Bayesian Networks for Omics Data Analysis

Anand K. Gavai

Promotoren:

Prof. dr. J.A.M. Leunissen
Hoogleraar Bioinformatica
Laboratorium voor Bioinformatica
Wageningen Universiteit

Prof. dr. M.R. Muller
Hoogleraar Voeding, Metabolisme en Genomics
Afdeling Humane Voeding
Wageningen Universiteit

Co-promotoren:

Dr. ir. G.J.E.J. Hooiveld
Universitair docent
Afdeling Humane Voeding
Wageningen Universiteit

Dr. P.J.F. Lucas
Universitair hoofddocent
Institute for Computing and Information Sciences
Radboud Universiteit, Nijmegen

Promotiecommissie:

Prof. dr. J.N. Kok (Universiteit Leiden)
Dr. ir. C.T. Evelo (Universiteit Maastricht)
Prof. dr. R.J. Bino (Wageningen Universiteit)
Dr. J.H.G.M. van Beek (Vrije Universiteit Amsterdam)

Dit onderzoek is uitgevoerd binnen de onderzoeksschool VLAG

Bayesian Networks for Omics Data Analysis

Anand K. Gavai

Proefschrift
ter verkrijging van de graad van doctor
op gezag van de rector magnificus
van Wageningen Universiteit,
Prof. dr. M.J. Kropff,
in het openbaar te verdedigen
op maandag 8 juni 2009
des voormiddags te elf uur in de Aula

Anand K. Gavai, 2009

Bayesian Networks for Omics Data Analysis

Thesis Wageningen University, Wageningen, The Netherlands.
With Summaries (English & Dutch).

ISBN: 978-90-8585-390-9

"This thesis is dedicated to my father Kamalnayan Vasantrao Gavai for his endless faith in my abilities"

Contents

Chapter 1

Introduction

"There is much more learning than knowing in this world."

Unknown

This thesis is about making sense of data obtained from measurements of biological processes using modern computational methods, in particular based on the theory of probabilistic graphical models with a major emphasis on Bayesian networks. Where there is a plethora of other methods for data interpretation available, both in literature and in computer-based tools, few of them are able to bridge the gap between the actual data and biological meaning directly. The more narrow this gap is, the easier it is for the biologist to draw conclusions from the experiments by means of which the data were generated. Probabilistic graphical models consist of a joint probability distribution, used to represent the uncertainty that is in the data modeled probabilistically, and an associated graph that qualitatively indicates which attributes in the dataset are conditionally dependent and independent of others. It is in particular the qualitative reading of probabilistic graphical models that is attractive to the biologists, and, thus a major part of this thesis is devoted to exploring these techniques in the analysis of biological data.

In this chapter, first the biological problems studied in this thesis are reviewed, which is followed by a review of the main reasons why probabilistic graphical models were chosen as the main vehicle of the research described. This chapter is rounded off by a brief summary of the remainder of the content of this thesis.

1.1 The biological data explosion

Lately, high throughput technologies like microarrays, mass spectrometry and protein chips have gained much interest in the biological domain. These techniques allow one to measure thousands of entities simultaneously. Often these experiments are conducted at genomic, metabolomic or proteomic levels. The values measured for genes, proteins or metabolites are examined in great detail to understand the mechanics behind functioning of a living organism. Often the data generated by these techniques come in the form of a matrix, where rows represent observations and columns represent entities (gene, metabolite or protein). These data sets are generally very large and often the number of observations is small compared to the entities themselves. This problem is often termed as *"large p and small n problem"*. Therefore it becomes a major challenge to find

significantly expressed entities in a biological context. Various analytical techniques have been used to look for these subtle changes in expression pattern to infer the biological meaning. Complex statistical models are employed to understand these processes. However standard statistical approaches cannot take into account non-linear relations and incorporation of prior knowledge. These bottlenecks are very well tackled by the use of probabilistic graphical models. These models allows one to discover *causal* relationships instead of mere *correlations*. The emphasis in this thesis is on Bayesian networks. Normally this type of analysis is preceded by pre-processing and normalization steps only later to be interpreted in a biological context. Various steps involved in this process can be chained together to form a complete life cycle of an experiment. Automating these steps requires a framework where data can be stored and managed; therefore databases play an important role to manage and analyze these data in a systematic fashion. The next step then is to combine the data from multiple experiments to look at the organism at systems level. These techniques are later applied to data from metabolomics. Due to general applicability of these methods they are used at all *-omics* levels.

The aims of the research underlying this thesis were:

- develop a framework where data generated from microarray experiments can be stored, managed and analyzed;
- explore probabilistic graphical models as a technique for data interpretation, in particular for data obtained from microarrays;
- use this technique to analyze microarray data generated from other microarray platforms in combination with metadata; and finally
- apply these techniques to data generated in other domains like metabolomics.

1.2 Microarray databases in nutrigenomics research

Microarrays are playing an important role in modern day genomics. This is a technique to measure transcript level of several genes from a certain experimental condition. These devices are made of glass slides containing thousands of small fragments of DNA from the organism under study. Microarrays are widely being used for analyzing the genetic material obtained from such experiments. The generated data sets are huge and therefore difficult to manage, maintain and interpret. There are four different types of data arising from a successful microarray experiment:

- The generation of data from hybridization of arrays,
- The extra information attached with these data (the metadata),
- The data generated after performing standard quality control checks and normalization procedure, and

- The data generated after high level analysis of these data to answer a biological question.

The main objective of these analysis is to answer a biological question for which standard procedures are available to perform analysis in Bioconductor (Gentleman et al. 2004). Bioconductor is an open source initiative to genomic analysis. It consists of several packages to perform analysis at various levels. Packages can be chosen to perform various tasks (normalization or high-level analysis) which can be hooked together to perform a specific task and are often termed as pipelines. These packages can be chained together using R (R Development Core Team 2008) which is a free software environment for statistical computing and graphics. Most *omics* analysis lately is performed in R. Therefore support for this platform determines the usability, credibility and extensibility of a system.

There are various challenges which arise at different level of storage and analysis. For e.g. Genetic material is hybridized on various microarray technologies (affymetrix, Agilent, cDNA etc), therefore the data generated is in different format. The metadata which plays an important role in answering a biological question is not recorded at all or partially recorded. There are various quality control checks and normalization procedures which are continuously evolving for various technologies and finally the high level analysis to answer a biological question has no standard techniques.

Various systems have been emerging over the last decade to address these core issues. A detailed comparison of some popular databases can be found in (Gardiner-Garden and Littlejohn 2001). Due to consistent change in the microarray technology and standards it has become an important challenge to built a system which can address these core issues together and minimize the life cycle of a successful microarray experiment from several months to days or even hours. There are several microarray systems being developed with varied features, these databases are used to store many different type of data format for e.g.two dye arrays and oligonucleotide arrays. There have also been systems commercially available to maintain these data (Biosoftware 2000). The question remains at what level data should be stored. The data generated from microarray experiments are the gene expression profiles which are calculated by statistical algorithms. Due to considerable development new techniques are being developed and new ways are being discovered to analyze these data, these techniques require raw data to be analyzed further. Many of the modern day systems if not all lack this. Because of which users are stuck with already processed data, this makes it even more difficult for others to reevaluate the results using new techniques which are being developed.

Most journals lately require data to be made publicly available in public repositories like GEO (Edgar et al. 2002) and ArrayExpress (Parkinson et al. 2007). These repositories require the experiments to fulfill the minimum criteria to measure the credibility of experiments. These criteria include proper LIMS (Laboratory Information Management System) information to be recorded; and is standardized using MIAME (Minimum information about a microarray experiment) (Brazma et al. 2001). To facilitate the exchange between various systems a standard xml

based format has been defined MAGE-TAB (Rayner et al. 2006) which is a simple spreadsheet based, MIAME-supportive format for microarray data.

These missing feature have already been taken care by some of the upcoming databases. A brief overview of all the popular databases is presented in table1.1. Our main focus here is on the functional aspects of a system rather than the technological aspects (viz. softwares and platforms).

1.3 Probabilistic graphical models

Graphical models have become very popular in recent times in biological research and in particular graphical models that allow encoding joint probability distributions, so-called *probabilistic graphical models*, attract a lot of attention. Although these types of probabilistic models have been successful, they only started to have a major impact on the research areas of uncertainty reasoning in artificial intelligence and statistical machine learning after the publication of the book *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* in 1988 by Judea Pearl. Although probabilistic graphical models were already known prior to the publication of Pearl's book, these models were associated with undirected graphs. Pearl's book introduced probabilistic directed graphical models, called originally *belief networks* by Pearl, whereas later the term *Bayesian networks* became predominant. He clearly shows in his book that these probabilistic graphical models are not only powerful for the efficient representation and reasoning with probabilistic information, but, moreover, they have a very strong intuitive basis. It is this latter feature which explains why Bayesian networks have become so popular in a relative short period of time.

Although probabilistic graphical models have dominated the area of reasoning under uncertainty in artificial intelligence since the end of the 1980s, it is only since 2000 that researchers have started to explore the use of these techniques for the interpretation of biological data. The reason why graphical models are attractive to the biological domain is that most biological processes are hierarchical in nature like cellular processes, gene regulatory networks and metabolic pathways. It is a branch of machine learning which overlaps very much with statistics and deals with representation of high dimensional data in the form of nodes and edges where nodes represent random variables and edges represent statistical dependencies between them. Historically both these fields have its own origins, the main difference between machine learning and statistical techniques is one concerns with testing hypotheses whereas other deals with development and examining the hypothesis; so where statisticians prefer theoretical soundness machine learners require both theoretical soundness and experimental effectiveness. Statistics are often good when it comes to describing data whereas machine learning helps not only to describe the data but also helps in predicting future outcomes. Machine Learning focuses on the question of how to get computers to program themselves (from experience plus some initial structure). Whereas statistics has focused primarily on what conclusions can be inferred from data (Mitchell 2006). In recent times statistics has benefitted a lot from machine learning and vice versa. In biology the

Database	Open source	Array types storage	Low level analysis	High level analysis	Web address
BASE	Yes	Affymetrix, Agilent, Genepix, illumina	Yes (Plugin based)	No	http://http://base.thep.lu.se/
Rosetta Resolver	No	Affymetrix, Agilent, Illumina, and GE Healthcare.	No (Workflow based in R)	No (Workflow based in R)	http://www.gosettabio.com/
ArrayExpress	Public Repository	Affymetrix, Agilent, illumina, c-DNA,	Some extent	Third party	http://www.ebi.ac.uk/arrayexpress/
AMAD	Yes	C-DNA	No	Yes (J-Express)	http://www.microarray.org/software.html
SNoMAD	Yes	Affymetrix, c-DNA	Yes (R usage)	No	http://pevsnerlab.kennedykrieger.org/sNomadinput.html
ExpressDB	Public Repository	Affymetrix, c-DNA	No	No	http://salt2.med.harvard.edu/ExpressDB/
SMD	Public Repository	c-DNA	Yes	No	http://geNome-www4.stanford.edu/Microarray/SMD/
GEO	Public Repository	Affymetrix, Agilent, illumina, c-DNA,	Yes	Third party	http://www.cbil.upenn.edu/RAD/php/index.php
J-Express	Public	No	No	Yes	http://www.ii.uib.No/ bjarted/jexpress/
Expression Profiler	Public	No	No	Yes	http://www.ebi.ac.uk/expressionprofiler/
MAGIC tool	Public	No	No	Yes	http://www.bio.davidson.edu/projects/magic/magic.html
MIMIR	Public Repository	Affymetrix, Agilent, illumina, c-DNA,	Third party	Third party	http://microarray.csc.mrc.ac.uk/subsection.html?id=28
MARS	Public Repository	Affymetrix, c-DNA	Yes	Yes	http://geNome.tugraz.at
SBEAMS	Public Repository	Affymetrix	Yes	No	http://www.sbeams.org/Microarray/
GEPAS	Yes	No	No	Yes	http://transcriptome.ens.fr/gepas/
Gecko	Yes	Affymetrix	No	Yes	http://sourceforge.net/projects/geckoe.
MicroGen	No	Affymetrix, c-DNA	Yes	Third party	http://microgen.ouhsc.edu/inbre/
PlasmoDB	Public Repository	Affymetrix	NA	NA	http://plasmodb.org/plasmo/
maxD	No	Affymetrix, Agilent, illumina, c-DNA	No	No	http://www.bioinf.manchester.ac.uk/microarray/
MADMAX	Yes	Affymetrix, Agilent, Illumina, c-DNA	Yes	Yes	https://madmax.bioinformatics.nl

Table 1.1: *Web addresses of popular microarray databases and analysis solutions in use*

core interest lies in understanding how cellular systems function and adapt to their environment; and it is believed that techniques like these can help to solve some of the issues in this domain. There are now many different types of probabilistic graphical models, such as undirected graphical models, called *Markov networks* or *Markov random fields*, directed graphical models, the Bayesian networks just mentioned, hybrid graphical models, which may include both undirected and directed edges, called *chain graph models*, and *maximal ancestral graphs*, which allow one to model reasoning with observations and hidden variables and include undirected, directed and bidirected edges.

Probabilistic graphical models are multivariate probabilistic distributions that offer a unified and simplified view on multivariate statistics (Whittaker 1990). Probabilistic graphical models allows us to talk about conditional dependencies and independencies between variables in an intuitive way by representing, e.g. family trees, genetic network, circuit diagrams etc. Graphs also allows us to be abstract about conditional independence relationships between variables, handles incomplete datasets and facilitates the combination of domain knowledge and data. Thus while conventional techniques rely completely on data, graphical models provide us with a more holistic approach where domain knowledge, for example in the form of a graph structure and *subjective* probabilities, can be efficiently combined with data, i.e. *objective* probabilities. Thus it allows us to answer questions like *Is X dependent on Y given that we already know the value of Z* just by looking at the graph. Moreover graphs provide us with an efficient way to perform computation over a network. Probabilistic graphical models essentially represent a marriage between statistics, graph theory and computer science.

$$\textit{probabilistic graphical models} = \textit{statistics} \times \textit{graph theory} \times \textit{computer science}$$

There are many applications in biology where classifications of data is required for e.g. gene function prediction. A computer program can be easily made by defining these rules by setting *if-else* statements, but in reality we cannot take all the possibilities into account as there are many exceptions to these rules. This is what is referred as *uncertainty* in a system which comes into play regularly in a biological domain. Graphical models allow us to model these uncertainties in practice. A graphical representation provides a simple way to visualize the structure of the model, and can provide new insights to make new models. In the rest of this thesis variables,nodes and entities (genes, metabolites or proteins) represent the same meaning, unless explicitly stated.

1.4 Outline of this thesis

The remainder of this thesis is organized as follows. The second chapter summarizes the basics of probability theory and probabilistic graphical models, Bayesian networks in particular; it lays the background of the techniques used in the rest of the chapters. The third chapter presents MADMAX (Management and Analysis Database for Multi-platform microArray Experiments),

a specialized database used for managing, storing and analyzing microarray data. The fourth chapter shows how Bayesian networks can be used to perform microarray analysis on feeding and fasting conditions in mice. The fifth chapter uses this technique on a smokers/non-smokers data set generated using Agilent arrays in combination with other experimental data to get significant results. Chapter six demonstrates the use of Bayesian networks in metabolomics experiments to show the applicability of this technique. Finally, chapter seven provides a general discussion on the work done in this thesis with the current trends in research in this domain.

Chapter 2

Preliminaries

"I have never met a man so ignorant that I couldn't learn something from him."

Galileo Galilei

This chapter is devoted to a review of some of the basic theory underlying the methods and tools used in the remainder of this thesis. In particular, the theory of Bayesian networks is reviewed.

2.1 Probability distributions

We start by giving a brief, informal review of the basic concepts used in probability theory.

2.1.1 Events and variables

The state of the world is often described in terms of the so-called *events* E that occur. As several events may occur alternatively or concurrently such joint events need to be described, which is done using Boolean operators such as negation ($\neg E$, also denoted $\bar{E}$), conjunction (($E \wedge E'$), also denoted ($E \cap E'$), or (E, E')), and disjunction (($E \vee E'$), also denoted ($E \cup E'$)). All events are defined as subsets of a so-called *sample space* $\mathcal{S}$, which includes all possible outcomes of an experiment. Often, it is more convenient to decompose a sample space into several, possibly disjoint, subsets, which introduces the notion of a *variable*. Variables may take elements of their associated domain as value, which is denoted as $X = x$, where X denotes the variable, or sets of variables, and x the value the variables take on.

In this thesis, a variable represents the state of a gene, metabolite, or protein. For example the state of a gene G can be up or down regulation, which can be denoted by $G = up$ and $G = down$, respectively.

2.1.2 Probability distributions

By attaching a number from the closed real interval $[0, 1]$ to events by means of a mapping P, a *probability distribution* can be defined by ensuring it fulfils the well-known axioms of probability theory first stated by A. Kolmogorov (Kolmogorov 1956). A probability distribution expresses

uncertainty about the occurrence of events, for example, due to measurement errors. For any set of variables X with associated domain $D(X)$, it is stated that the values from $D(X)$ are exhaustively, and mutually exclusive. This implies that it holds for a set of discrete variables that $P(X = x, X = x') = 0$, where $x \neq x'$, are two different values of X, and in addition that $\sum_{x \in D(X)} P(X = x) = 1$.

There is in general no difference in theory between whether X is a single variable, or singleton set, or a non-singleton set of variables. However, in a biological setting probability distributions become particularly interesting if X represents a non-singleton set as only then they can be used to represent and explore relations between variables. A probability distribution $P(X)$ where X is a set of variables is called a *joint* or *multivariate* probability distribution. Special techniques have been developed to cope with joint probability distributions including several variables. Coping with statistical independence information becomes particularly relevant when representing joint probability distributions, as the size of a discrete joint probability distributions is exponential in the number of variables. This storage requirement can be greatly reduced by exploiting independence information, as will be discussed below.

2.1.3 Conditional probability distributions

A conditional probability is denoted by $P(A \mid B)$, and is a way of calculating the probability of event A when we are absolutely certain that event B has occurred. By definition:

$$P(A \mid B) = \frac{P(A, B)}{P(B)}, \tag{2.1}$$

whenever $P(B) > 0$. This equation makes explicit the fact that the uncertainty with respect to event B is divided out in the uncertainty in the joint event $(A \wedge B)$, and, thus, only event A remains uncertain.

As an example, suppose that we observe that gene B is expressed 12 times under particular experimental conditions, and, of those 12, gene A is also expressed 10 times, then $P(A \mid B)$ would be estimated at $10/12 \approx 0.83$. In other words, the probability that gene A is expressed given the fact that gene B is already expressed can be estimated as being 83% by using a method that involves counting the occurrence of the events, which is translated into relative frequencies of the events A and B.

If A tends to occur when B occurs, then knowing that B has occurred allows us to assign a higher probability to A's occurrence than in a situation in which we did not know that B occurred. This idea can be expressed formally by comparing the probability distribution $P(B \mid A)$ and the prior distribution $P(A)$. More generally, if A and B systematically co-vary in some way, then $P(A \mid B)$ will not be equal to $P(A)$. Conversely, if A and B are independent events, then $P(A \mid B)$ would be expected to equal P(A). The need to compute a conditional probability thus arises any time we think the occurrence of some event that has a bearing on the probability of another event's occurring.

The most basic and intuitive method for computing $P(A \mid B)$ is "the set enumeration method". Using this method, $P(A \mid B)$ can be computed by counting the number of times A and B occur together, $N_{A,B}$, and dividing by the number of times B occurs, N_B, resulting in the formula:

$$P(A \mid B) = \frac{N_{A,B}}{N_B}, \tag{2.2}$$

i.e. resulting in the well-known relative frequency definition of probabilities due to Laplace (de Laplace 1812).

From the definition of conditional probability distributions, it is straightforward to derive the well-known Bayes' theorem. From Equation (2.1) it follows that:

$$P(A, B) = P(A \mid B)P(B). \tag{2.3}$$

Similarly, we also have that:

$$P(B, A) = P(B \mid A)P(A). \tag{2.4}$$

As conjunction $\wedge$, which is also represented as $P(A, B)$ is commutative and holds that $P(A, B) = P(B, A)$, and, thus, we can write:

$$P(A \mid B)P(B) = P(B \mid A)P(A). \tag{2.5}$$

By rewriting this equation, we arrive at *Bayes' theorem*:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}, \tag{2.6}$$

whenever $P(B) > 0$. Notice that this formula for computing a conditional probability is similar to the original formula with the exception that the joint probability $P(A, B)$ that used to appear in the numerator has been replaced with the equivalent expression $P(B \mid A)P(A)$.

Finally, Equation (2.3) is sometimes called the *factorization theorem*. This theorem can also be used recursively, such as in:

$$P(A, B, C) = P(A \mid B, C)P(B, C) = P(A \mid B, C)P(B \mid C)P(C),$$

which allow decomposing any joint probability distribution into factors.

2.1.4 Summarizing genetic data

Joint probabilities in broad terms describe everything about what the data has to say. Consider the following example of a dataset in Table 2.1, where gene A is a regulator and gene B is the target. Here 1 represents *down* regulation and 2 represents *up* regulation. It is generally agreed

upon that a fold change of less than 0.70 is down regulation of a gene and a fold change of 1.2 is up regulation of the gene. This process of labeling values into 1 and 2 is fairly common and is called discretization. This technique allows us to check and update states of the variables on the fly. Instead of discrete random variables, it is also possible to use *continuous* variables, resulting in continuous probability distributions. To convert this table of frequencies to a table of probabilities,

gene A	gene B
1	1
2	2
2	2
2	1

Table 2.1: *Discretized values for genes A and B.*

we divide each cell frequency by the total frequency. Note that dividing by the total frequency also ensures that gene $A \times$ gene B cell probabilities sum to 1 as can be seen from Table 2.2 below as the joint probability distribution of gene A and gene B.

		gene B	
		1	**2**
gene A	**1**	1.0	0
	2	0.33	0.67

Table 2.2: *Probabilities of genes for being* up *or* down *regulated.*

To summarize, we can compute a conditional probability $P(A = 2 \mid B = 2)$ from joint distribution data by dividing the relevant joint probability $P(A = 2, B = 2)$ by the relevant marginal probability $P(A = 2)$. As we see, it is often easier and more feasible to derive estimates of a conditional probability from summary tables like this, rather than expecting to apply more data-intensive enumeration methods.

From a statistical point of view, joint discrete probability distributions which will be explored in this thesis follow a multinomial probability distribution, or binomial in the binary case. The continuous probability distributions are assumed to be Gaussian or normal.

2.2 Conditional independence and dependence

Conditional independence is not only one of the central notions of probability theory: it is also central to using probability theory for data analysis and probabilistic model building. As an example, consider that a grandfather and a grandson are related to each other only because of the existence of a father. Knowing the characteristic of both grandfather and father does not convey

more information than only knowing the characteristics of the father: grandson and grandfather are conditionally independent given the father.

We already know that the conditional probability of A given B is represented by $P(A \mid B)$. Formally, the sets of variables A and B are said to be *conditionally independent given* the set C if

$$P(A \mid B, C) = P(A \mid C).$$

An alternative, but equivalent, definition is that A and B are conditionally independent given the set C if it holds that

$$P(A, B \mid C) = P(A \mid C)P(B \mid C).$$

This is symbolically denoted by $A \perp\!\!\!\perp_P B \mid C$. The $\perp\!\!\!\perp_P$ can be seen as a ternary relation between subsets of random variables. If the set C is the empty set, i.e. $C = \emptyset$, then A and B are called *marginally* or *unconditionally* independent. The independence relation has specific properties that can be used to derive new independence statements from given independence statements. For example, the simplest property is that from $A \perp\!\!\!\perp_P B \mid C$ it can be derived that $B \perp\!\!\!\perp_P A \mid C$ holds, which is known as the *symmetry* axiom (Pearl 1988). It is known that the independence relation cannot be fully characterized by a finite number of independence axioms. This is a clear indication of the complexity of this relation.

Finally, it is worth realizing that a set of variables A and B is either conditionally independent given a third set C or conditionally *dependent* given C, but never both. Conditional dependence is represented by the relation $\not\perp\!\!\!\perp_P$. Let V be the total set of variables, then we thus have that:

$$\wp(V) \times \wp(V) \times \wp(V) = \perp\!\!\!\perp_P \cup \not\perp\!\!\!\perp_P,$$

and $\perp\!\!\!\perp_P, \not\perp\!\!\!\perp_P \subseteq \wp(V) \times \wp(V) \times \wp(V)$, with $\perp\!\!\!\perp_P \cap \not\perp\!\!\!\perp_P = \emptyset$, and $\wp(V)$ the powerset of V, i.e. the set of all subsets of V.

2.3 Bayesian networks

2.3.1 Basic concepts

Many times in literature Bayesian learning and Bayesian statistics, which are specialized fields in themselves, are confused with Bayesian networks. Bayesian learning and statistic is about taking into account, and updating, *prior* uncertain knowledge when building and selecting statistical models. Bayesian networks, on the other hand, are structured joint probability distributions that have not necessarily a direct relationship to Bayesian learning or statistics. The adjective 'Bayesian' in Bayesian networks derived from the use of Bayes' theorem in reasoning with a given joint probability distribution. As Bayesian networks are instances of statistical models, they can in principle also be learnt using Bayesian learning methods. However, traditional maximum likelihood estima-

tor methods can also be used. Bayesian networks has its roots in the domain of machine learning then classical multivariate statistics. A detailed comparison on both these field can be found in (Cunningham 1995).

Bayesian networks are a type of a probabilistic graphical models. They are currently playing an increasingly role in analyzing data of high dimensionality. They consist of two parts, referred as the *quantitative part*, being a factorized joint probability distribution P on a set of random variables X, and the *qualitative part*, being an acyclic directed graph, *ADG*[1] for short, $G = (V, E)$. The ADG consists of a set of *nodes* or *vertices* V and a set of directed *edges*, or *arcs*, $E \subseteq V \times V$. An ADG is a graph, where if you start in one node and follow the directions of the edges on a path, you do not come back to the same node again, i.e. there are no *cycles*. The reason why Bayesian networks have an acyclic graph representation is due to the finite factorization of the variable, i.e. the factorization of the joint probability distribution $P(A, B, C) = P(A \mid B, C)P(B \mid C)P(C)$: it stops after 3 terms and from C we do not start again with A.

Formally a Bayesian network $\mathcal{B}$ is defined as a pair $\mathcal{B} = (G, P)$. There exists a one to one correspondence between the random variables in X and the vertices in V, i.e. for any random variable in X there exists exactly one and only one node in V and vice versa. It is common to use the same names for both nodes and variables, although formally, nodes and variables are different concepts. In this thesis, we will do so likewise.

An example Bayesian network can be found in figure 2.1. There are situations in biology where one wishes to model feedback loops; to model such loops dynamic Bayesian networks can be used (Murphy 2002).

Next, some of the terminology used in the thesis will be reviewed. Let again $G = (V, E)$ be an ADG. A pair $(u, v) \in E$ is denotes a directed edge or an arc $u \to v$ from u to v and it is said that u is a *parent* of v, denoted $\mathrm{pa}(v)$, and v is said to be a *child* of u, denoted $\mathrm{ch}(u)$. For undirected graphs $G = (V, E)$, the set of nodes that are connected by an undirected edge to a given node u are called its *neighbors*, denoted by $\mathrm{ne}(u)$. Undirected graphs are sometimes used in this thesis for the construction of an ADG. Often such a directed relationship between nodes also represents a *cause-effect* relationship. A path $v_1, v_2, \ldots, v_n$ is a set of distinct nodes such that $v_i \to v_{i+1}$ or $v_i \leftarrow v_{i+1}$, for each i. A path is called directed if $v_i \leftarrow v_{i+1}$ or $v_i \to v_{i+1}$, for each i. A node v is called a *descendant* of a node u if there is a directed path from u to v in the graph.

The ADG part of a Bayesian network is used to constraint its associated joint probability distribution. This is basically done by reading off independence relations from the graph, which then by definition also hold for the joint probability distribution. The method by means of which independence statements are derived from an ADG is called the *d-separation*, as mentioned in Pearl (Pearl 1988). As an example, consider figure 2.2, which depicts three types of subgraphs which are relevant in interpreting a Bayesian network. The subgraphs $X \to Y \to Z$ and $X \leftarrow Y \leftarrow$

[1] ADG is also often referred to as *directed acyclic graphs* (DAGS); however the term ADG is more precise, since the term DAG implies a *directed* version of an *acylic graph*, which is not well-defined.

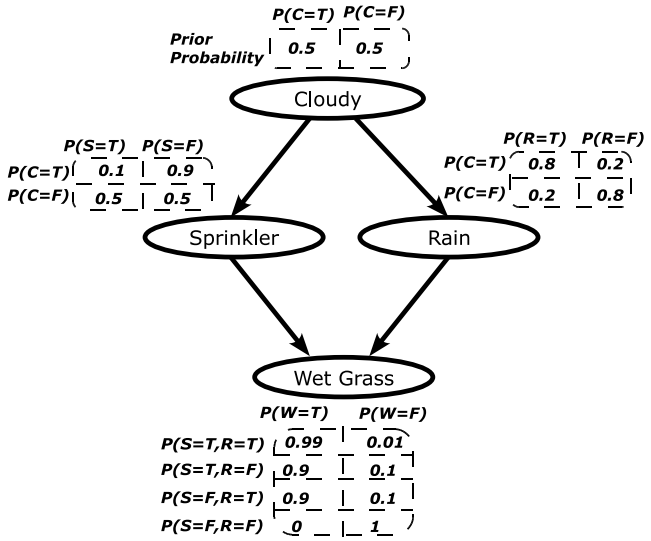


Figure 2.1: *A Bayesian network representation, where the nodes and edges represent the qualitative part and the tables represent the quantitative part.*

Z, called *serial connection* and $X \leftarrow Y \rightarrow Z$, called *diverging connection*, are equivalent. If node Y in a serial or diverging connection is known, then it blocks information flow on the path between node X and Z. It is also said that node Y *d-separates* node X from Z and the other way around. There is also the non-equivalent *converging connection*, also depicted in the figure. Here it holds that node X and Z are d-separated if Y is *not* observed; if Y or a descendant of Y is observed, then X and Y become *d-connected*, i.e. they are able to exchange information. Note the difference in interpretation between the diverging and serial connection, on the one hand, and the converging connection, on the other hand. One way to understand this somewhat puzzling interpretation of the converging connection is that node Y (or one if its descendants) can be seen as the *common consequence* of two *independent* causes X and Z. As soon as the common consequence has been observed, knowing also one of the two causes has an effect on our knowledge about the other cause, as one of the causes may already sufficient to explain the observed common consequence. This type of reasoning, called *explaining away* by Pearl (Pearl 1988), has a very natural intuition in the biological and medical sciences.

D-separation is represented by the ternary relation $\perp\!\!\!\perp_G$, i.e. $U \perp\!\!\!\perp_G V \mid W$ should be read as saying that any path between any node in U and V is blocked or d-separated by a node in W. D-connection is denoted by the ternary relationship $\not\perp\!\!\!\perp_G$.

All three of these causal situations give rise to different conditional independence of the asso-

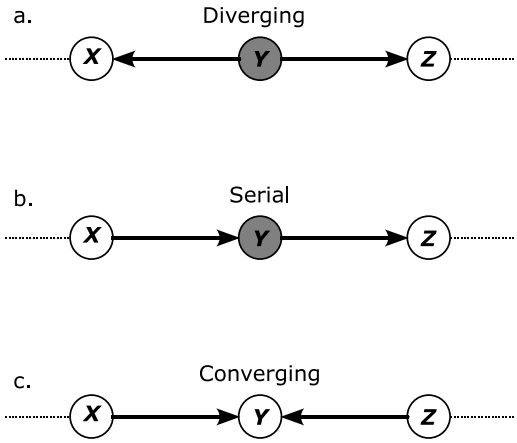


Figure 2.2: *(a) Y blocks (d-separates) X and Z: $X \perp\!\!\!\perp_G Z \mid Y$; (b) Y blocks (d-separates) X and Z: $X \perp\!\!\!\perp_G Z \mid Y$; (c) Y d-connects X and Z: $X \not\perp\!\!\!\perp_G Z \mid Y$ (the same holds for descendants of Y); it also holds that $X \perp\!\!\!\perp_G Z \mid \varnothing$, or X and Z are d-separated given nothing.*

ciated random. Formally d-separation is expressed as follows. Two nodes in an ADG $G = (V, E)$ are said to be d-separated by a set of nodes $S \subseteq V$ if there is a node w such that either:

- $w \in S$ and w does not have a converging connection on any path connecting the two nodes and w on this path, or
- $w \notin S$ and neither is any of the descendants of w in S.

The relationship between d-separation in an ADG G and conditional independence in a joint probability distribution P is that if two sets of nodes are *d-separated* given an third set of nodes in the ADG, then the corresponding sets of variables are conditionally independent given the set of variable corresponding to the third set of variables. Formally, it is said that the ADG G is an *independence map*, or *I-map* for short, of the joint probability distribution P. This can also be expressed as follows. For any set of node and identically named variables $A, B, C \subseteq V$ the I-map condition implies that it holds that

$$A \perp\!\!\!\perp_G B \mid C \Rightarrow A \perp\!\!\!\perp_P B \mid C.$$

Thus, all independence statements derived by using d-separation in a graph G are preserved in the associated probability distribution P; however, not necessarily all by d-separation derived depen-

dencies in the graph G are preserved in P.

Using these results, the joint probability distribution P of a Bayesian network $\mathcal{B} = (G, P)$ can be factorized as follows:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \text{pa}(X_i)), \tag{2.7}$$

where $\text{pa}(X_i)$ stands for the set of random variables associated with the parents of the node corresponding to the variable X_i. Hence, the structure of the ADG is used to decompose the joint probability distribution into factors. For example, the Bayesian network shown in figure 2.1 can be factorized as follows:

$$P(C, S, R, W) = P(W \mid S, R)P(S \mid C)P(R \mid C)P(C).$$

2.3.2 Learning Bayesian networks

As a Bayesian network consists of a qualitative and a quantitative part, there are also two aspects that need to be considered when discussing learning Bayesian networks from data. Learning the qualitative part, i.e. the graph, is called *structure learning*. Learning the quantitative part, i.e. learning the probability distributions, tables in the discrete case and densities in the continuous case, is called *parameter learning*.

In the context of analysis of biological data in particular structure learning is important, as it is the resulting graph that conveys information about how variable interact. Most of the discussion here will therefore be about structure learning.

There are two types of structure learning in Bayesian networks: score-based learning and constraint-based learning. Basically, *score-based learning methods*, also called search-and-score methods, explore the space of possible ADGs, which is superexponential in size, and score each element in this space with respect to the data. Many of the scoring methods used in score-based learning are based on the following principle. Let M stand for a Bayesian network *model* (thus, structure and probability tables), and D for a given dataset. The model that best fits the data can then be computed by using Bayes' theorem:

$$P(M \mid D) = \frac{P(D \mid M)P(M)}{P(D)}. \tag{2.8}$$

This Bayesian scoring regime assumes that there is prior information about the likelihood of the model, $P(M)$, whereas there is also knowledge about the likelihood of the data, $P(D)$. The likelihood of the model can be exploited to use background prior knowledge about particular models. If such prior knowledge is lacking, then usually the uniform probability distribution is used. Of

course, $P(D)$ may not be readily available either, even though it can be computed as follows:

$$P(D) = \sum_{M \in \mathcal{M}} P(D \mid M) P(M),$$

where $\mathcal{M}$ stands for the complete space of models. As mentioned above, this space is superexponential in size rendering this formula computationally infeasible. However, $P(D)$ only acts as a normalizing constant in Equation (2.8), which will be the same when comparing two models M and M'. Thus, knowing $P(D)$ is not crucial.

In practice it often suffices to compare $P(M \mid D)$ and $P(M' \mid D)$ for two different models $M, M' \in \mathcal{M}$, using the *Bayes factor*:

$$\mathrm{BF}(D) = \frac{P(D \mid M)}{P(D \mid M')},$$

as the ratio between $P(D \mid M)$ and $P(D \mid M')$ is the same as the ratio between $P(M \mid D)$ and $P(M' \mid D)$ if we assume that $P(M) = P(M')$, i.e. the uniform distribution just mentioned.

A disadvantage of the Bayes factor is that it does not take into account the complexity of the Bayesian network models. As Bayesian networks are I-maps, more complex models always fit the data better than simpler models; actually a complete graph, where each node is linked to all other nodes, will always give a perfect fit. This is because even though some arcs may be spurious, as will be reflected in the probability tables of the Bayesian network, representing them is still allowed, as the ADG may always include redundant arcs in comparison to the associated joint probability distribution. From the point of structure learning, this is clearly unacceptable, and normally a penalty term is added to the likelihood of the data given the model in order to take complexity into account. If we take the logarithm of Equation (2.8) and add the penalty term, we get a score of the following form:

$$\mathrm{score}(M, D) = \log P(D \mid M) + \log P(M) - \log P(D) - \mathrm{penalty}(M),$$

Again ignoring $P(D)$ and, possibly also $P(M)$, the penalty is then computed in terms of the complexity of the graph structure. Finding the right trade-off between goodness of fit, i.e. $\log P(M \mid D)$, and the penalty is now the major challenge. For data generated from a multinomial distribution, $P(D \mid M)$ is computed as follows:

$$P(D \mid M) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N_{ij})}{\Gamma(N_{ij} + M_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(a_{ijk} + s_{ijk})}{\Gamma(a_{ijk})}, \tag{2.9}$$

where Γ stands for the gamma function, n stands for the number of variables in the Bayesian network. A detailed explanation on this and its parameters can be found in Heckerman et al. (Heckerman 1995). The higher the score is, the better the data fits the model. A final issue, that is nor-

1. start with a complete undirected graph $G = (V, E)$;
2. let $i \leftarrow 0$;
3. **repeat**
4. **foreach** $X \in V$;
5. **foreach** $Y \in \text{ne}(X)$;
6. test whether exists a set $S \subseteq (\text{ne}(X) - \{Y\})$ with $|S| = i$ and $X \perp\!\!\!\perp Y \mid S$;
7. **if** this set exists, then
8. create triple $X \perp\!\!\!\perp Y \mid S$;
9. remove edge $X - Y$ from E;
10. let $i \leftarrow i + 1$;
11. **until** $\forall X : |\text{ne}(X)| < i$

Figure 2.3: *Algorithm used for the construction of the skeleton of an ADG.*

mally taken into account in designing a scoring scheme is that ADGs that are equivalent in terms of the d-separation relationship $\perp\!\!\!\perp_G$ yield the same score. For example, the networks $X \rightarrow Y$ and $X \leftarrow Y$ represent exactly the same d-separation relationship, namely that X and Y are d-connected. Score measures that take d-separation equivalence into account are called *score equivalent* (Heckerman 1995).

So far we have only discussed the design of a score measure for score-based learning methods. However, search is another characteristic of these type of methods. Given the enormous size of the search space, usually simple hill climbing (greedy search) is used, i.e. the space is explored by always selecting the best next state and never backtracking to an alternative state. Although hill climbing is a heuristic search method that is not guaranteed to find a global optimum, it is able to find local optima. These are often of sufficient quality, as experience learns.

Constraint-based learning methods are based on a different principle. Instead of exploring a search space, they act on the set of random variables and uses a statistical test of conditional independence in order to determine whether there should be an arc between the corresponding nodes in the network. As they have a more incremental nature, score-based methods seem in particular appropriate as analytic tools for data analysis. A well-known implementation of a constraint-based learning algorithm is the PC (Peter and Clark) algorithm (Sprites et al. 2000). This algorithm first constructs an undirected skeleton graph based on independence tests with an associated level of significance, which is used to accept or reject edges, as shown in figure 2.3. In the algorithm $\perp\!\!\!\perp$ represents the conditional independence test, which in practice is implemented using the G^2 test statistic. Subsequently the direction of the arcs is determined using orientation rules, where only the converging connection gives rise to unique directions of arcs. The major advantage of the PC algorithm is that it is possible to incorporate background knowledge into the algorithm, e.g. by using biological knowledge about cause-effect relationships in the orientation of the arcs.

When constructing Bayesian networks from data or from expert knowledge, or both, there are various scenarios which should be considered:

- The structure of the Bayesian network is known and the data is complete. Here, parameter learning, subjective assessment of probabilities, or both can be used.

- The structure of the Bayesian network is known and the data is incomplete. Now, special parameter learning techniques must be used, as discussed below. Clearly, expert assessments can also be used in this case.

- The structure of the Bayesian network is unknown and we have complete data. In this case, the structure of the network can be learned from the data, possibly exploiting background knowledge.

- The structure of the Bayesian network is unknown, and the data are incomplete. This case again involves dealing with the incompleteness of the data both in structure and parameter learning.

In the biological domain one is interested in finding significant relationships between variables of interest (where often not all data are completely observed); this scenario is represented as ”data being incomplete and the structure of the Bayesian network being unknown”. Clearly, this is also the most difficult case from a computational point of view. Incomplete data may involve *missing values*, and *imputation* techniques are employed to find appropriate values; these values are assumed to be missing at random. Sometimes particular crucial variables are missing, called *hidden* or *latent* variables. Using optimization algorithms such as the EM algorithm, which stands for the Expectation Maximization algorithm, it is sometimes possible to learn all parameters, including the hidden ones (Dellaert 2002). More details on this can be found by the work done by Friedman (Elidan and Friedman 2003).

There have already been quiet successful *search-and-score* methods implemented for gene expression data sets, but as the number of variables grow in size the computation time grows exponentially. However, there have been several successful *approximate methods* used in the biological domain to learn relations (structures) (Husmeier and Werhli 2007). As mentioned before, constraint-based methods have the advantage that they allow incorporation of prior biological knowledge about dependence of variables.

Finally, there are various packages in R which can be used for learning the structure of a Bayesian network; some of them are discussed in *Appendix B* of this thesis.

The graphical nature of the network combined with probability theory allows one to do data analysis in an intuitive way. It is important to understand the interaction between different variables but it is more important to understand the nature of these relationships.

2.3.3 Other learning scenarios

Not always we will have knowledge about relationships between genes or metabolites of interest in which case we would want to find these relationships from available experimental data. This approach is not uncommon when the number of variables is large and there is little or no knowledge available of the underlying process. Moreover, it can be a laborious task to construct networks of several hundred nodes just by hand. Therefore there has been considerable research in this area to do unsupervised learning of conditional dependence and independence relationships from data. The key assumptions used for this approach are that biological processes are hierarchical and sparse in nature. In the past correlation and clustering methods have been used successfully to identify groups of metabolites clustering together and reconstruct pathways (Yilmaz 2001, Ursem et al. 2008). Bayesian network allow us to model causal relationships by looking at the direction of the arrows in the directed graph. Causal relationships play a vital role when we want to find out when one variable causes a change in another variable. Often these relations are investigated by experimental research to determine if changes in one variable truly causes change in another variable. A detailed review on how to find causal relationship is described in (Cooper 1999). For Bayesian networks there is the restriction that arrows are not allowed to form directed cycles (paths that end at the node where they started)–these graphs are called *acyclic*. Of course, there can be feedback loops involved in a problem domain which can be perfectly modeled using other type of Bayesian networks, so-called *dynamic* Bayesian networks (Murphy 2002), which requires time-series data.

2.3.4 Bayesian networks as classifiers

Classification is a technique used in *-omics* to put entities into categories, these can be healthy or disease states of an organism, high or low concentration of metabolites, or up and down regulation of a genes in an experiment. There are various classification techniques in use like kNN, decision trees and ZeroR which operate on discretized datasets. The most popular among them is the naïve Bayesian classifier. A naïve Bayesian classifier is just a special Bayesian network, where a distinction is made between variables that are used for classification, called *class variables*, and variables that can be observed, called *feature variables* or *evidence*. If we for simplicity's sake assume that there is only one class variable, the Bayesian network $\mathcal{B} = (G, P)$ is used to compute:

$$c^* = \arg\max_{c \in D(C)} P(C = c \mid E), \tag{2.10}$$

where E stands for the set of observed values for some or all of the feature variables, also called *evidence*. A naïve Bayesian network has a structure where there is an arc from the class variable C to any feature variable F_i. Using d-separation, it follows that we have that $F_i \perp\!\!\!\perp_G F_j \mid C$, for any pair of feature nodes $F_i, F_j, i \neq j$. This implies that the corresponding feature variables

are conditionally independent given the class variable. As a consequence, the factorization of the naïve Bayesian network is particularly simple:

$$P(C, F_1, \ldots, F_n) = P(C) \prod_{i=1}^{n} P(F_i \mid C).$$

Thus only a linear number of probabilities $P(F_i \mid C)$ is required. Computation of the posterior probability $P(C \mid E)$ is then also easy as:

$$P(C \mid E) = \frac{P(C) \prod_{F_i \in E} P(F_i \mid C)}{P(E)}.$$

The joint probability distribution $P(E)$ referred as a normalizing constant which cancels out in the calculation, and is redundant in computing the value of C where $P(C \mid E)$ which attains the highest probability as in Equation (2.10). Thus it is sufficient to compute $P(C) \prod_{F_i \in E} P(F_i \mid C)$ for any value of C and a fixed set of evidence E. They exhibit high accuracy when applied to large datasets, but can also be used on datasets that are small, which is particularly attractive in microarray analysis. The highest accuracy is often explained in terms of *bias-variance decomposition*. Although the naïve Bayesian network has high bias, as its structure is seldom in accordance to the data, its variance is low, because it has little tendency to perfectly fit, and therefore overfit, to the data. Overfitting is one of the clear disadvantages of Bayesian networks in general, a problem well described for neural networks. This, in fact, explains the use of the penalty factor (see above) in learning Bayesian networks from data.

Despite its simplicity the naïve Bayesian classifier often outperforms other classification techniques. We chose this technique to depict a black box approach for benchmarking the other algorithms used in this thesis. Studies comparing classification techniques have found that naïve Bayesian classifiers are robust and yield result comparable to other state-of-the-art classifiers, such as classification trees, support-vector machines and neural network.

Submitted for publication as: *"MADMAX - Management and Analysis Database for Multi-platform microArray eXperiments"* – Anand K. Gavai, Philip J. de Groot, Ke Lin, Mark V. Boekschoten, Yannan Liu, Harm Nijveen, Pieter B.T. Neerincx, Guido Hooiveld, Michael Müller, and Jack A.M. Leunissen

Chapter 3

MADMAX - Management and Analysis Database for Multi-platform microArray eXperiments

"The belief that complex systems require armies of designers and programmers is wrong. A system that is not understood in its entirety, or at least to a significant degree of detail by a single individual, should probably not be built."

Niklaus Wirth

Abstract

The number of microarray experiments performed in the biomolecular life sciences is rapidly increasing. However, the proper handling, storage, analysis and exchange of this type of data is for many biologists still a complex task. Moreover, different microarray platforms usually require vendor specific software solutions, making the storage and analyses even more challenging. We developed MADMAX as an integrated repository for the management and analyses of multiplatform microarray experiments, primarily for use in the Netherlands Nutrigenomics Consortium. MADMAX organizes microarray data in experiments and allows the storage of the raw data, regardless the array platform used. Moreover, metadata exactly describing all expects of an experiment is stored using a controlled vocabulary (MIAME/Nut), which can be queried across experiments. Data is submitted on-line via secured web interfaces, and can be accessed only by authorized users. MADMAX is linked through web services to a computational pipeline that implements state-of-art methodologies, e.g. offered by the Bioconductor project, for biologists simplifying the complexities of quality control, advanced statistical analyses and therefore functional interpretation. Data can be exported in a standardized format (MAGE-TAB) prior to submission to public repositories such as ArrayExpress and the Gene Expression Omnibus. MADMAX is a flexible management and analysis database for multiplatform microarray experiments. MADMAX is primarily designed for use by biologists lacking detailed knowledge on biostatistics and bioinformatics, and provides a complete life cycle for microarray experiments. MADMAX currently hosts over 1000 Affymetrix hybridizations covering a variety of species, including Homo sapiens, Mus musculus, Rattus norvegicus, Arabidopsis thaliana, Lycopersicon esculentum and Medicago trunculata. A free account can be obtained upon request.

3.1 Introduction

Microarrays provide a practical method for measuring the expression level of thousands of genes simultaneously (Schena et al. 1995, Lockhart et al. 1996). Consequently, microarray experiments generate huge volumes of data and therefore pose major challenges, in particular to biologists, to manage, maintain and interpret. To enable rapid and simple access of microarray experiments, the data produced need to be stored in a reliable, flexible system that ideally handles multiple array platforms and integrates various analysis tools. Moreover, for a proper reanalysis of already stored data, descriptions of the sample from which the RNA has been derived and all details of its treatment, the so-called metadata, should be stored as well. In addition, a storage system should also enable easy sharing and exchange of both raw and processed data between researchers involved in the same project. Finally, biologists should obtain meaningful data from their array experiments in hours rather than weeks.

To serve these requirements within the Netherlands Nutrigenomics Consortium (Mueller and Kersten 2003), we developed MADMAX, a management and analysis database for multi-platform microarray experiments.

3.2 Modules of MADMAX

3.2.1 System architecture and implementation

MADMAX is built on Oracle technology (Oracle 2008), one of the most robust, scalable and widely used database platforms. The primary reason to use Oracle is that it is built to manage and store data of high magnitude and dimensionality, making it ideal for microarray experiments. The interface of MADMAX is built in Oracle Application Express (Oracle APEX), a hosted declarative development environment (built on Java) for developing and deploying database-centric web applications. It turns a single Oracle database into a shared service by enabling multiple workgroups to built and access applications as if they were running in separate databases. Likewise, other database management systems, including the public repositories ArrayExpress (Parkinson et al. 2007) and Gene Expression Omnibus (GEO) (Barrett et al. 2007), and e.g. MiMiR (Navarange et al. 2005), MARS (Maurer et al. 2005) and SMD (Demeter et al. 2007), use Oracle as back-end to store and manage data of this size. Users can login using username, password and a group name, which are provided by a system administrator. All users and groups are managed using Oracle's user management functionality. Access to data and information is password protected and secured, and for each experiment the 'owner' can set the 'visibility' of the experiment, granting (or not) other users from the same user group access to the data. MADMAX uses the raw experimental files from, Affymetrix, Agilent or Illumina platforms for performing analysis. All calculations are performed on a high-end computation server (32 GB main memory, 4 dual core Xeon processors)

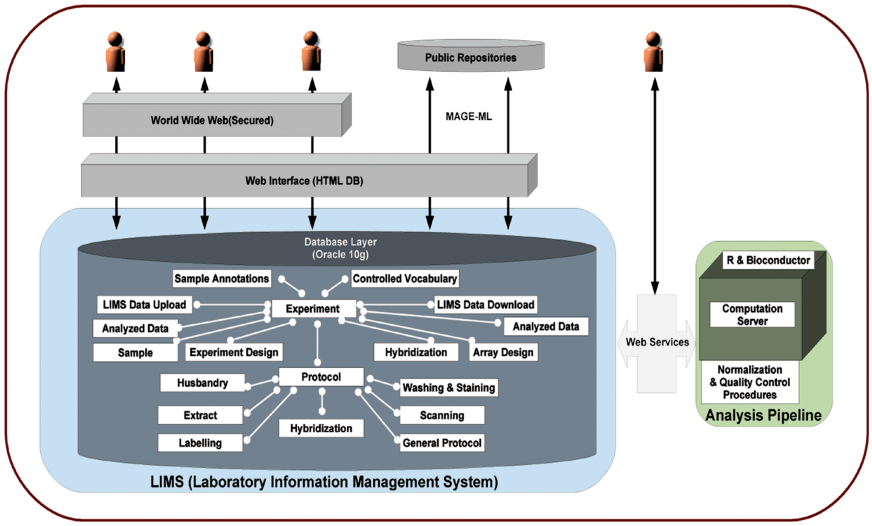


Figure 3.1: *Schematic representation of the system architecture of MADMAX: Conceptual overview of MADMAX system. Connections represent logical relationships between database components representing LIMS. The green box represents the analysis pipeline implemented as a web service on a high-end computational server.*

through a BioMoby-based web service (Neerincx and Leunissen 2005) interface. MADMAX implements various types of analysis, which are performed in R (Ihaka and Gentleman 1996) and Bioconductor (Gentleman et al. 2004). At initialization of an experiment, the raw files are transferred to this server, the requested pipeline is executed and the results are imported back to the user space after which an email message is sent to the user. To ensure smooth processing a scheduling mechanism is implemented using Sun Grid Engine (Gentzsch 2001).

There are basically two parts to MADMAX, i.e. *data storage & management* and *analysis features* figure 3.1. The data storage & management part provides fields to be filled-in by a user to describe the experiment, the so-called metadata, and allows the uploading of the raw array files. The analysis features provide the user with an interface where data can be analyzed using web services on a central computation server. One advantage of having this framework is that the storage & management part is completely isolated from the analysis capabilities, allowing an easy expansion of the system with updated analysis pipelines and integration of not yet supported array platforms. A list of key features of MADMAX are shown in table3.1.

No.	Features
1.	Storage and analysis of multiple platform microarray data
2.	Query across multiple experiments
3.	Accessable through web-based forms
4.	MIAME compliant; Controlled vocabulary to store LIMS information
5.	Data export using MAGE-TAB format for deposition to public repositories
6.	Biomoby based webservice access to analysis pipelines
7.	Log files give information of the steps performed in the analysis

Table 3.1: *Key features of MADMAX*

Data storage & management

The highest level of organization in MADMAX is the group to which a user belongs, subsequently every users is assigned to one or more more experiments being performed within the group, therefore experiments performed outside the group are not accessible to the users within the group unless the user is a member of other groups. Experiments are associated with protocols and raw array data. These data are entered in MADMAX through web-based interfaces that are encrypted using secure socket layer (SSL). The fields describing an experiment and array hybridizations, are compliant with the minimum information about a microarray experiment (MIAME), proposed and developed by the Microarray Gene Expression Data (MGED) society (Brazma et al. 2001, Ball et al. 2002). This standard outlines the core information that is common to most microarray experiments and establishes a foundation for the standardized annotation of microarray data. However, since MADMAX is intensively used by many researchers involved in molecular nutrition research, additional fields are provided to capture the minimal reporting requirements for nutrigenomics studies, called MIAME/Nut, which is an extension to the MIAME core standard (Rocca-Serra et al. 2004, Sansone et al. 2006). A controlled vocabulary is employed at most stages of the metadata of an experiment. Users can select terms from the web-based forms and progress through a series of simple steps to describe their experiment, analytical settings and upload the data files. A schematic representation of these steps is shown in figure 3.2 . To all users a set of protocols is available that describe the procedures that are usually always the same in an array facility, such as the hybridization, washing & staining, and scanning of arrays. However, new protocols can easily be created if none of the listed protocols suits the experimental procedures. Submissions are automatically checked for MIAME/Nut compliance, accuracy and completeness of biological information provided, as well as for data consistency. Array data, either individual array files or aggregated in a single archive like a ZIP or CAB archive (the latter is routinely generated by the Gene Chip Operating System (GCOS) from Affymetrix), can be uploaded directly into MADMAX. MADMAX's query interface provides the ability to query the metadata data by various attributes, including species, submitter and treatment. Finally, to facilitate data sharing, e.g. as a submission to public repositories like ArrayExpress and GEO, MADMAX exports experiments in

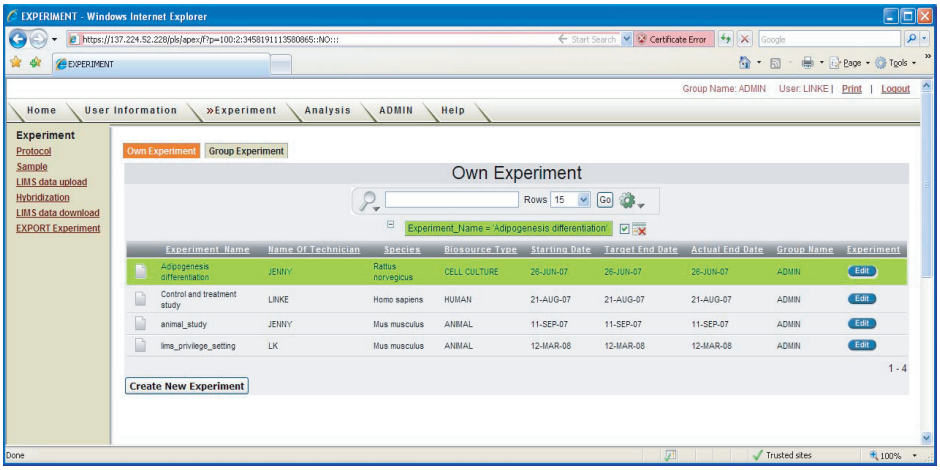


Figure 3.2: *Screen shot of the experiments in MADMAX:Experiments displayed across all hybridizations showing information about submitters and type of experiments, more detailed information can be accessed only on the users level of access privileg*

MAGE-TAB format (Rayner et al. 2006) that can be generated on the click of a button.

Analysis features

There is a plethora of libraries and packages available, e.g. in the R/Bioconductor project, to asses array quality and statistical significance. With the biologist in mind, MADMAX offers users standard quality control checks and analysis pipelines that are built specifically for Affymetrix, Agilent and Illumina arrays by chaining together various packages from Bioconductor. These pipelines are set up in such a way that extending them is relatively easy. A schematic overview of the pipelines is given in figure 3.3. After each run a log file is created which stores the parameters that were passed to the R/Bioconductor scripts from within the interface, allowing an easy re-run of any analysis. In general, one should evaluate the quality of the hybridizations of the arrays before performing subsequent analyses. Therefore, to ascertain only excellent quality arrays are used in the statistical and functional analysis, the statistical analysis pipeline is available only after a quality control check has been performed. The quality control procedures are implemented in a pipeline that utilizes various advanced quality metrics, diagnostic plots, pseudo-images and classification methods to indicate which arrays are of good or lesser quality. Figure 3.4 shows a collection of quality control metrics that are generated from within MADMAX. In our laboratory, the interpretation of the images is normally performed by experienced technicians that are responsible for uploading the array data to MADMAX, although this identification can also be done automatically (Heber and Sick 2006). Arrays of lesser quality are flagged as 'bad' in the metadata section

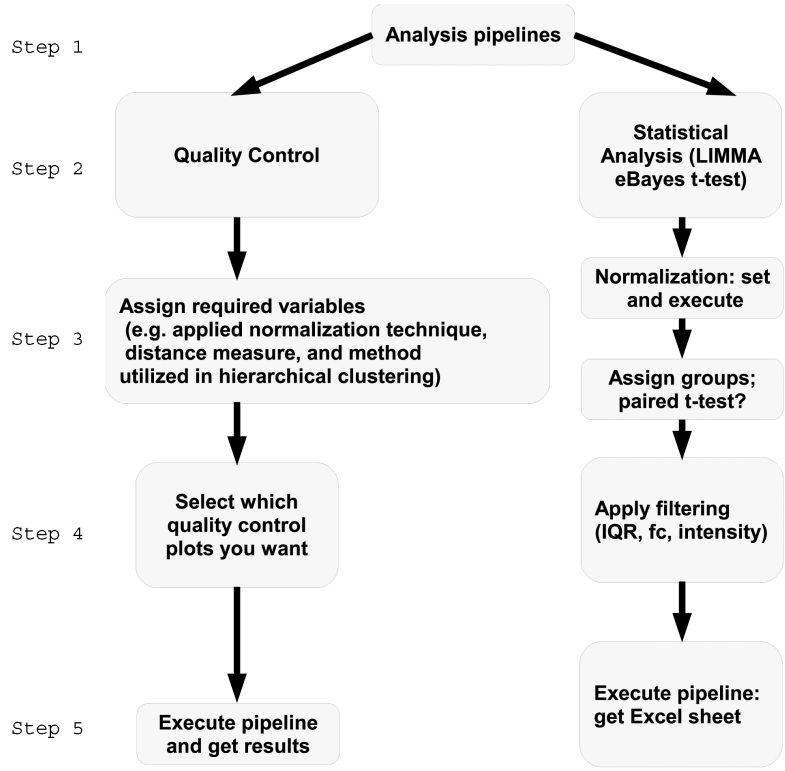


Figure 3.3: *Steps involved in data analysis pipeline: Steps involved in a typical analysis of an experiment in MADMAX.*

of the experiment, thereby eliminating these arrays from the subsequent statistical analyses, but without losing it.

To perform any analysis, the user first needs to select the experiment to be analyzed. Next, after selecting an appropriate method, the arrays are normalized. In addition to the chip definition files provided by the manufacturer of an array, MADMAX also enables the use of redefined definition files that utilize the current genome information (Dai et al. 2005). To increase statistical power, MADMAX offers the user the option to filter the expression data on intensity, interquartile range or fold-change before statistical testing (von Heydebreck et al. 2004). Statistical testing is then routinely performed using linear models that apply moderated t-statistics implementing empirical Bayes regularization of standard errors (Smyth 2004). To adjust for both the degree of independence of variances relative to the degree of identity and the relationship between variance and signal intensity, the moderated t-statistic is extended by a Bayesian hierarchical model to define a intensity-based moderated T-statistic (Sartor et al. 2006). P-values are corrected for multiple

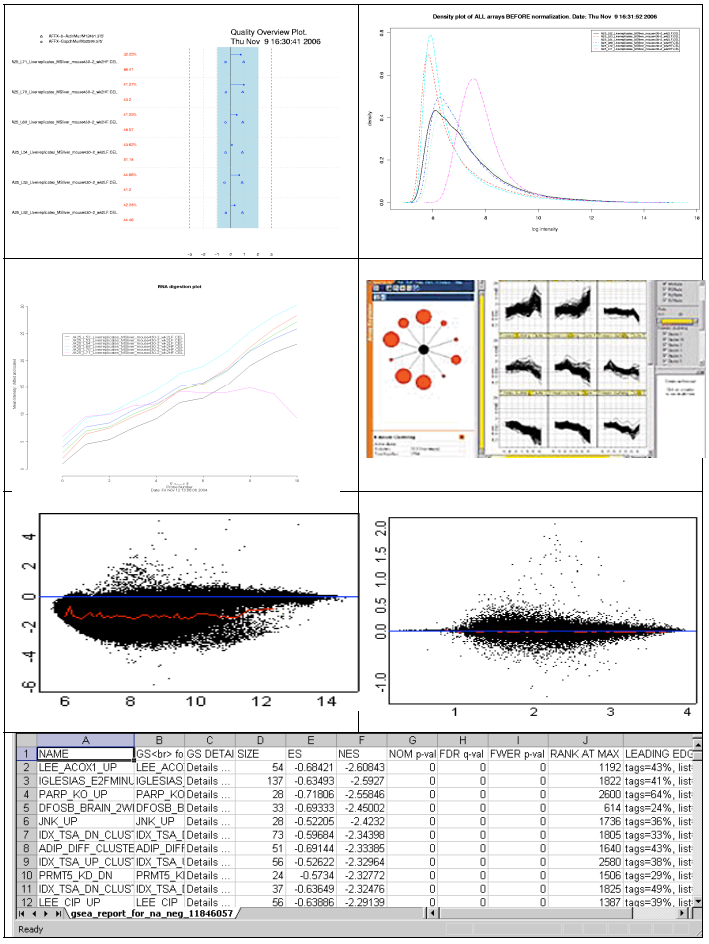
testing using false discovery rate methods (Storey and Tibshirani 2003, Dudoit and Ge 2005). To relate changes in gene expression to functional changes, two complementary methods are applied. One method is based on over-representation of Gene Ontology (GO) terms (Lee et al. 2005). Another approach, gene set enrichment analysis (GSEA), takes into account the broader context in which gene products function, namely in physically interacting networks, such as biochemical, metabolic or signal transduction routes (Subramanian et al. 2005). Both methods are applied on unfiltered data sets and have the advantage that they are unbiased, because no gene selection step is used. Moreover, since a score is computed based on all genes in a particular GO term or gene set, the signal-to-noise ratio is boosted allowing the detection of transcriptional programs that are distributed across an entire set of interacting genes yet are subtle at the level of individual genes. Results of each run are stored in the user space and can be downloaded for interpretation and further analyses. This whole procedure enables a user to obtain biologically relevant answers in hours rather than weeks. Features of both pipelines with their respective libraries from Bioconductor are summarized in tables 3.2 & 3.3.

Procedures	**Library from Bioconductor**
Quality Check	simpleaffy (Wilson and Miller 2005)
Raw Images (Before Normalization)	normalization (Gautier et al. 2004)
Images (After Normalization)	affyPLM (Bolstad et al. 2005)
Density Distribution Curves (Before)	affy (Gautier et al. 2004)
Boxplots (Before)	affy (Gautier et al. 2004)
Density Distribution Curves (After)	affy (Gautier et al. 2004)
Boxplots (After)	affy (Gautier et al. 2004)
RNA Degradation Plot	affy (Gautier et al. 2004)
Correlation Plot (After Normalization)	affyQCReport (Parman and Halling. 2006)
RLE plot	affyPLM (Bolstad et al. 2005)
NUSE plots	affyPLM (Bolstad et al. 2005)
Dendogram (Hierarchical Clustering)	stats(hclust)

Table 3.2: *Summary of libraries from Bioconductor for low-level analysis*

Procedures	**Library from Bioconductor**
Differential gene expression	LIMMA (Smyth 2004)
False discovery rate	multest (Dudoit and Ge 2005)
False discovery rate	qvalue (Storey and Tibshirani 2003)
Testing differentially expressed genes	IBMT (Sartor et al. 2006)
Gene Ontology	ErmineJ (Lee et al. 2005)
Gene set enrichment analysis	GSEA (Subramanian et al. 2005)

Table 3.3: *Summary of libraries from Bioconductor for high-level analysis.*



	A	B	C	D	E	F	G	H	I	J	
1	NAME	GS fo	GS DETAI	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDG
2	LEE_ACOX1_UP	LEE_ACO	Details ...	54	-0.68421	-2.60843	0	0	0	1192	tags=43%, list
3	IGLESIAS_E2FMINU	IGLESIAS_	Details ...	137	-0.63493	-2.5927	0	0	0	1822	tags=41%, list
4	PARP_KO_UP	PARP_KO	Details ...	28	-0.71806	-2.55846	0	0	0	2600	tags=64%, list
5	DFOSB_BRAIN_2WI	DFOSB_B	Details ...	33	-0.69333	-2.45002	0	0	0	614	tags=24%, list
6	JNK_UP	JNK_UP	Details ...	28	-0.52205	-2.4232	0	0	0	1736	tags=36%, list
7	IDX_TSA_DN_CLUS	IDX_TSA_	Details ...	73	-0.59684	-2.34398	0	0	0	1805	tags=33%, list
8	ADIP_DIFF_CLUSTE	ADIP_DIFF	Details ...	51	-0.69144	-2.33385	0	0	0	1640	tags=43%, list
9	IDX_TSA_UP_CLUS	IDX_TSA_	Details ...	56	-0.52622	-2.32964	0	0	0	2580	tags=38%, list
10	PRMT5_KD_DN	PRMT5_KI	Details ...	24	-0.5734	-2.32772	0	0	0	1506	tags=29%, list
11	IDX_TSA_DN_CLUS	IDX_TSA_	Details ...	37	-0.63649	-2.32476	0	0	0	1625	tags=49%, list
12	LEE_CIP_UP	LEE_CIP	Details ...	56	-0.63886	-2.29139	0	0	0	1387	tags=39%, list

gsea_report_for_na_neg_11846057

Ready

Figure 3.4: *Quality control plots: Different quality control images generated from the pipeline like NuSE plot, MVA plot, RNA degradation plot etc. Information on these plots can be found in the help section of MADMAX.*

3.2.2 Comparison with similar systems

A variety of similar systems have been developed to serve the purpose of storage and/or analysis of array experiments (Dysvik and Jonassen 2001, Kapushesky et al. 2004, Heyer et al. 2005, Navarange et al. 2005, Maurer et al. 2005, Marzolf et al. 2006, Herrero et al. 2003, Theilhaber et al. 2004, Burgarella et al. 2005, Hancock et al. 2005, Saal et al. 2002). Some of these systems make use of third-party tools to do low level analysis microarray data (Saal et al. 2002, Parkinson et al. 2007, Barrett et al. 2007), some are good in data storage and management (Saal et al. 2002), while others are good in performing high level analysis to find gene regulatory networks (Dysvik and Jonassen 2001, Kapushesky et al. 2004, Maurer et al. 2005). Some of these systems even provide storage and analysis capabilities for specific microarray platform where analysis is built on Microsoft technology and is not suitable for other array types for e.g.(affymetrix) (Saal et al. 2002). Recent evolution of these systems has managed to get over some of these obstacle. However, analysis is not as user friendly and the user need to rewrite their own scripts in R to hook it up to these systems in order to run their analysis pipelines (Biosoftware 2000, Saal et al. 2002). Most of these systems lack controlled vocabulary for recording LIMS information, which is very important aspect for annotating data while performing microarray experiments. However there is a need to have a system which combines storage and analysis (both low and high level) capabilities embedded together and at the same time be able to support multiple array types. The features implemented in MADMAX takes these issues into account and expands them to improve upon the existing features.

3.3 Conclusions

In this paper we describe the development and features of MADMAX, a repository for the management and analyses of array data, initially developed for use in the Netherlands Nutrigenomics Consortium. In contrast to other available databases and analysis systems, MADMAX provides a one stop solution to handle a complete life cycle (i.e. storage, annotation, quality control, analysis and submission) of arrays experiments. As microarray experiments will benefit greatly from proteomics and metabolomics data, support for these data is scheduled to be implemented in the future release of MADMAX.

Submitted for publication as: *"Comparing model based and black box approaches for estimating metabolic state of organisms from nutrigenomics experiments"* – Anand K. Gavai, Guido J.E.J Hooiveld, Sander Kersten, Michael Müller, Peter J.F. Lucas and Jack A.M. Leunissen

Chapter 4

Comparing model based and black box approaches for estimating metabolic state of organisms from nutrigenomics experiments

"Learning is not compulsory...neither is survival"

W. Edwards Deming

Abstract

Correlation and clustering techniques are popular methodologies for the analysis of microarray data. These approaches allow obtaining an overview of the interactions between objects of interest. Nonetheless these techniques have limitations, as they do not take non-linear relationship into account and lack the possibility to incorporate prior knowledge from biologist. Furthermore, these approaches cannot efficiently be used to construct gene regulatory networks or to classify conditions. Bayesian networks, on the other hand, allow one to model relationships in a graphically intuitive format; one can also use this technique to classify cases in a dataset based on different experimental conditions. Here we show how constraint-based learning of a Bayesian network, complemented by learning naïve Bayesian classifiers, can be used to identify important features (genes) which play a significant role in discriminating feeding and fasting states of mice. We also show that this way of analyzing data is intuitive and powerful to get in-depth understanding of the genes involved in the key processes. Constraint-based learning of Bayesian networks, complemented by learning a naïve Bayesian classifier, offers a useful combined technique when faced with the problem of analyzing data with small sample sizes, as is common in many studies. When applied to a nutrigenomics dataset, it is able to correctly find subtle changes in gene expression and to identify the metabolic state of the animal.

4.1 Background

Driven by the continuing and accelerating progress in 'omics'-technologies, such as microarrays and mass spectrometry, unique possibilities have emerged to investigate the genome-wide effects of nutrients at the molecular level. This research field is called nutritional genomics, or nutrigenomics, and it is widely recognized that nutrigenomics has the potential to increase our understanding of how nutrition influences metabolic pathways and homeostasis, how this

regulation is disturbed in a diet-related disease, and to what extent individual genotypes contribute to such diseases (Fogg-Johnson and Merolli 2000, Mueller and Kersten 2003, Ordovas and Mooser 2004, Kaput and Rodriguez 2004, Afman and Mueller 2006).

The identification of differentially expressed genes or metabolites is the usual, first step in these type of analyses. Usually various statistical filters are applied to pre-select the genes or metabolites of interest for further investigation, e.g. on their activities, reconstruction of affected biological networks, or potential to classify biological status (Stuart et al. 2003, Segal et al. 2005, Quackenbush 2006). In the past several approaches addressed the core issue of reconstructing gene regulatory network utilizing techniques like Boolean networks (Kervizic and Corcos 2008), neural networks (Toure and Basu 2001), support vector machines (Brown et al. 2000) and Bayesian networks (Friedman et al. 2000). Of these, approaches based on Bayesian networks have gained much popularity. Bayesian networks are network-based representations of joint probability distributions in which conditional independencies are represented as a graph. This is because the relationship between genes of interest are often represented in the form of a network revealing probable metabolic pathways. In the past, these techniques have been successfully used to construct gene regulatory network to help us understand the nature of relationships between genes of interest. Bayesian networks can be used efficiently in many scenarios when it comes to solving a biological problem involving expression data (Zou and Conzen 2005). A Bayesian network typically consists of nodes connected with edges, where each node is associated to a random variable and edges represent dependencies between them. Techniques such as these can be combined together with biological insight and data in various ways to obtain more biological insight into the system.

There is increasing evidence that complex diseases do not arise from an individual molecule or gene, but from complex interactions between the cell's different compartments and its environment over a period of time. However, significance analysis at the gene level may suffer from the typically low number of samples obtained from experiments in higher organisms, which is often a case in genomics experiments, and because of this the false discovery rate is high. One way to solve this problem is to use knowledge bases such as Gene Ontology (Ashburner et al. 2000) or metabolic maps [Kyoto Encyclopedia of Genes and Genomes (KEGG) maps] (Kanehisa et al. 2008) to find the annotation of genes in pathways of interest. The genes involved in various metabolic pathways are represented at several levels of abstraction, e.g. relationships between genes represented in the form of a network or a graph showing a chemical reaction. The choice of abstraction is based on the biological problem being addressed and the type of data available. Recently, there has been much work on selecting biological pathways from pathway databases that are closely related to a (*Class*) variable of interest or a phenotype using the gene expression data (Goeman et al. 2004, Liang et al. 2006, Tai and Pan 2007, Tomfohr et al. 2005). Recently Quanz et al. (Quanz et al. 2008) have shown that incorporating selected significant pathways into the classification process as features can reveal much insight into the biological process. With the few samples and

many features that typically arise in these experiments incorporating any additional information, such as known biological pathways, into the data analysis process becomes a key priority. To demonstrate the utility of the pathways-as features approach, we evaluate our approach to discriminate between normal tissue and metabolically affected tissue, exemplified resp. by the *fed* state and *fasted* state of murine liver. We would like to show how the outcome only depends on the gene expression values with respect to how they affect the state of the pathways. Although learning Bayesian network structures requires much data, less data is needed by considering local independence structure, in particular the Markov blanket of a variable. As the structure of a naïve Bayesian classifier, a special Bayesian network where a distinction is made between a class variables and feature variables, is fixed, it's structure does not need to be learned. Naïve Bayesian classifiers are also not very demanding with respect to the amount of data. In this paper, we explore each of the techniques, and, in addition, adopt constraint-based structure learning as a method of feature selection for the naïve Bayesian classifier. The results obtained are validated with online databases like the pathway express (Draghici et al. 2007) in order to deduce there biological meaning. Techniques such as these can be effectively used to generate hypothesis on genes involved in determination of metabolic state.

4.1.1 Feeding and fasting conditions

Fasting, the act of willingly abstaining from food, is a frequently occurring natural status in humans. Fasting is a popular strategy to manage overweight or obesity, it is a traditional habit in certain religions or societies, and it is an accepted pre-surgical procedure. The overall metabolic response to fasting operates at numerous levels and has been relatively well characterized (Owen et al. 1979, van den Berghe 1991). During fasting whole-body fuel utilization gradually shifts from carbohydrates and fat in the fed state to proteins and fat after a day of fasting. This adaptation is particularly striking in the brain, an obligate glucose utilizer in the fed state, which is able to acquire energy predominantly from ketone bodies after prolonged fasting. Most of the actual interconversions in energy substrates occur in the liver, which plays a central role in the adaptive response to fasting.

Fasting is characterized by low insulin concentrations and high glucagon, glucocorticoids, and (nor)epinephrine concentrations in plasma. This hormonal profile promotes the hydrolysis of triacylglycerols in adipose tissue, thereby increasing the concentration of free fatty acids (FFAs) in plasma. The fatty acids are taken up by the liver, where they are either re-esterified to triacylglycerol and secreted as VLDL or oxidized in the mitochondria via beta-oxidation. The majority of fatty acids are only partially oxidized to acetyl-coenzyme A (acetyl-CoA), which then condenses with itself to form ketone bodies, an important fuel for the brain. The energy released in the process of beta-oxidation is used by the liver to carry out gluconeogenesis from substrates such as glycerol, lactate, and amino acids.

The main metabolic processes involved in the post-prandial (=fed) state are *glycogenesis*, *gly-*

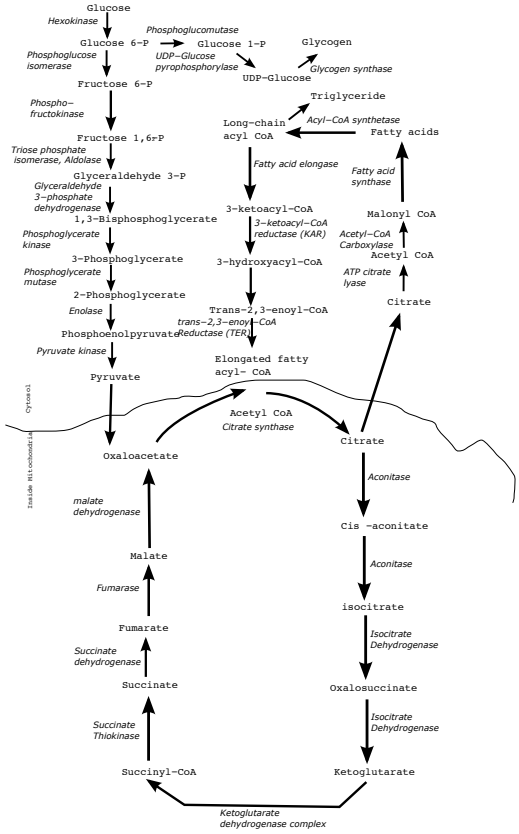


Figure 4.1: *Metabolic pathways involved in feeding process Glycolysis, Fatty acid synthesis and TCA cycle.*

colysis, *Krebs/TCA cycle*, and *fatty acid synthesis*, whereas during the fasting state *gluconeogenesis* and *fatty acid degradation* are active, in addition to the *Krebs/TCA cycle*. Under both metabolic conditions, the indicated pathways function similarly, although sometimes, one is a reverse flux of the other. However there are few differences in enzymes involved and their products in both of them. The figures 4.1 and 4.2 give a summary of these two pathways curated from experts in this domain.

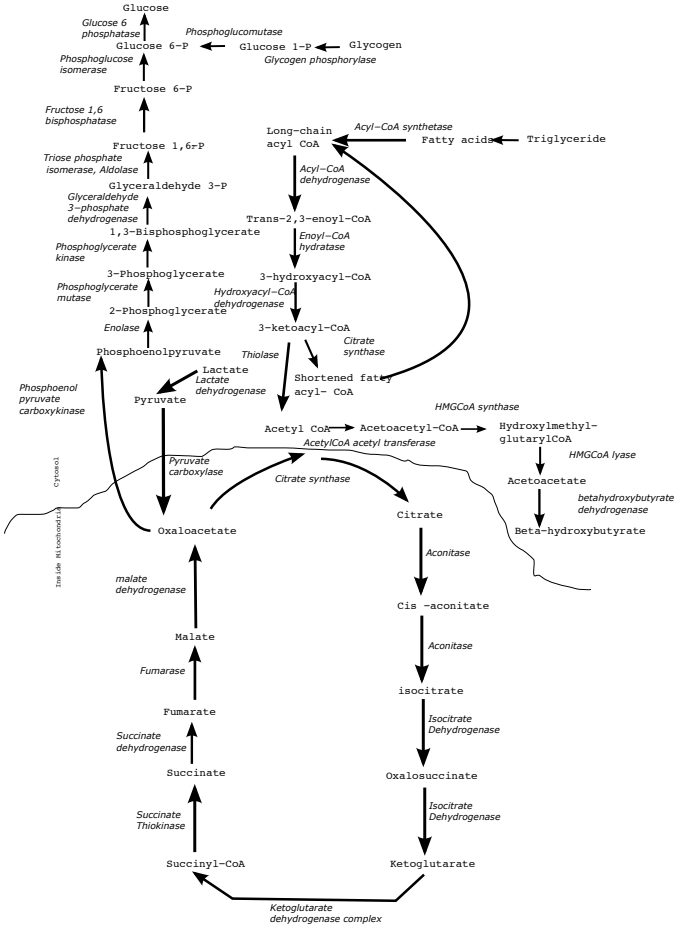


Figure 4.2: *Metabolic pathways involved in Gluconeogenesis, Fatty acid degradation and TCA cycle.*

4.2 Materials and Methods

4.2.1 Description of the data set

Data from two fasting experiments were used in this study. The experiments were originally designed to phenotype two different transgenic mice strains under metabolic stress conditions (S. Kersten et al, unpublished data). However, since in both experiments male C57BL/6J mice

of similar age (12 – 20 weeks) served as control, and combined with the identical experimental design, this allowed us to combine the array data from the two experiments. In total 19 arrays were included; n=9 for control (fed) and n=10 for fasted conditions. Briefly, mice had free access to water and standard laboratory chow (RMH-B, Hope Farms, Woerden, the Netherlands). At the start of the experiment, mice were killed, or fasted for 24 hours before killing. Fasting started at the onset of the light cycle. Mice were anaesthetized with a mixture of isofluorane (1.5%), nitrous oxide (70%), and oxygen (30%). Livers were excised, minced, frozen in liquid nitrogen, and stored at -80 degrees until RNA isolation. All experiments were approved by the Local Committee for Care and Use of Laboratory Animals.

4.2.2 RNA isolation and quality control

Total RNA was extracted from liver with TRIzol reagent (Invitrogen, Carlsbad, CA), treated with DNAse and purified on columns using the SV Total RNA Isolation System (Promega, Leiden, The Netherlands). RNA integrity was checked on an Agilent 2100 bioanalyzer (Agilent Technologies, Amsterdam, the Netherlands) with 6000 Nano Chips according to the manufacturer's instructions. RNA was judged as suitable for array hybridization only if samples showed intact bands corresponding to the 18S and 28S ribosomal RNA subunits, displayed no chromosomal peaks or RNA degradation products, and had a RIN (RNA integrity number) above 8.0.

Affymetrix GeneChip oligoarray hybridization, scanning and quality control

Total RNA (5 ug) was labeled using the Affymetrix One-cycle Target Labeling Assay kit (Affymetrix, Santa Clara, CA). The correspondingly labeled RNA samples were hybridized on Mouse Genome 430 2.0 Arrays (Affymetrix), washed, stained and scanned on an Affymetrix GeneChip 3000 7G scanner. Detailed protocols for the handling of the arrays can be found in the Genechip Expression Analysis Technical Manual, section 2, chapter 2 (Affymetrix; P/N 701028, revision 5), and are also available upon request. Packages from the Bioconductor project, integrated in an in-house developed on-line management and analysis database for multiplatform microarray experiments, were used for analysing the scanned Affymetrix arrays (Gentleman et al. 2004, Gavai et al. 2009). Various advanced quality metrics, diagnostic plots, pseudo-images and classification methods were applied to ascertain only excellent quality arrays were used in the statistical and functional analyses (Heber and Sick 2006). An extensive description of the applied criteria is available upon request.

4.2.3 Statistical analysis of microarray data

Probesets were redefined according to Dai et al. (Dai et al. 2005) because the genome information utilized by Affymetrix at the time of designing the arrays is not current anymore, which may result in unreliable reconstruction of expression levels. In this study probes were reorganized based on

the Entrez Gene database, build 36, version 2 (remapped CDF v11). Expression estimates were obtained by GC-robust multi-array (GCRMA) analysis, employing the empirical Bayes approach for background adjustment, followed by quantile normalization and summarization (Wu et al. 2004). Differentially expressed probesets were identified using linear models, applying moderated t-statistics that implement empirical Bayes regularization of standard errors (Smyth 2004). To adjust for both the degree of independence of variances relative to the degree of identity and the relationship between variance and signal intensity, the moderated t-statistic was extended by a Bayesian hierarchical model to define a intensity-based moderated T-statistic (IBMT) (Sartor et al. 2006). P-values were corrected for multiple testing using a false discovery rate method (Storey and Tibshirani 2003). Probesets that satisfied the criterion of $FDR < 5\%$, $q - value < 0.05$ and $fold - change > 1.2$ were considered to be significantly regulated.

4.2.4 Source of molecular interaction data

Recently research has been described for selecting significant pathways given a variable of interest or phenotype using gene expression data (Draghici et al. 2007). Since we would like to use pathways as features we used the KEGG (Kanehisa et al. 2008) pathway database as the source of information due to its popularity and online access. Molecular interaction data, i.e. interactions between proteins and (in)organic molecules, were extracted using the Bioconductor package KEGGgraph (Zhang and Wiemann 2008).

4.2.5 Bayesian networks

Bayesian networks have become a popular technique to model metabolic networks and pathways (Gavai et al. 2009). They consist of a qualitative part, a directed graph, or network, G that does not include directed cycles, called an acyclic directed graph, and a quantitative part that corresponds to a joint probability distribution P. The nodes of the network correspond to the variables in the joint probability distribution. The structure of the network reflects conditional independence assumptions that must be satisfied by the associated probability distribution. These assumptions are called *Markov properties*.

By applying the chain rule of probability theory, taking into the Markov properties, the joint probability distribution P can be obtained in the form as shown in equation:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \text{Parents}_G(X_i)), \tag{4.1}$$

where $\text{Parents}_G(X_i)$ represents the set of parents of the vertices corresponding to X_i in graph G.

Using the Markov properties, it can also be deduced that any variable X is conditionally independent of all the other variables given its parents, children, and children's parents. The parents, children, and children's parents are called the *Markov blanket* of the variable. Thus, the Markov

blanket of a variable X, MB(X), shields it off from the other variables:

$$P(X_i \mid X - X_i) = P(X_i \mid \mathrm{MB}(X_i)).$$

There are various ways in which one is abe to construct a Bayesian network for a specific problem. For example, if one knows the interaction of metabolites in a certain pathway, one could even construct a hypothetical network manually, based only on knowledge extracted from literature. Each of the metabolites in the network would then be associated with probabilities embedded in conditional probability tables often known as prior probabilities or a person's degree of belief. Changing prior probabilities of one node then allows us to explore and update probability values over the rest of the nodes. The graphical nature of the network combined with probability theory allows one to do data analysis in an intuitive way. If there are feedback loops involved in a domain, these can be perfectly modeled using *dynamic* Bayesian networks (Murphy 2002). Relationships between variables of interest are established from data using partial correlations as a measure to identify (conditional) independences between variables from the dataset at hand. Once constructed, direct and indirect relationships can be identified easily just by looking at the Bayesian networks. Genes missing edges indicate (conditional) independence. There are two aspects for representing data using this technique viz. qualitative and quantitative. The qualitative aspect includes representation of data using nodes and edges and these relationships can be quantified using a conditional probability distribution. The nodes represent variables and edges represent causal or influential relationships between variables. Each node of the graph is associated with a conditional probability table. Bayesian networks allow us to infer the relationships, and reason about them efficiently, from cause to effect, yielding a kind of *predictive reasoning*, from effect to cause, yielding a kind of *diagnostic reasoning*, and intercausal relationship, then called *intercausal reasoning*.

Learning the networks from data, i.e. finding the relationship between variables, can be achieved in two ways: 1) using search-and-score-based methods, and 2) using constraint-based method. There have already been quite successful search and score methods implemented for biological data sets, but the problem with these methods is that they are very data demanding and computationally inefficient. The search space that needs to explored is superexponential in the number of variables, and therefore, normally heuristic or approximation methods are used (Husmeier et al. 2005). However, this does not resolve the large requirements with respect to the amount of data. Constraint-based method, where local structure is obtained by carrying out conditional independence tests on sets of variables is then more feasible. These methods also have the advantage that background knowledge can be readily incorporated into the learning process.

As not always knowledge about relationships between metabolites of interest is available, one is forced to find these relationships from the experimental data. This approach is not uncommon when the number of variables is large and there is little or no knowledge available of the underlying process. Moreover, it can be a laborious task to construct networks of several hundred nodes just by hand. Therefore, there has been considerable research in this area to do unsupervised learn-

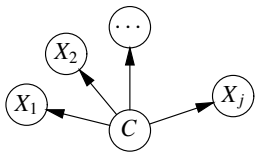


Figure 4.3: *Representation of relationship between class variable and the genes, there are no dependencies assumed between genes themselves.*

ing of conditional dependence and independence relationships from data, and learning Bayesian networks is an example of such research. The key assumptions used for this approach are that biological processes are hierarchical in nature and links between metabolites in metabolic processes are sparse.

4.2.6 Naïve Bayesian classifiers

Naïve Bayesian classifiers (Mitchell 1997) are special Bayesian networks in the sense that they have a special node, called the *class variable* C, that is surrounded by *feature variables* F_k; there is a directed edge from the class variable to each feature variable. The conditional independence assumption underlying the associated joint probability distribution $P(C, F_1, \ldots, F_n)$ is that each feature variable F_k is conditionally independent given the class variable C. A naïve Bayesian classifier is used to compute $P(C \mid \mathcal{F})$, with $\mathcal{F} \subseteq \{F_1, \ldots, F_n\}$. The value of C in the dataset is compared to the value C that surpasses a particular probabilistic threshold, often 0.5; if the values correspond, the case has been classified correctly (true positive); otherwise, its classification is incorrect (false negative). Despite its simplicity naïve Bayes classifier often outperform other classification techniques. We chose this technique to depict a "black box approach" for our analysis. There are two class variables on which the mice are being treated for *fed* and *fasted* conditions. A naïve Bayes classifier can be also thought upon as a Bayesian network as depicted in figure4.3.

4.2.7 Data representation in Bayesian networks

Data often come in a rectangular format consisting of columns X associated with the number of observations as rows. These columns are represented as a finite set of variables $X = \{X_1, \ldots, X_n\}$ of random variables, where each variable X_i can take a value x_i.

The main aim of many statistical models is to seek for a joint probability distribution that best fits the data; often this is a computationally expensive process. Bayesian networks have the advantage that they factorize this joint probability distribution, and by incorporating conditional independence assumptions the computations involved are reduced. As the independence assumptions of the naïve Bayesian classifier are particularly strong, computation is very efficient.

Given a fixed graph of a Bayesian network, the associated conditional probabilities $P(X_i \mid \text{Parents}(X_i))$ are learned either from experts or from the data at hand. If the data consists of discrete values a conditional probability table is constructed for each node given by $P(X_i \mid \text{Parents}(X_i))$. If the data is continuous, a Gaussian distribution is used which has the form

$$P(X \mid \text{Parents}(X_i)) \approx N(a_0 + \sum_{j=1}^{m} a_j x_j, \sigma^2), \tag{4.2}$$

where x_j are the values of the parent variables of X_i.

4.3 Results and Discussion

In the current study we investigated whether Bayesian networks could be applied to discriminate between two different metabolic conditions. To this end, a set of 19 arrays was used that were hybridized with RNA from livers derived from control (fed) mice, or mice that were fasted for 24 hours. Gene expression levels were estimated by applying GCRMA normalization.

After normalization and preprocessing of these data the next step is to perform feature selection. In microarray classification studies, typically a feature selection method is used to select a subset of genes from key genes present on the microarray. The criterion for this selection is mostly based on statistical procedures, such as the χ^2 test or t-test. Reducing the list of features is necessary since the majority of genes represented on the are not involved in not involved in sample classification. In addition, the problems of dimensionality and noise also have to be limited. Moreover, microarray data may be highly variable, and therefore the before mentioned procedures are often not robust enough. Instead, to perform this step we choose we choose 6 metabolic pathways which are known to be affected by the conditions of *feeding* and *fasting* viz. *glycolysis*, *fatty acid synthesis*, *TCA cycle*, *gluconeogenesis*, *fatty acid degradation* and *PPAR signaling pathway*. Genes involved in these pathways were extracted to map the features with pathways and these genes were selected from the experimental datasets. Merged pathways are presented in figure4.4 for illustration to show how complex these networks can be. Of the 238 genes represented by the merged pathways, 135 were represented on the microarray and their expression values were extracted from the dataset. Genes having missing values were removed. The *prior* for class variables were set to 0.48 and 0.52 for *feeding* and *fasting* respectively (due to uneven sample size for feeding and fasting states). Due to the small number of samples in two experiments, we used 19 fold cross validation to assess the classification performance. Various criteria are used for measuring classification accuracy including the total rate of correctly classified samples, the *sensitivity* and *specificity*, and the receiver operating characteristic (ROC) curve. Here we relied on the simple and intuitive measure on the correctly classified instances. The naïve Bayes classifier gave a classification accuracy for 135 genes to 89.47%. To further investigate these results we used

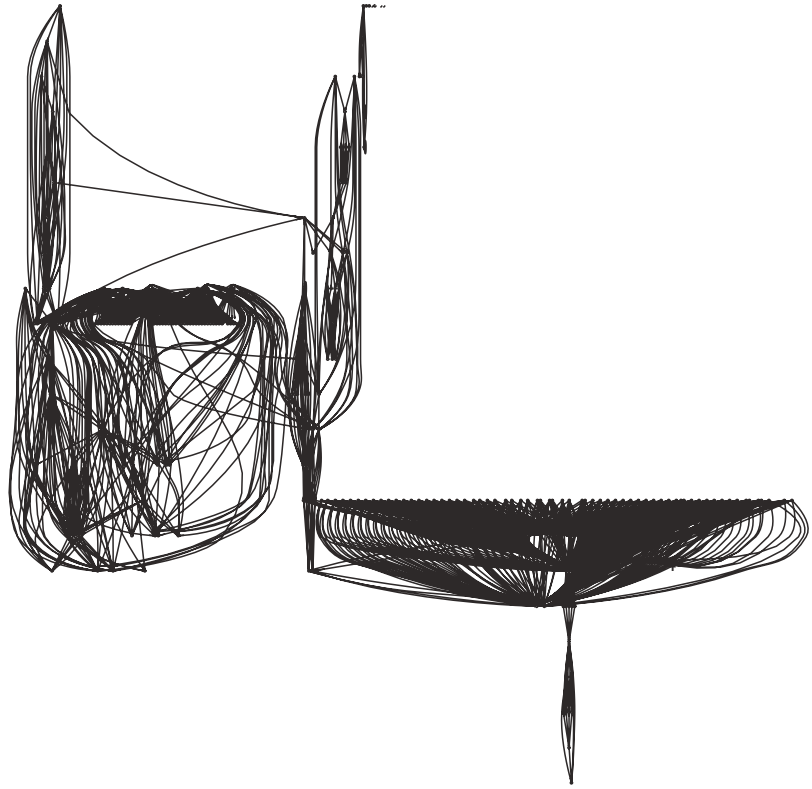


Figure 4.4: *Metabolic pathways involving* glycolysis/gluconeogenesis, fatty acid degradation/metabolism *and* tca *cycle.*

the PC algorithm for these genes. The learned network is presented in figure 4.5. This network represents 5 genes which fall in the Markov blanket of the class variable. These are *Hsd17b10*, *Acsl3*, *Acox1*, *Fasn* and *Scd1*. Classifying only on these 5 genes gives 100% accuracy. For cross validation purpose removing these genes from the complete data set of 135 genes the classification accuracy 84.21%. Given the fact that nutritional effects on gene expression are rather subtle these are good results, illustrating the utility of selecting pathways as features as a preliminary step. This approach, complemented with the PC algorithm, helps to identify important features needed for classification. This methodology acts as a proof-of-concept, motivating further development of the idea of pathways-as-features giving a focussed approach to the problem.

The genes involved in the Markov blanket of the class variable are involved in key pathways

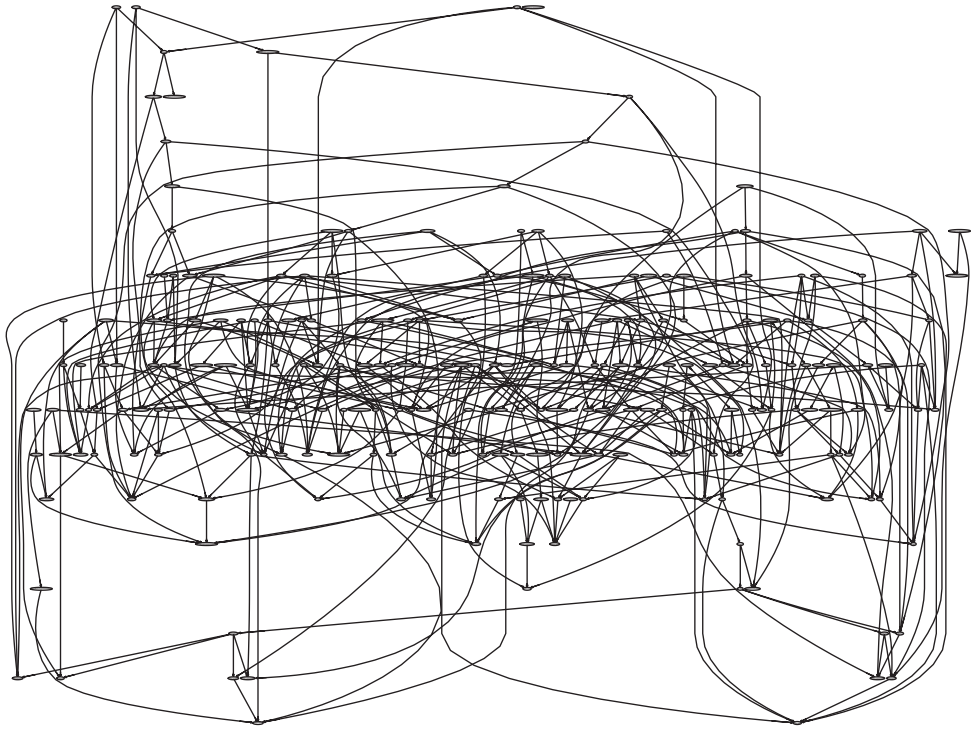


Figure 4.5: *Bayesian network consisting of 135 genes and Class variable. The Bayesian network encodes the dependence and independence information of the Class variable with other genes. Also relationship with other genes can be estimated in this way.*

like *biosynthesis of unsaturated fatty acids*, *PPAR signaling pathway*, *adipocytokine signaling pathway* and *insulin signaling pathway*. Our purpose here was to show a combined approach by incorporation of background knowledge combined together with constraint based learning to improve classification accuracy.

4.4 Conclusions

This study presented the importance of feature selection based to estimate feeding and fasting conditions in mice. The features selected using this technique adds valuable source of input to the constraint based learning algorithm. As classification accuracy is determined based on the noise in data. This can be improved by selecting the features involved in the Markov blanket of the *Class* variable of a constraint based algorithm. Naïve Bayes classifier can then be used to perform classification on these selected features showing its relevance in feature selection. Use

of naïve Bayes classifier combined together with state-of-art PC algorithm proved useful to perform classification task effectively and do exploratory data analysis. We have also explored the possibility of feature selection based upon pathways provides intuitive and focused approach to the classification problem. Bayesian network assumes independence among variables of interest and class variables. Because of these, Bayesian networks can also be used to represent knowledge in a graphical format. The learned network from such a technique could show some interesting relation among genes which could further be investigated by conducting experimental research.

In Preparation: *"Estimating the effect of cigarette smoke on gene expression - a Bayesian approach"* – D.M. van Leeuwen, Anand K. Gavai, M.H.M. van Herwijnen, R.W.H. Gottschalk, Jack A.M Leunissen, J.H.M. van Delft and J.C.S. Kleinjans

Chapter 5

Estimating the effect of cigarette smoke on gene expression - a Bayesian approach

"Smoking is one of the leading causes of statistics."

Unknown

Abstract

Chemical carcinogenesis induced by life style factors such as cigarette smoking is a major research area. Genome-wide transcriptomic analysis holds the opportunity to study the effects of such an exposure at the genome level yielding more mechanism-based information on the effects. An important issue in whole genome transcriptomics, yielding tremendous amounts of data, is adequate data analysis. Since genes do not function individually but more like a cascade of interactions, analyzing and visualizing these networks of genes would provide important mechanistically relevant information. Therefore, the aim of our study was to investigate the transcriptomic response to cigarette smoking using a Bayesian network approach. Bayesian networks are powerful tools as it allows us to model cause-effect *relationships between genes and clinical parameters. The population under study consists of* 9 *monozygotic twin pairs consisting each smoker and a non-smoker,*13 *independent smokers and* 13 *independent non-smokers. Gene expression from blood was analyzed using Agilent* $4 \times 44k$ *oligonucleotide microarrays. Other parameters, such as plasma cotinine levels and DNA adduct were chosen to be the variables of interest. The biological plausibility of the learned networks, together with the new observations provides us an opportunity to generate new hypotheses making the Bayesian network approach a very powerful analysis method for toxicogenomics data.*

5.1 Introduction

Cigarette smoke is a well known and thoroughly studied example of xenobiotic exposure. It is a rich source of chemicals and carcinogens, comprising a complex mixture of over 60 proven, probable and possible carcinogenic agents (Smith et al. 2003). It is known to induce many genotoxic, pre-carcinogenic effects, such as the formation of DNA adducts, micronuclei, chromosome aberrations and mutations, which are detectable in blood cells long before health effects

appear (Bonassi and William 2002, van Delft et al. 1998). Many of these can be measured or detected in both target tissue and surrogate target tissue like blood cells (Schooten et al. 1998, Flora et al. 2003). Transcriptomic analysis comprises a promising tool to gain more mechanistically relevant knowledge on the effects of xenobiotic exposures on the genome level (Aardema and James T 2002, Hamadeh et al. 2002, Toraason et al. 2004). The increased availability of whole genome microarrays has provided the opportunity to analyze effects in exposed populations. This in turn enables the study of biological responses to such exposures in an integrative manner, and provides a platform for e.g. new biomarker development.

Differential gene expression in human peripheral blood cells in vivo has been reported in a few studies (Wu et al. 2003, Lampe et al. 2004, Wang et al. 2005, Forrest et al. 2005). In a study on human leukocyte gene expression in smokers versus non-smokers, particular gene expression modulations have been found to correlate significantly with plasma cotinine levels and these genes accurately distinguished smokers from non-smokers (Godschalk et al. 1998).

Various methods in classical multivariate statistics are used to discover and visualize gene regulatory networks using supervised and unsupervised clustering methods (Ursem et al. 2008, Opgen-Rhein and Strimmer 2007). Clustering and correlation techniques provide good summarization of data which might indicate functional relations between genes of interest and clinical parameters but as these measures are global one cannot expect to find relations which are relevant for data where sample size is small. The interaction between genes and clinical parameters have to be viewed from a biological perspective. Knowledge which comes from experts and literature often plays an important role unfortunately it cannot be incorporated in these techniques as the relationships can be of different nature. Therefore one cannot estimate this just by accounting for linear relationships alone and one need to look beyond standard approaches to take into account the non-linear relationships as well as nature of these relationships (Husmeier et al. 2005).

To obtain an optimal yield of such studies, thorough mining of these data for information is crucial. Data mining of whole genome expression studies is currently a hot topic and many techniques exists like clustering, correlation and decision trees. However, the challenge is to simultaneously analyze the data comprising of gene expressions and clinical parameters. Moreover relevant information can only be found doing this type of analysis in an exploratory fashion to get the complete pictures of the key biological process. Various forms of graphical models could be used for this task for example Boolean networks (Kervizic and Corcos 2008), neural networks (Toure and Basu 2001) and Bayesian networks (Friedman et al. 2000). Approaches based on Bayesian networks have gained much popularity in recent years. This is because the relationship between genes of interest is often represented in the form of a network revealing probable metabolic pathways. Bayesian networks can be used efficiently in many scenarios when it comes to solving a biological problem involving expression data (Zou and Conzen 2005, Friedman et al. 2000). A Bayesian network which is a form of a probabilistic graphical model (Cooper 1999) consists of nodes connected with edges, where each node is associated with a random variable and edges

represent dependencies between them. The representation in the form of nodes and edges is often referred as the *qualitative* part and each node in this network is associated with a probability distribution which is referred as *quantitative* part. An example bayesian network is presented in figure 5.1. Techniques such as these can be combined together with biological insight and data in various ways to obtain more biological insight into the system. There is increasing evidence that complex diseases do not arise from an individual molecule or gene, but from complex interactions between the cell's different compartments and its environment over a period of time. Therefore the aim of this study was to use this technique and show *causal* relations instead of mere *correlations* from the dataset at hand. These methods are able to take the background knowledge (in the form of clinical parameters) into account making the results intuitive and exploratory data analysis can be performed efficiently.

The advantage of Bayesian networks over alternative techniques is that entering known information over a node in a network enables us to update probabilities over the rest of nodes. Analyzing data like this can result in some exceptionally powerful analysis technique that is not possible when using other forms of reasoning in classical analysis tools. For e.g. regression models are standard approaches and take into account data alone and produce dependence and independence within variables. But when there is insufficient data it cannot accommodate expert judgment and explanation to the relationships obtained by it. Moreover it cannot accommodate the impact of *process changes* on variables. These scenarios can be perfectly handled by Bayesian network models with high level of accuracy. In short, standard analysis techniques are good in describing the nature of data but fail in future prediction. Our primary focus here is to show how a graphical modeling approach based on Bayesian networks can be used to find relationships by combination of transcriptomics data coupled together with clinical parameters obtained from a monozygotic twin smoking population. The study by van Leeuwen et al(2007) (van Leeuwen et al. 2007) provides relevant background information on this dataset. We present here how the results obtained by it show *dependence* and *independence* information among variables of interest. In this study our main focus is on two important clinical parameters; DNA adducts and plasma cotinine levels. These variables are known to have strong effects from gene expression influenced by cigarette smoke.

5.2 Materials and Methods

5.2.1 Study population and blood collection

The study protocols were approved by the Local Committee of Medical Ethics of the Catholic University of Leuven (Belgium) and by the Institutional Review Board/Ethical Committee of Antwerp University (Belgium). Prior to venous blood collection, all study participants gave informed consent. The study population was composed of 13 independent smokers and 13 independent non-smokers, and nine monozygotic twin pairs discordant for cigarette smoking (van Leeuwen

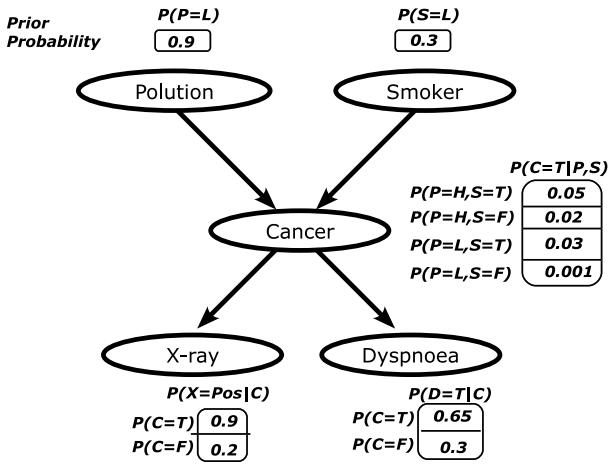


Figure 5.1: *A Bayesian network representation, where nodes and edges represent* qualitative part *and tables represent* quantitative part

	Current study
Study population	- 9 pairs of smoking-discordant MZ twins - 13 independent smokers - 13 independent non-smokers
Blood collection	- PAXgene for RNA (whole blood) - EDTA for cotinine and DNA adducts (tot. WBC)
Microarray platform	Agilent human 4x 44k oligonucleotide microarrays (44.000 genes)

Table 5.1: *Description of the dataset*

et al. 2007). Study population characteristics are summarized in Table 1. Blood was collected first into one tube containing EDTA and then into five PAXgene Blood RNA tubes (PreAnalytix, Qiagen, Hilden, Germany; 2.5 ml each) which instantly stabilize RNA. Tubes were kept at room temperature for at least two hours but maximally one day, and thereafter stored at 4 degree celcius until DNA and RNA isolation.

5.2.2 Nucleic acid isolation

Total RNA was isolated within a week after blood collection using the PAXgene Blood RNA Kit (PreAnalytix, Qiagen, Hilden, Germany) according to the manufacturer's instructions and as described previously (van Leeuwen et al. 2007). RNA quantity and purity was assessed spectrofotometrically and integrity was analyzed with the BioAnalyzer (Agilent, Palo Alto, CA, USA). Leukocyte fractions were isolated from EDTA anti-coagulated whole blood, after removal of the plasma layer, through erythrocyte lysis (lysis buffer containing 155 mM NH4Cl , 10 mM KHCO3, 10 mM EDTA:blood = 3:1) for 30 min at 4 degree celcius. Leukocytes were stored at -80C until DNA isolation. DNA was isolated by standard phenol extraction (Godschalk et al. 1998).

5.2.3 Gene expression analysis

Aliquots of 0.5 micrograms of total RNA were prepared for gene expression analysis on Agilent 4x44k human oligonucleotide microarrays, according to the manufacturer's instructions using the Agilent Low RNA Input Linear Amplification system (Agilent Technologies, Palo Alto, CA, USA). A common reference pool was composed of aliquots of total RNA from all smokers. Each microarray was hybridized with a mix of a Cyanine 5-labeled cRNA from a study individual and a Cyanine 3-labeled cRNA from the common reference pool, resulting in 44 hybridizations. Microarrays were scanned on a ScanArrayExpress scanner (Perkin Elmer, Boston, MA) with a fixed laser power and a variable photomultiplier gain, such that no spots were saturated. Scan images were imported into Imagene (BioDiscovery, Marina del Rey, CA) and translated into raw data files, which were transformed (data preparation) using GeneSight (BioDiscovery) as described previously (van Leeuwen et al. 2007).

5.2.4 Cotinine measurements

Plasma was collected from EDTA blood by centrifugation. Plasma cotinine was measured by high performance liquid chromatography (HPLC) according to the protocol described by Van Vunakis et al. (Vunakis et al. 1993) with the exception of the presence of EDTA in the standard.

5.2.5 DNA adduct measurements

DNA adduct levels were measured by 32P-postlabeling with Nuclease-P1 enrichment as originally described by Reddy and Randerath (Reddy and Randerath 1986) with minor modifications as described by Godschalk et al. (Godschalk et al. 1998).

5.2.6 Learning Bayesian Networks

Learning in Bayesian networks refer to finding relationships as well as nature of these relationships among variables of interest from a dataset. Bayesian networks can be learned in two ways

either by using the search and score technique (Friedman et al. 2000) or using constrain-based technique(Gavai et al. 2009). There have already been quite successful search-and-score methods implemented for biological data sets, but as the number of variables grows in size the computation time grows exponentially. Constraint-based methods have the advantage that they allow incorporation of prior biological knowledge and are computationally tractable; therefore this has been taken here as the method of choice. Moreover to our knowledge this is the first attempt to explore constraint based learning techniques on toxicogenomics data. PC-algorithm (Peter and Clark) (Sprites et al. 2000) is a constraint-based method, many variations of it exists and are widely used. There are certain assumptions to this approach, like the independence relationships have a perfect representation by a *ADG*, under this assumption the PC algorithm will discover an equivalent Bayesian network or hierarchical relationship among variables of interest. Another assumption is that it performs better for sparse networks, which is often the case in toxicogenomics data. A formal overview on this technique can be found in chapter 2 of this thesis.

Statistical analyses of differences in gene expression between smokers and non-smokers were performed in the Gene Expression Pattern Analysis Suite GEPAS *http://www.gepas.org* (Herrero et al. 2003) after applying preprocessing features within the program; merging replicates, imputing missing values with K-nearest neighbor imputation (KNN, n=15) and omitting genes with less than 70% present data. From these dataset features were selected for three different populations, chosen based upon their p-value. Probe sets having only unique identifiers have been kept for the analysis. A Bayesian network was constructed first for all the 300 genes and later for the top 25 genes combined together with clinical parameters (DNA adducts and plasma cotinine levels) of the subjects. The dependence and independence of genes and parameters were obtained by detecting the Markov blankets of equivalent network structures also called the (*I-map*) independence map for these variables (Tsamardinos et al. 2003). The Markov blanket of a node contains all the variables that shield the node from the rest of the network. This means that the Markov blanket of a node is the only knowledge needed to predict the behavior of that node. In our case the nodes of interest are DNA adducts and plasma cotinine levels. This suggests that a joint probability distribution of genes around these nodes are the only ones necessary to explain their behavior. Later these 300 genes were analyzed using pathway express (Draghici et al. 2007) which is tool to identify key pathways the variables of interest are involved in. This tool explores available pathways from KEGG (Ogata et al. 1999) database and genes are ranked based on its location in the pathway to show its significance. The analysis was performed using R (R Development Core Team 2008) and bioconductor (Gentleman et al. 2004), which are open source initiative for genomic data analysis. The PC algorithm was used from pcalg (Kalisch and Buhlmann 2007) package from bioconductor and the Markov blanket for the networks were calculated from the bnlearn (Scutari 2008) package from bioconductor.The visualization was obtained using Pajek[http://vlado.fmf.uni-lj.si/pub/networks/pajek/] which is a free tool for network visualization. The results were cross checked with the Comparative Toxicogenomics Database

http://ctd.mdibl.org/ (Davis et al. 2008).

5.3 Results and Discussion

The 22 twins population was divided into three groups. The complete population is referred to as *all*, the monozygotic twin pairs referred to as *twins* and the intersection of *all* and *twins* referred to as *independent*. The main purpose of this study was to investigate whether cigarette smoking causes elevated levels of gene expression and its effects on peripheral blood cells, like the DNA adducts and plasma cotinine levels.

Genes were selected based upon their differential expression and p-values obtained from a t-test for the *independent* population; these criteria were further extended for highest fold change for the *twin* and *all* population. The reason for selection of genes based only upon p-values for the *independent* population was that it contained more noise in the dataset as compared to the *twin* population.

These datasets were fed into the learning algorithm. Networks generated from these are depicted in figures 5.3, 5.4 & 5.5 showing relationships of genes only among the top 25 genes selected on p-value. Similarly networks were also constructed for the top 300 genes. Figure 5.2 gives an illustration on one of the networks showing how complex these networks can be. As the analysis focussed on the two variables of interest, i.e. DNA adducts and plasma cotinine levels, these were taken into account for the network construction algorithm. Our primary focus was to first estimate which genes interact with these variables and secondly how strong these interactions are. The strength of the interaction is shown by the thickness of the arrow, indicating strong statistical dependence among variables. From these network graphs we can also see which genes have a *direct* or *indirect* influence on the variables of interest. These *direct* and *indirect* effects can be found using the notion of the Markov blanket as mentioned in the materials and methods section. The Markov blanket of these networks were calculated and the result consisting of gene list, gene description and key pathways these genes are involved is shown in tables 5.2 - 5.7).

Genes having *causal* influence on DNA adducts and plasma cotinine may not always have a direct causal interaction, because of the fact that not all genes are profiled on the microarray. Such relationships indicate latent variables (hidden variables) and therefore indicate indirect effects on DNA adducts and plasma cotinine levels. Further investigation of these genes was performed using the Comparative Toxicogenomics Database (Davis et al. 2008). The intersection from the genes-disease relationship from this database and the Markov blanket of all the networks reveal the following genes to be common among them: *ASPH*, *ADM*, *ALPPL2*, *ASGR1 & ASGR2*. Gene *ASPH* is known to be involved in diseases like hepatocellular and transitional cells carcinoma. Gene *ADM* is involved in embryonal, hepatocellular, squamous cell, and transitional cell carcinoma as well cardiovascular diseases. Gene *ALPPL2* is also involved in hepatocellular carcinoma and gene *ASGR1* is involved in hepatocellular, non-small-cell lung & transitional cell carcinoma

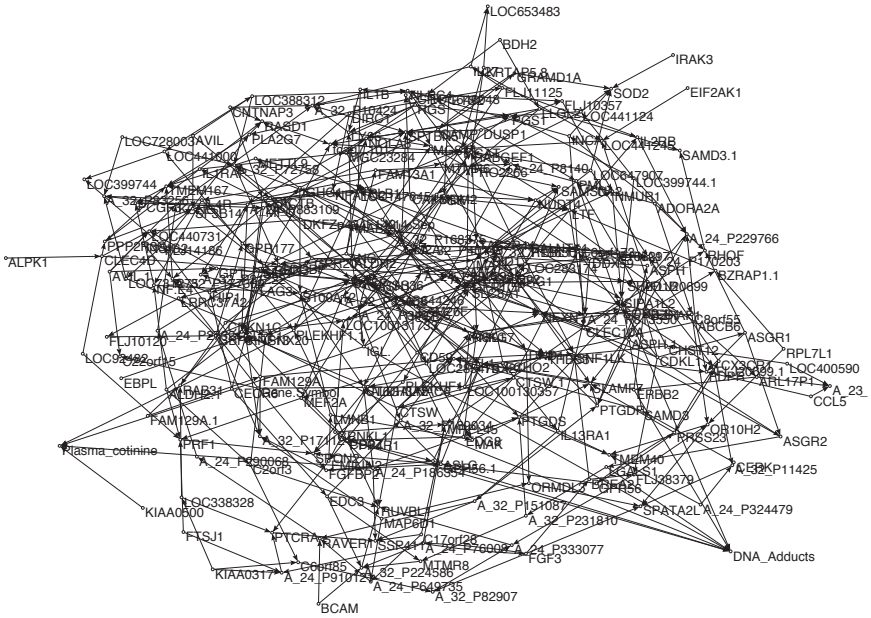


Figure 5.2: *Interaction of top* 300 *genes with clinical parameters for* independent *subjects*

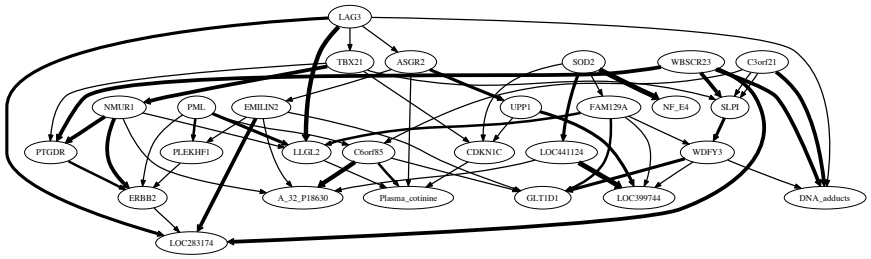


Figure 5.3: *A Bayesian network for* 25 *selected features from* all *subjects affecting DNA adducts and plasma cotinine levels*

along with gene *ASGR2*. The reason to find only these genes from this database may be the fact that cigarette smoke is composed of many carcinogens and chemicals that can cause DNA adducts and plasma cotinine levels. Not all these genes can be recovered using a single experiment like this. Apart from this, even if the effects are observed they are rather subtle and cannot be found easily using the standard hypothetical tests, which is used as a feature selection criterion in the current analysis. The most significant gene found from the network is the *ERBB2* and its relationship has been established by Amann et al. (Amann et al. 2005), where it is found to be active

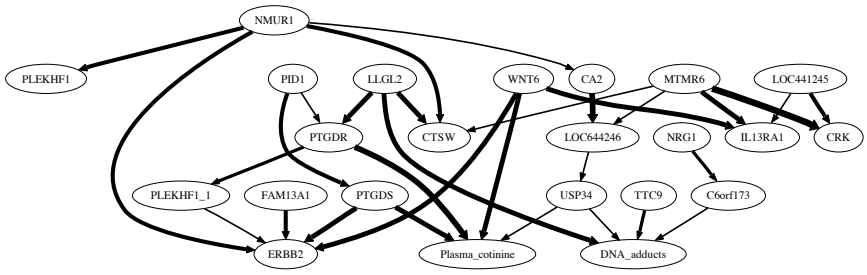


Figure 5.4: *A Bayesian network for* 25 *selected features from* independent *subjects affecting DNA adducts and plasma cotinine levels*

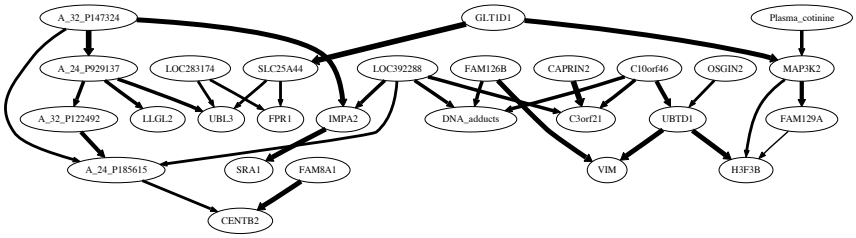


Figure 5.5: *A Bayesian network for* 25 *selected features from* twin *subjects affecting DNA adducts and plasma cotinine levels*

in diseases leading to bladder cancer, endometrial cancer, non-small cell lung cancer, pancreatic cancer and prostate cancer. Gene *FASLG* also plays an important role and leads to diseases like lung cancer (Okouoyo et al. 2004) and autoimmune thyroid disease as well as apoptosis. Besides these studies, smoke-induced effects on animal models like *SKH-1* mice show a strong influence on *MPK* expression (Izzotti et al. 2003, Izzotti et al. 2004). These results are also confirmed by the presence of *MAP3K2*. Gene *WNT6* is involved in diseases like basal cell carcinoma and gene *CDKN1C* is involved in cell cycle and is responsible for tumor suppression. These results taken together provide compelling evidence.

5.4 Conclusion

The present study is unique in two ways: first the use of graphical models is explored in this domain and secondly techniques based on constraint based learning have been exploited in molecular epidemiology. Approaches based on Bayesian networks reveal the interaction of various entities in an intuitive way and at the same time *causality* can be established among the variables of interest. Further use of constraint based learning will not only identify key genes responsible for

Subjects		GeneSymbol	Gene Description	KEY PATHWAYS
All 300	**DNA adducts**			
		ASPH	Aspartate beta-hydroxylase	
		IGHA1	Immunoglobulin heavy constant alpha 1	
		KIR2DS4	Killer cell immunoglobulin-like receptor, two domains, short cytoplasmic tail, 4	Antigen processing and presentation - Homo sapiens (human)
		LOC644246	Hypothetical protein LOC644246	
		SASP	Skin aspartic protease	
		ADM	Adrenomedullin	
		ALPPL2	Alkaline phosphatase, placental-like 2	gamma-Hexachlorocyclohexane degradation; Folate biosynthesis
		C20orf199	Chromosome 20 open reading frame 199	
		CNTNAP3	Contactin associated protein-like 3	
		SDCBP	Syndecan binding protein (syntenin)	
		CD160	CD160 molecule	
		CSDA	Cold shock domain protein A	Tight junction;
		PROK2	Prokineticin 2	
	Plasma cotinine			
		ASGR1	Asialoglycoprotein receptor 1	
		FAM126B	Family with sequence similarity 126, member B	
		HPSE	Heparanase	Glycosaminoglycan degradation Glycan structures - degradation
		LOC552891		
		SASP	Skin aspartic protease	
		SPON2	Spondin 2, extracellular matrix protein	

Table 5.2: *Top* 300 *genes influencing DNA adducts and Plasma cotinine levels for* all *subjects*

Subjects		GeneSymbol	Gene Description	KEY PATHWAYS
All 25				
	DNA adducts			
		WDFY3	WD repeat and FYVE domain containing 3	
		WBSCR23	Williams-Beuren syndrome chromosome region 23	
		C3orf21	Chromosome 3 open reading frame 21	
		LAG3	Lymphocyte-activation gene 3	
		C10orf46	Chromosome 10 open reading frame 46	
		LOC392288		
		FAM126B	Family with sequence similarity 126, member B	
	Plasma cotinine			
		ASGR2	Asialoglycoprotein receptor 2	
		CDKN1C	Cyclin-dependent kinase inhibitor 1C (p57, Kip2)	Cell cycle
		LLGL2	Lethal giant larvae homolog 2 (Drosophila)	Tight junction
		C6orf85	Chromosome 6 open reading frame 85	
		MAP3K2	Mitogen-activated protein kinase kinase kinase 2	MAPK signaling pathway; Gap junction;GnRH signaling pathway
		GLT1D1	Glycosyltransferase 1 domain containing 1	

Table 5.3: *Top 25 genes influencing DNA adducts and Plasma cotinine levels for* all *subjects*

Subjects		GeneSymbol	Gene Description	KEY PATHWAYS
Indep 300				
	DNA adducts			
		ERBB2	V-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	ErbB signaling pathway;Calcium signaling pathway;Focal adhesion; Adherens junction;Pancreatic cancer; Endometrial cancer; Prostate cancer; Bladder cancer; Non-small cell lung cancer;
		FASLG	Fas ligand (TNF superfamily, member 6)	MAPK signaling pathway; Cytokine-cytokine receptor interaction; Apoptosis ; Natural killer cell mediated cytotoxicity; Type I diabetes mellitus - Homo sapiens (human);
		IGL		
		LGALS1	Lectin, galactoside-binding, soluble, 1 (galectin 1)	
		A_32_P189034		
		PPP4R1	Protein phosphatase 4, regulatory subunit 1	
		SIPA1L2	Signal-induced proliferation-associated 1 like 2	
	Plasma cotinine			
		IGL		
		KIAA0500	KIAA0500 protein	
		A_24_P186354		
		PLEKHF1		

Table 5.4: *Top* 300 *genes influencing DNA adducts and Plasma cotinine levels for* independent *subjects*

Subjects		GeneSymbol	Gene Description	KEY PATHWAYS
Indep 25				
	DNA adducts			
		LLGL2	Lethal giant larvae homolog 2 (Drosophila)	Tight junction;
		TTC9	Tetratricopeptide repeat domain 9	
		USP34	Ubiquitin specific peptidase 34	
		C6orf173	Chromosome 6 open reading frame 173	
	Plasma cotinine			
		WNT6	Wingless-type MMTV integration site family, member 6	Wnt signaling pathway; Hedgehog signaling pathway; Melanogenesis; Basal cell carcinoma;
		PTGDS	Prostaglandin D2 synthase 21kDa (brain)	path:hsa00590 Arachidonic acid metabolism - Homo sapiens (human)
		PTGDR	Prostaglandin D2 receptor (DP)	path:hsa04080 Neuroactive ligand-receptor interaction - Homo sapiens (human)
		USP34	Ubiquitin specific peptidase 34	
		C6orf173	Chromosome 6 open reading frame 173	

Table 5.5: *Top* 25 *genes influencing DNA adducts and Plasma cotinine levels for* independent *subjects*

Subjects		GeneSymbol	Gene Description	KEY PATHWAYS
Twins 300				
	DNA adducts			
		DDEF1	Development and differentiation enhancing factor 1	
		A_24_P602168		
		NEURL	Neuralized homolog (Drosophila)	
		OR11A1	Olfactory receptor, family 11, subfamily A, member 1	
		RFX2	Regulatory factor X, 2 (influences HLA class II expression)	
	Plasma cotinine			
		CPSF2	Cleavage and polyadenylation specific factor 2, 100kDa	
		KIAA0319	KIAA0319	
		A_23_P251196		
		PTOV1	Prostate tumor overexpressed gene 1	
		UBTD1	Ubiquitin domain containing 1	

Table 5.6: *Top* 300 *genes influencing DNA adducts and Plasma cotinine levels for* twin *subjects*

Subjects		GeneSymbol	Gene Description	KEY PATHWAYS
Twins_25				
	DNA adducts			
		LOC392288		
		FAM1268		
		C10orf46	Chromosome 10 open reading frame 46	
	Plasma cotinine			
		MAP3K2	Mitogen-activated protein kinase kinase kinase 2	MAPK signaling pathway; Gap junction;GnRH signaling pathway.
		GLT1D1	Glycosyltransferase 1 domain containing 1	
		CAPRIN2	Caprin family member 2	
		OSGIN2	Oxidative stress induced growth inhibitor family member 2	

Table 5.7: *Top* 25 *genes influencing DNA adducts and Plasma cotinine levels for* twin *subjects*

the overall effect but will also serve in predicting the behavior of these genes. Models like these have allowed us to understand the dependence and independence relationships between genes and other parameters. The approach described in the present paper helps recover key genes and their influence on clinical parameters. The results are confirmed the literature and the Comparative Toxicogenomics database. As more and more data becomes available these methods can outperform classical multivariate statistical techniques and be used to find novel relationships. The resulting hypothesis can then be used to form the basis of subsequent research which can learn from data, take prior inputs from molecular biologist and update probabilities in the light of ”new evidence”.

Metabolomics (2009): *"Constraint-based probabilistic learning of metabolic pathways from tomato volatiles"* – Anand K. Gavai, Yury Tikunov, Remco Ursem, Arnaud Bovy, Fred van Eeuwijk, Harm Nijveen, Peter J.F. Lucas, Jack A.M. Leunissen - In Press

Chapter 6

Constraint-based probabilistic learning of metabolic pathways from tomato volatiles

"I am learning all the time. The tombstone will be my diploma."

Eartha Kitt

Abstract

Clustering and correlation analysis techniques have become popular tools for the analysis of data produced by metabolomics experiments. The results obtained from these approaches provide an overview of the interactions between objects of interest. Often in these experiments, one is more interested in information about the nature of these relationships, e.g. cause-effect relationships, than in the actual strength of the interactions. Finding such relationships is of crucial importance as most biological processes can only be understood in this way. Bayesian networks allow representation of these cause-effect *relationships among variables of interest in terms of whether and how they influence each other given that a third, possibly empty, group of variables is known. This technique also allows the incorporation of* prior knowledge *as established from the literature or from biologists. The representation as a directed graph of these relationship is highly intuitive and helps to understand these processes. This paper describes how constraint-based Bayesian networks can be applied to metabolomics data and can be used to uncover the important pathways which play a significant role in the ripening of fresh tomatoes. We also show here how this methods of reconstructing pathways is intuitive and performs better than classical techniques. Methods for learning Bayesian network models are powerful tools for the analysis of data of the magnitude as generated by metabolomics experiments. It allows one to model* cause-effect *relationships and helps in understanding the underlying processes.*

6.1 Introduction

Metabolomics plays an increasingly important role in the research area of drug discovery, food & nutrition, plant and animal biology and many other applications. Where transcriptomic and proteomic analysis does not tell the complete story, metabolic profiling can add significantly to the picture of what is happening inside a living cell. Statistical and mathematical techniques are commonly used to correlate changes in metabolic composition with changes in biological conditions ((Suizdak 2003, Eiceman and Karpas 2005, Gohlke 1959, Weckwerth 2003, Kopka

et al. 2004)). Chromatography coupled to mass spectrometry based methods, e.g. *GC-MS* and *LC-MS*, have been the most popular metabolic profiling techniques over the past decade. Hundreds of new metabolites have been identified in plants ((Fiehn et al. 2000, Moco et al. 2006, Schauer et al. 2006)) and the improved sensitivity of modern methods has led to an increased amount of metabolic information. Techniques such as these have enabled identification of metabolites at much higher resolutions than previously possible. Interesting relationships can thus be found by integrating different types of data *(omics)* from various analytical sources. In this paper, Bayesian network learning methods are explored to uncover molecular pathways of tomato metabolism. This is done by using constraint-based learning methods. The related research is reviewed in the next section.

6.1.1 Related research

Various methods in classical multivariate statistics have been used in the past to discover and visualize complex metabolic networks using supervised and unsupervised clustering methods. Clustering techniques ((Opgen-Rhein and Strimmer 2007)) provide good summarization of data concerning functional relations between metabolites but as these methods are global one cannot expect to find relations which are relevant for small subsets of data. These techniques are good for capturing whether or not variables influence each other; however the nature of interaction between metabolites is complex and cannot be estimated using only linear correlations ((Husmeier et al. 2005)). Furthermore, domain knowledge, which often plays a vital role to find novel relationships and which can be obtained from literature and experts, cannot be incorporated in these traditional techniques with the exception of choosing the proper parametric form of the functional interaction between variables, e.g. linear or exponential. Learning logistic regression models from data is the standard approach for capturing the statistical interactions among a set of input variables to predict the value of a dependent, or output, variable. However, it is difficult to establish the impact of process changes among the variables using only regression models. Moreover, when there are insufficient data it cannot accommodate background knowledge (expert judgement) and causal explanation to the relationships obtained.

We present here a constraint-based Bayesian network approach, which is a specialized form of graphical models ((Jordan 2004)). Use of Bayesian networks to analyze biological datasets in various *genomics* domain has been growing in the last decade. Different types of Bayesian network learning methods, e.g. search and score, have been used to recover target-regulator pairs from a yeast cell cycle microarray datasets ((Murphy 2002, Zou and Conzen 2005)), and also significant work has been done by Friedman et al. to reconstruct gene regulatory networks from microarray datasets ((Friedman et al. 2000)). Similar techniques have also been used for pathway identification to understand the underlying biological processes. In this paper we demonstrate how Bayesian network learning techniques can be used to uncover a very important metabolic pathway (oxylipin pathway) which plays an important role in the ripening of fresh tomatoes. We also show

why this technique is better than other statistical techniques used in this domain.

In principle a Bayesian network is a graphical representation of a multivariate probability distribution and is an example of a probabilistic network. A probabilistic network typically consists of nodes connected by edges, where each node corresponds to a random variable and edges represent dependence between them. Absence of edges between nodes represents conditional independence. Bayesian networks, a special type of a probabilistic network, contain directed edges, also called as arcs or arrows. The advantage of Bayesian networks over alternative techniques (e.g. logistic regression) is that they allow explicit representation of the mutual interactions among variables and groups of variables. They take into account the explanatory power of known variables in an intuitively simple graphical format. As a Bayesian network is a multivariate probability distribution with statistical independence assumptions, it is possible to reason probabilistically with this representation *(For an example see the next section)*. The other advantage lies in the fact that probability distributions can be updated in the light of new, known information. This is why Bayesian networks can be used to support decision making. The results obtained using this technique are exceptionally intuitive and this type of analysis is not possible by classical analysis tools. Nonetheless, there are few limitations; example as the number of variables increase in size the computational complexity increases and is referred as NP-hard problem. Apart from this Bayesian networks are also sensitive to sample size as regression models. However, on the positive side most biological processes are hierarchical in nature and there are more variables then relationships between the variables (the graphs are sparse).

Example

In figure 6.1 we present an example of the urea and citric acid cycles. These two cycles are linked by the synthesis of *fumarate*. To highlight the basics of the Bayesian network method, the reactions leading *to* and *from fumarate* are shown in figure 6.2, where both the probability distribution and the graph are shown. The graph encodes rather subtle information about statistical dependence and independence between sets of variables. For example, according to the graph in figure 6.2, the concentrations of both *argininosuccinate* and *succinate* are independent as there are no arrows connecting these nodes. Both production of *fumarate* and *malate* are a common consequence of presence of *argininosuccinate* and *succinate* (as there are directed paths going from these nodes to both *fumarate* and *malate*). The semantics attached to Bayesian networks implies that if either *fumarate* or *malate* or *oxaloacetate* are observed in high levels, then *argininosuccinate* and *succinate* become dependent given the fact that we know levels of *fumarate*. Finally, *argininosuccinate* (or *succinate*) and *oxaloacetate* are conditionally independent given the concentration levels of *fumarate*. It also means that if one has observed high or low levels of *fumarate* then this does not convey any new information about concentrations of *malate* or *oxaloacetate*, and vice versa. Given a Bayesian network, any conditional probability involving any of the variables included in the model can be computed.

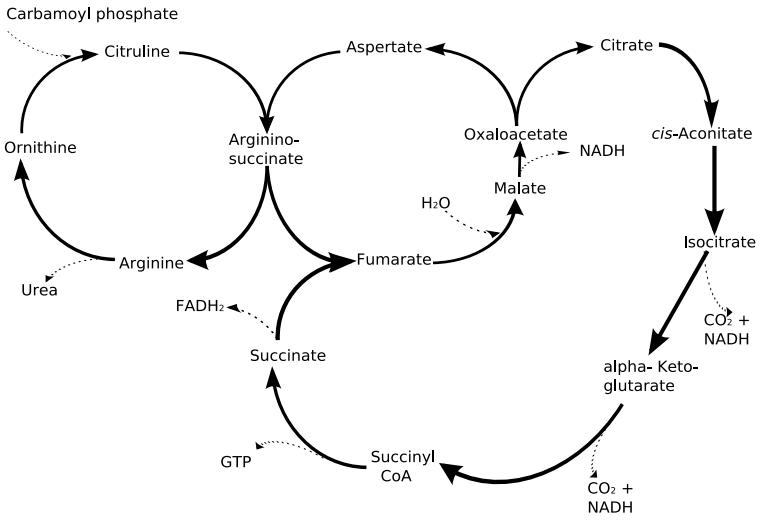


Figure 6.1: *Urea/Citric Acid cycle.*

It is a standard practice to compute probabilities of individual variables from a set of variables. These probabilities are referred as marginal or conditional and are updated by fixing observations over one or more variables. Figure 6.3 shows the bar graphs associated with the individual variables when the frequency of states of these variables are computed; in figure6.4 the variables have been conditioned on the assumption that oxaloacetate = high to display that for the case when the concentration is higher for a certain metabolite. In both figures simply looking at the shape of the bar graphs already conveys much information on the concentration levels of each metabolite.

6.1.2 Overview of Bayesian networks

Bayesian networks can be learned from data as a standard practice in multivariate statistics, but as they are easily understood they can also be manually constructed based on expert knowledge in a particular problem domain. For example, if one knows the interaction of metabolites in a certain pathway, one can even make a hypothetical network based on literature without data. Each of the metabolites in the network would be associated with a probability embedded in a contingency table (also known as a conditional probability table), expressing an expert's degree of belief. As was illustrated above, if one has observed a level of one or more compounds it is possible, using the Bayesian network, to predict the likelihood or concentration levels of the other compounds. Thus the likelihood represents the concentration levels of metabolites. The important aspects of understanding Bayesian networks lies in the fact that the graph structure of network is separate

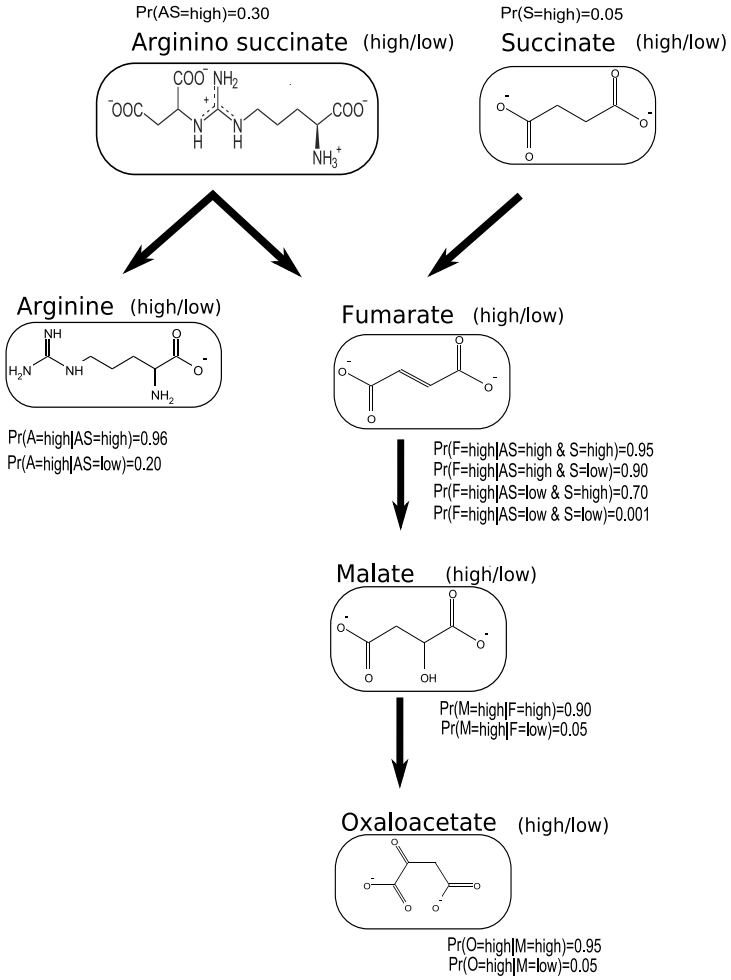


Figure 6.2: *Example of a simple Bayesian network consisting of a probability distribution Pr and a directed graph. The probability distribution Pr is specified using conditional probability distribution associated to the individual nodes, such as* $\Pr(\mathrm{A} = \mathrm{high} \mid \mathrm{AS} = \mathrm{high}) = 0.96$.

from the probability distribution associated with it.

The graphical nature of the network combined with probability theory allows one to do data analysis in an intuitive way. It is important to understand the interaction between different vari-

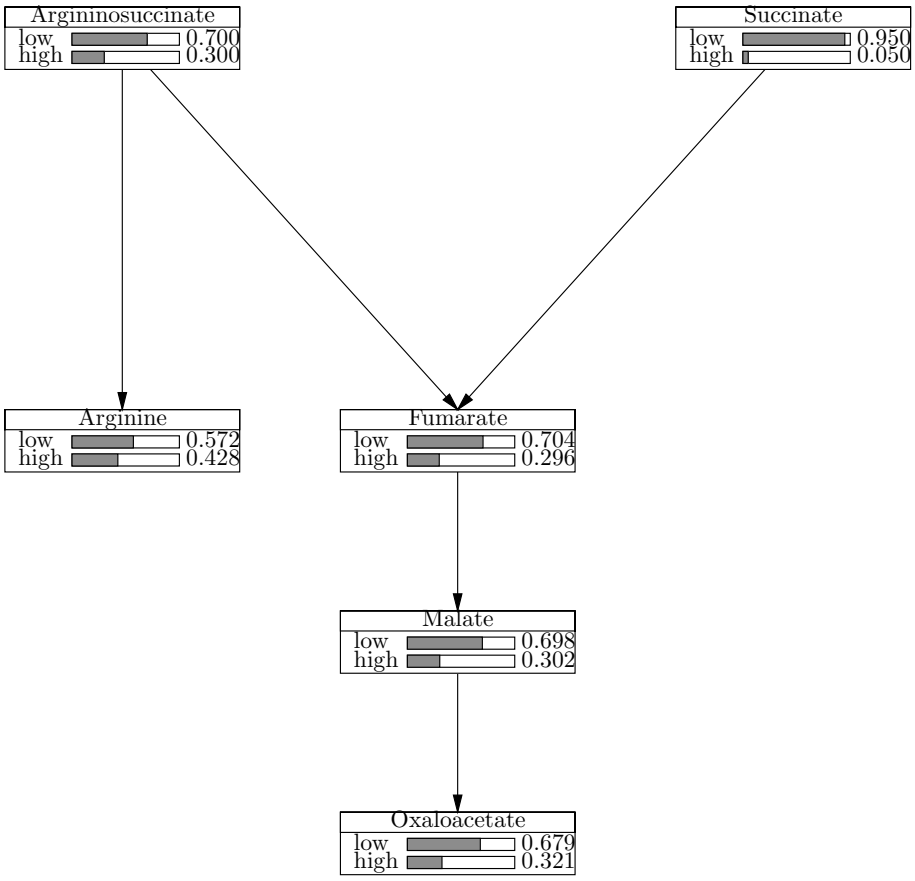


Figure 6.3: *Prior marginal probability distributions for the Bayesian belief network shown in Figure 2.*

ables but it is more important to understand the nature of these relationships. From example in the previous section representing *urea/citric* acid cycle in figure 6.1 *argininosuccinate*, *fumarate* and *malate* represent a serial connection, *arginine*, *argininosuccinate* and *fumarate* represent a diverging connection and *argininosuccinate*, *fumarate* and *succinate* represent a converging connection. As mentioned before knowing information about the concentrations levels of *fumarate* makes *argininosuccinate* and *malate* independent in a serial connection, knowing concentration levels of *argininosuccinate* make *arginine* and *fumarate* independent in a diverging connection and knowing concentration levels of *fumarate* makes *argininosuccinate* and *succinate* dependent in a converging connection. A formal overview on Bayesian networks can be found in chapter 2

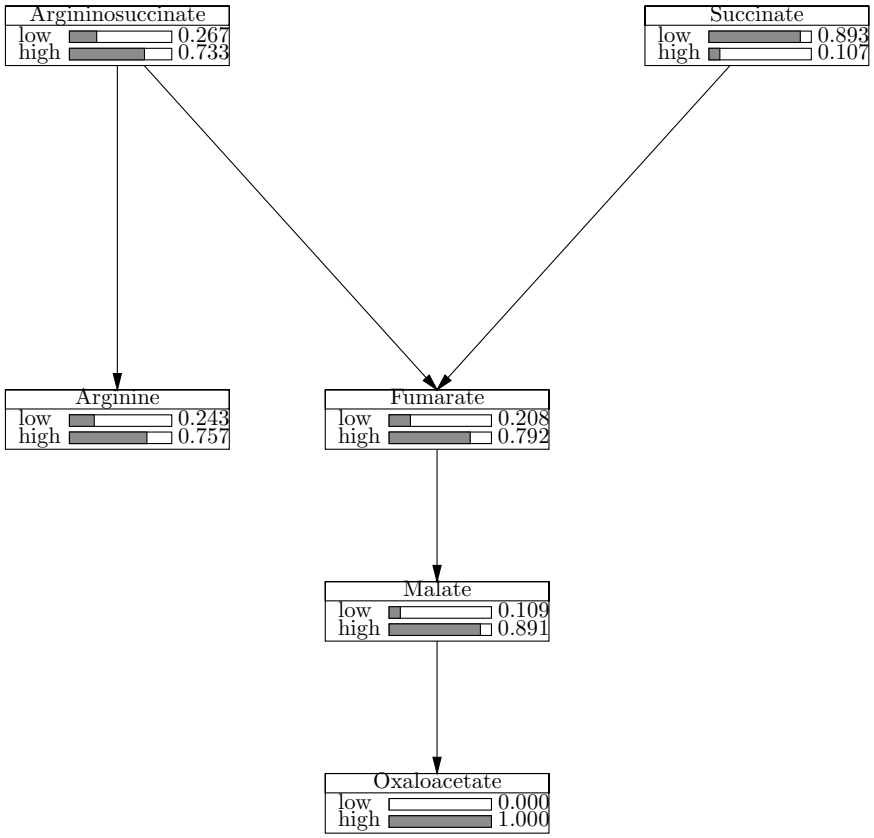


Figure 6.4: *Posterior marginal probability distributions for the Bayesian belief network after entering evidence on concentration levels of oxaloacetate . Note the increase in probabilities of the levels of concentrations of both oxaloacetate and argininosuccinate compared to Figure 3. It also predicts that it is more likely that the concentration levels of argininosuccinate to be high.*

of this thesis.

Not always we will have knowledge about relationships between metabolites of interest in which case we would want to find these relationships from available experimental data. This approach is not uncommon when the number of variables is large and there is little or no knowledge available of the underlying process. Moreover, it can be a laborious task to construct networks of several hundred nodes just by hand. Therefore there has been considerable research in this area to do unsupervised learning of conditional dependence and independence relationships from data.

The key assumptions used for this approach are that biological processes are hierarchical in nature and links between metabolites in metabolic processes are sparse in nature. In the past, correlation and clustering methods have been used successfully to identify groups of metabolites clustering together and reconstruct pathways ((Yilmaz 2001, Ursem et al. 2008)). Bayesian networks allows us to model causal relationships by looking at the direction of the arrows in the directed graph. Causal relationships play a vital role when we want to find out when one variable causes a change in another variable. Often these relations are investigated by experimental research to determine if changes in one variable truly cause change in another variable. However to generate an equivalent network is still possible using Bayesian network. For Bayesian networks there is the restriction that arrows are not allowed to form directed cycles (paths that end at the node where they started)–these graphs are called *acyclic*. Of course, there can be feedback loops involved in a problem domain which can be perfectly modeled using another type of Bayesian networks, so-called *dynamic* Bayesian networks ((Murphy 2002)), which require time-series data.

Learning in Bayesian networks

Learning the graph of a Bayesian network is done by exploring data using partial correlations as a means to distinguish dependent from independent relationships. To put it simple, direct and indirect relationships can be identified easily from the constructed networks: metabolites missing arrows indicate (conditional) independence. There are two aspects of representing data using this technique viz. qualitative and quantitative. The qualitative aspect includes representation of data using nodes and arrows and these relationships can be quantified using a conditional probability distribution.Constraint-based methods have the advantage that they allow incorporation of prior biological knowledge about dependence of variables, and therefore this has been taken as the method of choice for the present research. We consider here the PC-algorithm (Peter and Clark) ((Sprites et al. 2000)) which is a constraint-based method. There are certain assumptions to this approach such as the independence between nodes has a perfect representation by an *ADG*; under this assumption the PC algorithm will discover an equivalent Bayesian network. Another assumption is that networks are sparse, i.e. have few relationships between metabolites as shown in figure 1. There are several ways to verify conditional independence relationships which include reducing size of the database, finding correlations, direct query from experts and finding clusters in a causal network.

The algorithm is based on asking true independence relationship between sets of variables of the form $X_i \perp\!\!\!\perp X_j \mid S$), where S is a subset of variables. An overview of the steps are as follows:

- Construct an undirected graph.
- Find converging connections, by testing for independence and
- Give directions to the links without producing cycles.

Here we consider an imaginary oracle as our expert which tells us if two nodes are conditionally independent given a subset of nodes S (later this oracle will be replaced by statistical test to find partial correlations, e.g. G^2 test or Fisher's Z transformation). If the oracle (domain expert) says two nodes are conditionally independent given a third node S then we remove the edge between two nodes and make them independent. Asking questions like this for all the nodes involved and recursively deleting edges between nodes based on the answers will result in an undirected graph also known as the skeleton of the network. The next step then is to give directions to the edges based on rules ((Meek 1995b)) to generate a ADG, sometime there might exist bi-directional arrows which is not a part of Bayesian networks, but its presence indicate hidden nodes, which might have been missed from the experiment or not being observed.

In this paper, we demonstrate how learning and reasoning with Bayesian networks can be used to reconstruct the oxylipin pathway found in the synthesis of fresh tomato volatiles. We show how the results obtained by running Bayesian network analysis on experimental data of biochemical compounds of beef, round and cherry tomatoes supports the elucidation of the nature of the biological processes. For brevity, our focus is only on learning of structures and not on parameter estimation, which is often used to learn the probability distribution of the dataset. Subsequently, we discuss how to interpret the graphs generated by Bayesian network learning (which is an important aspect of data analysis). Prior knowledge obtained from literature is also taken into account in the form of metabolite selection for the analysis. The dataset used includes data of tomato volatile metabolite profiling, as described in ((Tikunov et al. 2005)). A detailed description of this dataset can be found below in the materials sections of the article.

6.2 Materials and Methods

6.2.1 Description of the Metabolite Dataset

Flavors in tomatoes are important targets for plant breeders to improve the quality of fresh tomatoes. Therefore it has become a popular area of research among molecular biologists to study the pathways involved in biosynthesis, of the *oxylipin (lipoxygenase)* pathway of volatile compounds (VOC, volatiles) in tomatoes. In plants the substrates of these pathways are *linoleic* and *linolenic acid*, while their mammalian equivalents are *arachidonic* and *eicosapentaenoic* acids. Tomato volatiles are generally divided into six groups ((Yilmaz 2001)) lipid derived, carotenoid related , amino acid related , terpenoids , lignin related and miscellaneous. Each group participates in different pathways involved in the biosynthesis of the aroma volatiles. Figure6.6 shows formation of lipid-derived volatiles through one of these pathways. There has been substantial research in this area, however the exact nature of the relationship between volatile compounds involved is still unknown.

Volatiles have been analyzed using gas chromatography mass spectrometry in ripe fruits of 94 tomato (Solanum lycopersicum L.) varieties as described in Tikunov et al. ((Tikunov et al. 2005)).

The varieties selected represent a considerable collection of genetic and therefore phenotypic variation. 322 VOC have been detected and 69 VOC identified most reliably have been chosen for the present study. This set of 69 VOC contains metabolites of 7 biochemical groups: volatiles derived from lipids, two phenylalanine derived groups, leucine and/or isoleucine derived volatiles, open-chain carotenoid derivatives, cyclic carotenoid derivatives and terpenoids. The last three groups are biochemically related and called isoprenoids.

6.2.2 Analysis

We consider here finding relationships between volatile metabolites of lipid derivatives involved in the *oxylipin / lipoxygenase (LOX)* pathway which occurs during ripening of tomato fruit ((Yilmaz 2001)) as depicted in figure 6.6. We used the package *pcalg* ((Kalisch and Buhlmann 2007)) which is an implementation of the PC algorithm in R ((Gentleman et al. 2004)) which is an open source environment for statistical computing to perform the analysis on a real-life dataset as mentioned in the results and discussion section of the article. The inputs were the tomato dataset and the algorithm allows to set a threshold to find significant conditional (independent) relationships between these metabolites. The algorithm generates a graph object which is an ADG. A bidirectional arrow in such network implies presence of hidden variables from the experiment being conducted (hidden factors or latent variables), e.g. when a metabolite is missing from the experiment ((Beal et al. 2005)). Here we show how this algorithm performs on a real-life dataset and finds relationships among metabolites of interest which are biologically meaningful and by taking into account the prior knowledge the generated network is compared to the established knowledge found in literature. This can be done by counting missing edges (false negatives) and extra edges (false positives) which were computed for these metabolites by the algorithm. Finally the structure Hamming distance metric ((Tsamardinos et al. 2006)) a measure to calculate the number of substitutions required to transform one graph to another, is used to calculate the distance (difference) of the computed network from the actual network (pathway). The lower this score is, the better the PC algorithm has performed on the dataset.

The visualization of complex networks is not easy and the knowledge represented by them is sometimes not obvious just by looking at these complex networks. The key to solving these issues is to make use of an interactive graph which allows looking at the chemical structures and getting relevant information from online repositories using well-established visualization techniques. We therefore implemented a framework which handles these issues, and the networks shown in this paper have been created using the tools. The tool was developed using open source software and standards, such as Graphviz (http://www.graphviz.org) and SVG (http://www.w3.org/Graphics/SVG/). All the scripts used for construction of networks are available from the authors upon request.

6.3 Results and Discussion

A primary interest in biology is finding novel biochemical pathways describing relationships between metabolites and their dependence on environmental factors. In this study we show how parts of a plant metabolic system can be reconstructed and visualized by applying Bayesian networks. We focused on 69 volatile compounds and the choice of these metabolites was based upon prior knowledge obtained from the domain experts and relevant literature ((Tikunov et al. 2005, Yilmaz 2001, Yilmaz et al. 2001)). We used constraint-based learning of Bayesian networks with a very low significance level α of 0.0001 on this dataset. The test statistic used to find the relationship and their strength are based on Fisher's Z transformation ((Kalisch and Buhlmann 2007)). A Bayesian network estimates a ADG and so relationships do not form a cycle, observing all such relationships indicate hidden variables which might have been missed by chance or not being observed in the experiment. As Bayesian networks generate equivalent structures, considering the example from figure 6.5, subgraphs $A \rightarrow B \rightarrow C$, $A \leftarrow B \leftarrow C$ and $A \leftarrow B \rightarrow C$ are equivalent. Therefore the methods described in the analysis section estimated 13 arcs; when comparing this estimated network with the relationships found in the literature we were able to find 66 % true positives, 7 % false positives and a structure Hamming distance of 9, meaning that it would take 9 operations of adding, deleting and changing the direction of arrows to reach the true graph. Analysis of the experiment using the *search-and-score* (Heckerman 1995) method produced a graph with 16 arcs, and contained only 25% true positives and 56% false positives (the corresponding network graph can be found in the supplementary material) and a structure Hamming distance of 15. The true positives and the structure hamming distance are influenced by the fact that not all metabolites are assigned, absence of metabolites from the experiment and unknown relationships. Nevertheless, these figures still are useful to measure the performance of these techniques. Quantitatively analyzing techniques like these are common practice, but the real advantage lies in the graphical representation which is much more intuitive than standard statistical tests. From figure6.6 it can be seen that the enzymes involved in these pathways are generally known to oxidize certain fatty acids containing a *cis*, *cis-1*, *4-pentadine* structure. The main substrate is therefore *linoleic acid (C18:2)* and *linolenic acid (C18:3)* as shown in figure 6.6. The upper part in figure 6.6 which consist of *phospholipids*, *galactolipids* and *triacylglcerols* has not been taken into account in the experiment in question as these metabolites are large chemical structures and therefore are not volatile.

In principle the relationships (correlation and causal) generated, can be compared to the relationships found by Tikunov et al. ((Tikunov et al. 2005)). We consider here 13 metabolites *1-pentene-3-ol* , *1-penten-3-one*, *E-2-pentenal*, *1-pentanol* , *Z-2-penten-1-ol* , *Z-3-hexenal*, *hexanal*, *E-2-hexenal*, *Z-3-hexenol*, *1-hexanol*, *heptanal*, *E-2-heptenal*, *n-pentanal* which are involved in the substrate formation of free fatty acids *(lower section of figure 6.6)* of the oxylipin pathway. As the exact nature of these relationship is not known ((Yilmaz et al. 2001)) a plausible expla-

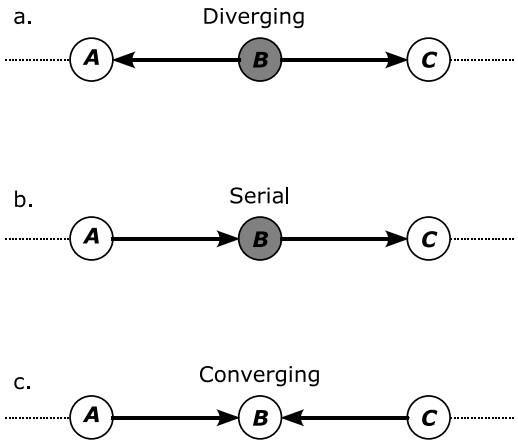


Figure 6.5: *a. B blocks (d-separates) A and C:* $\{A\} \perp\!\!\!\perp \{C\} \mid \{B\}$. *b. B blocks (d-separates) A and C:* $A \perp\!\!\!\perp \{C\} \mid \{B\}$. *c. B d-connects A and C:* $\{A\} \not\perp\!\!\!\perp \{C\} \mid \{B\}$. *(same holds for successors of B); note* $\{A\} \perp\!\!\!\perp \{C\} \mid \phi$

nation is still possible by looking at the chemical structures of these metabolites. Figure 7 indicates such a network constructed using Bayesian approach showing metabolites *1-pentene-3-ol*, *1-pentene-3-one*, *2-hexenal*, *E-2-pentenal*, *heptanal*, *E-2-heptanal* show significant causal relationships. *E-2-hexenal* is derived by isomerization of *Z-3-hexenal* ((Baldwin et al. 2000)) and the relationship of these two compounds cannot be observed in the graph; the reason for this could be absence of isomerization factor or less number of samples. A relationship between *1-penten-3-one* and *1-pentan-3-ol* also makes sense since the first is a dehydrogenation product of the second. From figure 7 correct relationships from *1-pentene 3-ol* $\rightarrow$ *1-pentene-3-one*, *E-2 pentenal* $\rightarrow$ *Z-2-penten-1-ol* and *1-pentanol* $\rightarrow$ *n-pentanal* can be deduced. There are certain relationships which may not make sense just by looking at them, e.g. *1-hexanol* $\rightarrow$ *Z-3-hexenol*, but the advantage is such relationships could be easily explained using Bayesian networks which may indicate present of hidden variables (latent variables) as not all the metabolites were observed in the experiment. Similarly, bi-directional arrows also indicate presence of such variables. To deduce exact relationships is difficult but the advantage lies in searching for equivalent relationships which can be easily deduced such as *Z-3-hexenal* $\rightarrow$ *Z-3-hexenol* can also be seen in figure 7.

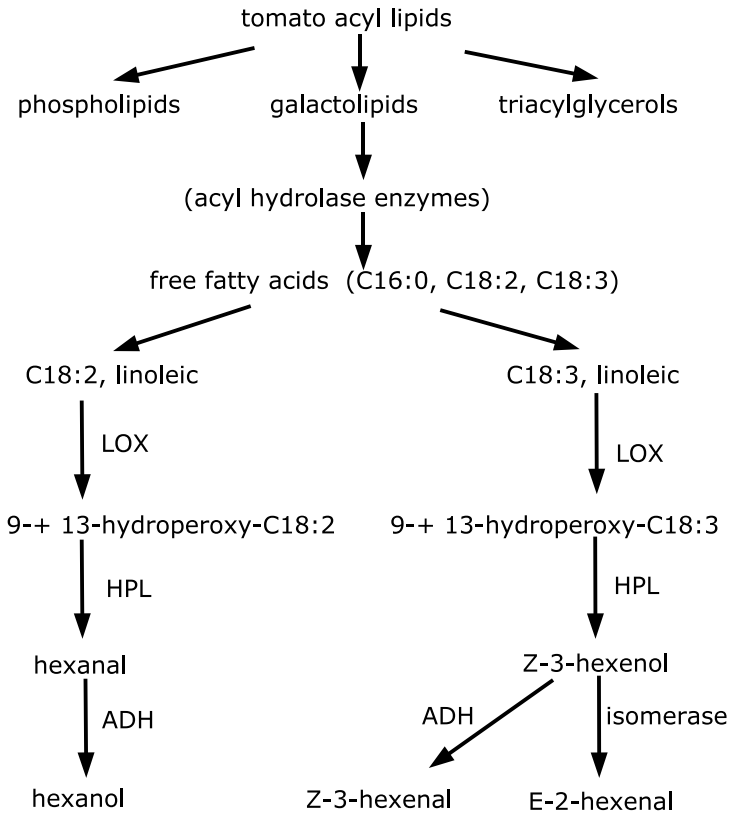


Figure 6.6: *Formation of lipid-derived volatiles through biosynthesis in oxylipin pathway.*

6.4 Concluding Remarks

Constructing graphically intuitive models has become a popular technique in metabolomics experiments ((Morgenthal et al. 2006, Beal et al. 2005)). Models like this allow us to understand the underlying biological processes involved in metabolic networks and reconstruct pathways. This knowledge is normally visualized by means of a directed graph, where the nodes of the graph

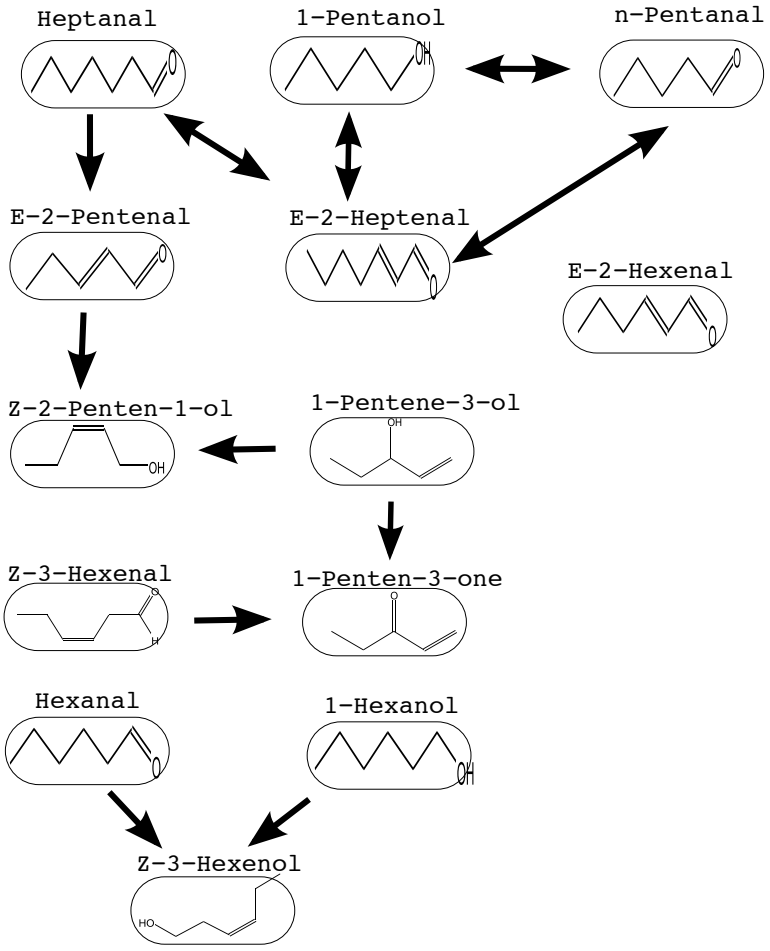


Figure 6.7: *Constructed Bayesian network for 13 plant-derived compounds for beef, round and cherry tomatoes $\alpha = 0.0001$.*

correspond to variables and arrows in the graph are used to express statistical dependence and independence information. The approach described in the present study proved useful to discover causal biochemical relationships in complex metabolomics data. The results are confirmed by previous observations on the same data as well as information found in literature. Methods such as Bayesian networks which are used for causal modeling of high-dimensional data are powerful tools in modeling of complex systems, since these approaches do take into account correlation methods before constructing an equivalent or exact causal relationship. Here we show how a Bayesian network can be used to analyze metabolomics data which is a powerful technique and

helps us to get indepth understanding of the biological process. This method can be exploited further by coupling it with pathway databases in order to get exact and more plausible information to understand the *process changes* at hand. As more and more data become available these methods can outperform classical statistical techniques and be used to find novel biochemical pathways. We have shown here how this approach can be used for exploratory data analysis in searching for causal relationships in metabolomics. The resulting hypothesis can then be used to form the basis of subsequent analysis which can learn from data, take prior inputs from molecular biologist and update probabilities in the light of ”new information” and/or ”data”.

Chapter 7

General discussion

Biological experiments often generate very large data sets. The purpose of these experiments is to try to understand the underlying biological process of a living organism. Various technologies are used to conduct experiments of this kind. This includes genomics, transcriptomics, metabolomics and proteomics and are generally referred as *-omics*. The outline of this thesis is to make sense of these data generated from *-omics* experiments to understand key biological processes involved in a problem domain. Computation techniques such as Bayesian networks provides a solid foundation in understanding these biological processes. Bayesian networks have played a major role in understanding biological problems (Friedman et al. 2000, Murphy 2002). Bayesian networks represent data in a graphically intuitive format where nodes represent variables and edges between nodes represent statistical dependence and independence information between them. In this thesis we have explored the applicability of this technique in other areas of high throughput technology. However, before these techniques can be used, data generated from experiments need to be managed. Therefore the aim of this thesis is to first construct a system which can be used to store and manage these data and later use Bayesian network techniques to analyze the data resulting from nutrigenomics and metabolomics experiments.

Microarrays are used to conduct nutrigenomics experiments. One of the most frequently encountered questions in the data analysis is the lack of a consistent framework to store, manage and analyze these data from various types of microarrays. There is a plethora of database and data mining systems to tackle these issues. However, there is not a single software system which addresses all the issues arising from different microarray platforms in a consistent way. Considerable amounts of information related to the experiment need to be recorded as well. In most laboratories analysis is performed using the R (R Development Core Team 2008) and Bioconductor (Gentleman et al. 2004) packages. This takes considerable amount of time and effort for the biologist to write their own scripts. These analysis procedures are well defined in Bioconductor, however, they require to be formulated together in the form of a pipeline. There are very few systems which support R and Bioconductor; Rosetta Resolver is one of them which supports R and Bioconductor fully and analysts can write their own analysis methods and implement them in the form of a pipeline. These pipelines are then implemented on a central server where users can use them. Chapter 1 gave an overview of the most popular systems in this domain and Chapter 3 focussed on MADMAX which is a specialized database to store and analyse data generated

from multiple microarray platforms. Current trend shows transcriptomics data becoming precursor for conducting metabolomics experiments to get the complete picture at systems (biological) level. The next step in this regard is to accommodate data generated from other *-omics* technologies as well. Due to the flexible architecture of MADMAX this can be easily extended for metabolomics as well. Every analysis of experiment generates new knowledge, this knowledge needs to be recorded. Various analysis performed on the same data generates new knowledge as well. Different type of biological processes can be understood only by exploiting this knowledge. To serve this purpose there is a need to develop a Comparative Nutrigenomics Database (*CND*) which would store information at *gene - disease*, *gene - nutrient* & *nutrient - disease* levels. The next step in this respect would be to do *data mining* on these information to establish correlation between experiments to get the complete picture at systems level.

Chapter 4 uses this technique and shows how simple classifiers viz. naïve Bayes classifiers (Mitchell 1997) can be used to find the metabolic state of an organism. The domain being nutrigenomics, the emphasis is to find subtle nutritional effect on gene expression data. Previous research in this area has been devoted to constructing gene regulatory networks based on search and score techniques (Friedman et al. 2000); and as the number of genes increase in size compared to their observations, approximate methods (Husmeier and Werhli 2007) are used. Nonetheless this previously conducted research shows good performance on cancer and tumor samples where the effects are strong compared to nutrition. Often studies like this are meant to identifying key pathways given a subset of genes, in chapter 4 we also explored the possibility by utilizing a reverse approach where genes were selected based on key pathways.

Feedback loops which represent gene regulating itself can be easily modeled using other forms of Bayesian networks, considerable work in this area is provided by Kevin Murphy (Murphy 2002). These techniques have also been modified and used efficiently by Zou et al (Zou and Conzen 2005). However not much research has been done to construct gene regulatory network using constraint based technique like the *PC* algorithm (Sprites et al. 2000). Chapter 4 explored this technique for its comparison with the standard naïve Bayes classifier and found that simple models are equally comparable to complex models like the PC algorithm. Other constraint based algorithms like *Grow-Shrink* (Margaritis May 2003), *Incremental Association* (Tsamardinos et al. 2003), *Fast Incremental Association* (Yaramakala and Margaritis 2005) and *Interleaved Incremental Association* (Yaramakala and Margaritis 2005) have not been explored so far in the area of nutrigenomics research.

Chapter 5 uses a constraint based technique to analyze gene expression data combined with clinical parameters. The focus was to find the dependence of gene expression on clinical parameters. The gene expression dataset and the clinical parameters from a smoking population were combined together. The focus of this research was to seek the dependence of certain genes on dna adducts and plasma cotinine levels (clinical parameters) of the samples. There is limited research done in this area in epidemiology and more research is done finding linear relations rather

than non-linear. Use of graphical models is rare. However recent trend shows Bayesian networks becoming popular in this domain as search and score methods are being used to find SNP information (Rodin et al. 2005). The work performed in chapter 5 is unique in two ways, first, gene expression dataset was combined together with clinical parameters and secondly, the constraint based technique based on PC algorithm was used. Further research on the use of graphical models is required in this area and the use of constraint based techniques should be considered.

Chapter 6 show how constraint based techniques based on PC algorithms can be used in a different domain like metabolomics. Mass spectrometry based experiments are being used as high throughput technologies become cheaper. These experiments generate more data than usual microarray experiments. Chapter 6 deals with samples were extracted from ripening tomatoes and the aim was to construct *Oxylipin* pathway which is a key process involving volatile compounds from tomatoes. The paper presented here show construction of of this metabolic network. Although more information on this specific pathway can be found by performing time series experiment at different stages of ripening of tomatoes. For this analysis dynamic bayesian networks (Murphy 2002) can then be used to construct a metabolic network identifying key metabolites which are responsible in the transition phase in between time points from raw to ripped tomatoes.

In this thesis Bayesian networks focussed on constraint based learning techniques have been applied to gene expression and metabolomics data for the study and discovery of biological networks. The reason for using these techniques is because biological pathway embeds hierarchy among metabolites, enzymes, genes and/or proteins. The probabilistic nature of the concentration of metabolites/genes requires one to design experiments with more observations to get relevant results, which is often a bottleneck. However effective analysis can still be performed by combining observations from multiple experiments focussing on biological question under similar conditions. Further research in this area should focus on incorporation of background knowledge from KEGG pathway database (Kanehisa et al. 2008) and be used to improve estimation of experimental results. KEGG should serve as a good source and has been widely used as a knowledge reference for the biological pathways and cellular processes. Various pathways are stored and represented at different levels of abstraction. Pathways are represented as graphs, where nodes represent molecules (protein, compound, enzymes etc) and edges represent relation types between these nodes, e.g. activation or phosphorylation. Often these edges represent catalytic activity by enzymes which are encoded by one or more genes depending on the environmental conditions an experiment is performed. Thus it becomes interesting to find out how the genes involved in a pathway provide information about the state of metabolites present in them. As these pathways are represented in graphical format gene network can be easily extracted; similarly compound and protein networks can be extracted as well and represented as a graph object. There are various uncertainties involved in metabolic pathways represented and various isoforms of certain enzymes are required for activation or phophorylation. These enzymes play key role and determine the fate of certain

metabolites which cannot be ignored. These relationships are very unique in analyzing data at pathway level. Bayesian networks however cannot take these relationships into account. To model such relationships *chain graph* (Lauritzen and Wermuth 1989) which can combine qualitative and quantitative information can be used. The knowledge from pathway databases can therefore be used to construct networks of various forms. The networks from pathway databases can be automatically curated using R which is an open source software environment for statistical computing and Bioconductor which is an open source environment built on R for analysis of genomic data; using package libraries such as *KEGGgraph* (Zhang and Wiemann 2008). The next step in this regard is to look at multiple pathways simultaneously to get a global view at systems level of an organism.

Summary

This thesis focuses on two aspects of high throughput technologies, i.e. data storage and data analysis, in particular in transcriptomics and metabolomics. Both technologies are part of a research field that is generally called *omics* (or '-omics', with a leading hyphen), which refers to genomics, transcriptomics, proteomics, or metabolomics. Although these techniques study different entities (genes, gene expression, proteins, or metabolites), they all have in common that they use high-throughput technologies such as microarrays and mass spectrometry, and thus generate huge amounts of data. Experiments conducted using these technologies allow one to compare different states of a living cell, for example a healthy cell versus a cancer cell or the effect of food on cell condition, and at different levels. The tools needed to apply omics technologies, in particular microarrays, are often manufactured by different vendors and require separate storage and analysis software for the data generated by them. Moreover experiments conducted using different technologies cannot be analyzed simultaneously to answer a biological question. Chapter 3 presents MADMAX, our software system which supports storage and analysis of data from multiple microarray platforms. It consists of a vendor-independent database which is tightly coupled with vendor-specific analysis tools. Upcoming technologies like metabolomics, proteomics and high-throughput sequencing can easily be incorporated in this system. Once the data are stored in this system, one obviously wants to deduce a biological relevant meaning from these data and here statistical and machine learning techniques play a key role. The aim of such analysis is to search for relationships between entities of interest, such as genes, metabolites or proteins. One of the major goals of these techniques is to search for causal relationships rather than mere correlations. It is often emphasized in the literature that "correlation is not causation" because people tend to jump to conclusions by making inferences about causal relationships when they actually only see correlations. Statistics are often good in finding these correlations; techniques called linear regression and analysis of variance form the core of applied multivariate statistics. However, these techniques cannot find causal relationships, neither are they able to incorporate prior knowledge of the biological domain. Graphical models, a machine learning technique, on the other hand do

not suffer from these limitations. Graphical models, a combination of graph theory, statistics and information science, are one of the most exciting things happening today in the field of machine learning applied to biological problems (see chapter 2 for a general introduction). This thesis deals with a special type of graphical models known as probabilistic graphical models, belief networks or Bayesian networks. The advantage of Bayesian networks over classical statistical techniques is that they allow the incorporation of background knowledge from a biological domain, and that analysis of data is intuitive as it is represented in the form of graphs (nodes and edges). Standard statistical techniques are good in describing the data but are not able to find non-linear relations whereas Bayesian networks allow future prediction and discovering nonlinear relations. Moreover, Bayesian networks allow hierarchical representation of data, which makes them particularly useful for representing biological data, since most biological processes are hierarchical by nature. Once we have such a causal graph made either by a computer program or constructed manually we can predict the effects of a certain entity by manipulating the state of other entities, or make backward inferences from effects to causes. Of course, if the graph is big, doing the necessary calculations can be very difficult and CPU-expensive, and in such cases approximate methods are used. Chapter 4 demonstrates the use of Bayesian networks to determine the metabolic state of feeding and fasting mice to determine the effect of a high fat diet on gene expression. This chapter also shows how selection of genes based on key biological processes generates more informative results than standard statistical tests. In chapter 5 the use of Bayesian networks is shown on the combination of gene expression data and clinical parameters, to determine the effect of smoking on gene expression and which genes are responsible for the DNA damage and the raise in plasma cotinine levels of blood of a smoking population. This study was conducted at Maastricht University where 22 twin smokers were profiled. Chapter 6 presents the reconstruction of a key metabolic pathway which plays an important role in ripening of tomatoes, thus showing the versatility of the use of Bayesian networks in metabolomics data analysis. The general trend in research shows a flood of data emerging from sequencing and metabolomics experiments. This means that to perform data mining on these data one requires intelligent techniques that are computationally feasible and able to take the knowledge of experts into account to generate relevant results. Graphical models fit this paradigm well and we expect them to play a key role in mining the data generated from *omics* experiments.

Samenvatting

Dit proefschrift belicht twee aspecten van 'high-throughput' technologien, namelijk de dataopslag en data-analyse, in het bijzonder in de transcriptomics en metabolomics. Beide technologien vormen een onderdeel van het onderzoeksveld dat *omics* (of "-omics", met streepje) genoemd wordt, wat een verzamelterm is voor genomics, transcriptomics, proteomics en metabolomics. Hoewel deze technologien verschillende objecten bestuderen (namelijk genen, genexpressie, eiwitten en metabolieten) gebruiken ze allen high-throughput technieken zoals microarrays en massaspectrometrie, en genereren derhalve massale hoeveelheden data. Experimenten die met dergelijke techniek uitgevoerd worden stellen ons in staat om de verschillende toestanden waarin een levende cel kan verkeren te bestuderen, bijvoorbeeld het verschil tussen een gezonde en een kanker cel, of het effect van voeding op de conditie van een cel. De hulpmiddelen die nodig zijn om omics technologie toe te kunnen passen, in het bijzonder de microarrays, worden door verscheidene bedrijven geproduceerd, en behoeven elk verschillende software pakketten voor opslag en analyse van de gegevens. Bovendien kunnen experimenten die met verschillende platforms uitgevoerd zijn niet tegelijkertijd geanalyseerd worden. Hoofdstuk 3 presenteert daarom MADMAX, een nieuw software systeem dat de opslag en analyse van meerdere microarrays platforms mogelijk maakt. Het bestaat uit een databank waarin de microarray gegevens onafhankelijk van de producent opgeslagen kunnen worden, gekoppeld aan array-specifieke analyse software. Daarnaast kan het systeem relatief eenvoudig uitgebreid worden voor nieuwe technologien zoals metabolomics, proteomics en high-throughput sequencing. Eenmaal opgeslagen in MADMAX wil de onderzoeker uiteraard een zinvolle biologische betekenis aan deze data toekennen, en hierbij spelen statistische en 'machine learning' technieken een centrale rol. Het doel van een dergelijke analyse is het zoeken naar relaties tussen de gegevens (genen, eiwitten en/of metabolieten), waarbij men meer (nog) in causale verbanden dan in correlaties genteresseerd is. In de wetenschappelijke literatuur wordt vaak onderstreept dat "correlatie geen oorzaak [is]", omdat men er toe neigt om oorzakelijke verbanden af te leiden terwijl men eigenlijk alleen correlaties ziet. De statistiek is goed in het vinden van dergelijke verbanden; technieken zoals lineaire regressie en variantie

analyse vormen de basis van de toegepaste multivariate statistiek. Deze technieken kunnen echter geen causale verbanden vinden, noch kunnen zij gebruik maken van bestaande biologische kennis. Graphical models, een machine learning techniek, kent daarentegen deze beperkingen niet. Graphical models, een combinatie van wiskunde (grafentheorie), statistiek en informatica, zijn een van de meest opwindende moderne ontwikkelingen op het gebied van de machine learning toegepast op biologische problemen (zie hoofdstuk 2 voor een algemene inleiding). Dit proefschrift behandelt een speciale klasse graphical models, die bekend staan als probabilistic graphical models, belief networks of Bayesiaanse netwerken (BN). Het voordeel van Bayesiaanse netwerken ten opzichte van de klassieke statistische technieken is dat zij het integreren van achtergrondkennis uit het biologische domein toestaan, en dat de analyse van de gegevens intutief is omdat deze in de vorm van grafen (netwerken) gepresenteerd worden. Standaard statistische technieken zijn goed in het beschrijven van de gegevens, maar zijn alleen in staat om lineaire verbanden te vinden, terwijl Bayesiaanse netwerken het mogelijk maken om voorspellingen te doen en niet-lineaire verbanden te ontdekken. Bovendien staan BN een hirarchische representatie van de data toe, wat ze uitermate geschikt maakt voor het representeren van biologische gegevens, omdat de meeste biologische processen hirarchisch van aard zijn. Wanneer we eenmaal een causaal netwerk gemaakt hebben (hetzij handmatig, hetzij via een computerprogramma) kunnen we de invloed van een object op het netwerk voorspellen door de status van andere objecten in het netwerk te veranderen, en kunnen de oorzaken van effecten afleiden. Wanneer het netwerk erg groot kunnen de benodigde berekeningen erg moeilijk en kostbaar wat betreft rekentijd worden, en in dat soort gevallen neemt men meestal de toevlucht tot benaderingsmethoden. Hoofdstuk 4 laat het gebruik van Bayesiaanse netwerken zien om de metabole staat van gevoerde en vastende muizen te bepalen, met als uiteindelijk doel om de invloed van een vetrijk dieet op de genexpressie te bepalen. Dit hoofdstuk laat ook zien hoe een selectie van genen gebaseerd op de voornaamste biologische processen meer informatieve resultaten geeft dan de standaard statistische tests. In hoofdstuk 5 worden BN gebruikt op een combinatie van genexpressie data (microarray data) en klinische parameters, om het effect van roken op de genexpressie te bepalen en na te gaan welke genen er verantwoordelijk zijn voor de DNA schade en de toename van cotinine in bloedplasma in een populatie van rokers. Dit onderzoek is uitgevoerd aan de Universiteit Maastricht, waarbij 22 tweelingparen bestudeerd werden. Hoofdstuk 6 toont de toepasbaarheid van BN in metabolomics aan de hand van de reconstructie van een metabool pad dat een hoofdrol speelt bij het rijpen van tomaten. De algemene trend in onderzoek is dat een vloedgolf aan gegevens afkomstig van sequencering en metabolomics experimenten beschikbaar komt. Dit betekent dat om op deze gegevens datamining toe te kunnen passen er slimme technieken nodig zijn die rekenkundig haalbaar zijn en die het mogelijk maken om de kennis van experts mee te nemen in de analyse. Graphical models passen uitstekend in dit model en wij verwachten dan ook dat ze een steeds belangrijker rol in het 'minen' van *omics* data zullen gaan spelen.

Bibliography

Aardema, M. J. and James T, M.: 2002, Toxicology and genetic toxicology in the new era of "toxicogenomics": impact of "-omics" technologies., *Mutat Res* **499**(1), 13–25.

Afman, L. and Mueller, M.: 2006, Nutrigenomics: from molecular nutrition to prevention of disease., *J Am Diet Assoc* **106**(4), 569–576.

Amann, J., Kalyankrishna, S., Massion, P. P., Ohm, J. E., Girard, L., Shigematsu, H., Peyton, M., Juroske, D., Huang, Y., Salmon, J. S., Kim, Y. H., Pollack, J. R., Yanagisawa, K., Gazdar, A., Minna, J. D., Kurie, J. M. and Carbone, D. P.: 2005, Aberrant epidermal growth factor receptor signaling and enhanced sensitivity to egfr inhibitors in lung cancer., *Cancer Res* **65**(1), 226–235.

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

Baldwin, E., Scott, J., Shewmaker, C. and Schuch, W.: 2000, Flavor trivia and tomato aroma: Biochemistry and possible mechanisms for control of important aroma components, *HortScience* **35**, 1013–1022.

Ball, C. A., Sherlock, G., Parkinson, H., Rocca-Sera, P., Brooksbank, C., Causton, H. C., Cavalieri, D., Gaasterland, T., Hingamp, P., Holstege, F., Ringwald, M., Spellman, P., Stoeckert, C. J., Stewart, J. E., Taylor, R., Brazma, A., Quackenbush, J. and Society, M. G. E. D. M.: 2002, Standards for microarray data., *Science* **298**(5593), 539.

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M. and Edgar, R.: 2007, Ncbi geo: mining tens of millions of expression profiles; database and tools update., *Nucleic Acids Res* **35**(Database issue), D760–D765.

Beal, M. J., Falciani, F., Ghahramani, Z., Rangel, C. and Wild, D. L.: 2005, A bayesian approach to reconstructing genetic regulatory networks with hidden factors., *Bioinformatics* **21**(3), 349–356.

Biosoftware, R.: 2000, Rosetta resolver system for gene expression data analysis., *Technical report*, Rosetta.

Bolstad, B., Collin, F., Brettschneider, J., Simpson, K., Cope, L., Irizarry, R. and Speed, T.: 2005, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.*, Statistics for Biology and Health, Springer, chapter 3: Quality Assessment of Affymetrix GeneChip Data., pp. 33–47.

Bonassi, S. and William, W.: 2002, Biomarkers in molecular epidemiology studies for health risk prediction., *Mutat Res* **511**(1), 73–86.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. and Vingron, M.: 2001, Minimum information about a microarray experiment (miame)-toward standards for microarray data., *Nat Genet* **29**(4), 365–371.

Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M. and Haussler, D.: 2000, Knowledge-based analysis of microarray gene expression data by using support vector machines., *Proc Natl Acad Sci U S A* **97**(1), 262–267.

Burgarella, S., Cattaneo, D., Pinciroli, F. and Masseroli, M.: 2005, Microgen: a miame compliant web system for microarray experiment information and workflow management., *BMC Bioinformatics* **6 Suppl 4**, S6.

Cooper, G. F.: 1999, *An overview of the representation and discovery of causal relationships using Bayesian networks*., MIT Press.

Cunningham, S. J.: 1995, Machine learning and statistics: A matter of perspective, *Technical report*.

Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., Bunney, W. E., Myers, R. M., Speed, T. P., Akil, H., Watson, S. J. and Meng, F.: 2005, Evolving gene/transcript definitions significantly alter the interpretation of genechip data., *Nucleic Acids Res* **33**(20), e175.

Davis, A. P., Murphy, C. G., Rosenstein, M. C., Wiegers, T. C. and Mattingly, C. J.: 2008, The comparative toxicogenomics database facilitates identification and understanding of chemical-gene-disease associations: arsenic as a case study., *BMC Med Genomics* **1**, 48.

de Laplace, P.: 1812, *Théorie analytique de Probabilités*, Paris.

Dellaert, F.: 2002, The expectation maximization algorithm, *Technical report*, College of Computing, Georgia Institute of Technology.

Demeter, J., Beauheim, C., Gollub, J., Hernandez-Boussard, T., Jin, H., Maier, D., Matese, J. C., Nitzberg, M., Wymore, F., Zachariah, Z. K., Brown, P. O., Sherlock, G. and Ball, C. A.: 2007, The stanford microarray database: implementation of new analysis tools and open source release of software., *Nucleic Acids Res* **35**(Database issue), D766–D770.

Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., Georgescu, C. and Romero, R.: 2007, A systems biology approach for pathway level analysis., *Genome Res* **17**(10), 1537–1545.

Dudoit, S. and Ge, Y.: 2005, Bioconductors multtest package, Bioconductor Documentation.

Dysvik, B. and Jonassen, I.: 2001, J-express: exploring gene expression data using java., *Bioinformatics* **17**(4), 369–370.

Edgar, R., Domrachev, M. and Lash, A. E.: 2002, Gene expression omnibus: Ncbi gene expression and hybridization array data repository., *Nucleic Acids Res* **30**(1), 207–210.

Eiceman, G. and Karpas, Z.: 2005, *Ion Mobility Spectrometry*, CRC Press.

Elidan, G. and Friedman, N.: 2003, The information bottleneck em algorithm, *In Proceedings of UAI*, Morgan Kaufmann, pp. 200–208.

Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R. N. and Willmitzer, L.: 2000, Metabolite profiling for plant functional genomics., *Nat Biotechnol* **18**(11), 1157–1161.

Flora, S. D., D'Agostini, F., Balansky, R., Camoirano, A., Bennicelli, C., Bagnasco, M., Cartiglia, C., Tampa, E., Longobardi, M. G., Lubet, R. A. and Izzotti, A.: 2003, Modulation of cigarette smoke-related end-points in mutagenesis and carcinogenesis., *Mutat Res* **523-524**, 237–252.

Fogg-Johnson, N. and Merolli, A.: 2000, Nutrigenomics: the next wave in nutrition research, *Nutraceuticals World* **3**, 86–95.

Forrest, M., Lan, Q., Hubbard, A., Zhang, L., Vermeulen, R., Zhao, X., Li, G., Wu, Y., Shen, M., Yin, S., Chanock, S., Rothman, N. and Smith, M.: 2005, Discovery of novel biomarkers by microarray analysis of peripheral blood monocuclear cell gene expression in benzene-exposed workers., *Environ Health Perspect* **113**, 801–7.

Friedman, N., Linial, M., Nachman, I. and Pe'er, D.: 2000, Using bayesian networks to analyze expression data., *J Comput Biol* **7**(3-4), 601–620.

Gardiner-Garden, M. and Littlejohn, T. G.: 2001, A comparison of microarray databases., *Brief Bioinform* **2**(2), 143–158.

Gautier, L., Cope, L., Bolstad, B. M. and Irizarry, R. A.: 2004, affy–analysis of affymetrix genechip data at the probe level., *Bioinformatics* **20**(3), 307–315.

Gavai, A. K., Tikunov, Y., Ursem, R., Bovy, A., van Eeuwijk, F., Nijveen, H., Lucas, P. J. F. and Leunissen, J. A. M.: 2009, Constraint-based probabilistic learning of metabolic pathways from tomato volatiles. Under Review.

Gentleman, R. C., Carey, V., Bates, D. M., Bolstad, B., MarcelDettling, Dudoit, S., andLaurent Gautierand Yongchao Ge, B. E., Gentry, J., Hornik, K., TorstenHothorn, Huber, W., Iacus, S., Leisch, R. F., Li, C., Maechler, M., Sawitzki, A. J. R. G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H. and Zhang, J.: 2004, Bioconductor: Open software development for computational biology and bioinformatics, *Genome Biology* **5**, R80.

Gentzsch, W.: 2001, Sun grid engine: Towards creating a compute power grid, *Proceedings of the 1st International Symposium on Cluster Computing and the Grid*.

Godschalk, R. W., Maas, L. M., Zandwijk, N. V., van 't Veer, L. J., Breedijk, A., Borm, P. J., Verhaert, J., Kleinjans, J. C. and van Schooten, F. J.: 1998, Differences in aromatic-dna adduct levels between alveolar macrophages and subpopulations of white blood cells from smokers., *Carcinogenesis* **19**(5), 819–825.

Goeman, J. J., van de Geer, S. A., de Kort, F. and van Houwelingen, H. C.: 2004, A global test for groups of genes: testing association with a clinical outcome., *Bioinformatics* **20**(1), 93–99.

Gohlke, R.: 1959, Time-of-flight mass spectrometry and gas-liquid partition chromatography, *Anal. Chem* **31**, 535–41.

Hamadeh, H. K., Amin, R. P., Paules, R. S. and Afshari, C. A.: 2002, An overview of toxicogenomics., *Curr Issues Mol Biol* **4**(2), 45–56.

Hancock, D., Wilson, M., Velarde, G., Morrison, N., Hayes, A., Hulme, H., Wood, A. J., Nashar, K., Kell, D. B. and Brass, A.: 2005, maxdload2 and maxdbrowse: standards-compliant tools for microarray experimental annotation, data management and dissemination., *BMC Bioinformatics* **6**, 264.

Heber, S. and Sick, B.: 2006, Quality assessment of affymetrix genechip data., *OMICS* **10**(3), 358–368.

Heckerman, D.: 1995, A tutorial on learning with bayesian networks, *Technical report*, Microsoft Research.

Herrero, J., Al-Shahrour, F., Diaz-Uriarte, R., Mateos, A., Vaquerizas, J. M., Santoyo, J. and Dopazo, J.: 2003, Gepas: A web-based resource for microarray gene expression data analysis., *Nucleic Acids Res* **31**(13), 3461–3467.

Heyer, L. J., Moskowitz, D. Z., Abele, J. A., Karnik, P., Choi, D., Campbell, A. M., Oldham, E. E. and Akin, B. K.: 2005, Magic tool: integrated microarray data analysis., *Bioinformatics* **21**(9), 2114–2115.

Husmeier, D., Dybowski, R. and Roberts, S.: 2005, *Probabilistic modeling in bioinformatics and medical informatics*, New York: Springer.

Husmeier, D. and Werhli, A. V.: 2007, Bayesian integration of biological prior knowledge into the reconstruction of gene regulatory networks with bayesian networks., *Comput Syst Bioinformatics Conf* **6**, 85–95.

Ihaka, R. and Gentleman, R.: 1996, R: A language for data analysis and graphics, *Journal of Computational and Graphical Statistics* **5**(3), 299–314.

Izzotti, A., Balansky, R. M., Cartiglia, C., Camoirano, A., Longobardi, M. and Flora, S. D.: 2003, Genomic and transcriptional alterations in mouse fetus liver after transplacental exposure to cigarette smoke., *FASEB J* **17**(9), 1127–1129.

Izzotti, A., Cartiglia, C., Longobardi, M., Balansky, R. M., D'Agostini, F., Lubet, R. A. and Flora, S. D.: 2004, Alterations of gene expression in skin and lung of mice exposed to light and cigarette smoke., *FASEB J* **18**(13), 1559–1561.

Jordan, M. I.: 2004, Graphical models, *Statistical Science* **19**, 140–155.

Kalisch, M. and Buhlmann, P.: 2007, Estimating high-dimentional directed acyclic graphs with the pc-algorithm, *Journal of Machine Learning Research* **8**, 613–636.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. and Yamanishi, Y.: 2008, *KEGG* for linking genomes to life and the environment., *Nucleic Acids Res* **36**(Database issue), D480–D484.

Kapushesky, M., Kemmeren, P., Culhane, A. C., Durinck, S., Ihmels, J., Korner, C., Kull, M., Torrente, A., Sarkans, U., Vilo, J. and Brazma, A.: 2004, Expression profiler: next generation–an online platform for analysis of microarray data., *Nucleic Acids Res* **32**(Web Server issue), W465–W470.

Kaput, J. and Rodriguez, R. L.: 2004, Nutritional genomics: the next frontier in the postgenomic era., *Physiol Genomics* **16**(2), 166–177.

Kervizic, G. and Corcos, L.: 2008, Dynamical modeling of the cholesterol regulatory pathway with boolean networks., *BMC Syst Biol* **2**(1), 99.

Kolmogorov, A.: 1956, *Foundations of the Theory of Probability*, Chesea, New York. 2nd edition of English translation.

Kopka, J., Fernie, A., Weckwerth, W., Gibon, Y. and Stitt, M.: 2004, Metabolite profiling in plant biology: platforms and destinations., *Genome Biol* **5**(6), 109.

Lampe, J. W., Stepaniants, S. B., Mao, M., Radich, J. P., Dai, H., Linsley, P. S., Friend, S. H. and Potter, J. D.: 2004, Signatures of environmental exposures using peripheral leukocyte gene expression: tobacco smoke., *Cancer Epidemiol Biomarkers Prev* **13**(3), 445–453.

Lauritzen, S. and Wermuth, N.: 1989, Graphical models for association between variables, some of which are qualitative and some quantitative., *Annals of Statistics,* **17**, 31–57.

Lee, H. K., Braynen, W., Keshav, K. and Pavlidis, P.: 2005, Erminej: tool for functional analysis of gene expression data sets., *BMC Bioinformatics* **6**, 269.

Liang, L. R., Lu, S., Wang, X., Lu, Y., Mandal, V., Patacsil, D. and Kumar, D.: 2006, Fm-test: a fuzzy-set-theory-based approach to differential gene expression data analysis., *BMC Bioinformatics* **7 Suppl 4**, S7.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E. L.: 1996, Expression monitoring by hybridization to high-density oligonucleotide arrays., *Nat Biotechnol* **14**(13), 1675–1680.

Margaritis, D.: May 2003, *Learning Bayesian Network Model Structure from Data*., PhD thesis, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA.

Marzolf, B., Deutsch, E. W., Moss, P., Campbell, D., Johnson, M. H. and Galitski, T.: 2006, Sbeams-microarray: database software supporting genomic expression analyses for systems biology., *BMC Bioinformatics* **7**, 286.

Maurer, M., Molidor, R., Sturn, A., Hartler, J., Hackl, H., Stocker, G., Prokesch, A., Scheideler, M. and Trajanoski, Z.: 2005, Mars: microarray analysis, retrieval, and storage system., *BMC Bioinformatics* **6**, 101.

Meek, C.: 1995b, Causal inference and causal explanation with background knowledge., *Uncertainty in Artificial Intelligence* **11**, 403–410.

Mitchell, T.: 1997, *Machine Learning*, McGraw-Hill Companies.

Mitchell, T. M.: 2006, The discipline of machine learning. CMU-ML-06-108.

Moco, S., Bino, R. J. and et al., O. V.: 2006, A liquid chromatography-mass spectrometry-based metabolome database for tomato., *Plant Physiol* **141**(4), 1205–1218.

Morgenthal, K., Weckwerth, W. and Steuer, R.: 2006, Metabolomic networks in plants: Transitions from pattern recognition to biological interpretation., *Biosystems* **83**(2-3), 108–117.

Mueller, M. and Kersten, S.: 2003, Nutrigenomics: goals and strategies., *Nat Rev Genet* **4**(4), 315–322.

Murphy, K. P.: 2002, Dynamic bayesian networks.

Navarange, M., Game, L., Fowler, D., Wadekar, V., Banks, H., Cooley, N., Rahman, F., Hinshelwood, J., Broderick, P. and Causton, H. C.: 2005, Mimir: a comprehensive solution for storage, annotation and exchange of microarray data., *BMC Bioinformatics* **6**, 268.

Neerincx, P. B. T. and Leunissen, J. A. M.: 2005, Evolution of web services in bioinformatics., *Brief Bioinform* **6**(2), 178–188.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M.: 1999, Kegg: Kyoto encyclopedia of genes and genomes., *Nucleic Acids Res* **27**(1), 29–34.

Okouoyo, S., Herzer, K., Ucur, E., Mattern, J., Krammer, P. H., Debatin, K.-M. and Herr, I.: 2004, Rescue of death receptor and mitochondrial apoptosis signaling in resistant human nsclc in vivo., *Int J Cancer* **108**(4), 580–587.

Opgen-Rhein, R. and Strimmer, K.: 2007, From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data., *BMC Syst Biol* **1**, 37.

Oracle: 2008, Oracle application express and oracle real application clusters creating a highly available environment for apex applications, White Paper.

Ordovas, J. M. and Mooser, V.: 2004, Nutrigenomics and nutrigenetics., *Curr Opin Lipidol* **15**(2), 101–108.

Owen, O. E., Reichard, G. A., Patel, M. S. and Boden, G.: 1979, Energy metabolism in feasting and fasting., *Adv Exp Med Biol* **111**, 169–188.

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans, U. and Brazma, A.: 2007, Arrayexpress–a public database of microarray experiments and gene expression profiles., *Nucleic Acids Res* **35**(Database issue), D747–D750.

Parman, C. and Halling., C.: 2006, A package to generate qc reports, *Bioconductor Documentation* .

Pearl, J.: 1988, *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Morgan Kaufmann.

Quackenbush, J.: 2006, Microarray analysis and tumor classification., *N Engl J Med* **354**(23), 2463–2472.

Quanz, B., Park, M. and Huan, J.: 2008, Biological pathways as features for microarray data classification, *Proceeding of the 2nd international workshop on Data and text mining in bioinformatics*.

R Development Core Team: 2008, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rayner, T., Rocca-Serra, P., Spellman, P., Causton, H., Farne, A., Holloway, E., Irizarry, R., Liu, J., Maier, D., Miller, M., Petersen, K., Quackenbush, J., Sherlock, G., Stoeckert, C., White, J., Whetzel, P., Wymore, F., Parkinson, H., Sarkans, U., Ball, C. and Brazma, A.: 2006, A simple spreadsheet-based, miame-supportive format for microarray data: Mage-tab., *BMC Bioinformatics* **7**(1), 489.

Reddy, M. V. and Randerath, K.: 1986, Nuclease p1-mediated enhancement of sensitivity of 32p-postlabeling test for structurally diverse dna adducts., *Carcinogenesis* **7**(9), 1543–1551.

Rocca-Serra, P., the European Nutrigenomics Organization (NuGO) and the RSBI Nutrigenomics working group: 2004, Nutrigenomics: Minimal reporting requirements.

Rodin, A., Mosley, T. H., Clark, A. G., Sing, C. F. and Boerwinkle, E.: 2005, Mining genetic epidemiology data with bayesian networks application to apoe gene variation and plasma lipid levels., *J Comput Biol* **12**(1), 1–11.

Saal, L. H., Troein, C., Vallon-Christersson, J., Gruvberger, S., Borg, A. and Peterson, C.: 2002, Bioarray software environment (base): a platform for comprehensive management and analysis of microarray data., *Genome Biol* **3**(8), SOFTWARE0003.

Sansone, S.-A., Rocca-Serra, P., Tong, W., Fostel, J., Morrison, N., Jones, A. R. and Members, R. S. B. I.: 2006, A strategy capitalizing on synergies: the reporting structure for biological investigation (rsbi) working group., *OMICS* **10**(2), 164–171.

Sartor, M. A., Tomlinson, C. R., Wesselkamper, S. C., Sivaganesan, S., Leikauf, G. D. and Medvedovic, M.: 2006, Intensity-based hierarchical bayes method improves testing for differentially expressed genes in microarray experiments., *BMC Bioinformatics* **7**, 538.

Schauer, N., Semel, Y. and et al., U. R.: 2006, Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement., *Nat Biotechnol* **24**(4), 447–454.

Schena, M., Shalon, D., Davis, R. W. and Brown, P. O.: 1995, Quantitative monitoring of gene expression patterns with a complementary dna microarray., *Science* **270**(5235), 467–470.

Schooten, F. J. V., Hirvonen, A., Maas, L. M., Mol, B. A. D., Kleinjans, J. C., Bell, D. A. and Durrer, J. D.: 1998, Putative susceptibility markers of coronary artery disease: association between vdr genotype, smoking, and aromatic dna adduct levels in human right atrial tissue., *FASEB J* **12**(13), 1409–1417.

Scutari, M.: 2008, *bnlearn: Bayesian network structure learning*. R package version 1.0.

Segal, E., Friedman, N., Kaminski, N., Regev, A. and Koller, D.: 2005, From signatures to models: understanding cancer using microarrays., *Nat Genet* **37 Suppl**, S38–S45.

Smith, C. J., Perfetti, T. A., Garg, R. and Hansch, C.: 2003, Iarc carcinogens reported in cigarette mainstream smoke and their calculated log p values., *Food Chem Toxicol* **41**(6), 807–817.

Smyth, G. K.: 2004, Linear models and empirical bayes methods for assessing differential expression in microarray experiments., *Stat Appl Genet Mol Biol* **3**, Article3.

Sprites, P., Glymour, C. and Scheines, R.: 2000, *Causation, Prediction and Search*, The MIT Press.

Storey, J. D. and Tibshirani, R.: 2003, Statistical significance for genomewide studies., *Proc Natl Acad Sci U S A* **100**(16), 9440–9445.

Stuart, J. M., Segal, E., Koller, D. and Kim, S. K.: 2003, A gene-coexpression network for global discovery of conserved genetic modules., *Science* **302**(5643), 249–255.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. and Mesirov, J. P.: 2005, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles., *Proc Natl Acad Sci U S A* **102**(43), 15545–15550.

Suizdak, G.: 2003, *The Expanding Role of Mass Spectrometry in Biotechnology*, MCC Press, San Diego, CA.

Tai, F. and Pan, W.: 2007, Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data., *Bioinformatics* **23**(23), 3170–3177.

Theilhaber, J., Ulyanov, A., Malanthara, A., Cole, J., Xu, D., Nahf, R., Heuer, M., Brockel, C. and Bushnell, S.: 2004, Gecko: a complete large-scale gene expression analysis platform., *BMC Bioinformatics* **5**, 195.

Tikunov, Y., Lommen, A., de Vos, C. H. R., Verhoeven, H. A., Bino, R. J., Hall, R. D. and Bovy, A. G.: 2005, A novel approach for nontargeted data analysis for metabolomics. large-scale profiling of tomato fruit volatiles., *Plant Physiol* **139**(3), 1125–1137.

Tomfohr, J., Lu, J. and Kepler, T. B.: 2005, Pathway level analysis of gene expression using singular value decomposition., *BMC Bioinformatics* **6**, 225.

Toraason, M., Albertini, R., Bayard, S., Bigbee, W., Blair, A., Boffetta, P., Bonassi, S., Chanock, S., Christiani, D., Eastmond, D., Hanash, S., Henry, C., Kadlubar, F., Mirer, F., Nebert, D., Rapport, S., Rest, K., Rothman, N., Ruder, A., Savage, R., Schulte, P., Siemiatycki, J., Shields, P., Smith, M., Tolbert, P., Vermeulen, R., Vineis, P., Wacholder, S., Ward, E., Waters, M. and Weston, A.: 2004, Applying new biotechnologies to the study of occupational cancer–a workshop summary., *Environ Health Perspect* **112**(4), 413–416.

Toure, A. and Basu, M.: 2001, Application of neural network to gene expression data for cancerclassification, *Neural Networks*, Vol. 1, pp. 583 – 587.

Tsamardinos, I., Aliferis, C. F. and Statnikov, A.: 2003, Algorithms for large scale markov blanket discovery., *In Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference*, pp. 376–381.

Tsamardinos, I., Brown, L. E. and Aliferis., C. F.: 2006, The max-min hill-climbing bayesian network structure learning algorithm., *Machine Learning* **65**, 31–78.

Ursem, R., Tikunov, Y., Bovy, A., van Berloo, R. and van Eeuwijk, F.: 2008, A correlation network approach to metabolic data analysis for tomato fruits, *Euphytica* **161**, 181–193.

van Delft, J. H., Baan, R. A. and Roza, L.: 1998, Biological effect markers for exposure to carcinogenic compound and their relevance for risk assessment., *Crit Rev Toxicol* **28**(5), 477–510.

van den Berghe, G.: 1991, The role of the liver in metabolic homeostasis: implications for inborn errors of metabolism., *J Inherit Metab Dis* **14**(4), 407–420.

van Leeuwen, D. M., van Agen, E., Gottschalk, R. W. H., Vlietinck, R., Gielen, M., van Herwijnen, M. H. M., Maas, L. M., Kleinjans, J. C. S. and van Delft, J. H. M.: 2007, Cigarette smoke-induced differential gene expression in blood cells from monozygotic twin pairs., *Carcinogenesis* **28**(3), 691–697.

von Heydebreck, A., Huber, W. and Gentleman, R.: 2004, Differential expression with the bioconductor project., *Bioconductor Project Working Papers* **Working Paper 7**.

Vunakis, H. V., Gijka, H. B. and Langone, J. J.: 1993, Radioimmunoassay for nicotine and cotinine., *IARC Sci Publ* (109), 293–299.

Wang, Z., Neuburg, D., Li, C., Su, L., Kim, J. Y., Chen, J. C. and Christiani, D. C.: 2005, Global gene expression profiling in whole-blood samples from individuals exposed to metal fumes., *Environ Health Perspect* **113**(2), 233–241.

Weckwerth, W.: 2003, Metabolomics in systems biology., *Annu Rev Plant Biol* **54**, 669–689.

Whittaker, J.: 1990, *Graphical Models in Applied Multivariate Statistics*, Wiley.

Wilson, C. L. and Miller, C. J.: 2005, Simpleaffy: a bioconductor package for affymetrix quality control and data analysis., *Bioinformatics* **21**(18), 3683–3685.

Wu, M., Chiou, H., Ho, I., Chen, C. and Lee, T.: 2003, Gene expression of inflammatory molecules in circula ting lymphocytes from arsenic-exposed human subjects., *Environ. Health Perspect* **111**, 1429–38.

Wu, Z., Irizarry, R. A., Gentleman, R., Murillo, F. M. and Spencer., F.: 2004, A model based background adjustment for oligonucleotide expression arrays, *Journal of the American Statistical Association* **99**, 909–917.

Yaramakala, S. and Margaritis, D.: 2005, Speculative markov blanket discovery for optimal feature selection., *In Proceedings of the Fifth IEEE International Conference on Data Mining*.

Yilmaz, E.: 2001, Oxylipin pathway in the biosynthesis of fresh tomato volatiles, *Turk J Biol* **25**, 351–360.

Yilmaz, E., Tandon, K. S., Scott, J. W., Baldwin, E. A. and Shewfelt, R. L.: 2001, Absence of a clear relationship between lipid pathway enzymes and volatile compounds in fresh tomatoes, *Plant Physiol* **158**, 1111–1116.

Zhang, J. D. and Wiemann, S.: 2008, *KEGGgraph: KEGGgraph: A graph approach to KEGG PATHWAY in R and Bioconductor*. R package version 0.8.7.

Zou, M. and Conzen, S. D.: 2005, A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data., *Bioinformatics* **21**(1), 71–79.

List of Symbols and Abbreviations

- $\perp\!\!\!\perp$: Independent.
- $\perp\!\!\!\perp_P$: Independent with respect to distribution P.
- $\not\perp\!\!\!\perp$: Dependent.
- $\not\perp\!\!\!\perp_P$: Dependent with respect to distribution P.
- $\perp$: d-separated.
- $\perp_G$: d-separated in graph G.
- $\not\perp$: d-connected.
- $\not\perp_G$: d-connected in graph G.
- $\rightarrow$: connected by directed link.
- $-$: connected by undirected link.
- $|$: "given" (e.g. $a \mid b$ means *a given b*)
- V : Set of nodes of a model.
- E : Set of edges of a model.
- TP : True Positives.
- FP : False Positives.
- TN : True Negatives.
- FN : False Negatives.
- $\frac{TP}{TP+FP}$: (*Precision*) the percentage of positive predictions that are correct.
- $\frac{TP}{TP+FN}$: (*Recall or Sensitivity*) the percentage of positive labeled instances that were predicted as positive.
- $\frac{TN}{TN+FP}$: (*Specificity*) the percentage of negative labeled instances that were predicted as negative.
- $\frac{TP+TN}{TP+TN+FP+FN}$: (*Accuracy*) the percentage of predictions that are correct.

- *Confusion Matrix* : A 2×2 table showing number/proportion of examples from one class classified in to another (or same) class.
- $\in$: *an element of* or *in* or *belongs to*.
- $u \perp_G v$: $u \in V$ and $v \in V$ are d-separated in graph $G = (V, E)$.
- $U \perp_G V$: Each $u \in V$ and each $v \in V$ are d-separated in graph G or U and V are d-separated in G.
- $X \perp\!\!\!\perp_P Y$: X and Y *marginally* independent with respect to probability distribution P.
- $X \perp\!\!\!\perp Y$: X and Y *marginally* independent probability distribution P understood from the context.
- $X \perp\!\!\!\perp Y \mid S$: X and Y *conditionally independent* given S .
- $X \not\perp\!\!\!\perp Y \mid S$: X and Y *conditionally dependent* given S .

Ambiguous terminologies

Glossary	
Machine Learning	**Statistics**
Network, Graphs	Model
Weights	Parameters
Learning	Fitting
Confusion Matrix	Contingency table
Generalization	Test set performance
Supervised learning	Regression/Classification
Unsupervised learning	Density estimation, Clustering
Attributes	Variables
Instances	Observations
Cause	Independent variables,Covariates
Effect	Dependent variables

List of R packages used for Graphical modeling

- SNA : Fully documented collection of R routines for social network analysis
- BNArray : Constructing gene regulatory networks from microarray data by using Bayesian network
- G1DBN : Dynamic Bayesian networks for time series data analysis
- Deal : Learning score based Bayesian networks
- gR : Graphical modeling initiative
- Rgraphviz : Interface to Graphviz engine
- igraph : Library to create and manipulate directed and undirected graphs
- RBGL : A fairly extensive and comprehensive interface to the graph algorithms contained in the BOOST library
- graph : A package to handle graph data structures
- Grappa : Junction tree implementation for inference in Bayesian network
- bnlearn : Learning constraint based Bayesian networks
- mimR : R interface to MIM package
- ggm : Functions of fitting Gaussian Markov models
- GeneNet : Learning dependency network from genomic data
- RWeka : R interface data mining and Bayesian network construction for Weka
- pcalg : Constraint based learning using PC algorithm
- network : A Package for Managing Relational Data in R
- gRain : Graphical Independence Network

Acknowledgements

No words can convey, how much I appreciate the support and help extended to me in all these years of my MSc and PhD to the inhabitants of Wageningen and the people of Netherlands in general.

The first and foremost I would like to thank to my promoter Prof. *Jack A.M. Leunissen*, it was a rare opportunity for me to learn as much as I could. The most important thing I learned from him was to keep things in its simplest form. I still keep seeking his guidance and suggestion in my professional career and will keep doing so. I am thankful to him for introducing me to various technologies and his indepth knowledge in the field of Bioinformatics. I really appreciate his last minute effort during finalization of this thesis.

Prof. *Michael Müller*, I thank you for putting lot of faith in me and giving me ample opportunities to explore things out. I also remember the NuGO conferences I had been visiting and enjoying the evening sessions with you. I thank you a lot for this and will always remember the discussions we had in these meetings.

Asst. Prof. *Guido Hooiveld*, it was a nice experience that I have met you and got the opportunity to closely work with you. I appreciate the time you spend with me explaining biological processes involved in experiments. Thankyou for being patient!. You had also been a guiding force in my professional work as well as dealing with the immigration department and your help had been instrumental throughout these years.

Prof. *Peter J.F. Lucas*, I have to admit if you were not there probably it would have been difficult for me to finish this project. I have to say I have never met someone with the knowledge you have on graphical models and at the same time able to explain it in its simplistic form. I consider myself to be lucky and fortunate to have worked with you and if need arise would like to do so in future as well. Working with you was an eye opening experience in every aspect of this research area. Though I could meet you only once a week we that was sufficient to get the most from it.

I would also like to thank Prof. *Terry Speed* from Berkeley, California, for being instrumental in explaining me the basics of Bayesian networks when I met him for the first time in NuGo Week in italy and successively happen to see him in other conferences as well. Thanks to Prof. *Fred van Eeuwijk* and *Cajo ter Baak* for all the inputs and feedback you had been providing me throughout my project. I would also like to thank *Ben van ommen* and *Chris Evelo* for giving suggestions and advice on my research on all the NuGO meetings and conferences we met. All this work would not have been possible without the excellent administrative support from *Marie Jensen*, *Marie-Jose*, *Lidwien vander Heyden*,*Gea Brussen* and

Lous Dyom. Thanks *Lidwien* for being available throughout whenever I had questions :-). Thanks *Sander Kersten* for helping me on one project, though we hardly spend much time but I learned a great deal from your vast knowledge on PPAR's.

When I first came to netherlands, I had gone through a lot of cultural shocks; of-course taking some and giving some :-) and survived and enjoyed every single moment of it. To start with I like to thank *Albert Ballast*, my former corridor mate for introducing me to wageningen and guiding me. I still enjoy evening discussion with him. I would like to also thank *Monique van der Wind*, who not only helped me in going through all the administrative procedures but also being a very close friend over the past six years. I thank *Monique* for all the help and support she has provided me over these years and I cherish our friendship being intact. *Maarten Paul* and *Rachel* are friends who were always there for me, and I enjoyed fishing and bbq'ing with almost every weekend in summer. And I will not forget those friday evening we spend drinking beer in front of the fire in your garden or in the Vlaam.

My utmost thanks to *Harm Nijveen*, for being always open for discussion, and I thank him for being my *babble fish* when it came to translation. His open minded nature and availability throughout has helped me tremendously in understanding many biochemistry concepts. I am indebt to him for translating the summary of my thesis in dutch.

I would also like to thank my squash partners *Pieter Neerincx* and *Mark Bouwens* for being available for squash and giving me a break from work once every week. I look forward to play with you again whenever we have the opportunity. I also remember my nice skiing experience in Germany with *Mechteld*, *Menno* and *Eric Peters*, it was fun to learn skiing and getting damaged, thanks for the nice experience!!!. I owe all my skiing techniques to *Eric* and *Menno* thanks. *Philip de groot* and *Lin ke* it was great knowing you and working with you, we formed a great team during the venture challenge thanks for all the support you have extended to me all these years. Writing about venture challenge it is difficult not to mention *Hans le Fever* and *Math Khonen*, I have to admit without their help and support it would have be impossible to get through. It was a fun filled experience for me and am thankful for your efforts. This was also a perfect opportunity for me to meet *Ruben Kok*, *Victor de Jager*, Prof. *Kees Koster* and *Marc van Driel* thank you all for the help and support extended.

Thankyou *Judith Risse* for geting me involved into inline skating, I guess I am improving, I go 5 km/hr now !!!. Thankyou *Arnold Kuzniar* for those nice research discussions we use to have at 10:00pm in the evening, I will definitely miss those discussions. Thanks to *Marieke* and *Carla* for all the nice times in the Phd Tours we had been travelling. *Linda Sanderson*, it was great to have known you and worked with you, it was fun and I admire your directness :-). I look forward to be in touch with you. *Mark Boekschoten*, it was nice to have casual discussions once in a while when I was dropping by to your office. *Lydia Afman and Floris* it was nice to have met you and thanks for all the casual discussions we had once in a while on special occasions. *Shohreh*, parties are no fun without you being around, I appreciate your's and *Harro's* friendship. Thanks *Maryam Rakhshandehroo* and *Saskia van Cruchten* for always being available for a break. Thanks *Natasha* for the overkant experience and the weekends (Greek) parties at your place. Thanks *Meike Bunger* though we never worked on any project together it was nice to talk with you in breaks having milk with coffee flavor. Thanks a lot Marjolijn van Stokkom for being a good friend since all these years, its a pity we hardly happen to see each other since you moved, but am looking to see you more often.

Klemen and *Andreja* it was a nice experience visiting your wedding in slovenia, I would also like to thank your mom as well for all the nice food and the experience, I will always remember your wedding. *Vujadin* I do not have any words to say thanks to you man, all the stuff we did from bbqing the complete

lamb, the late night parties and the corridor meetings. I will definitely miss them. *Laeticia Lichtenstein* thanks for being a good buddy and a great support all these years, I really appreciate it a lot. It was a pleasure to have met you and your mom and I look forward to be in touch with you in future as well. Thanks *Zee* for being a close friend, it was always fun to have a drink with you and discuss, we still need to catchup though!!. Thanks Wouter Kuit for introducing me to fencing, am pretty sure next time we play foil you have to watch out. It was a nice experience meeting your parents and brother on easter evening. Thanks *Ernest* for reminding me again and again when the schedule for fencing is :-), and I look forward to play someday with you. Thanks *Hong luo* for being an excellent collegue, though we did not work on similar project it was nice to have known you. Thanks *Pravin Sharma* for comming over to wageningen, it was good to see you after years and hope to see you more often. Thanks *Trudy and Ruud* for all the help and support you provided me since my days in Oss, thanks for everything and am sure we can meetup more often in future. I dare not forget my friends the twin (three) towers *Tim Hulsen*, *David Lutje Hulsik* and *Rick Hovens* for all the poker and casino evenings :-). I guess we are not doing that bad especially the stuff with betting!!!..?.

I would also like to thank *Fredric* and *Emeline*, I enjoyed coming over to Marsielle and meeting your parents, I cherish your friendship a lot. Thanks for sending me those post cards from wherever you go, even though I forget to respond back :-). Many thanks to *Rosalie Rutten* and *Pankaj* for being friends, and *Pankaj* for dropping by on the weekends. *Batian Nieuwerth* and *Ondine* thanks a lot for being available for me throughout my stay here in wageningen, your support and friendship had been excellent throughout these years. Thanks *Ajay Awati*, *Swati*, *Anjal*, *Ganesan Palaniswamy* and *Karishma Palande* for all the time we spend when you were in Wageningen. I am sure we will definitely have an opportunity to meetup again in future. I would also like to thank my ex-salsa partners *Annelies Hommersom*, *Katarina Kozantova* and *Annelotte Coolje* for letting me step on there feet time and again. Many thanks to *Jorge Mauricio Vivas Galvez* for helping out designing the cover of this thesis.

Susann Christine Bellmann my sweet girlfriend, I would like to thank you for all the help and support you extended to me during my thesis and also for traveling with me to all the beautiful countries in Scandinavia from Sweden to Iceland.

My parents had be a guiding force to me throughout my life and career and without their unconditional faith and support it would have been impossible to achieve this feat. I would also like to thank my brothers *Chandan Gavai*, and *Swapnil Gavai* for being always supportive. I would also like to take this opportunity to thank my uncles *Pratap* (MAMA and MAMI) and *Raju KAKA* for being available for my family throughout these years. Also I would like to thank my cousin brother *Vishal Thorat* and sister *Vaishali Thorat* for the support they extended during my stay in the Netherlands.

And last but not least I would like to thank all the people who are not mentioned in this list, who had been directly or indirectly involved in achieving my goals.

Anand K. Gavai

Wageningen, The Netherlands 8^{th} June 2009

List of Publications

- MADMAX - Management and Analysis Database for Multi-platform microArray eXperiments" - Anand K. Gavai, Philip J. de Groot, Ke Lin, Mark V. Boekschoten, Yannan Liu, Harm Nijveen, Pieter B.T. Neerincx, Guido Hooiveld, Michael Müller, and Jack A.M. Leunissen - *(2008) Submitted*
- Comparing model based and black box approaches for estimating metabolic state of organisms from nutrigenomics experiments" - Anand K. Gavai, Guido J.E.J Hooiveld, Sander Kersten, Michael Müller, Peter J.F. Lucas and Jack A.M. Leunissen - *(2009) Submitted*
- Estimating the effect of cigarette smoke on gene expression - a Bayesian approach" - D.M. van Leeuwen, Anand K. Gavai, M.H.M. van Herwijnen, R.W.H. Gottschalk, Jack A.M Leunissen, J.H.M. van Delft and J.C.S. Kleinjans - *(2009) In preparation*
- Constraint-based probabilistic learning of metabolic pathways from tomato volatiles" - Anand K. Gavai, Yury Tikunov, Remco Ursem, Arnaud Bovy, Fred van Eeuwijk, Harm Nijveen, Peter J.F. Lucas, Jack A.M. Leunissen - *Metabolomics 2009 - In Press*

About the author

Anand K. Gavai was born on 25th December 1975, near Bombay (now Mumbai), in India. After finishing his education in computer science from Amravati University, Maharashtra, India, he worked as a software professional for two years before he chose to further his studies in Bioinformatics. He secured admission in Wageningen University, Wageningen, the Netherlands in 2002. During this period he did his internship at Organon N.V. (now Merck) in Oss, the Netherlands, where he developed annotation methods for microarray data using the Gene Ontology. After finishing his MSc in Bioinformatics in 2004 he worked as a research assistant on microarray databases at the Laboratory of Bioinformatics and the Human Nutrition department of the Wageningen University. In 2005 he continued in these two groups starting a PhD project entitled "Bayesian Networks for Omics Data Analysis". His project was partly funded by *CBSG* (Center for BioSystems Genomics) and *TIFN* (Top Institute Food and Nutrition). He currently works as a post-doctoral researcher at Vrije Universiteit Amsterdam, in the Netherlands.

Overview of completed educational activities

Discipline specific activities

Courses

- Basic Statistics, PE&RC, Wageningen 2006
- Molecular Biology for Statisitics, Italy, NuGO (2006)
- Advance Statistics, PE&RC, Wageningen 2007
- Statistical methods in Bioinformatics 2007
- Data Mining, SIKS, 2007

Meetings

- NuGO Week, Italy 2005,(Poster)
- NuGO Week, Oxford 2006, (Poster, Meeting)
- Benelux Bioinformatics Conference (BBC), Wageningen 2006
- 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 6th European Conference on Computational Biology (ECCB) Vienna, Austria (Poster and Presentation)
- NuGO meeting, Venice, Italy 2007

General courses

- 7th International Masterclass Nutrigenomics, Wageningen, 2007 From molecular nutrition to prevention of disease
- Venture Challenge, NBIC (Netherlands Bioinformatics Center), 2007
- Spotfire workshop organized by NBIC, 2006
- Writing Grant Proposals, Wageningen University, 2008

Optionals

- PhD excursion UK and Ireland, oral and poster presentation
- Literature Discussion program Bioinformatics, HN group, 2005-2009, Wageningen
- PhD excursion USA, oral and poster presentation
- Preparing PhD research proposal

The research described in this thesis was financially supported by the Wageningen Center for Food Sciences, currently known as Top Institute Food and Nutrition and Center for BioSystems Genomics.

Cover design and lay-out
The cover was adapted by Jorge Mauricio Vivas Galvez,(www.ozilatin.com) from the TikZ PGF example (http://www.texample.net/tikz/examples/) and the lay-out was done by the author using LATEX.

Printing
GVO Ponsen en Looijen B.V., Wageningen, the Netherlands.