be influenced by no more than 8 other genes, the biologically interpretable results were obtained only with the in-degree n = 3.

A data-driven approach to the reconstruction of genetic regulatory interactions is particularly difficult because of the combinatorial nature of the problem. Another severe obstacle is the so called dimensionality problem: there are too many variables (many genes) and too few conditions where the gene transcription is measured.

The major limitation of standard Boolean networks is their inherent determinism, contradicting with the stochastic nature of the biological process of gene transcription and with the noisy character of the experimental measurements of mRNA. The Boolean networks demonstrated the drastic drop in performance, when even slight amount of measurement noise was introduced. To be able to reliably extract the complex gene interactions, the modelling systems must be robust against noise.

Friedman *et al.* (2000), Linial and Pe'er (2000) proposed to employ Bayesian networks. Bayesian networks as a modelling tool for the genetic regulatory network have certain advantages because:

- Bayesian networks are stochastic models, i.e. they deal with probability as a mean to express uncertainty about the modelling variables and their dependencies.
- they describe global processes as composed of locally interacted components;
- these local interactions are being described probabilistically;
- there are statistically based foundations for learning Bayesian networks from observations.

In Bayesian network model each gene is considered a random variable and is represented as a node. Assume, gene A is a transcription factor of gene B. Measuring high expression levels of gene A might imply that gene B is overexpressed as well. Alternatively, if gene A is an inhibitor of gene B, overexpression of A likely implies underexpression of B. The levels of expression of gene A and B are dependent.

As already introduced in Section 2.4, learning a Bayesian network consists of induction of the graphical structure of the model and, when the model structure is known, of estimation the parameters (i.e. the conditional probability distributions). As described in Sections 2.3 and 2.4, foundations for learning Bayesian networks from data were developed based on Bayesian statistics (K2 algorithm of Cooper and Herskovits, 1992).

Generally, a notion, that facilitated the Bayesian networks employment, was a notion of *causal independence*: given the value of its parents, the variable is causally independent of other variables in the network except its descendants. Causal independence allows for compact representations of probabilistic relationship among variables in the network via the conditional probability tables. But still the Bayesian network formalism offers too much freedom, modelling *arbitrary* interactions between parents $X_1, \ldots, X_n$ of a node Y. This leads to high computational costs and decrease in parameters reliability while learning the model from data. Besides this, it is highly difficult to infer the regulatory relations from conditional probability tables obtained by learning Bayesian network. I propose a model for the genetic regulatory interactions that combine the simple and biologically motivated boolean logic semantics of Boolean networks and the possibility of dealing with uncertainty offered by Bayesian networks. In contrast to Bayesian networks, the parents interactions of variables in the model is defined with logical functions, clearly describing the art of gene regulation.

I address the problem of reconstructing genetic regulatory pathways from data as follows: for each gene and a certain logical function, identify a set of its candidate regulatory genes.

In Section 3.5 I introduce my model of genetic regulatory interactions which originates from the field of probabilistic graphical models. I have developed an approach for learning the structure and parameters of the model from data. I employed the methodology of Bayesian model selection which was introduced in Section 2.3. I describe my method in 3.6. Since there is no closed form solution for the problem of calculating the posterior distribution of a candidate model given data, the central quantity in Bayesian model selection, I turn to the Markov Chain Monte Carlo simulation technique, in particular, to Gibbs sampling. I introduced an additional parameter into the model, so that the problem of model selection transformed into the variable selection task. I review briefly the different Gibbs sampling algorithms developed for solving the problem of variable selection, and introduce my model definition for the Gibbs sampling in the Section 3.6.3. I also consider issues of checking a convergence of Markov chain and checking goodness-of-fit of a model in Section 3.6.5. I tested my approach on the dataset of budding yeast cell cycle (Spellman, 1998) and present the results in Section 3.7. I address previous models for genetic regulatory networks, relevant to ours in that they attempt to insert "noise" into the Boolean networks models, and discuss the perculiarities of my method in Section 3.8.

3.5 The model of genetic interactions

Bayesian network formalism exploits independencies among variables in the network and achieves more compact representations of the joint probability distribution of the variables by expressing them with conditional probability tables. The Bayesian network formalism allows for modelling arbitrary interactions between parents $X_1, \ldots, X_n$ of a node Y. The CPT expresses the multinomial distribution. Such a modelling freedom has its price. In this representation, the complete CPT for a binary variable with n parents requires the specification of $2^n - 1$ independent parameters (one parameter for each parents' state configuration). The number of parameters associated with a variable is exponential in the number of parents of the variable. This exponential explosion of the parameter space makes the learning of the network model computationally expensive. Conditional probability distributions are obtained from relative counts of various outcomes in those data cases, that fulfill the conditions described by a given combination of the outcomes of the parents. In small datasets there might be no sufficient cases for learning conditional probabilities. Learning distributions with fewer parameters is more reliable. Besides the computational problems, the general, combinatorial semantics of the parents interaction in Bayesian network make it difficult to interpret the results of Bayesian network learning and to read out the "true" functional relationships among the variables covered in this presentation.
Because of the limitations mentioned, there has been interest in learning Bayesian network models with more parsimonious representations for the conditional probability of variables given their parents (*local structure*), allowing for reductions in the dimension of the parameter space and, hence, in learning effort. Friedman and Goldszmidt (1996 a) used decision trees.
One can further exploit the independencies between parents of a variable in a Bayesian network to get more compact representations of CPTs. In the past, there were models introduced with special types of causal interaction (see Heckerman and Breese, 1994, Meek and Heckerman, 1997, Srinivas, 1993). One kind of such models is the *causal independence model* which uses the notion of independence of parents of each variable in the model. The variables $X_1, \ldots, X_n$, which are parents of the variable Y, can effect Y

through independent "mechanisms". The results of these effects are combined by a rule represented with a boolean-logic function. Such models were introduced originally by J.Pearl (1998) and called "noisy OR-Gate" ("noisy AND-Gate").
I employ such kind of models for modelling the genetic regulatory interactions. I assume that the variable X_i (regulator) can execute its influence on the variable Y (regulatee) independently of other possible regulators $X_1, \ldots, X_n$ of the gene Y. The biological mechanism underlying this modelling assumption is the binding of protein transcribed by the regulator to the DNA of the regulatee. This process is not deterministic, rather each gene X_i can regulate the gene Y with probability θ_i and can fail to do this with probability $1-\theta_i$. The general structure of the genes interaction in my model is represented as a directed graph as shown in Figure 3.5. In this graphical representation intermediate variables $I_1, \ldots, I_n$ are introduced, through which the variables $X_1, \ldots, X_n$ execute their influence on a given common effect variable Y. Each intermediate variable I_i has only one parent, a variable X_i. Its probability distribution is defined as follows: given that $X_i = 1$, I_i takes the value 1 with probability θ_i and the value 0 with probability $1 - \theta_i$, respectively. Given that $X_i = 0$, I_i takes the value 0 with probability 1. The combined regulatory influence on the the variable Y is calculated as the boolean function F on the input variables $I_1, \ldots, I_n$. If $X_1, \ldots, X_n$ are activators, then the state of the variable Y is $F(I_1, \ldots, I_n)$; if $X_1, \ldots, X_n$ are inhibitors, the state of Y is $1 - F(I_1, \ldots, I_n)$. The boolean function F ("interaction function") defines in which way the intermediate effects I_i, and indirectly, the variables X_i interact. I consider two interaction functions: "AND" and "OR". The semantics of the "OR"-function, for example, implies that the variables X_i are each assumed to be sufficient to influence Y. In the case of "AND"-function the variables X_i all need to execute their own

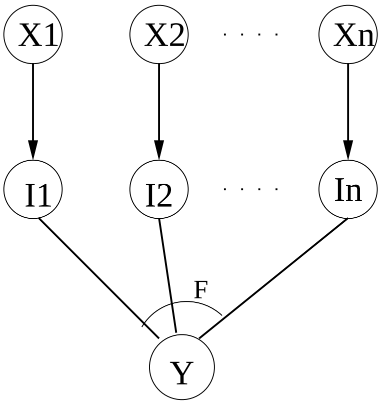

Figure 3.5: Model of the gene interactions, F - Boolean function

Y			
X1	X2	0	1
0	0	1	0
1	0	1	0
0	1	1	0
1	1	$1-\theta_1\theta_2$	$\theta_1\theta_2$

Table 3.1: Conditional probability table of regulatee Y, ”AND”-regulation

Y			
X1	X2	0	1
0	0	1	0
1	0	$1-\theta_1$	θ_1
0	1	$1-\theta_2$	θ_2
1	1	$(1-\theta_1)(1-\theta_2)$	$1-(1-\theta_1)(1-\theta_2)$

Table 3.2: Conditional probability table of regulatee Y, ”OR”-regulation

influence on the variable Y, so that Y will be active.

Introduction of the hidden state variables I_i allows to insert ”noise” into the Boolean-logic based models. It enables to express that the biological mechanism of the regulation of one gene by another could be inhibited by unknown reasons. Thus, the input variables can be considered as observables from which we make our noisy measurements and the hidden variables have the ”true” latent biological values.

The conditional probability distribution for the regulatee Y that is activated by the combined action of two regulators (”AND”-regulation) is presented in Table 3.1. If none or only one of the regulators X1 and X2 is active, the probability of the regulatee Y to be active is 0. The regulatee Y takes the value 1 with probability $\theta_1\theta_2$, i.e. only in the case when both X1 and X2 are active. The regulator X1 (X2) is active and executes its influence on Y with probability θ_1 (θ_2). Table 3.2 represents the case when the regulatee Y can be activated by two possible activators (”OR”-regulation). In this case, the regulatee Y can be activated by X1 or X2 with probability

Y			
X1	X2	0	1
0	0	0	1
1	0	θ_1	$1-\theta_1$
0	1	θ_2	$1-\theta_2$
1	1	$1-(1-\theta_1)(1-\theta_2)$	$(1-\theta_1)(1-\theta_2)$

Table 3.3: Conditional probability table of regulatee Y, ”NOR”-regulation

θ_1 or θ_2, respectively. If both of regulators are active, the regulatee Y takes the value 0 with probability $(1-\theta_1)(1-\theta_2)$, i.e. when both regulators failed. Table 3.3 represents the repression of Y by two possible repressors ("NOR"-regulation). The regulatee Y takes the value 0 (inactive) with probability not equal to 0, if one or both repressors X1 and X2 are active and have executed their influence on Y. One can see that the values for the conditional probability table of regulatee Y in case of repression can be obtained as one minus the respective values of the conditional probability table of Y in case of activation.

Note that the model with the Boolean logic-based interaction of parent variables allows to specify the entire conditional probability distribution for a variable with only n parameters $\theta_1, \ldots, \theta_n$, hence polynomial on number of parents, in contrast to the general Bayesian network model prescribing the combinatorial interaction of parents, and therefore demanding the exponential number of parameters.

In the present work I consider simple models with activatory regulation ("OR", "AND") and inhibitory regulation ("NOR", "NAND"), as well as complex models: "AND-NAND", "AND-NOR", "OR-NAND" and "OR-NOR". In the complex models the regulatory influences of multiple activators and multiple inhibitors are combined with "AND"-function as presented in Figure 3.6. These models are sufficient to represent the biological mechanism of gene regulation.

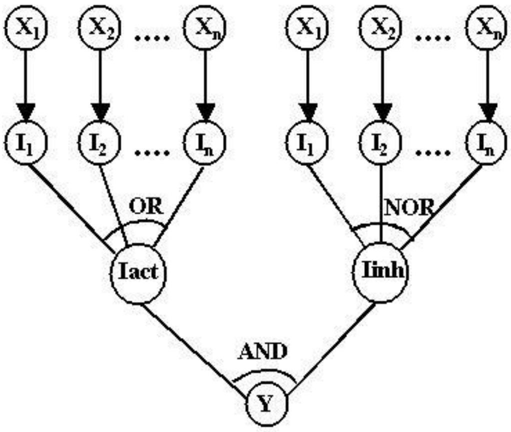

Figure 3.6: Complex model of gene regulatory interactions with activators and inhibitors ("OR-NOR" regulation).

3.6 Bayesian model selection

I employ Bayesian approach to learning the structure and parameters of a model from data (i.e., model selection). As already described in Section 2.3, the Bayesian approach addresses the problem as calculating the posterior probability of a model given data for a collection of candidate models and selecting the most probable model. Bayesian statistics differs from classical statistical theory in that it considers any unknown parameter as random variable and, for this reason, each of the parameters should have a prior distribution. The introduction of a prior converts statistical inference into an application of probabilistic inference based on Bayes' theorem. Suppose that the data D have been generated by a model m, one of a set M of candidate models, $m \in M$. If $p(m)$ is the prior probability of model m, then the posterior model probability by Bayes rule is $p(m|D) \propto p(D|m)p(m)$. The marginal likelihood $p(D|m)$ is calculated as $p(D|m) = \int p(D|m, \theta_m)p(\theta_m|m)d\theta_m$, where $p(\theta_m|m)$ is the prior distribution of model parameters θ_m for model m. The calculation of the marginal likelihood is the major computational bottleneck of Bayesian methods since the integral is analytically tractable only in certain restricted examples, when there exists a conjugate prior distribution for the parameters of the model, so that the integral will have a closed form solution. Let us consider the "local" model of genetic interactions. Assume the variable Y is commonly influenced by the variables $X_1, \ldots, X_n$. The probability distribution of Y given the values of its parents in the model with "OR"-activation function can be written as:

$$P(Y = 0|\theta) = \prod_{i=1}^{n}(1 - \theta_i)^{X_i}$$

and

$$P(Y = 1|\theta) = 1 - \prod_{i=1}^{n}(1 - \theta_i)^{X_i}.$$

Assume we have a sample of N cases corresponding to the states of the variables $X_1, \ldots, X_n$ and the variable Y. Denote with Y_j the state of the variable Y in case j, and with X_{ij} the state of the variable X_i in case j. The likelihood function is then:

$$L(\theta) = \prod_{j=1}^{N}(\prod_{i=1}^{n}(1 - \theta_{ij})^{X_{ij}})^{1-Y_j}(1 - \prod_{i=1}^{n}(1 - \theta_{ij})^{X_{ij}})^{Y_j},$$

If we substitute ψ_{ij} by $-log(1 - \theta_{ij})$, then

$$\prod_{i=1}^{n}(1 - \theta_{ij})^{X_{ij}} = e^{-\sum_{i=1}^{n}\psi_{ij}X_{ij}}.$$

Denote $\eta_j = \sum_{i=1}^{n} \psi_{ij} X_{ij}$ (i.e. *linear predictor*). Then the likelihood function transforms into:

$$L(\psi) = \prod_{j=1}^{N} (e^{-\eta_j})^{1-Y_j} (1 - e^{-\eta_j})^{Y_j}.$$

This is the *generalized linear model* (see McCullagh and Nelder 1983, Myers *et al.* 2002). There is no conjugate prior for such model, because it does not have the form of the general exponential family parametric models (for introduction into conjugate analysis see Bernardo and Smith, 1994). The "AND" model as well as complex models like "OR-NOR" etc. are intractable analogously.

The "global" model with such "local" structures as presented will not have the closed form solution as well. Thus, we will not have a global criteria like e.g. Bayesian scoring for the estimation of the posterior probability of the whole genetic network.

Possible solutions of evaluating multi-dimensional integrals involved in the Bayesian methods are asymptotic approximations (see Lindley 1980, Kass 1988). The asymptotic approximations are based on the observation that, as the number of cases in the dataset increases, the posterior on the parameters will be strongly peaked and can be approximated with a multivariate-Gaussian (normal) distribution. For relatively small datasets, as the biological datasets generally are, the assumption of asymptotic normality might be inaccurate (see Berger 1993). Another approach for the estimation of posterior probabilities involved in the Bayesian modelling are the variational methods (see Jaakkola and Jordan 1996). The authors propose the lower and upper bounds approximations of the posterior distributions of the generalized linear models.

Instead of using approximation techniques, in view of the lack of empirical data, I turn to the stochastic simulation techniques - Markov Chain Monte Carlo (MCMC) methods, recently broadly used in many Bayesian modelling applications (see Gilks 1996). MCMC techniques generate samples from the joint posterior distribution of the unknown quantities in a model allowing to make estimates on them. MCMC sampling from the joint posterior distribution $p(m, \theta_m|D)$ allows to estimate the posterior model probability $p(m|D)$ and $p(\theta_m|D)$.

I perform MCMC simulations for selecting the local model of interactions of one gene with its regulators. In the following, m stays for the local model introduced in this section. In the next subsection I provide a brief introduction into Markov Chain Monte Carlo methodology.

3.6.1 Markov Chain Monte Carlo

Markov Chain Monte Carlo methods were first developed for applications in statistical physics, and were used in spatial statistics and image analysis (Metropolis *et al.* 1953, Hastings 1970). In recent years MCMC methods have had a great effect on Bayesian statistics (Gilks 1996). Bayesian framework allows for the flexible specification of complex models. However, this comes with a certain price. The probability distributions arising in the Bayesian modelling can be very complex, with probabilities varying over a high-dimensional space. Bayesians need to integrate over these high-dimensional probability distributions to make inference about model parameters or to make predictions. Often, however, a sample of points drawn from such a distribution can provide a satisfactory picture of it.

Markov Chain Monte Carlo draws a sample of points from a required distribution, and then calculates sample averages to obtain expectations of various functions of the variables.

Suppose $X = X_1, \ldots, X_n$ is the set of random variables taking on values $x_1, \ldots, x_n$. These variables might be, for example, parameters of the model. The *expectation of a function* $a(X_1, \ldots, X_n)$ - it's average value with respect to the distribution over X - can be approximated by

$$\begin{aligned}\langle a \rangle &= \sum_{\widetilde{x}_1} \cdots \sum_{\widetilde{x}_n} a(\widetilde{x}_1, \ldots, \widetilde{x}_n) P(X_1 = \widetilde{x}_1, \ldots, X_n = \widetilde{x}_n) \\ &\approx \frac{1}{N} \sum_{t=0}^{N-1} a(x_1^{(t)}, \ldots, x_n^{(t)}), \qquad (*)\end{aligned}$$

where $x_1^{(t)}, \ldots, x_n^{(t)}$ are the values for the t-th point $X^{(t)}$ in a sample $X^{(0)}, X^{(1)}, \ldots, X^{(N)}$ of size N.

Problems of prediction can be formulated in terms of finding such expectations.

One way of generating $X^{(t)} = X_1^{(t)}, \ldots, X_n^{(t)}$ for the set of variables at step t is through a *Markov chain* having $P()$ as its *stationary distribution.* The Markov chain is defined by giving an initial distribution for $X^{(0)}$ and the transition probabilities for $X^{(t)}$ given the value for $X^{(t-1)}$. These probabilities are chosen so that the distribution of $X^{(t)}$ converges to that for X as t increases. The Markov chain is simulated by sampling from the initial distribution and then, in succession, from the conditional (transition) distributions.

If Markov chain is sufficiently long, the output of it can be used to estimate expectations as defined in equation $(*)$.

Generally, the output of Markov chain simulation is used to summarize the posterior distribution of the variable of interest in terms of means, standard deviations, etc. For example, if $P(x)$ is a Gaussian distribution, then $\langle x \rangle$ is the estimate of the mean of this distribution, and $\langle (x - \langle x \rangle)^2 \rangle$ is the estimate of the variance. The square root of the variance is the standard deviation. Usually, while doing estimations from the Markov Chain Monte Carlo output, a particular number of starting iterations of the Markov chain are being discarded. This is needed for the chain to "forget" its starting values (M iterations is the so called *burn-in time*):

$$\langle a \rangle \approx \frac{1}{N-M} \sum_{t=M+1}^{N} a(x_1^{(t)}, \ldots, x_n^{(t)}).$$

In Bayesian modelling, MCMC methods are used to obtain expectations $\langle p(m|D) \rangle$ and $\langle p(\theta_m|D) \rangle$.
Typically, the Markov chain explores the space in a "local" fashion. In some methods $x^{(t)}$ differs from $x^{(t-1)}$ in only one component of the state, for example, it may differ with respect to x_i for some i, but have $x_j^{(t)} = x_j^{(t-1)}$ for $j \neq i$. Other methods may change all components at once, but usually by only a small amount. Locality is often crucial to the feasibility of these methods. In the Markov chain framework it is possible to guarantee that such step-by-step local methods eventually produce a sample of points from the globally-correct distribution.
One of the MCMC approaches is Gibbs sampling (Geman and Geman 1984). I introduce this method briefly in the next section.

3.6.2 Gibbs sampling

Gibbs sampler reduces the problem of dealing simultaneously with a large number of unknown parameters in a joint distribution into a much simpler problem of dealing with one variable at a time, iteratively sampling each from its full conditional distribution given the current values of all other variables in the model.
Suppose, we wish to sample from the joint distribution for $X = X_1, \ldots, X_n$ given by $P(x_1, \ldots, x_n)$. The Gibbs sampler does this by repeatedly replacing each component with a value generated from its distribution conditional on the current values of all other components. The algorithm proceeds as follows: choose initial values $x_1^{(0)}, \ldots, x_n^{(0)}$ and generate a value $x_1^{(1)}$ from the conditional density

$$p(x_1|x_2^{(0)}, \ldots, x_n^{(0)}).$$

Similarly, generate a value $x_2^{(1)}$ from the conditional density

$$p(x_2|x_1^{(1)}, x_3^{(0)}, \ldots, x_n^{(0)})$$

and continue up to the value $x_n^{(1)}$ from the conditional density

$$p(x_n|x_1^{(1)}, x_2^{(1)}, \ldots, x_{n-1}^{(1)}).$$

With the new realization $X^{(1)}$ of X, the above process is iterated, simulating a homogeneous Markov chain $X^{(0)}, X^{(1)}, X^{(2)}, \ldots, X^{(t)}$. For large t, the state of the Markov chain will converge to the desired distribution, $X^{(t)}$ can be regarded as a simulated observation from P(X).

Performing stochastic simulation, especially Gibbs sampling, is particularly appropriate with a probabilistic graphical model (Pearl, 1987). Due to the factorization of the joint probability distribution, the full conditional for a given node in the DAG involves only a subset of nodes which participate in its Markov blanket (that is the set of node's parents, children and parents of the children). The Gibbs sampler generates samples for unobserved nodes while fixing (or clamping) the observed nodes with the data. From the generated samples one estimates the quantities of interest, e.g. the posterior distributions of the unknown parameters given data.

3.6.3 Gibbs Variable Selection

The "OR" model introduced in Section 3.5 can be written as:

$$Y \sim Bernoulli(1 - \prod_{i=1}^{n}(1-\theta_i)^{X_i})$$

(The operator $\sim$ stands for 'is distributed as'.) Now consider the complex model "OR-NOR". Assume the variable Y is influenced by a set of activators $X_1^{act}, \ldots, X_n^{act}$ and a set of inhibitors $X_1^{inh}, \ldots, X_k^{inh}$. The variable Y takes the value 1, if the activators executed their influence *and* the inhibitors failed, otherwise Y is 0. The "OR-NOR" model then can be defined as:

$$Y \sim Bernoulli((1 - \prod_{i=1}^{n}(1-\theta_{ij}^{act})^{X_{ij}^{act}}) \prod_{i=1}^{k}(1-\theta_{ij}^{inh})^{X_{ij}^{inh}})$$

Consider the "OR" model. Our problem of model selection is formulated as: given the data on the gene Y and its potential regulators $X_1, \ldots, X_p$, identify a subset $X_1, \ldots, X_n$ of actual regulators of Y.

Standard Markov Chain Monte Carlo techniques, such as Gibbs sampler, cannot be directly applied to the problem of model selection because the candidate models have different number of parameters (n). One must take a particular care by setting the probabilities of jumps between different models. Green (1995) have developed the *reversible jump* MCMC algorithm which can account for this.

Another approach for MCMC exploration of the model spaces is "the model composition" approach of Carlin and Chib (1995). The authors introduce the joint model-parameter space, *composite model space*, which is created by considering the product space of the model indexing variable m and parameters from all possible candidate models: $M \times \prod_{k \in M} \Theta_k$, a parameter space for

$(m, \theta_k : k \in M)$. This allows to consider only one joint space of model indexing variable m and model parameters, keeping the dimensionality constant across all possible models. They use Gibbs sampler to generate from the posterior distribution $p(m, \theta_k|D)$. Since the parameter space has changed, a prior distribution for $(m, \theta_k : k \in M)$ is no longer completely specified by $p(m)$ and $p(\theta_m|m)$, rather the use of *pseudopriors* $p(\theta_k|m \neq k), k \in M$ is required. The main drawback of this method is the unavoidable specification of, and generation from, many pseudoprior distributions.

Gibbs sampling approaches applicable for the model selection problems were further considered in the works of George and McCulloch (1996), Kuo and Mallick (1998), and by Dellaportas *et al.* (2000, 2002). It was proposed to substitute the model indicator $m \in M$ with a *variable indicator* γ, binary vector, representing which of the parameters are included or excluded from the model. This allows to consider only one joint space of variable indicators and model parameters, keeping the dimensionality constant across all possible models. By introducing the variable indicator the "OR" model may be written as:

$$Y \sim Bernoulli(1 - \prod_{i=1}^{n}(1 - \theta_i)^{\gamma_i X_i})$$

In this representation model choice problem is referred to as the *variable selection problem.*

Let D denote the observed data (for the variables $X_j, j = 1, \ldots, p$ and Y).

A Bayesian approach to variable selection requires setting up a joint probability distribution (*full probability distribution*) over all the observables and parameters. The sampling procedure samples from the full probability distribution conditional on the observed data, that is $p(\theta, \gamma|D)$. The Gibbs sampling procedure samples successively from univariate conditional distributions, generating a sequence of values

$$\theta^{(0)}, \gamma^{(0)}, \theta^{(1)}, \gamma^{(1)}, \ldots, \theta^{(t)}, \gamma^{(t)}, \ldots$$

which constitutes a Markov chain. The subsequence of values

$$\gamma^{(0)}, \gamma^{(1)}, \ldots, \gamma^{(t)}, \ldots$$

converges to $p(\gamma|D)$. This sequence can be used to identify the high probability values of γ_j. These are the values that appear most frequently in the sequence. (empirical frequency estimate of γ). The sequence of values

$$\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(t)}, \ldots$$

gives an estimate of θ. If the estimate for γ_j is close to 1, the respective X_j should be included in the desirable "true" model, other X_j should be excluded.

Consider a partition of θ into $(\theta_\gamma, \theta_{-\gamma})$ corresponding to those components of θ which are included and not included in the model. Then $p(\theta|\gamma, D)$ may be partitioned into $p(\theta_\gamma|\theta_{-\gamma}, \gamma, D)$ and $p(\theta_{-\gamma}|\theta_\gamma, \gamma, D)$. From the model definition it is obvious that the components of the vector $\theta_{-\gamma}$ do not affect the model likelihood. The full conditional posterior distributions for the sampling procedure are given by:

$$p(\theta_\gamma|\theta_{-\gamma}, \gamma, D) \propto p(D|\theta, \gamma)p(\theta_\gamma|\gamma)p(\theta_{-\gamma}|\theta_\gamma, \gamma),$$

$$p(\theta_{-\gamma}|\theta_\gamma, \gamma, D) \propto p(\theta_{-\gamma}|\theta_\gamma, \gamma),$$

the last relation is true since the components of the vector $\theta_{-\gamma}$ do not affect the model likelihood.

Hence, the Gibbs variable selection procedure requires the specification of the model likelihood $p(D|\theta, \gamma)$, the specification of the model prior $p(\theta_\gamma|\gamma)$ and the *pseudoprior* $p(\theta_{-\gamma}|\theta_\gamma, \gamma)$.

In our model the terms γ_j of the variable indicator γ are independent. Each γ_j can be sampled from a Bernoulli distribution with success probability $O_j/(1 + O_j)$, where

$$O_j = \frac{p(\gamma_j = 1|\gamma_{-j}, \theta, D)}{p(\gamma_j = 0|\gamma_{-j}, \theta, D)} =$$

$$\frac{p(D|\theta, \gamma_j = 1, \gamma_{-j})}{p(D|\theta, \gamma_j = 0, \gamma_{-j})} \frac{p(\theta|\gamma_j = 1, \gamma_{-j})}{p(\theta|\gamma_j = 0, \gamma_{-j})} \frac{p(\gamma_j = 1, \gamma_{-j})}{p(\gamma_j = 0, \gamma_{-j})}$$

The methods on Gibbs variable selection differ in their approaches on specifying the prior distributions on the model parameters. The most simple is the "unconditional prior" approach of Kuo and Mallick where the prior distributions of model parameters θ is defined independent of variable indicator γ.

In the *Stochastic Search Variable Selection* method of George and McCulloch (SSVS), the priors for θ_j depend on γ_j:

$$p(\theta_j|\gamma_j) = \gamma_j p(\theta_j|\gamma_j = 1) + (1 - \gamma_j)p(\theta_j|\gamma_j = 0).$$

By Stochastic Search Variable Selection, the parameters priors are defined as mixtures of Normal distributions for $\gamma_j = 0$ and $\gamma_j = 1$. If $\gamma_j = 0$, the parameters (pseudopriors) are kept close to 0 by defining the mean of the Normal distribution equal 0.

The method of Dellaportas *et al.* (2000, 2002) differs from SSVS in that the pseudopriors may not be distributed around 0, rather they may be chosen in a way to help to increase the efficiency of the sampling procedure. The pseudoprior distribution does not effect the model likelihood and is only relevant to the behaviour of the Markov chain. Efficient performance can be achieved, if the moves between models would be "local" (Dellaportas *et al.*

2000, 2002). In variable selection problems, where the new sampled value of γ differs from the current value in a single component, it is reasonable to retain the parameter values for those terms which are present in both the current and new models. The method of Dellaportas *et al.* (2000, 2002) use the so called *proposal* densities for pseudopriors, which can be estimated using a *pilot run* of the MCMC for the *saturated* model, i.e. the model where all terms $\gamma_j = 1$ for all j. In the present thesis I adopt the method of Dellaportas *et al.* (2000, 2002).

Bayesian modelling allows for the hierarchical formulation of the model: the distributions for the parameters can be formulated, in turn, with the help of hyperparameters. I defined the parameter priors with Beta distribution with hyperparameters a_j and b_j:

$$\theta_i \sim Beta(a_j, b_j).$$

(The notation $\sim$ stands for 'is distributed as'.) The choice of Beta distribution was required because the parameters θ_i had to be constrained to the [0,1]-interval.

The specification of the distributions for the hyperparameters a_j and b_j is further required, namely for the cases when $\gamma_j = 1$ and when $\gamma_j = 0$ (pseudoprior). If $\gamma_j = 1$, I defined the hyperparameters equal to 1, therefore making the prior noninformative: Beta(1,1).

If $\gamma_j = 0$, I used the proposal distribution, following Dellaportas *et al.*. I calculated the hyperparameters a_j and b_j by the *method of moments*:

$$a_j + b_j = \frac{mean_j(1 - mean_j)}{var_j} - 1,$$

$$a_j = (a_j + b_j)mean_j,$$

$$b_j = (a_j + b_j)(1 - mean_j),$$

where $mean_j$ and var_j, the mean and the variance of the parameters θ_j, were estimated by the pilot run of the saturated model.

Next, one must define the prior distribution for the model indicator γ. Since the terms γ_j are independent, the prior can be decomposed into independent Bernoulli distributions for each term: $\gamma_j \sim Bernoulli(\pi_j)$, where π_j is the prior probability to include term j into the model. A simple and popular choice in variable selection problems is the uniform prior on γ, assuming that models are a priori equally probable, i.e $\pi_j = \pi = 0.5$. This prior is noninformative in the sense of favoring all models equally, but is not noninformative with respect to the model size. If p is the number of potential regulators, and n is the number of actual regulators, then $E(n) = 0.5p$ and $var(n) = 0.25p$. For example, if $p = 19$ (as in the test study described bellow), then n lies in the range 5 to 14 with prior probability close to 1, and thus it is possible that sampling procedure will not sample models with less than 5 regulators. This may be crucial for "AND" models, since there might be a sparse number of regulators of a gene combined with AND-function. To favor more parsimonious models, one can set the probability π so to restrict n *a priori* to lie in a short range by setting $E(n)$ and $var(n)$ to the desired values and using:

$$E(n) = \pi * p,$$

$$var(n) = \pi(1-\pi)p.$$

A more flexible approach is to place a hyperprior on π:

$$\pi \sim Beta(\alpha, \beta),$$

then the prior for the number of actual regulators n is Beta-binomial:

$$n \sim Betabin(p, \alpha, \beta)$$

The values for α and β can be choosed by setting $E(n)$ and $var(n)$ to the desired values and solving the following equations (see Kohn *et al.* 2001):

$$p\frac{\alpha}{\alpha+\beta} = E(n)$$

$$\frac{\alpha+1}{\alpha+\beta+1} = \frac{var(n) - E(n)(1-E(n))}{(p-1)E(n)}$$

While performing Gibbs variable selection with the complex models like "OR-NOR" I considered the same set of variables (genes) as potential activators and inhibitors. I used two variables indicators: γ^{act} and γ^{inh}, representing that a particular variable is included in the model as activator or inhibitor, respectively. To ensure that terms γ_j^{act} and γ_j^{inh} cannot be 1 at the same time, I specified γ_j^{inh} as:

$$\gamma_j^{inh} \sim Bernoulli((1-\gamma_j^{act})\pi_j^{inh}),$$

where π_j^{inh} is the prior probability to include the term j into the set of "true" inhibitors.

I have implemented Gibbs variable selection by utilizing the software BUGS (Bayesian Updating with Gibbs Sampling) (see Spiegelhalter *et al.* 1996, Gilks 1996, also Ntzoufras 2002). This is the general purpose software for Gibbs sampling on graphical (DAG) models. BUGS uses a specially designed high level language to describe a graphical model. That is, for each node in a graphical model its probability distribution must be specified, if the node is a stochastic variable, or, if the node is a deterministic variable, a functional expression must be specified for it. BUGS provides a certain number of distributions and functions. BUGS automatically constructs the necessary full conditional distributions for the sampling procedure. It exploits the factorization of the joint probability distribution in the graphical model, namely that the full conditional for a given node in the DAG involves only a subset of nodes participating in its Markov blanket. BUGS contains a small expert system for deciding the best method of sampling from full conditionals. Firstly, it will be attempted to identify conjugacy, where the full conditional reduces to a well-known distribution, and to sample accordingly. BUGS also applies the *adaptive rejection sampling* (Gilks and Wild 1992) which relies on the log-concavity of the underlying function.

The BUGS code for the "OR"-model is presented in Appendix. The "AND", "NOR", "NAND", "AND-NOR", "OR-NAND" and "OR-NOR" models were specified in analogous fashion.

The runs of the MCMC can be monitored using the package CODA implemented in R-language (see http://cran.r-project.org).

As described in Section 3.6.1, the output of Markov chain simulation is used to obtain the estimate on the variable of interest. For example, the marginal mean of the variable γ_j from the Monte Carlo output $\{\gamma_j^{(t)}, t = M+1, \ldots, N\}$ can be estimated by:

$$\gamma_j = \frac{1}{N-M} \sum_{t=M+1}^{N} \gamma_j^{(t)}.$$

Here, M is the burn-in time of the Markov chain, i.e. iterations being discarded while doing estimations from the MCMC output.

After the burn-in time of 2000 iterations, I use 10000 Markov chain simulations to make estimations on the parameters. For the complex models, like "OR-NOR" I used 5000 iterations for the burn-in, and 10000 iterations to make estimations. I obtained the final statistics on the parameters of interest γ_j (its mean value). If the mean value of γ_j was substantially close to 1 (here, higher than 0.7), I assumed $\gamma_j = 1$, otherwise $\gamma_j = 0$. The example of the trace of Markov Chain Monte Carlo simulation for parameter γ_j (for some j) is shown in Figures 3.7. The left part of the figure displays the 12000 values sampled for this variable, and the right part presents the density estimate for the γ_j. Figure 3.8 presents the MCMC sampled values and density estimate for the parameter θ_j.

3.6.4 Monitoring convergence of the Markov chain

The fundamental problem of inference from Markov chain simulation is the lack of convergence or slow convergence of the Markov chain. There will always be areas of the target distribution that have not been covered by the finite chain. As the simulation progresses, the ergodic property of the Markov chain causes it eventually to cover all the target distribution but,

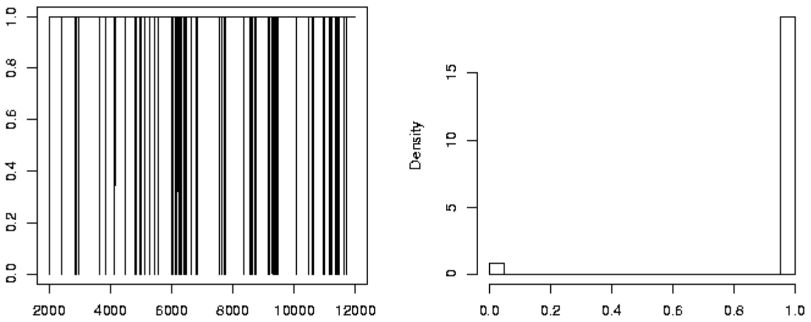

Figure 3.7: Example of the trace of the MCMC sampled values and density estimate for the parameter γ_j.

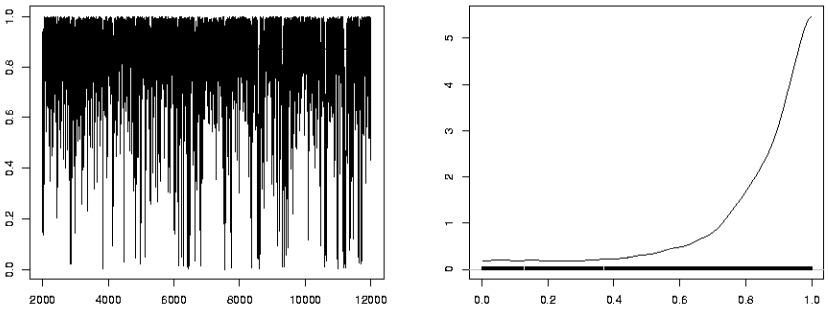

Figure 3.8: Example of the trace of the MCMC sampled values and density estimate for the parameter θ_j.

in general, the simulations cannot tell us about areas they have not been. The simulations might move too slow or stick in separate places in the target distribution. The simulations can remain for many iterations in a region heavily influenced by the starting distribution. In particular, the Gibbs sampler depends on local properties of the model, and it is hardly possible to understand the large-scale features of the joint distribution. The slow convergence of the Markov chain can be also due to an inappropriate model. For these reasons, the Markov chain must be monitored for diagnosing slow convergence or lack of convergence.

As proposed by Gelman and Rubin (1992), a number of parallel runs of Markov chains should be carried out from different starting points. The method of Gelman and Rubin is based on the comparison of within-chain and between-chain variances, and is similar to the classical analysis of variances. Approximate convergence is diagnosed when the output from different Markov chains is indistinguishable, i.e., the variance between the different chains is no larger than the variance within each individual chain. That is, the two sequences are much farther apart than we could expect, based on their internal variability.

Assume, we have m parallel Markov chain simulations, each of length n, for the quantity of interest ψ: (ψ_{ij}), $j = 1, \ldots, n$, $i = 1, \ldots, m$. I compute two quantities, the between-chain variance B and the within-chain variance W:

$$B = \frac{n}{m-1} \sum_{i=1}^{m} (\overline{\psi_{i.}} - \overline{\psi_{..}})^2,$$

where

$$\overline{\psi_{i.}} = \frac{1}{n} \sum_{j=1}^{n} \psi_{ij},$$

$$\overline{\psi_{..}} = \frac{1}{m} \sum_{i=1}^{m} \overline{\psi_{i.}},$$

and

$$W = \frac{1}{m} \sum_{i=1}^{m} s_i^2,$$

where

$$s_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (\psi_{ij} - \overline{\psi_{i.}})^2.$$

The between-chain variance B contains a factor of n because it is based on the variance of the within-chain means, $\overline{\psi_{i.}}$, each of which is an average of n values ψ_{ij}.

From the two quantities two estimates of the variance of ψ in the target distribution are obtained. First,

$$\widehat{var(\psi)} = \frac{n-1}{n}W + \frac{1}{n}B$$

is an *overestimate* under the more realistic assumption that the starting points are overdispersed. The within-chain variance W should *underestimate* the variance of ψ because the individual chains have not had time to range over all of the target distribution and, as the result, will have less variability. In the limit as $n \to \infty$, both $\widehat{var(\psi)}$ and W approach $var(\psi)$, but from opposite directions.

The convergence of the Markov chain can be estimated by the ratio between the estimated upper and lower bounds for the standard deviation of ψ, which is called *estimated potential scale reduction*:

$$\sqrt{\widehat{R}} = \sqrt{\frac{\widehat{var(\psi)}}{W}}.$$

(This is $\widehat{R}$ (estimate) rather than R because the numerator and denominator are merely estimated upper and lower bounds on the variance). As the simulation converges, the potential scale reduction declines to 1, meaning that the parallel Markov chains are essentially overlapping. If the potential scale reduction is high, then we have reason to believe that proceeding with further simulations may improve our inference about the target distribution. According Gelman and Rubin, it is desirable to choose starting points that are far apart in the parameter space. I used different initial values of the parameters indicators γ ($\gamma_j = 0$ for all j and $\gamma_j = 1$ for all j).

3.6.5 Bayesian model checking

In previous sections I described the model selection approach for the models with different Boolean-logic semantics. After the execution of the Gibbs variable selection method described above and the estimation of the variable indicator γ, the check of goodness-of-fit of the model to data is required, to check whether the model assumptions were appropriate. In context of Bayesian modelling it is called *Bayesian model checking.* Bayesian model checking uses *posterior predictive distributions* (Gelman and Meng 1993, Gelman 2004). The goal is to perform posterior predictions under the model and to assess the discrepancy between the predicted and the observed data. If the model is reasonably accurate, the generated prediction data should look similar to the observed data.

The aspect of the inferred regulatory model that is reasonable to check is its ability to predict the state of the gene Y from the states of its regulators.

Let y be the observed data on Y and θ be the vector of parameters. Denote y^{rep} the *replicated* data generated under the model with parameters θ. The posterior predictive distribution is

$$p(y^{rep}|y) = \int p(y^{rep}|\theta)p(\theta|y)d\theta.$$

The posterior predictive distribution can be computed by simulation: simulate parameters θ from their posterior distribution, and simulate y^{rep} from the sampling distribution $p(y^{rep}|\theta)$ conditioning on values of the simulated parameters. An advantage of using BUGS is that the generation of the replicate data can be easily incorporated into the model inference procedure. Based on the current simulated values of the parameters θ obtained at each iteration of the MCMC, I generate replicate dataset $\{y^{rep}\}$ from the sampling distribution of Y.

My model checking strategy is based on examination of individual observations of Y $y_i, i = 1, \ldots, N$ (N is the number of data samples) and comparison of them to the posterior predictive distributions. For the comparison I use the residual function $r_i = y_i - E(y_i)$, where the expectation $E(y_i)$ is estimated based on the replicate dataset. Observations for which the residual is not close to 0 indicate some lack-of-fit of the model and should be regarded as outlier. I regarded the residual as not close to 0 if in its absolute value it exceeded one estimated standard deviation. I calculate the model prediction accuracy as the percent of non-outliers.

My approach for learning the model of gene interactions from expression data is presented in Figure 3.9. The general workflow in this framework is represented with rectangles containing the descriptions of the procedures and connected with solid arrows. Rectangles with rounded corners contain the descriptions of the data obtained after performing the procedures and serving as input into further procedures. The dashed arrows represent the dataflow.

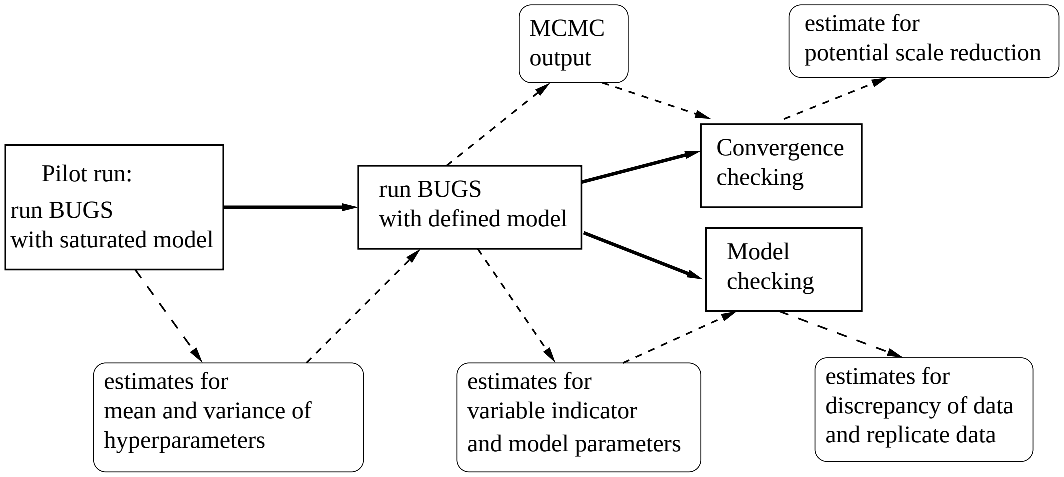
MCMC output
estimate for potential scale reduction
Pilot run: run BUGS with saturated model
run BUGS with defined model
Convergence checking
Model checking
estimates for mean and variance of hyperparameters
estimates for variable indicator and model parameters
estimates for discrepancy of data and replicate data

Figure 3.9: Method for inferring the model of gene interactions from expression data.

3.7 Inferring the interactions of the cell cycle regulated genes of *S. cerevisiae*

To test my approach I used the microarray data from Spellman *et al.* (1998) and Cho *et al.* (1998). The microarrays were produced for *S. cerevisiae* cell cultures that were synchronized by three different methods. I used the cdc15 dataset (arrest of cdcd15 temperature-sensitive mutant) which has the largest number of samples (25).

I discretized the continuous gene expression values into two states (0 - not expressed, 1 - expressed) using vector quantization. I.e. for each gene, I applied the standard k-means clustering algorithms with two initial values: 0 and the maximum expression value of the gene.

The mitotic division of the yeast cell is a series of periodic events (Mendenhall and Hodge 1998). Such events as DNA replication and chromosome segregation by the mitotic spindle are promoted with the actions of specific cyclin-dependent kinases (CDKs). The activity of CDKs are dependent on the binding of a cyclin subunit. Besides the cyclins activity, the cycle periodicity demands the degradative, proteolytic processes, that eliminate cyclically acting proteins at stages, when they are no longer required. Some cell cycle transitions are negatively regulated by specific inhibitors which must be eliminated in a timely fashion to initiate cell cycle transition. The gene transcription in the mitotic division of the yeast is coordinated in a periodic manner according to the consecutive phases of the cell cycle: G1, S, G2, M, M/G1.

I considered the group of 20 genes known to be involved in the cell-cycle regulation. The same set of genes was used by Chen *et al.*, 2000 and Soinov *et al.*, 2003. I applied my approach for learning the models ”AND”, ”OR”, ”NOR”, ”NAND”, ”AND-NAND”, ”AND-NOR”, ”OR-NAND” and ”OR-NOR” from data, for each gene in the dataset considering all other genes in the dataset as candidate regulators. After the vector of variable indicators was obtained from Gibbs variable selection procedure, I performed model checking.

	OR		**NOR**	
Genes	Activators	Accuracy	Inhibitors	Accuracy
CLN1	CLB6, CDC28	48	CLB2	76
CLN2	CLB5	88	CDC20	76
CLN3	MBP1	48	CLB4, MCM1	88
CLB1	CLB2	96	CLB6, SIC1, SWI4	88
CLB2	CLB1, SWI5	88	CLB6, SIC1, SWI4	92
CLB4	CDC34	80	CLN3, CLB6	64
CLB5	CLN2	88	CDC20	72
CLB6	CLN1	80	CLB2	80
MCM1	SKP1, CLN2	40	CDC20	52
SIC1	SWI4	60	CLB2	68
SWI6	SWI4	68	CDC20	36
CDC28	no	no	CLB5	60
CDC53	no	no	no	no
MBP1	SKP1	44	CDC34	84
CDC34	CLB4	80	MBP1	84
SWI5	CLB2	92	SWI4	80
SKP1	MBP1	44	CLB6, CDC34	60
SWI4	SIC1	80	CLN3, CDC53, CDC34	20
CDC20	SIC1	68	CLN2, CLN3	68
HCT1	MBP1	40	no	no

Table 3.4: Regulators of the genes found by learning ”OR” and ”NOR” models from data.

The results for the models I applied are displayed in Tables 3.4, 3.5. I have experimented with different settings of the prior for the variable indicator γ. I tried the Bernoulli distribution with parameters $\pi = 0.5$ and $\pi = 0.1$, and also the setting with Beta distribution described previously. I tried $Beta(16, 133)$ that keeps expectation and variance of the number of actual regulators $E(n) = 2$, $var(n) = 2$, and also $Beta(0.8, 14.4)$ with $E(n) = 1, var(n) = 2$. The results of the ”OR” and ”OR-NOR” models with these different prior settings appeared to be the same, but for the ”AND” model, which is apparently more restrictive, I found only few regulatory relations for some genes with $Bernoulli(0.1)$ and Beta distribution settings.

	OR-NOR			**OR-NAND**		
Genes	Activators	Inhibitors	Accuracy	Activators	Inhibitors	Accuracy
CLN1	CLB6	no	56	CLB6	no	60
CLN2	CLB5	CDC20	80	CLB5	no	56
CLN3	no	no	no	MBP1	no	72
CLB1	CLB2, SWI5	no	88	CLB2	no	60
CLB2	CLB2, SWI5	no	88	CLB1, SWI5	no	56
CLB4	CDC34	CLN3	88	CDC34	no	56
CLB5	CLN2	no	88	CLN2	no	60
CLB6	CLN1, CLN2	no	64	CLN1	no	56
MCM1	CLN2, SKP1	CDC20	48	CLN2, SKP1	no	52
SIC1	no	no	no	SWI4	no	76
SWI6	CLB5, SWI4	no	52	CLB5, SWI4	no	52
CDC28	no	no	no	no	no	no
CDC53	no	no	no	no	no	no
MBP1	SKP1	CDC34	80	CLN3, SWI5, HCT1	CDC34	92
CDC34	CLB4, CDC20	MBP1	92	CLB2, CLB4, SIC1	MBP1	92
SWI5	CLB1, CLB2	no	92	CLB2	no	48
SKP1	MBP1	no	44	MBP1	no	60
SWI4	SIC1	no	80	SIC1	no	64
CDC20	SIC1, CDC34	no	52	SIC1	no	52
HCT1	MBP1	no	40	no	no	no

Table 3.5: Regulators of the genes found by learning "OR-NOR" and "OR-NAND" models from data.

Genes	Activators	Inhibitors	Accuracy
CLN1	CLB6	CLB2	84
	CDC28	CLB2	80
CLN2	CLB5	CDC20	80
CLN3	no	CLB4, MCM1	88
CLB1	CLB2, SWI5	CLB6, SIC1, SWI4	92
CLB2	CLB1, SWI5	CLB6, SIC1, SWI4	96
CLB4	CDC34	CLN3, CLB6	80
	CDC34	CLN3	88
CLB5	CLN2	CDC20	80
	CLN2	no	88
CLB6	CLN1, CLN2	CLB2	84
MCM1	CLN2	CDC20	72
SIC1	SWI4	no	76
SWI6	CLB5	CDC20	72
CDC28	no	no	no
CDC53	no	no	no
MBP1	CLN3, SWI5, HCT1	CDC34	92
CDC34	CLB2, CLB4, SIC1	MBP1	92
SWI5	CLB1, CLB2	SWI4	92
SKP1	MBP1	CLB6, CDC34	68
SWI4	SWI6	CLB2	88
CDC20	SIC1, CDC34	CLN2	80
HCT1	no	no	no

Table 3.6: Final result: possible activators and inhibitors of the genes.

The results were covered by the results of "OR" model.
Anylyzing the tables 3.4, 3.5, one can see that for some genes the "NOR"-model suggest more inhibitors that "OR-NOR"-model. Learning the "NOR"-model identifies only the inhibitors of a gene, the model "explains" the non-activity of the gene with the activity of its regulators. By the "OR-NOR"-model the non-activity of the regulatee can be also "explained" with the failure of its activators.
For the final result (Table 3.6), I tested the accuracy of the "OR-NOR"-model with the activators and inhibitors from the Tables 3.4 and 3.5, and selected the highest accuracy results. The resulting genetic regulatory interactions are presented in Figure 3.10. The relationship between genes regulating one common gene is described with "OR"-function. Note, that this is not a graphical model, because it is not a directed acyclic graph, rather this graph contains cycles: for some genes there were symmetric interactions found. Note, that we searched for the local models, i.e., for each gene we considered all other genes as possible regulators, and did not test any global criteria for the whole network. The symmetric activatory relations between the genes A and B represent the following: "if A is active, then B is active, and vice versa". The symmetric inhibitory relations represent: "if

A is active, then B is inactive, and if B is active then A is inactive". The "true" direction of the regulatory relation can be reconstructed by using further biological knowledge, e.g., that one gene lies upstream from the other.
My results are consistent with previous biological knowledge: the interrelationships between the genes reflect the coincidence with different phases of the cell cycle. The genes CLN1 and CLN2, transcribing the so called G1 cyclins, are expressed in G1-phase (an interval between mitosis and DNA replication). The genes CLB5 and CLB6, transcribing the B-cyclins Clb5, Clb6, are also expressed in G1. Note the activatory connections between the genes CLN1, CLN2, CLB5, CLB6. The genes CLB1, CLB2, and SWI5 are expressed in G2. Note the activatory connections between the genes CLB1, CLB2, and SWI5. The negative connections were found between the genes related to G1- and G2-phases confirming that G1 and G2 cyclins are separated in time. It is known that Clb5 stimulate some Start-specific transcripts and initiate the S phase. The results show that CLB5 can activate the gene SWI6 which encodes the Swi6, the regulatory component of SBF and MBF transcription factor complexes important for Start-specific gene expression. Note that the gene SWI6 has the activatory connection to SWI4. The protein Swi4 is a component of the SBF complex which is a transcription factor controlling the expression of genes in G1-phase. Swi4 forms the complex with Swi6. Our method found negative connections of the gene SWI4 to the genes expressed in G2-phase.
The results suggest that the transcription of SIC1 depends on the activity of SWI4. The gene SIC1 is known to be an inhibitor of the Clb complexes and is active in the G1-phase maintaining the genes CLB1 and CLB2 in inactive state. Note the inhibitory connections of SIC1 to the G2 cyclins CLB1 and CLB2. The protein Sic1 degrades at G1-S boundary in the process of ubiquitin-mediated proteolysis triggering the initiation of the DNA synthesis. The Sic1 must be phosphorylated in order to be recognized by the ubiquitinating machinery. The CDK complexes CDC20 and CDC34 are needed for the phosphorylation of Sic1, this explains the positive connection from SIC1 to CDC20 and CDC34 found by our method.
The gene CDC34 encodes the protein Cdc34, that is the E2 ubiquitin conjugating enzyme. Both genes CDC20 and CDC34 are required for proteolytic degradation of G1 regulators. This explains the negative connection of CDC20 to SWI6, which encodes the component of SBF and MBF transcription factors and the negative connection of CDC20 to MCM1, that is the transcription factor. CDC20 is transcribed in late S/G2 phase, whereas CLN2 and CLB5 have their transcription peak in G1. This explains the negative connection of CDC20 to these genes. In our results, the genes CDC34 and MBP1 negatively influence each other, likely because the activity of CDC34 as part of the SCF ubiquitinating complex, and the activity of MBP1 as part of MBF transcription factor complex, are completely separated in time.

In the Figure 3.10 there is a negative connection from the gene CDC34 to the gene SKP1, whereas Skp1 is the E3 ubiquitin ligase which is needed for Cdc34 essential function. However, if I used the "time delayed" samples of the gene SKP1, I found the positive influence of the gene CDC34 on the gene SKP1. This suggests that a certain time interval is needed between the transcription of these genes to achieve their function.
The gene CLB4 is expressed in S- and G2-phases. Its product can initiate S-phase, if Clb6 is lacking (note the inhibitory connection from CLB6 to CLB4). CDC34, is required for the proteolysis of Clb proteins Clb2 and Clb4 at the border of G2-M (positive connections from CLB2 and CLB4 to CDC34. The gene CLN3 is expressed at the M/G1 border. The MCM1 gene encodes the transcription factor and is active during G2/M transition. The time delay in the activities of MCM1 and CLN3 implied the negative connection from MCM1 to CLN3.
Obviously, most of the regulatory interactions coordinating the cell division cycle of the yeast occur at protein level. Such events cannot be measured by microarray experiments. The genetic interactions reconstructed from the gene expression data can only give a hint towards the genetic regulatory pathways. Many events in this chain of events will likely remain hidden.

3.8 Discussion

In this thesis I present a model for the genetic regulatory interactions and an automated approach for learning the structure and parameters of the model from gene expression data. The model represents the Boolean logic semantics of gene interactions which is a biologically plausible assumption. In contrast to the standard Boolean networks, my model has a probabilistic nature representing probabilistic dependencies between a gene and its regulators. This stochasticity is more suitable for modelling noisy biological process and experimental measurements. The model is a graphical model that explicitly represents the (in)dependencies among variables. It can be seen as an intermediate model between the structures of gene interactions represented by Boolean networks and by Bayesian networks. The model is not fully observable, rather it contains hidden variables allowing for the representation of factors that could not be measured.
Due to the statistical context of the model, unlikely to Boolean networks, I could employ the methodology of Bayesian statistics for learning the model from data. Bayesian approach treats the uncertainty on model structure and parameters in a unified fashion, defining the priors on these quantities. The Bayesian modelling deals with complex models with many parameters giving the possibility of the hierarchical formulation of the model: the prior on the model parameters can be defined with the help of further parameters. It was

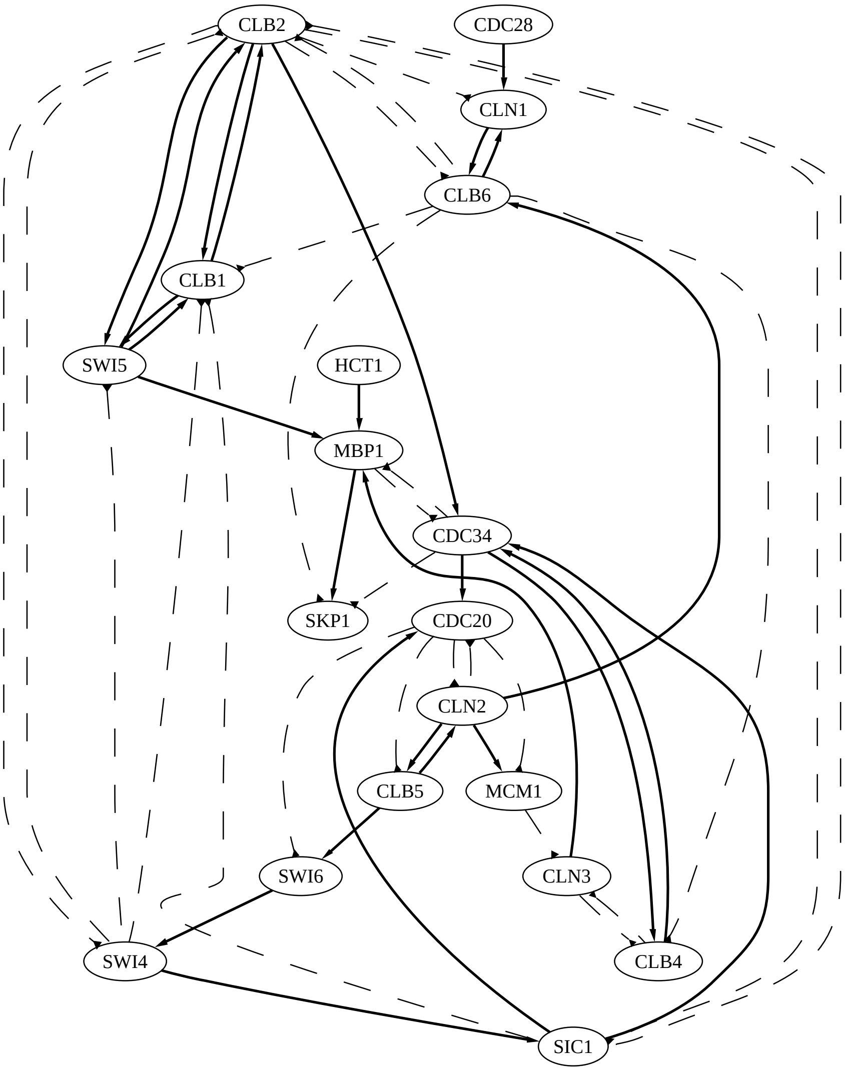

Figure 3.10: Gene regulatory interactions inferred for 20 genes of *S. cerevisiae*. The full arcs represent activatory regulation, the dashed arcs represent inhibitory regulation. The relationship between genes influencing one common gene is described by "OR"-function.

possible to insert into the model a new parameter (variables indicator). The learning of the resulting complex model was facilitated with sampling-based methods.

Bayesian networks are acyclic graphs. However, there is an evidence that genetic networks can have circular dependencies - feedback loops (gene A affects gene B but gene B also affects gene A). In contrast to the Bayesian networks, in my approach I do not have a global criteria to score the whole genetic network with respect of its fitness to the data, rather I investigate local relationships between the genes.

In the past, there were some models suggested with the same attempt to extend Boolean networks to make them robust against noise. In the *noisy Boolean networks* of Akutsu (2000), they defined a probability with which a certain number of input/output patterns of gene expression will not be discarded by an inference algorithm, even if a certain Boolean function is not satisfied. In contrast, my approach considers the probability of "noise" as parameters of the model, giving a possibility to apply statistical learning for model inference.

Shmulevich *et al.* (2002) present *Probabilistic Boolean networks.* They insert "noise" into the model by accomodating more than one possible Boolean functions for each node in the network. They introduced a probability with which a certain Boolean function is selected from the set of possible functions for calculating the output of the target gene. In contrast, I see the source of uncertainty of the model not in the realizations of different Boolean functions, but in the fact that independent basic elements of the gene regulatory modules could fail. The authors investigate the global dynamical properties of the model, and do not present any algorithms for learning the model from data. In contrast, my work focuses on the data mining task of learning the structure of these regulatory modules from data.

The identification of genetic regulatory interactions and gene regulatory pathways is an important part of bioinformatics and biological research. Cellular responses and actions are often a result of coordinated activity of a group of genes. There is a growing indication that most single-gene mutations do not have marked phenotypes. Most phenotypes are the result of the collective response of a group of genes. Genes act 'in concert' to achieve certain phenotypic characteristics. Reconstructing genetic regulatory interactions help rationalize how these complex traits arise and which genes are responsible for them.

Recent estimates on the number of genes in the human genome suggest that there are about 35000 human genes, only about twice that of the worm *Caenorhabditis elegans*, and about five times more than by the *Saccharomyces cerevisiae.* This relative "simplicity" of the human genome can be explained by several hypotheses. First, the proportion of regulatory genes (encoding signalling proteins, transcription factors, etc.) could be higher than in other

genomes. Second, the human genetic network could have a higher mean number of connections per gene, which implies that the encoded proteins contain more binding sites. Both hypotheses could be tested by determining and comparing genetic networks of different organisms (comparative genomics).
The topology of gene networks might be responsible for the robustness shown by organisms. A particular gene network topology might have been selected during evolution to permit the system robustness against drastic perturbations at the genetic level (40% of the genes of *Saccharomyces cerevisiae* can be removed without causing noticeable phenotypes).
The robust inferring of genetic regulatory relations can help for further identification of transcription factor binding sites, promoter prediction, identification of "target" genes regulated by a particular regulatory element. Earlier methods based on computational screening of the genomic sequences were less advanced than, for example, methods for predicting coding regions of the genes, because the regulatory elements are very diverse and comprized of short motifs. There were a lot of false positives obtained by such kind of analysis, and only a small fraction of the predicted binding sites were functionally significant. The robust reconstruction of genetic interactions from data can provide reliable sets of co-regulated genes, which might be possibly regulated by similar mechanisms, i.e. by common transcription factors, and therefore, should have transcription factor binding sites in common. Knowing a set of transcription factors can allow to predict new, previously unknown target genes that are responsive to this factors.
Knowledge about gene regulatory interactions might provide valuable clues and lead to new ides for treating complex diseases like cancer. Biomolecules that affect transcription (either inducers or inhibitors of transcription) become high-priority targets for pharmaceutical research and drug development. By introducing compounds known to affect transcription and then by studying the actual transcription profiles with microarrays, it is possible to identify critical steps in regulatory and other cellular pathways.

Chapter 4

Conclusions and future perspectives

The presented thesis focuses on the application of probabilistic graphical models in bioinformatics. I was dealing with two kinds of models:

- Bayesian networks, that represent multivariate probabilistic dependencies between variables, and
- networks with the dependencies between variables defined by Boolean-logic functions.

I applied these models for the analysis of two different kinds of biological data:

- the cytogenetic data about the allelic losses in tumors (losses of heterozygosity LOH) and
- the genetic data about expression of the genes obtained with microarray experiments.

Learning the models from data allows to determine and quantify stochastic relationships between biological entities. In the first application I was able to reveal primary and secondary allelic losses and to suggest pathways of progression of these abnormalities which are possibly associated with tumor pathogenesis. In the second application I have learned the model from microarray data to uncover the regulatory interactions of the genes.

Both models are probabilistic, hence being capable to deal with the substantial amount of noise present in the biological data. Both models are parametric models with many parameters. Learning probabilistic graphical models from data is facilitated by the modern statistics approach - Bayesian modelling. Bayesian modelling is characterized by the conversion of the statistical inference into the probabilistic inference based on the Bayes theorem. Therefore, priors on model parameters must be inserted into the model formulation. Bayesian approach is well suited to the problems with many variables and many parameters due to the possibility to define the model hierarchically, i.e. the prior on model parameters can be formulated with the help of further parameters.

One of advantages of the Bayesian approach is that it enables to include "subjective" prior information into the model. In this study I used the subjective prior specification to enforce the number of gene regulators to lie in the desired range. Potentially, one could define priors aiming to incorporate previous biological knowledge into the model learning making the model more biologically plausible.

The major obstacle to the broad employment of the Bayesian methodology was that integrals involved in the Bayesian problems like model estimation have no analytical solution (conjugate prior for model likelihood for many models does not exist). Recently developed computer intensive approaches based on Markov Chain Monte Carlo algorithms such as Gibbs sampler have revolutionized the application of Bayesian methods. I have applied MCMC simulation for the model selection in the second application, because the problem had no closed form solution. Due to the flexibility of the Bayesian modelling, I could introduce an additional parameter into the model, so that the problem of model selection transformed into the variable selection task, and performed Gibbs variable selection.

The idea behind the whole Bayesian approach was to extend classical statistics with a form of decision-making. Bayesian approach allows to fit a model to the data and to perform predictive inference under the model. It allows to make estimates on the variables of interest given the observed values of some other variables. This enables to use models as predictors, or even as simulating systems and to bring together two kinds of analysis: the *data-mining* analysis, which is an explorative, knowledge discovery task and the *simulation-based* analysis.

Biological systems (like cell or tumor cell) are complex and heterogeneous. They are comprized of a very large number of elements, which are frequently multifunctional, and different functions emerge from the specific interactions of the elements. The systems exhibit complex behaviour that is usually not predictable from the properties of individual components alone. To understand complex biological systems requires not only to uncover a multitude of biological facts, but to gain insight into their various dependencies. Developing methods for establishing cause-effect relationships between biological entities on the basis of observed data is an important part of bioinformatics research.

With the advancement of high-throughput methods the amount of quantitative biological data will increase. But it will never be possible to discover all biological relations experimentally, and to explore the complex behaviour of interacting cellular components by focusing on single molecules and re-actions. Insights into the functioning of biological systems will not result

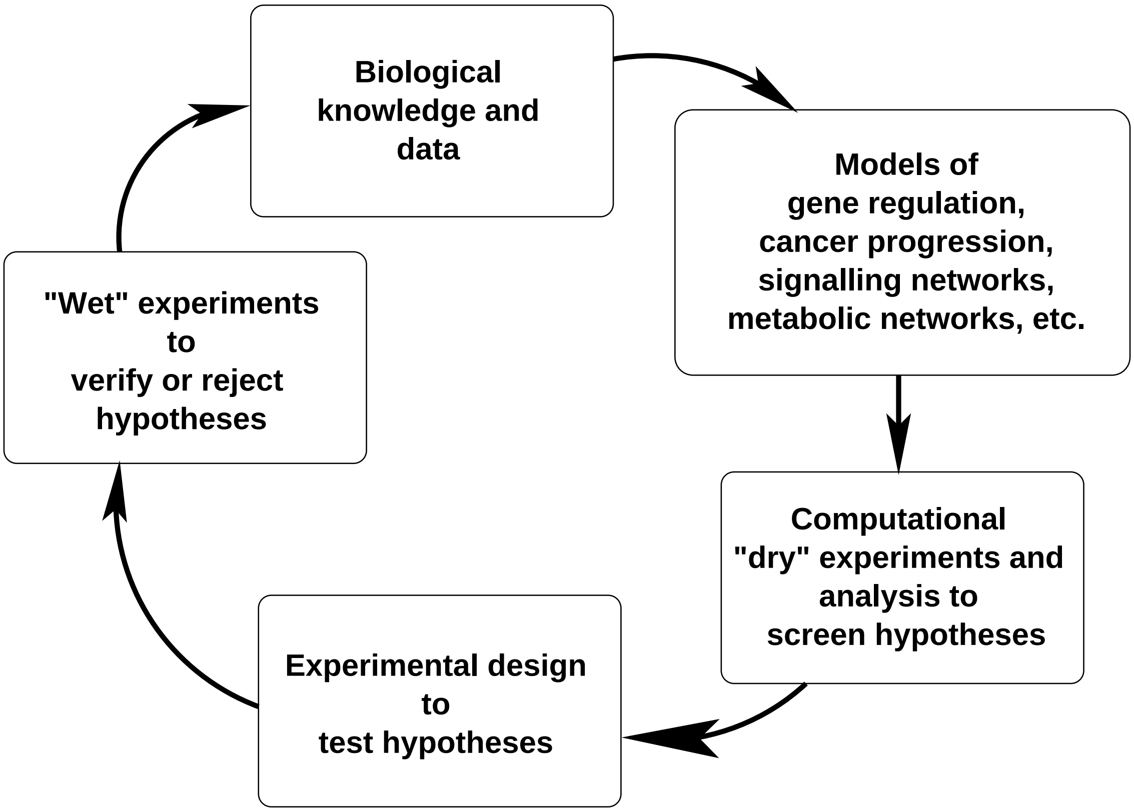
Biological knowledge and data
Models of gene regulation, cancer progression, signalling networks, metabolic networks, etc.
Computational "dry" experiments and analysis to screen hypotheses
Experimental design to test hypotheses
"Wet" experiments to verify or reject hypotheses

Figure 4.1: Systematic approach to hypotheses testing and knowledge discovery in biology.

from purely intuitive approach, but a lot of modelling work and theoretical exploration will be required. The systematic integration of experimental and computational research will determine the future perspectives of the biological research. A new discipline *systems biology* arised with the focuse on this systematic view. This can be represented as in Figure 4.1. Data-mining approaches and modelling are needed to fit the computer-executable models based on current biological knowledge and data. Simulations and predictions made under the models ("dry" experiments) allow to verify and validate assumptions underlying the model. Inconsistency at this stage means that the assumptions representing our knowledge are at best incomplete. Models that survive validation can then be used to make predictions to be tested by biological ("wet") experiments.
In this work I developed a general computational framework enabling to define a model of gene interactions with a particular regulatory function and to perform Bayesian learning of this model from data. The main advantage

of the model is that the relationships found by the model have a clear logical semantics and do not require laborious analysis for their interpretation. Therefore, the results obtained with this approach can be further utilized in an automatic system for discovering transcription factor binding sites, new regulatory elements and pathways.

The regulatory pathways of the cell rely not only on the transcriptional regulation, but to a great extent on the post-translational and external signalling events. "Genetic networks" are phenomenological, because they do not explicitly represent the proteins and metabolites mediating interactions in the cell. They provide a system view at the level of gene activities, when the expression level of one gene affects the expression level of some others. Often "pathways" rather then "networks" are referred to when one is interested in a particular series of interactions (e.g. cellular pathways regulated by specific transcription factors). The reconstruction of the genetic regulatory interactions from gene expression data can give only hypotheses on the cellular pathways.

Unobserved events on protein level can be represented in a probabilistic model by introducing hidden variables. When more detailed proteomics data will be available, it can be also handled by the approach introduced here.

There are other networks being considered in bioinformatics. *Metabolic networks* represent the chemical transformations between metabolites. *Protein networks* (also known as *signalling networks*) represent protein-protein interactions, such as formation of complexes and protein modification by signalling enzymes. The ultimate goal and challenge of biology and bioinformatics is to combine the genomic, transcriptomic and proteomic information and to link the genes and their products into global functional networks. When more detailed proteomics or other kinds of data will be available, this can be easily incorporated into the framework of my approach.

To summarize, the present thesis demonstrates the novel application of Bayesian network model in the domain of molecular genetics data. My further contribution to the bioinformatics and informatics research is the development of a Bayesian model with Boolean-logic based semantics representing the genetic interactions. I have applied the Bayesian modelling approach and developed a novel method for learning my model based on Gibbs sampling.

Bayesian models like Bayesian networks and the second Boolean-logic based model presented in this work can find a broad application in bioinformatics due to the modelling freedom they provide, the ability to incorporate prior knowledge and the possibility of predictive inference. The well grounded theoretical foundations of Bayesian modelling and the development of Markov Chain Monte Carlo techniques will facilitate, that Bayesian models would meet further challenges of bioinformatics, systems biology, in general, and oncological research, in particular.

Cellular processes as for example regulatory processes and gene expression have a dynamical nature. When the biological quantitative methods will be able to capture the dynamics of the underlying biological processes (i.e. to make measurements with an appropriate time resolution), the *time-series* analysis methods will be required. Dynamic Bayesian networks is an extension of Bayesian networks to represent the statistical dependencies evolving in time. One of my future goals is the investigation of Bayesian models capturing the system's dynamics.

Publications

1. Bulashevska S., Szakacs O., Brors B., Eils R., Kovac G., Pathways of urothelial cancer progression suggested by Bayesian network analysis of allelotyping data, *International Journal of Cancer*, in press, 2004

2. Bulashevska S., Groll A., Eils R., ”ISCN Parser”: a Software Tool for Interpreting Cytogenetic Data Notated in The International System for Human Cytogenetic Nomenclature (ISCN 1995), in: *Proc. of NETTAB ”Network Tools and Applications in Biology” Conference*, Genoa, Italy, 2002

3. Berrar D., Dubitzky W., Solinas-Toldo S., Bulashevska S., Granzow M., Conrad C., Kalla J., Lichter P. and Eils R. Design and Implementation of a Database System for Comparative Genomic Hybridization Analysis, *IEEE Eng. Med. Biol.* 20(4): pp. 75-83, 2001.

4. Bulashevska S., Dubitzky W., Eils R., Mining Gene Expression Data using Rough Set Theory, in: *Proc. of CAMDA ”Critical Assessment of Techniques for Microarray Data Analysis” Conference*, Duke University, 2000

5. Dubitzky W., Granzow M., Berrar D., Bulashevska S., Conrad C., Gerlich D., Eils R., Symbolic and subsymbolic machine learning approaches for molecular classification of cancer and ranking of genes, in: *Proc. of CAMDA ”Critical Assessment of Techniques for Microarray Data Analysis” Conference*, Duke University, 2000

Acknowledgements

I would like to thank Prof.Dr.Roland Eils, the head of the Division ”Intelligent Bioinformatics Systems” at the German Cancer Research Center (DKFZ), for bringing me in the excellent environment of an active biological research and making it possible to satisfy my scientific interest and creativity. Thanks for his constant engagement to meet the challenges of bioinformatics and systems biology, for his encouragement and support.

I want to thank Prof.Dr.Peter Lichter, the head of the Department ”Molecular Genetics”, DKFZ, and his colleagues, especially Dr.Stefan Joos, and also Dr.Sabina Solinas-Toldo for their collaboration while working with cytogenetic data.

Special thanks to Prof.Dr.Gyula Kovacs, Laboratory of Molecular Oncology, Department of Urology, University Heidelberg, and his colleagues for providing the excellent data and our fruitful collaboration.

I am grateful to Ioannis Ntzoufras, Department of Business Administration, University of the Aegean, Greece, for his valuable suggestions and comments on the application of Gibbs variable selection method.

Thanks to Dr.Peter Lucas, Institute for Computing and Information Sciences, University of Nijmegen, Netherlands for his comments on my work in the field of Bayesian networks.

I enjoyed the discussions with Dr.Lucas Toldo, pharmaceutical company Merck, Darmstadt.

I want to thank my colleagues Dr.Benedikt Brors and Armin Groll for their collaboration and help. I am grateful to my sister and colleague Alla Bulashevska for our discussions on bioinformatics problems.

The work relied on permanently running computer systems, which were administered by Karlheinz Gross and Rolf Kabbe, whose support I highly appreciate.

Many thanks to Prof.Dr.Roland Vollmar, Faculty of Informatics, University Karlsruhe for his great motivation to supervise the work on application of informatics in the biological research.

Bibliography

[Akutsu *et al.* 2000] Akutsu T., Miyano S., Kuhara S., Inferring qualitative relations in genetic networks and metabolic pathways, *Bioinformatics*, 16: 727-734, 2000.

[Arkin *et al.* 1998] Arkin A., Ross J., McAdams H., Stochastic kinetic analysis of developmental pathways bifurcation in phage lambda infected escherichia coli cells, *Genetics*, 149, 1275-1279, 1998.

[Andersen *et al.* 1989] Andersen S.K., Olesen K.G., Jensen F.V., Jensen F., HUGIN - a shell for building Bayesian belief universes for expert systems, *11th International Joint Conf. on Artificial Intelligence*, vol I, pp1128-1133, 1989.

[Berger 1993] Berger J.O., Statistical decision theory and Bayesian analysis, *Springer series in statistics*, 2. ed., New York, Springer, 1993.

[Bernardo and Smith 1994] Bernardo J.M. and Smith A.F.M., Bayesian theory, Willey Series in Probability and Mathematical Statistics, John Willey and Sons, Chichester, 1994.

[Blader *et al.* 2001] Blader I.J., Manger I.D., Boothroyd J.C., Microarray analysis reveals previously unknown changes in Toxoplasma gondii-infected human cells, *J. Biol. Chem.*, 276(26):24223-31, 2001.

[Soinov *et al.* 2003] Soinov L.A., Krestyaninova M.A., Brazma A., Towards reconstruction of gene networks from expression data by supervised learning, *Genome Biology*, Vol. 4: R6, 2003.

[Carlin and Chib 1995] Carlin B.P. and Chib S., Bayesian model choice via Markov chain Monte Carlo, *J. Roy. Statist. Soc.*, B57, 473-484, 1995.

[Castillo 1997] Castillo E., Gutierrez J.M., Expert systems and probabilistic network models, New York, Heidelberg: Springer, 1997.

[Chen *et al.* 1999] Chen T., He H.L., Church G.M., Modeling gene expression with differential equations, *Proc. Pacific Symp. on Biocomputing*, 4, 29-40, 1999.

[Chickering 1996 a] Chickering D.M., Learning Bayesian Networks is NP-complete, in: Fisher D. and Lenz H. J. (eds.), *Learning from Data: Artificial Intelligence and Statistics 5.*, Berlin: Springer Verlag, 1996.

[Chickering 1996 b] Chickering D.M., A transformational characterization of equivalent Bayesian Network structures, in: *Proc. of the 11th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers, Inc., San Francisco, 1996.

[Cho *et al.* 1998] Cho *et al.*, A genome-wide transcriptional analysis of the mitotic cell cycle, *Mol. Cell*, 2, 65-73, 1998.

[Chu *et al.* 1998] Chu S., DeRisi J., Eisen M., Mulholland J., Botstein D., Brown P.O., Herskowitz I., The transcriptional program of sporulation in budding yeast, *Science*, Oct. 23;282(5389):699-705, 1998.

[Congdon 2001] Congdon P., Bayesian statistical modelling, *Wiley Series in Probability and Statistics*, John Wiley & Sons LTD, Chichester, 2001.

[Cooper and Herskovits 1992] Cooper G.F. and Herskovits E., A Bayesian method for the induction of probabilistic networks from data, *Machine Learning*, 9: 309-347, 1992.

[Cowell 1999] Cowell R.G., Probabilistic networks and expert systems, *Statistics for engineering and information science*, New York, Berlin, Heidelberg: Springer, 1999.

[Dalbagni *et al.* 1993] Dalbagni G., Presti J., Reuter V., Fair W. R., Cordon-Cardo C., Genetic alterations in bladder cancer, *Lancet*, 342: 469-471, 1993.

[Diez *et al.* 1997] Diez F., Mira J., Iturralde E., Zubillage S., DIAVAL: a Bayesian expert system for echocardiography, *Articial Intelligence in Medicine*, 10, 1997.

[D'Haeseleer *et al.* 1999] D'Haeseleer P., Wen X., Fuhrman S., Somogyi R. Linear modeling of mRNA expression levels during CNS development and injury, *Pacific Symp. on Biocomputing*, 4, 41-52, 1999.

[Dellaportas *et al.* 2000] Dellaportas P., Forster J.J., Ntzoufras I., Bayesian Variable Selection using the Gibbs Sampler, in: *Generalized Linear Models: a Bayesian Perspective*, Dey D.K., Ghosh S., Mallick B. (eds.), New York: Marcel Dekker, 271-286.

[Dellaportas *et al.* 2002] Dellaportas P., Forster J.J., Ntzoufras I., On Bayesian model and variable selection using MCMC, *Statistics and Computing*, 12, 27-36, 2002.

[Dempster 1967] Dempster A.P., Upper and lower probabilities induced by multi-valued mapping, *Annals of Mathematical Statistics*, 28, 325-39, 1967.

[DeRisi *et al.* 1997] DeRisi J.L., Iyer V.R., Brown P.O., Exploring the metabolic and genetic control of gene expression on a genomic scale., *Science*, 24; 278(5338):680-6, 1997.

[Desper *et al.* 1999] Desper R., Jiang F., Kallioniemi O.P., Moch H., Paradimitriou C.H., Schäffer,A. Inferring Tree Models for Oncogenesis from Comparative Genome Hybridization Data, *J. Comp. Biol.*, 6: 37-51, 1999.

[Desper *et al.* 2000] Desper R., Jiang F., Kallioniemi O.P., Moch H., Paradimitriou C.H., Schäffer A., Distance-based reconstruction of tree models for oncogenesis, *J. Comp. Biol.*, 7:789-803, 2000.

[Efron and Tibshirani 1998] Efron B., Tibshirani R.J., An introduction to the bootstrap, Chapman & Hall, 1998.

[Eisen *et al.* 1998] Eisen M.B., Spellman P.T., Brown P.O., Botstein D., Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA*, 8, 95(25):14863-8, 1998.

[Fearon and Vogelstein 1990] Fearon E.R., Vogelstein B., A genetic model for colorectal tumorigenesis, *Cell*, 61: 759-767, 1990.

[Friedman and Goldszmidt 1996 a] Friedman N., Goldszmidt M., Learning Bayesian networks with local structure, in: *Proc.of the Conf. on Uncertainty in Artificial Intelligence*, 1996.

[Friedman and Goldszmidt 1996 b] Friedman N., Goldszmidt M., Building classifiers using Bayesian networks, *Proc. of AAAI 96*, 1996.

[Friedman *et al.* 1999 a] Friedman N., Goldszmidt M., Wyner A., On the application of the bootstrap for computing confidence measures on features of induced Bayesian networks, in: *AI&STAT* VII, 1999.

[Friedman *et al.* 1999 b] Friedman N., Goldszmidt M., Wyner A., Data analysis with Bayesian Networks: a Bootstrap approach, in: Laskey K.B., Prade H. (eds.), *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, Stockholm, Morgan Kaufmann, 1999.

[Friedman *et al.* 2000] Friedman N., Linial M., Nachman I., Peer D., Using Bayesian network to analyze expression data, *J. Comput. Biol.*, 7, 601-620, 2000.

[Gasch *et al.*] Gasch A.P., Spellman P.T., Kao C.M., Carmel-Harel O., Eisen M.B., Storz G., Botstein D., Brown P.O., Genomic expression programs in the response of yeast cells to environmental changes, *Mol. Biol. Cell*, 11(12):4241-57, 2000.

[Gelman 2004] Gelman A., Carlin J.B., Stern H.S., Rubin D.B. (eds.), Bayesian data analysis, 2nd. edition, Chapman & Hall, 2004.

[Gelman and Meng 1993] Gelman A., Meng X.-L., Model checking and model improvement, in: *Markov Chain Monte Carlo in Practice*, Gilks W.R., Richardson S., Spiegelhalter D. (eds.), Chapman and Hall, London, UK, 190-201, 1993.

[Gelman and Rubin 1992] Gelman A. and Rubin D.B., Inference from iterative simulation using multiple sequences, *Statist. Sci.*, 7, 457-511, 1992.

[Geman and Geman 1984] Geman S. and Geman D., Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 6, 721-741, 1984.

[George and McCulloch 1996] George E.I., McCulloch R.E., Stochastic Search Variable Selection, in: *Markov Chain Monte Carlo in Practice*, Gilks W.R., Richardson S., Spiegelhalter D. (eds.), Chapman and Hall, London, UK, 203-214, 1996.

[Gilks 1996] Gilks W.R., Richardson S., Spiegelhalter D.J. (eds.), Markov Chain Monte Carlo in practice, Chapman & Hall, London, 1996.

[Gilks and Wild 1992] Gilks W.R. and Wild P., Adaptive rejection sampling for Gibbs sampling, *Appl. Statist.*, 41, 337-348, 1992.

[Green 1995] Green P.J., Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, 82, 711-732, 1995.

[Habuchi *et al.* 1993] Habuchi T., Takahashi R., Yamada H., Ogawa O., Kakehi Y., Ogura K., Hamazaki S., Toguchida J., Ishizaki K., Fujita J., Influence of cigarette smoking and schistosomiasis on p53 gene mutation in urothelial cancer, *Cancer Res.* 53: 3795-3799, 1993.

[Hastings 1970] Hastings W.K., Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, 57, 97-109, 1970.

[Heckerman 1986] Heckerman D., Probabilistic interpretations for MYCIN's certainty factors, in: Kanal L.N., Lemmer J.F. (eds.), *Uncertainty in Artificial Intelligence*, North Holland, Amsterdam, pp.298-312, 1986.

[Heckerman 1998] Heckerman D. A tutorial on learning with Bayesian networks, in: Jordan M.I. (ed.), *Learning in Graphical Models*, Dordrecht: Kluwer Academic Publishers, 1998.

[Heckerman and Breese 1994] Heckerman D. and Breese J.S., Causal independence for probability assessment and inference using Bayesian networks, *IEEE Trans. on Systems, Man and Cybernetics*, 26(6), 826-831, 1994.

[Hoeglund *et al.* 2001] Hoeglund M., Sall T., Heim S., Mitelman F., Mandahl N., Fadl-Elmula I., Identification of cytogenetic subgroups and karyotypic pathways in transitional cell carcinoma, *Cancer Res.*, 61: 8241-8246, 2001.

[Hughes *et al.* 2000] Hughes T.R., Marton M.J., Jones A.R., Roberts C.J., Stoughton R., Armour C.D., Bennett H.A., Coffey E., Dai H., He Y.D., Kidd M.J., King A.M., Meyer M.R., Slade D., Lum P.Y., Stepaniants S.B., Shoemaker D.D., Gachotte D., Chakraburtty K., Simon J., Bard M., Friend S.H., Functional discovery via a compendium of expression profiles, *Cell*, Jul 7;102(1):109-26, 2000.

[Jaakkola and Jordan 1996] Jaakkola T.S. and Jordan M.I., Computing upper and lower bounds on likelihoods in intractable networks, in: *Proc. of 12th Conference on Uncertainty in Artificial Intelligence*, Portland, OR, pp.340-348, Morgan Kaufmann, 1996.

[Jensen 1996] Jensen F.V., Introduction to Bayesian Networks, New York: Springer, 1996.

[Kass *et al.* 1988] Kass R., Tierney L., Kadane J., Asymptotics in Bayesian computation, in: *Bayesian Statistics 3*, Bernardo J., DeGroot M., Lindley D., Smith A., 261-278, Oxford University Press, 1988.

[Kauffman 1993] Kauffman S.A., The origins of order: Self-organization and Selection in Evolution, Oxford University Press, New York, 1993.

[Knowles 2001] Knowles M.A., What we could do now: molecular pathology of bladder cancer, *Mol. Pathol.*, 54: 215-221, 2001.

[Kohn 2001] Kohn R., Smith M., Chan D. Nonparametric regression using linear combinations of basis functions, *Statistics and Computing*, **11**,313-322, 2001.

[Kuo and Mallick 1998] Kuo L. and Mallick B., Variable Selection for regression models, *Sankhya*, B, 60, Part1, 65-81, 1998.

[Langbein *et al.* 2002] Langbein S., Szakacs O., Wilhelm M., Sükösd F., Weber S., Jauch A., Lopez-Beltran A., Alken P., Kälble T., Kovacs G., Alteration of the LRP1B gene region is associated with high grade of urothelial cancer, *Lab. Invest.*, 82: 639-643, 2002.

[Lauritzen 1996] Lauritzen S.L., Graphical models, Oxford University Press, United Kingdom, 1996.

[Liang *et al.* 1998] Liang S., Fuhrman S., Somogyi R., REVEAL, a general reverse engineering algorithm for inference of genetic network architectures, *Proc. Pacific Symp. on Biocomputing*, 3, 18-29, 1998.

[Lindley 1980] Lindley D.V., Approximate Bayesian methods, in: *Bayesian statistics*, Bernardo, J.M., DeGroot M.H., Lindley D.V. (eds.), 223-245, Valencia: University Press, 1980.

[Logothetis *et al.* 1992] Logothetis C. J., Xu H. J., Ro J. Y., Hu S. X., Sahin A., Ordonez N., Benedict W.F., Altered expression of retinoblastoma protein and known prognostic variables in locally advanced bladder cancer, *J. Natl. Cancer Inst.*, 84: 1256-1261, 1992.

[Lucas *et al.* 2001] ucas P., van der Gaag L., Abu-Hanna A. (eds.), Bayesian Models in Medicine, *The European Conference on Artificial Intelligence in Medicine* AIME'01, Portugal, 2001.

[McCullagh and Nelder 1983] Generalized linear models, Chapman & Hall, London, 1983.

[Meek and Heckerman 1997] Meek C., Heckerman D. Structure and parameter learning for causal independence and causal interaction models, in: *Proc. of the Conf. on Uncertainty in AI*, 366-375, 1997.

[Mellon *et al.* 1996] Mellon J.K., Lunec J., Wright C., Horne C.H., Kelly P., Neal D.E., C-erbB-2 in bladder cancer: molecular biology, correlation with epidermal growht factor receptors and prognostic value, *J. Urol.*, 155: 321-326, 1996.

[Mendenhall and Hodge 1998] Mendenhall M.D. and Hodge A.E., Regulation of Cdc28 Cyclin-Dependent Protein Kinase Activity during the Cell Cycle of the Yeast *Saccharomyces cerevisiae*, *Microbiol. and Mol. Biol. Rev.*, 62, 1191-1243, 1998.

[Metropolis *et al.* 1953] Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H., Teller E., Equations of state calculations by fast computing machine, *J. Chem. Phys.*, 21, pp. 1087-1091, 1953.

[Mitelman, F. 1995] Mitelman F., ISCN (1995): An International System for Human Cytogenetic Nomenclature, Basel, Karger, 1995.

[Muschek *et al.* 2000] Muschek M., Sükösd F., Pesti T., Kovacs G., High density deletion mapping of bladder cancer localizes the putative tumor suppressor gene between D8S504 and D8S264 at chromosome 8p23.3, *Lab Invest.*, 80: 1089-1093, 2000.

[Myers 2002] Myers R.H., Douglas C., Montgomery G., Vining G., Generalized linear models: with applications in engineering and the sciences, New York: Wiley, 2002.

[Neapolitan 1990] Neapolitan R.E., Probabilistic reasoning in expert systems: theory and algorithms, New York, Wiley, 1990.

[Ntzoufras 2002] Ntzoufras I., Gibbs Variable Selection using BUGS, Technical report, http://www.ba.aegean.gr/ntzoufras/tr.htm, 2002.

[Pearl 1987] Pearl J., Evidential reasoning using stochastic simulation of causal models, *Artificial Intelligence*, 32: 245-257, 1987.

[Pearl 1998] Pearl J., Probabilistic Reasoning in Intelligent Systems, San Francisco, California: Morgan Kaufmann, 1998.

[Pearl and Verma 1991] Pearl J. and Verma T., A theory of inferred causation, in: Allen J., Fikes R., Sandewall E. (eds.), *Knowledge Representation and Reasoning: Proc. of the 2nd International Conference*, Morgan Kaufmann, New York, 1991.

[Ptashne and Gann 2002] Ptashne M. and Gann A., Genes and Signals, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2002.

[Reznikoff *et al.* 1996] Reznikoff C.A., Belair C.D., Yeager T.R., Savelieva E., Blelloch R.H., Puthenveettil J.A., Cuthill S. A molecular genetic model of human bladder cancer pathogenesis, *Semin.Oncol.*, 23: 571-584, 1996.

[Schaeffer *et al.* 2001] Schaeffer A.A., Simon R., Desper R., Richter J., Sauter G., Tree models for dependent copy number changes in bladder cancer, *Int. J. Oncol.*, 18: 349-354, 2001.

[Shafer 1976] Shafer G., A mathematical theory of evidence, Princeton University Press, Princeton, NY, 1976.

[Shmulevich *et al.* 2002] Shmulevich I., Dougherty E.R., Seungchan K., Zhang W., Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks, *Bioinformatics*, 18, 261-274, 2002.

[Shortliffe 1976] Shortliffe E.H., Computer-based medical consultations: MYCIN, Elsevier, New York, 1976.

[Somogyi and Sniegosky 1996] Somogyi R., Sniegosky C.A., Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation, *Complexity*, 1(45).

[Spellman *et al.* 1998] Spellman P.T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccaharomyces cerevisiae* by microarray hybridization, *Mol.Biol.Cell*, 9, 3273-3297.

[Spiegelhalter *et al.* 1996] Spiegelhalter D.J., Thomas A., Best N.G., Computation on Bayesian graphical models, in: *Bayesian Statistics* 5, Bernardo J.M., Berger J.O. Dawid A.P., Smith A.F.M. (eds.), pp. 407-425, Oxford University Press, 1996.

[Spruck *et al.* 1994] Spruck C.H., Ohneseit P.F., Gonzalez-Zulueta M., Esrig D., Miyao N., Tsai Y. C., Lerner S.P., Schmutte C., Yang A.S., Cote R., Dubeau L., Nichols P.W., Hermann G.G., Steven K., Horn T., Skinner D.G., Jones P.A., Two molecular pathways to transitional cell carcinoma of the bladder, *Cancer Res.*, 54: 784-788, 1994.

[Srinivas 1993] Srinivas S., A generalization of the noisy-or model, in: *Proc.of the Conf. on Uncertainty in Artificial Intelligence*, 1993.

[Strachan and Read 1996] Strachan T. and Read A.P., Human molecular genetics, BIOS Scientific Publishers, Oxford, UK, 1996.

[van Someren *et al.* 2000] van Someren E.P., Wessels L.F.A., Reinders M.J.T., Linear modeling of genetic networks from experimental data, in: *Intelligent Systems for Molecular Biology (ISMB 2000)*, August 19-23, San Diego, CA, 2000.

[von Knobloch *et al.*, 2000] von Knobloch R., Bugert P., Jauch A., Kälble T., Kovacs G., Allelic changes at multiple regions of chromosome 5 are associated with progression of urinary bladder cancer, *J Pathol.*, 190: 163-168, 2000.

[Wahde and Hertz 2000] Wahde M. and Hertz J., Coarse-grained reverse engineering of genetic regulatory networks, *Biosystems*, 55, 129-136, 2000.

[Weaver *et al.* 1999] Weaver D.C., Workman C.T., Stormo G.D., Modeling regulatory networks with weight matrices, *Pac. Symp. on Biocomputing*, 4, 112-123, 1999.

[Zadeh 1983] Zadeh L.A., The role of fuzzy logic in the management of uncertainty in expert systems, *Fuzzy Sets and Systems*, 11, 199-228, 1983.

Appendix A

Microsatellite loci

chromosome 1:	D1S-214, 2635, 2844, 2799
chromosome 2:	D2S-2200, 119, 2215, 2313, 335, 369, 128, 2204, 172, 248, 259
chromosome 3:	D3S-1560, 1289
chromosome 4:	D4S-412, 3032, 2974, 408, 426
chromosome 5:	D5S-502, 419, 695, 2055
chromosome 6:	D6S-470, 1665, 1639, 1633
chromosome 7:	D7S-817, 2847
chromosome 8:	D8S-504, 264, 1806, 1824, 1781, 262, 518, 1819, 1469, 1109, 549, 261, 282, 1739, 1114, 1758, 593, 198, 1753
chromosome 9:	D9S-163, 288, 937, 921, 274, 156, 157, 925, 1684, 162, 1870, 1748, 171, 161, 1788, 1876, 153, 1815, 1689, 1809, 53, 154, 1872, 195, 1830, 1838.
chromosome 10:	D10S-1744, 541
chromosome 11:	D11S-922, 4088, 1338, 902, WT1, 4083, 903, 4174, 1344, 4117, 1313, 1357, 4191, 1908, 971, 1314, 901, 917, 1339, 4111, 924, 1328, 4131.
chromosome 12:	D12S-375, 391
chromosome 13:	D13S-168, 153
chromosome 14:	D14S-1039, 267
chromosome 16:	D16S-539, 3253
chromosome 17:	D17S-796, 1353, 786, 806, 787
chromosome 18:	D18S-474, 1119, 64, 42, 61, 58, 461, 70

Table A.1: Microsatellite loci analysed with LOH

Appendix B

BUGS code

Example of the BUGS code for the model OR. The operator ∼ means ’is distributed as’, and ← corresponds to ’is logically defined by’.

```
model OR;
const
    N=25, # number of samples in dataset
    P=19; # number of parents of the variable Y
var
    Y[N], X[N,P], # data
    gamma[P], # variable indicators
    theta[P], # model parameters
    a[P], # a and b - hyperparameters of Beta distribution for model parameters
    b[P],
    aprop[P], # proposal for hyperparameters a and b
    aprop[P],
    priormean[P], # mean and variance of the parameters obtained by pilot run of the
model
    priorvar[P],
    I[N,P], # intermediate states of the variables X
    s[i,j], # help variables to calculate intermediate states
    sum[N],
    constraint[N];
data X, Y in ”Spellman.dat”;

    {
    # model definition
    for (i in 1:N) {
       for (j in 1:P) {
          s[i,j] ∼ dbern(theta[j]);
          I[i,j] ← X[i,j]*s[i,j]*gamma[j];
       } #end j

          sum[i] ← sum(I[i, ]); # sum over j
          constraint[i] ← step(sum[i]-1);
          # constraint is 1, if at least one of the components of the sum is 1, otherwise 0
          Y[i] ∼ dbern(constraint[i]);
    }# end i

        for (j in 1:P) {
          # definition of priors for variable indicators
          gamma[j] ∼ dbern(0.5); # noninformative uniform
          # gamma[j] ← 1; # for pilot run of the model

              # definition of priors for model parameters
          theta[j] ∼ dbeta(a[j],b[j]);
```

```
                # definition of priors for hyperparameters
            # if model parameters do not depend on variable indicators,
            # noninformative uniform:
            a[j] ← 1;
            b[j] ← 1;

                # Gibbs Variable Selection
            # if gamma[j]=1, noninformative uniform
            # if gamma[j]=0 (pseudoprior), proposal distributions:
            a[j] ← gamma[j]*1+(1-gamma[j])*aprop[j];
            b[j] ← gamma[j]*1+(1-gamma[j])*bprop[j];

                # calculate proposal distributions, use mean and variance obtained
            # by pilot run of the model:
            aprop[j] ← (priormean[j]*(1-priormean[j])/pow(priorvar[j],2)-1)*priormean[j] ;
            bprop[j] ← (priormean[j]*(1-priormean[j])/pow(priorvar[j],2)-1)*(1-priormean[j]);
      }#end j
}
```

Zusammenfassung

Die vorliegende Dissertation ist das Ergebnis meiner Arbeit im Deutschen Krebsforschungszentrum (DKFZ) auf dem interdisziplinären Gebiet Bioinformatik. Dieses Gebiet entwickelt sich zusammen mit den Fortschritten der molekularen Genetik, deren experimentellen Techniken grosse Datenmengen produzieren. Die in den letzten Jahren entstandene *Microarray*-Technik ermöglicht die globale parallele Erfassung von Expressionswerten von Tausenden von Genen. Das Studium der Genexpression und Genregulation ist von grosser Bedeutung für die bio-medizinische Forschung, denn z.B. die Krebsentstehung und Progression ist grundsätzlich mit den Veränderungen der Genexpression und Regulation verbunden. Zu der Methodenreservoir der molekularen Genetik und Zytogenetik gehört auch die Allelotypisierung (LOH), die die Ermittlung von allelischen Verlusten in Tumorzellen ermöglicht. Eine weitere Methode ist die komparative genomische Hybridisierung (CGH), die erlaubt, die genomische Aberrationen in Tumorzellen zu erfassen. Die genomische Abnormalitäten sind mit Tumorphänotypen assoziiert und spielen eine wichtige Rolle für das Verständnis der Kanzerogenese.

Effektive Verarbeitung und Auswerten von molekulargenetischen Daten verlangen die entsprechend angepassten computerbasierten Methoden. Das charakteristische Merkmal der biologischen Daten ist ihre Ungenauigkeit, bedingt durch die stochastische Natur der biologischen Prozesse und ein erhebliches Messrauschen. Um diese Ungenauigkeit zu tolerieren und zuverlässig zu sein, sollen die Analysemethoden auf dem statistischen Ansatz basieren. Die grundsätzliche Fragestellung bei dem Auswerten von biologischen Daten besteht darin, neue Muster, neue Zusammenhänge zwischen Ereignissen zu gewinnen, diese zu evaluieren und zu interpretieren, um bestehende Hypothesen prüfen oder gar neue stellen zu können, und Vorhersagen zu machen. Hierbei spielt die statistische Modellierung eine grosse Rolle, denn ein statistisches Modell erlaubt, den grundlegenden biologischen Prozess abstrakt, mit Hilfe von Parametern und eventuell von verborgenen Variablen, zu repräsentieren. Algorithmen zu entwickeln, die fähig sind, ein Modell aus den Daten zu lernen, ist eine Herausforderung für Informatiker.

Im Kernpunkt dieser Arbeit steht die Anwendung der probabilistischen grafischen Modellen. Diese Modelle sind für die Modellierung der biologischen

Prozesse vorteilhaft aufgrund ihrer Fähigkeit, komplexe Abhängigkeiten zwischen den Zufallsvariablen zu repräsentieren. Das Lernen eines probabilistischen grafischen Modells besteht aus der Suche nach einer optimalen Graphenstruktur und nach einem Parametersatz, die am besten an die gegebenen Daten angepasst sind. Die Analyse der grafischen und der quantitativen Struktur des Modells ermöglicht den Einblick in die Beziehungszusammenhänge von Variablen zu gewinnen.
In der vorliegenden Arbeit befasse ich mich mit den Daten von zwei Arten: LOH-Daten und Mikroarray-Daten, daher ist diese Arbeit in zwei Teile aufgeteilt. Im ersten Teil sollte man die mögliche Progression der allelischen Verluste in urothelialen Karzinomen rekonstruieren. Man sollte die Hypothesen aufbauen darüber, welche Abnormalitäten für die Entwicklung des Karzinoms primäre sind, und welche Abnormalitäten sich in Folge der vorherigen genetischen Veränderungen akkumulieren.
Bei dieser Fragestellung habe ich das probabilistische grafische Modell "Bayes'sche Netzwerke" angewandt. Ein Bayes'sches Netzwerk ist ein gerichteter azyklicher Graph, der die multivariaten Abhängigkeiten von Zufallsvariablen (hier - von allelischen Verlusten) darstellt. Dabei lässt sich die gemeinsame Verteilung der Variablen über das Produkt bedingter Wahrscheinlichkeiten definieren. Die bedingten Wahrscheinlichkeiten, d.h. Parameter des Modells, quantifizieren die Abhängigkeit von einer Variable von ihren Vätern in der Graphenstruktur.
Der Vorteil des Bayes'schen Netzwerk-Modells gegenüber den vorherigen Baum-basierten Ansätzen ist, dass das Modell die allgemeinste Abhängigkeitsstruktur erfassen lässt und erlaubt, die Heterogenität der Tumorentwicklung darzustellen. Die mathematischen Grundlagen für das Lernen von einem Bayes'schen Netzwerk aus den Daten ist die Bayes'sche Modellierung, ein moderner Zweig der Statistik. Ich habe das Lernverfahren für Bayes'sche Netzwerke angewandt, um die Graphen- und Parameterstruktur zu induzieren, die die Abhängigkeiten zwischen den allelischen Verlusten repräsentieren. Mit Hilfe des induzierten Modells kann man die probabilistische Inferenz durchführen, d.h. ermitteln, welche Wahrscheinlichkeit eine Variable hat wenn den Zustand von einer oder mehreren anderen Variablen bekannt ist. Die Analyse der Graphenstruktur und die probabilistische Inferenz ermöglichten, die interessanten Muster und Zusammenhänge zwischen den allelischen Verlusten zu entdecken, und die Progression der allelischen Veränderungen entlang der möglichen Tumorentwicklungswege zu beschreiben.
Der zweite Teil der vorliegenden Arbeit konzentriert sich auf die regulatorischen Beziehungen der Gene auf der Ebene der Expression. Die Daten über die Expression der Gene bei unterschiedlichen Zellzuständen und unter unterschiedlichen experimentellen Bedingungen werden in den Mikroarray-Experimenten gemessen. Transkriptionsfaktoren oder andere Signalproteine,

die von bestimmten Genen exprimiert werden, haben eine aktivierende oder inhibierende Wirkung auf mehrere Zielgene. Dabei wirken die Faktoren im Verbund miteinander. In Folge dieser kombinatorischen Interaktion kommt die Genregulation zustande. Man spricht hier von genetischen Regulationswegen. Die regulatorischen Beziehungen von Genen aus den Mikroarray-Daten zu rekonstruieren, war die weitere Fragestellung meiner Arbeit. Dazu habe ich das Modell definiert, das die Wirkung der Regulatoren auf das Zielgen mit der Boole'schen Logik festlegt. Im Unterschied zu den vorherigen Ansätzen von deterministischen Boole'schen Netzwerken, hat das Modell eine probabilistische Semantik, indem die Wirkung eines aktiven Regulators mit einer Wahrscheinlichkeit ausfallen kann. Somit enthält das Modell verborgene Variablen und Parameter. Ich habe die Bayes'sche Methodologie angewandt, um das Lernverfahren zum Lernen des Modells aus den Daten zu entwickeln. Da das Problem keine exakte Lösung hat, habe ich das *Markov Chain Monte Carlo* Simulationsverfahren angewandt, nämlich *Gibbs sampling.*

Diese Arbeit demonstriert die neuen Möglichkeiten der Anwendung der probabilistischen grafischen Modelle und der Bayes'schen Modellierung in der Bioinformatik und in der onkologischen Forschung.