Deriving Genetic Association Networks from Gene Expression Data and Prior Knowledge

Angelica Lindlöf

Department of Computer Science
University of Skövde, Box 408
S-54128 Skövde, Sweden

HS-IDA-MD-01-201

Deriving Genetic Association Networks from Gene Expression Data and Prior Knowledge

Angelica Lindlöf

Submitted by Angelica Lindlöf to the University of Skövde as a dissertation towardsr the degree of M.Sc. by examination and dissertation in the Department of Computer Science.

June 2001

I certify that all material in this thesis which is not my own work has been identified and that no material is included for which a degree has previously been conferred on me.

_____

Angelica Lindlöf

Abstract

In this work three different approaches for deriving genetic association networks were tested. The three approaches were Pearson correlation, an algorithm based on the Boolean network approach and prior knowledge. Pearson correlation and the algorithm based on the Boolean network approach derived associations from gene expression data. In the third approach, prior knowledge from a known genetic network of a related organism was used to derive associations for the target organism, by using homolog matching and mapping the known genetic network to the related organism. The results indicate that the Pearson correlation approach gave the best results, but the prior knowledge approach seems to be the one most worth pursuing.

Key words: genetic networks, homology, gene expression data, correlation measurement, Boolean network

Acknowledgements

Finally, I have reached the closure of this chapter in my life, ending with this thesis. It has been four years of hard work and intensive studying, but also lots of fun.

I would like to thank my supervisor Björn Olsson for guiding me through this work and Magnus L Andersson at AstraZeneca for the original idea of this work and for continuing providing me helpful suggestions on pursuing the work.

I would like to thank my fiancé Zlatan Hodzic for patiently standing by me during these years. The many hours I have spent with the books have many times tested our relationship, but I can truly never have made it this far without you.

I would like to thank my parents, and my brother with family for your support and for never have doubted that I would make it, as I so often have.

Last, but not least, I would like to thank my friends at the University for many interesting and enjoyful conversations during this period, both related and unrelated to this work.

Table of Contents

1 Introduction

1.1 Motivation

In the years to come a large amount of gene expression data will be produced as more organisms' genomes are characterized, the cost of such experiments decreases and the methods to derive gene expressions improve (Chen *et al.*, 1999). This will require efficient theoretical and computational tools to analyze the data from gene expression experiments (Thieffry and Thomas, 1998; Chen *et al.*, 1999).

Researchers have proposed several different computational methods for this purpose, such as reverse engineering methods, forward modeling methods and clustering techniques (Thieffry and Thomas, 1998; Chen *et al.*, 1999; Somogyi *et al.*, 1997; Akutsu *et al.*, 1999; Matsuno *et al.*, 2000; Akutsu *et al.*, 2000; Weaver *et al.*, 1999; D'haeseleer *et al.*, 2000).

The aim of these methods is to retrieve biological information from the expression data, for example, discovery of new genes, detection of mutations and polymorphism, mapping genomic libraries and deriving genetic networks (Ramsay, 1997; D'haeseleer, 2001). This will be useful information in areas such as disease treatment and improvement of agriculture (Weaver and Hedrick, 1997; D'haeseleer *et al.*, 2000).

1.2 Problem definition

Methods for reverse engineering have mostly concerned the genetic regulatory network, probably because the regulatory interactions are the most interesting ones in finding drug targets and genetic engineering. Experimental studies and analyses have

shown the possibilities and the constraints for each of the methods, with different level of performance and so far no method seems to perform well enough to be stated to solve this problem.

The aim of the proposed methods for reverse engineering (chapter 2.2.2) is to derive the genetic regulatory network. However, so far, none of those seems to fulfill this task. Since the methods for deriving the genetic regulatory network have only concentrated on regulatory interactions, other types of interactions will be missed. Ignoring these interactions could lead to errors in the derived network, which reflects the performance of the method.

However, the genetic regulatory network is not the only way of representing the genetic network and is in fact a rather narrow definition of a genetic network. There are also other types of representations, such as the genetic association network and the genetic hybrid network (see chapter 2.1). The genetic association network gives the overall architecture of the genetic network, while the genetic regulatory network holds more specific information of the regulatory interactions between genes.

Most methods have also been tested on hypothetical networks containing only regulatory interactions. Testing on genetic networks containing only regulatory interactions could be very misleading, since “real” networks also contain other types of interactions, such as interactions in protein complexes (Weaver and Hedrick, 1997). Even if the methods are tested on hypothetical regulatory networks their performance are often not well enough. This could be a result of only considering regulatory interactions. Genes with other types of interactions should also be reflected in the gene expression data, since the gene expression data is thought to reflect the underlying genetic network. Ignoring these interactions could lead to misinterpreting the data and thereby leads to incorrect derivation of the genetic network.

This implies that we may have to deal with this problem using another approach. Such an approach can be to first develop a method for deriving the genetic association network, either from existing methods for deriving the genetic regulatory network or by developing a novel method. The genetic association network reflects all kinds of interactions between genes. Once the genetic association network is known, the next step is to identify more specific interactions, such as regulatory interactions. If there is a method for inferring the genetic association network with a high performance, then the probability for inferring more specific interactions correctly increases.

1.3 Hypothesis

The hypothesis is that a correlation measurement, the Boolean network approach or prior knowledge can be used for deriving the genetic association network. Methods for reverse engineering are developed because it is thought that gene expression data reflects the underlying genetic regulatory network. If the data reflects the genetic regulatory network, then it should also reflect the genetic association network and therefore are methods for reverse engineering possible candidates for deriving the genetic association network. Once the associations between genes are known, more specific interactions can be derived.

1.4 Aims and objectives

The aim is to test three different approaches for deriving the unknown genetic association network for an organism, either from an existing method for reverse engineering or by developing a novel method. The tested methods will also be evaluated and compared with each other. A method that performs well in deriving the

genetic association network will be considered a very good starting point in deriving more complex genetic networks.

The objectives are:

- Develop a method for the prior knowledge approach.
- Choose a correlation measurement and a method based on the Boolean network approach.
- Test the three different methods.
- Gather data needed for deriving the genetic association network by the methods.
- Implement the methods.
- Extract the genetic association network using the methods.
- Define a measure for evaluation of the methods.
- Evaluate the derived genetic association network using the defined measurement.
- Make a comparison between the methods.
- Propose and test extensions or improvements of the methods.

1.5 Structure of the thesis

In chapter 2 the definition of a genetic network is discussed and definitions of different types of genetic networks are suggested. In this chapter gene expression techniques are also presented, and in addition the concepts of reverse engineering and forward modelling are explained.

In chapter 3 related and previous works in this area are presented and discussed, as well as different measurements for comparing and evaluating the performance of reverse engineering methods. Related work includes clustering of gene expression data and the Boolean network approach. Previous work is presented under ‘Other methods’ in this chapter.

In chapter 4 the three conceivable methods are introduced and described. The testing of the three methods is described in chapter 5. In chapter 6 the results from the testing and analysis of the results are presented. In chapter 7 the performance of the methods is discussed and in chapter 8 the conclusions of the testing and the hypothesis are presented.

2 Background

Lately a variety of genome projects are characterisating the genomes of diverse organisms, both prokaryotes and eucaryotes (Smolen *et al.*, 2000). Research on genes has focused a great deal on the genes' function, localization in the cell and protein product (Weaver and Hedric, 1997). Proteins are the products of genes and have a variety of functions in the cell, for example they provide the structure of the cell, carry signals between cells, control gene activity, catalyze chemical reactions as enzymes and much more (Weaver and Hedric, 1997; Somogyi *et al.*, 1997).

The process where a protein is produced from a DNA gene is known as the central dogma and includes several steps, where two major steps are identified (figure 1). The first step involves the transcription of a gene into a messenger RNA (mRNA), which is a complementary copy of the gene. The next step is the translation of the mRNA to produce a protein. In this way the information of the gene is carried in the mRNA and then translated into an amino acid sequence, which folds to make a protein.

This linkage between genes and proteins (figure 1) is important information in the treatment of diseases. When a gene is defective, i.e. an error has occurred in the gene, it is reflected in the protein. The defective gene gives rise to a defective protein, which means the protein cannot fulfill its function properly and a disease may develop

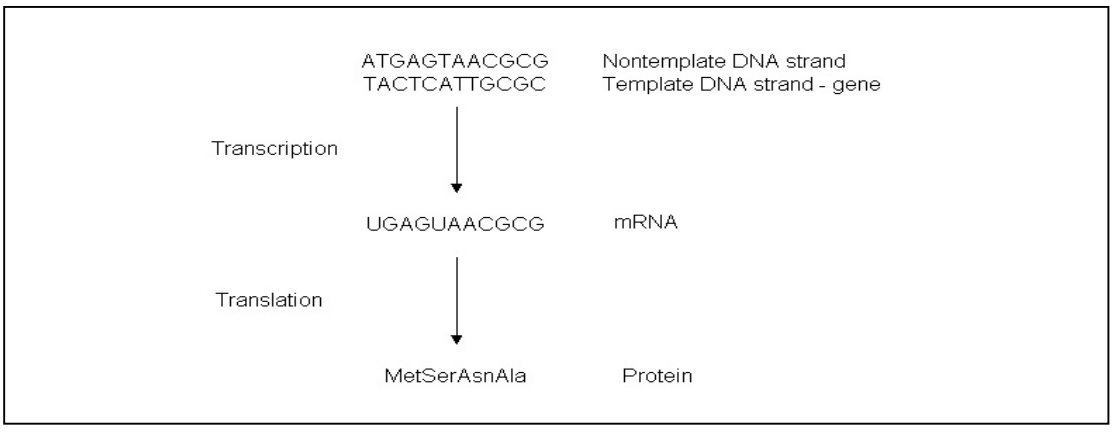


Figure 1. Two major steps in the process of protein synthesis.

in the organism. Defective genes are involved in a variety of diseases, such as cystic fibrosis, Huntington's disease and cancer (Weaver and Hedric, 1997).

Information about genes does not only concern diseases. It is also useful information in improvement of agriculture (Weaver and Hedric, 1997). For example, genes that confer herbicide resistance are useful because an herbicide-resistant plant can survive treatment while weeds around them die (Weaver and Hedric, 1997). The genes for herbicide resistance can be transferred to a plant that does not have this trait and in this way also become herbicide resistant (Weaver and Hedric, 1997).

2.1 Genetic networks

The process of producing proteins includes several steps, which are all regulated (Alberts *et al.*, 1994). The most important regulated step in this process is the transcription. This step regulates how often and when a gene is transcribed into an mRNA. The transcription control in eucaryotes has the process shown in figure 2:

A stimulus, such as a hormone, often activates a certain type of protein in the cell, a so called transcription factor (Alberts *et al.*, 1994). Activated transcription factors bind to specific sites on the DNA sequence and thereby allow a specific enzyme, called RNA polymerase, to bind to the DNA sequence (figure 2). The RNA polymerase executes the transcription of the nearby gene. The transcription factors regulate the transcription of the nearby gene on the DNA sequence, either by activating or suppressing the transcription of the gene. The activity of transcription factors, which control the transcription and thereby the regulation of genes, is adjusted by phosphorylation and other intermolecular interactions (Smolen *et al.*, 2000). In addition, some transcription factors have shown to regulate their own transcription

and there are also other genes that have shown to regulate their own transcription (Smolen *et al.*, 2000).

A gene that is expressed is translated into a protein, the gene product, which affects the state of the cell (Weaver *et al.*, 1999). The protein could affect the expression of other genes or its own expression level by changing the conditions in the cell, such as when the hormone activates the transcription factors and thereby affects the state of the cell. The expression of one or several genes will lead to a different state of the cell, where other genes will be expressed or repressed as a response (Weaver *et al.*, 1999). The effect a gene's expression has on other genes is termed gene regulation, which can be visualised in a conceptual model as a genetic network. The general definition of a genetic network is that it describes the regulatory interactions between genes (Szallasi, 1999).

This general definition of a genetic network is rather narrow, since it is known that

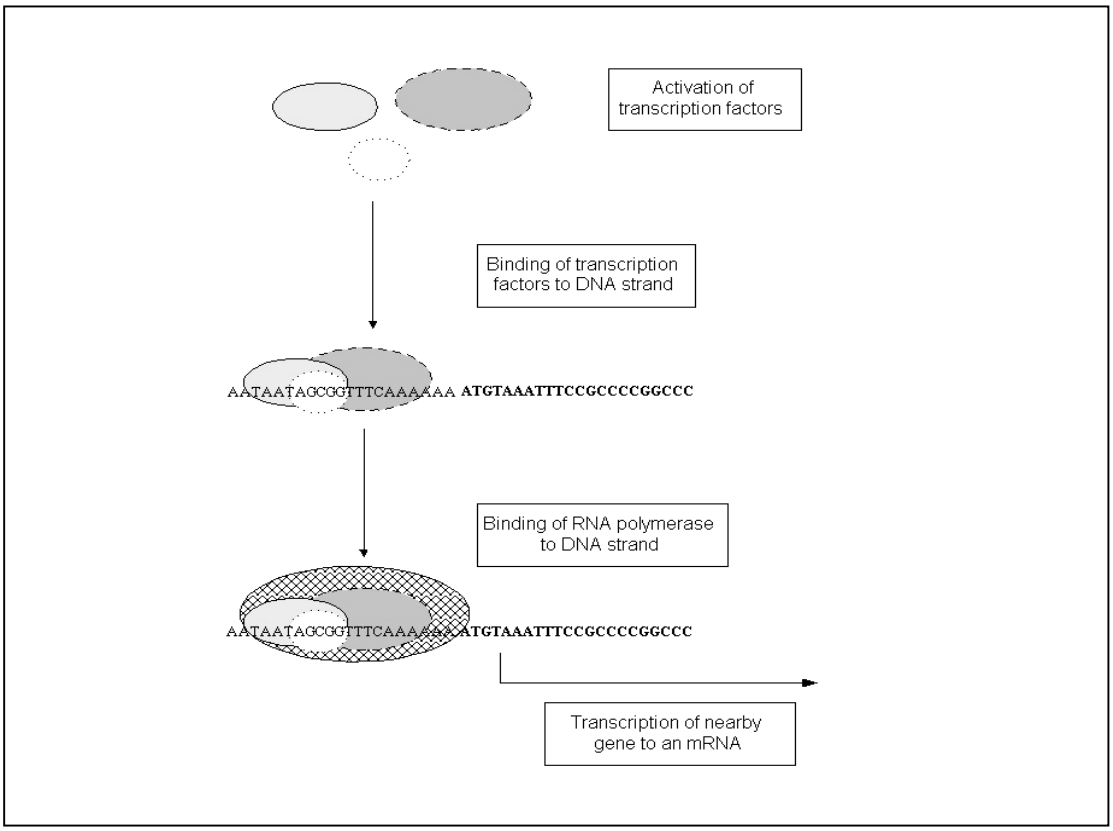


Figure 2. Basic concepts in the transcription of a gene into an mRNA

there are also other interactions between genes (Weaver *et al.*, 1999). For example consider the transcription of a gene described in figure 2. Here, the transcription factors interact with each other and with the RNA polymerase in a complex. There are also regulatory interactions between the transcription factors and the hormone. A more accurate general definition of a genetic network would rather be that it describes all the interactions between genes, without narrowing it down to one specific type of interaction. Thereafter different types of genetic networks could be defined. For example, the general definition of the genetic network that is used could instead be defined as the genetic regulatory network, since it contains purely regulatory interactions.

Another type of genetic network could be the genetic association network. In this type of network genes that interact with each other, such as the different transcription factors interact with each other, with the hormone and with the RNA polymerase, are considered to have an association with each other. In this representation regulatory interactions are treated simply as associations. The genetic association network also represents the topology of the genetic network, showing only which genes are connected. Then there could also be a hybrid type of genetic network. For example, a genetic hybrid network could contain regulatory interactions, complex interactions and associations for those interactions the specific type is unknown. In this thesis the focus will be on genetic association networks.

The genetic regulatory network holds more specific information about the regulation of the expression of the genes in the cell, while the genetic association network gives the overall architecture of the genetic network. It is important to realise that none of these types of genetic networks holds sufficient information for a total

understanding of the biological processes between genes and that one representation is not better than the other.

A conceptual model of the genetic regulatory network can be visualised as in figure 3a, where a box is a gene and the directed edges connecting the boxes represent the effect one gene has on another gene, activation or degradation of that affected gene. The genes in the genetic association network can also be visualised as boxes, as in the visualisation of the genetic regulatory network, but where an undirected edge between the boxes represents the association, see figure 3b. A conceptual model of a genetic hybrid network could be as in figure 4, which is a visualisation of the transcription of a gene. The stimulus, such as a hormone, regulates the activation or the repression of transcription factors and is considered a regulatory interaction. The activated transcription factors bind to the DNA strand and forms a complex with each other. This could be visualised as complex interactions between the transcription factors. The complex of transcription factors promotes the RNA polymerase to bind to the DNA strand, which transcribes the nearby gene. The interactions between the RNA polymerase and the transcription factor complex, and the RNA polymerase and the transcribed gene could be considered as associations, if no general definition of those kinds of interactions exists.

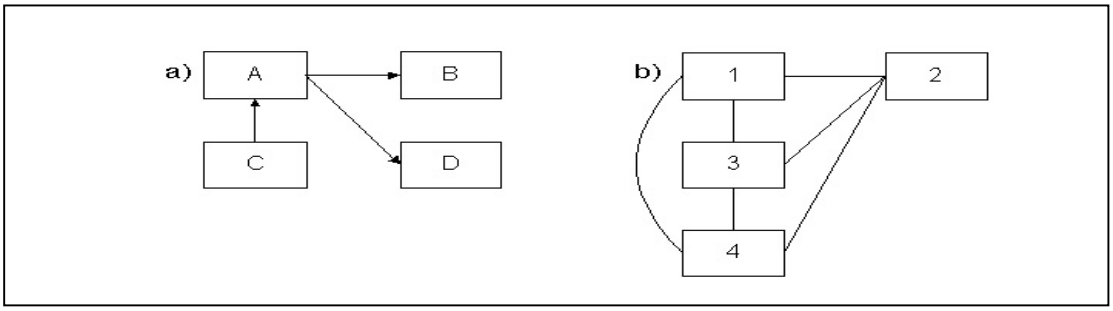


Figure 3. Two different types of a genetic network represented as boxes and edges. In a) a genetic regulatory network, where a box is a gene and a directed edge represents how a gene affects another gene, and in b) a genetic association network, where a box is a gene and an association between two genes is represented by an undirected edge.

2.2 Gene expression data

The level of expression of a gene can be estimated by measuring the protein level or the mRNA level of the gene in the cell (Duggan *et al.*, 1999; D'haeseleer *et al.*, 1999; Somogyi, 1999). There are several techniques developed for this purpose, such as northern blotting and micro arrays. Gene expression patterns are derived in response to specific stimuli or during the development of the cell (Smolen *et al.*, 2000). The expression levels of the genes are measured simultaneously for thousands of genes at a time (D'Haeseleer *et al.*, 1999). The aim of gene expression data gathering is to gain information of how single genes or groups of genes control cellular responses to stimuli from the environment and how genes interact with each other, which can be described as a genetic network (Smolen *et al.*, 2000). The gene expression patterns are assumed to reflect this network (Szallasi, 1999).

Ramsay (1997) reviewed a number of experiments where DNA micro arrays had been applied. The experiments reviewed concerned gene discovery, detection of mutations and polymorphism, and mapping genomic libraries. For example, gene expression data was used to explore differences in expression between *Arabidopsis*

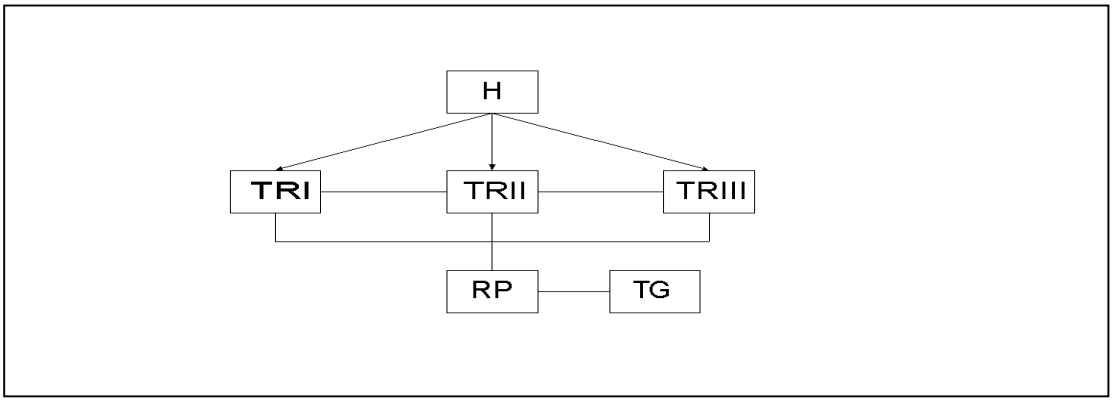


Figure 4. A possible representation of a genetic network containing both regulatory interactions and associations, where a directed edge represents a regulatory interaction and an undirected edge represents an association. The hormone (H) activates three transcription factors (TRI, TRII and TRIII), which thereafter associate with each other and the RNA polymerase (RP). The RNA polymerase then transcribes the nearby gene (TG), which is considered an association.

thaliana root and leaf, human T cells were examined under heat shock and exposure to phorbol ester. The experiments also concerned genome-wide sequence recognition in the *Escherichia coli* genome and mapping the *S. cerevisisae* genomic library by determining the order of overlapping clones. And in addition, detection of possible heterozygous mutations of the *BRCA1* breast and ovarian cancer gene, detecting mutations in the reverse transcriptase and protease genes in the HIV-1 virus.

2.2.1 Gene expression techniques

Two technologies for measuring gene expression have been widely accepted and used, the cDNA micro array and the *in situ* synthesised oligonucleotide array (Duggan *et al.,* 1999; Gerhold *et al.,* 1999; Dutilh, 2001). The methods are based on the same principle, they make use of (c)DNA clones attached to coated glass surfaces, a silicon chip or a nylon filter (Dutilh, 2001). They differ in the way of attaching the nucleotide sequences on the glass (Dutilh, 2001). The cDNA micro array method was developed at Stanford University (Gerhold *et al.*, 1999). In this method many copies of amplified DNA strands are attached to the chip by robots, that spot the strands onto the solid (Duggan *et al.*, 1999; Gerhold *et al.*, 1999). In the synthesised oligonucleotide array method smaller DNA strands, so called oligonucleotides, up to 25 nucleotides are directly synthesised onto the chip (Dutilh, 2001). On the chip 3'-OH ends are attached (sticking out), to which the oligonucleotides can be attached (Dutilh, 2001; Gerhold *et al.*, 1999.

The cDNA that are of interest, are attached onto a chip and the DNA chip is then available for hybridisation with mRNA (figure 5). Total mRNA from both the target and a reference are labelled with fluorescent dyes. Optimally, mRNA from single cells would be used, but there is a difficulty in purifying and amplifying mRNA from single cells (Dutilh, 2001). In practice, mRNA is purified from a specific tissue,

which has the disadvantage in increasing the error rate in the measurement (Dutilh, 2001). The extracted mRNA levels from an amount of cells are considered an average mRNA level in the cell population. The fluorescent labelled mRNAs are then allowed to hybridise with the clones on the DNA chip. Laser excitation of the hybridised mRNA yields an emission with a characteristic spectrum. The spectrum makes it possible to measure the amount of fluorescent marker of the hybridised clones. This spectrum is measured with a laser microscope, which yields an image of the DNA chip with different intensities depending on how much of the mRNA has hybridised with the cDNA.

The advantage of these methods is that many sequences (genes) can be measured in a single experiment and with a minimum of material, with up to 10 000 genes on one chip (Dutilh, 2001; Gerhold *et al.*, 1999). The disadvantage is that they have a higher error rate compared to traditional methods, such as Northern blot or quantitive

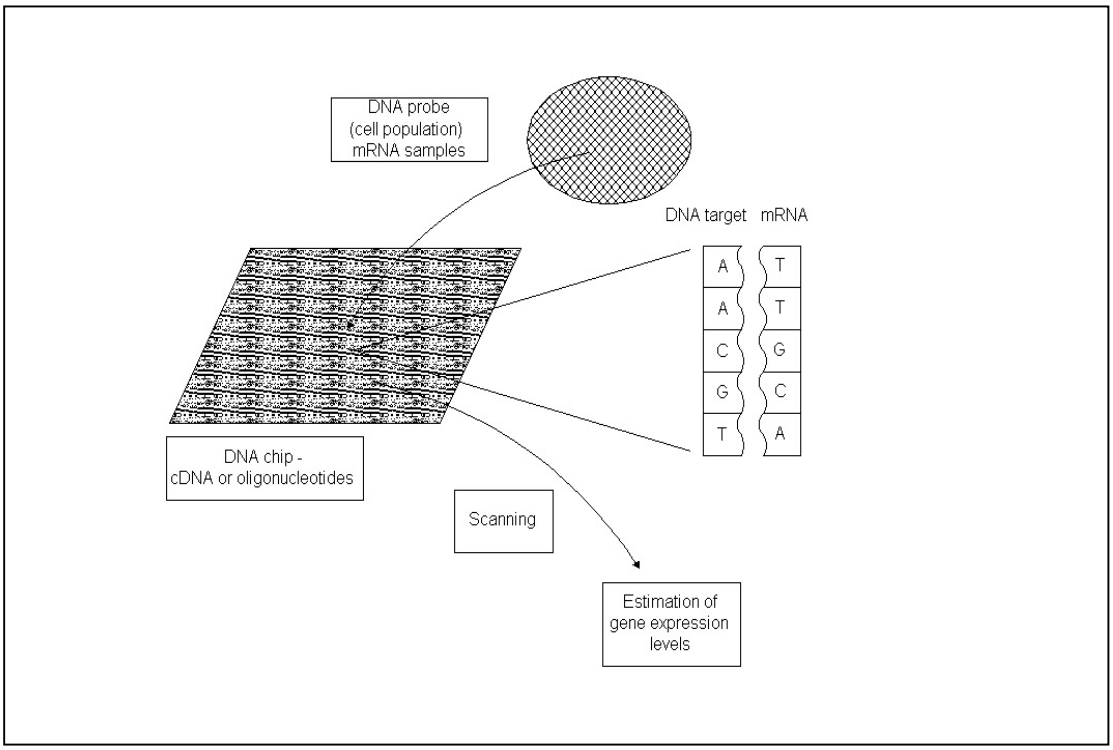


Figure 5. Illustration of the micro array technique. mRNA samples from a cell population hybridise with attached cDNA clones on a chip. The gene expression levels are estimated by scanning the surface.

PCR (Dutilh, 2001). For example, the error of quantitation can reach up to 50% compared to 20% in the traditional methods (Dutilh, 2001). Another important aspect is that only known genes can be analysed in the methods, which limits the potential of the experiment (Dutilh, 2001). This aspect is however useful in finding knew, unknown genes (Ramsay, 1997).

2.2.2 Reverse engineering and forward modeling

Szallasi (1999) states that analysis of gene expression data will support experimental biology in at least two ways, namely:

- Reverse engineering:

  The gene expression measurements are the results from an underlying genetic network. Reverse engineering methods are used to derive information of the underlying genetic network from the gene expression data. The purpose of current reverse engineering methods is to identify regulatory interactions in the genetic network, which thereafter can be experimentally tested and validated.

- Forward modelling:

  Forward modelling is used for simulation of the genetic network based on gene expression data. Empirically determined gene expression data are used as an initial set of parameters. This set together with a thoroughly analysed genetic network are expected to produce gene expression matrixes and accurately predict time dependent gene expression measurements.

Szallasi (1999) presented a number of factors, which will limit the amount of information contained in gene expression measurements and therefore will have an effect on the applicability of genetic network analysis. Some aspects of the factors that are discussed in the paper will be repeated here:

1. *The prevailing nature of the genetic network*

Szallasi states that the genetic networks can be visualized in two different ways, either as deterministic or as stochastic systems. In a deterministic system one gene expression state can only lead to one other gene expression state and cannot have two or more different successive outcomes. In a stochastic system a gene expression state can lead to more than one successive gene expression state, which means that similar cells can follow a different gene expression path between gene expression states. Stochastic systems are supported in reality and describe the kinetic of gene regulation more accurately than deterministic systems. Gene expression measurement always gives an average of a population of cells, which means changes in single cells will be missed and thereby the stochastic system (Szallasi, 1999).

2. *The effective size of the network*

In modelling a genetic network it is often treated as a deterministic network, where every state of a gene is unequivocally determined by the expression state of its input genes. However, there are several steps from a gene being activated to the effect the gene has on another gene, and for example regulatory interactions are not deterministic at the mRNA level (Szallasi, 1999). There are many regulatory factors in a genetic network and in modelling the network one must considered other factors than only genes, such as mRNA, proteins, co-factors, etc, for generating a network that behaves deterministic. This will yield about 10 more parameters in the network, which means a genetic network of a certain size will grow about 10 times if all these parameters are added (Szallasi, 1999).

3. *The compartmentalization of the genetic network*

The level of compartmentalization in the network affects the number of regulatory interactions between subgroups in the network and the numbers of regulatory

interactions that need to be tested by reverse engineering algorithms. A high level of compartmentalization means fewer interactions to test (Szallasi, 1999).

4. *The information content of gene expression matrices*

Because of the expected stochastic nature of the genetic network there is an upper limit to gene expression measurements. This means there are a maximum number of measurement points in the gene expression matrix that have to be covered (Szallasi, 1999). As examples, consider yeast where the limit is considered to be every 5 minutes and for mammalian cells every 15-30 minutes (Szallasi, 1999). More measurement points are not expected to give more information about the gene regulation.

3 Related and previous work

In this chapter related and previous work is presented. In chapter 3.1 different methods for reverse engineering and forward modelling are described, where related work is the clustering technique and the Boolean network appraoch. Previous work for reverse engineering and forward modelling is presented under 'Other methods'. Related work is also the suggested comparison measurements for reverse engineering methods by Wessels *et al.* (2001) and Ideker *et al.* (2000), which are presented in chapter 3.2.

3.1 Methods for reverse engineering and forward modeling

Since data from gene expression experiments can be abundant (an experiment can contain thousands of genes) computational algorithms and methods have been developed in order to infer and model the underlying genetic network. The aim of these methods is to find a universal, single method, which can infer and model the underlying genetic network (D'Haeseleer *et al.*, 1999; Thieffry and Thomas, 1998; Somogyi 1999; D'haeseleer *et al.*, 2000).

In this chapter some developed methods will be presented, along with known advantages and disadvantages in these. Most methods for reverse engineering (see chapter 2.2.2) have mainly concerned the genetic regulatory network, while methods for forward modeling (see chapter 2.2.2) have also incorporated other types of interactions between genetic components (i.e. genes, mRNAs and proteins).

3.1.1 Clustering of gene expression data

Clustering of data is a general technique for finding patterns of similarity in the data and has been applied to gene expression data. Clustering of gene expression patterns is often thought of as a way of retrieving biological information underlying the gene expression profiles (D'haeseleer *et al*., 2000; Heyer *et al*., 1999). For example, it has been used to retrieve information on genes that show a significant change in expression level depending on a certain condition (D'haeseleer *et al.*, 2000). It is also stated that genes that share similar functions and regulation should show similar gene expression profiles, and that clustering can be used to group these genes together (D'haeseleer *et al.*, 2000; D'haeseleer *et al*., 1999; Michaels *et al.*, 1998; Heyer *et al*., 1999). This is supported in several studies, but it is important to know that one can find functionally related genes that are not co-expressed as well, and that those genes do not typically end up in the same cluster (D'haeseleer *et al*., 2000; Heyer *et al*., 1999). Clustering of gene expression data will yield groups of genes that are tightly co-expressed over some specific time or experiment (D'haeseleer *et al*., 2000; Heyer *et al*., 1999).

The clustering is also said to reveal information about the underlying regulatory network, since genes that are regulated by a common gene are thought to be co-expressed and therefore share the same gene expression pattern (D'haeseleer *et al*., 2000; Heyer *et al*., 1999). It is important to realise that clustering may give information about which genes are co-regulated, but not the exact regulation (D'haeseleer *et al*., 2000).

Clustering of gene expression data involves four different steps, 1) pre-processing of the data, 2) choosing a similarity measure, 3) choosing a clustering technique, and 4) analysing the clusters (Heyer *et al*., 1999). Pre-processing could, for example, be

removal of data that has no relevance to the experiment, removal of data containing errors due to problems in gathering the data, recalculating the data into logarithmic values and normalisation of the data (Heyer *et al*., 1999; D'haeseleer *et al*., 2000).

Most clustering techniques use a matrix of pairwise distances between genes as input, which holds the difference between gene expression profiles as a distance measure (D'haeseleer *et al*., 2000). The pairwise measurement should assign high similarity scores to genes with related expression patterns (Heyer *et al*., 1999). Examples of similarity measurements are Pearson correlation, Spearman rank correlation, Jack-knife correlation and Euclidean distance (Heyer *et al*., 1999).

There are several different methods of clustering and they can be divided into two groups, hierarchical and non-hierarchical (D'haeseleer *et al*., 2000). A hierarchical clustering method group genes together and order the groups (clusters) into a hierarchical structure, whereas a non-hierarchical clustering method cluster genes into a number of groups according to some optimisation criterion in an iterative way until the criterion is reached (D'haeseleer *et al*., 2000). Examples of hierarchical clustering algorithms are FITCH, average-linkage analysis and divisive hierarchical algorithm, and of non-hierarchical algorithms are K-means, Self-Organising Maps (SOM) and the quality cluster algorithm (D'haeseleer *et al*., 2000; Heyer *et al*., 1999).

As an example of an application of clustering of gene expression data Zhu and Zhang (2000) clustered a set of gene expression data using three different approaches. The clustering was based on genes sharing similar expression profiles, function categorisation and promoter elements. Expression data from yeast sporulation was used in the experiment. The clustering based on expression profiles yields clusters containing genes with similar expression patterns, clusters based on function categorisation gives information about protein interactions and pathways, and clusters

based on promoter elements gives information about gene co-regulation on the transcriptional level. The clustering method was based on the density search method, termed largest-first since the largest cluster always is reported first. The results from clustering based on gene expression profiles show that genes in the same cluster may have totally different functions and promoter elements. Clustering based on function categorisation indicates that a transcription factor may play different roles in different clusters, and clustering based on promoter elements shows that the MSE regulatory element end up in several different clusters. Zhu and Zhang (2000) states, as a conclusion of the experiment, that clustering analysis gives an overview of the gene expression data, that no single method performs well enough, and that it is important to combine different approaches. Heyer *et al*. (1999) also point out that the clustering do not give the final answers and should be used as an exploratory tool for identifying candidate solutions for further analysis.

3.1.2 Boolean network approach

In the Boolean network approach the state of a gene is modeled as either ON or OFF. The ON and OFF state can be modeled in two different ways (Smolen *et al.*, 2000; D'haeseleer *et al.*, 2000)

1. when there is a gene expression of the gene it is ON and when there is not the gene is OFF
2. ON means the gene expression has increased from a steady-state expression level and OFF means the gene expression has decreased from a steady-state expression level of that gene in the cell

This approach means the model applies to gene expression patterns at steady-state or stationary-state (Maki *et al.*, 2000). The effect one gene has on another gene is

expressed by Boolean logical rules and the gene expression patterns are used as restricted conditions to the Boolean network (D'haeseleer *et al.*, 2000; Maki *et al.*, 2000; Smolen *et al*., 2000). An example of a logical rule:

gene A is ON if gene B *and* C is OFF

and

gene A is ON if gene D is ON

The genes are often represented as nodes in the network and the logical rules as edges, which are also represented in a separate matrix (figure 6).

An advantage of the model is that it is said to handle a large amount of gene expression data (Maki *et al.*, 2000). But the performance of the model depends on the

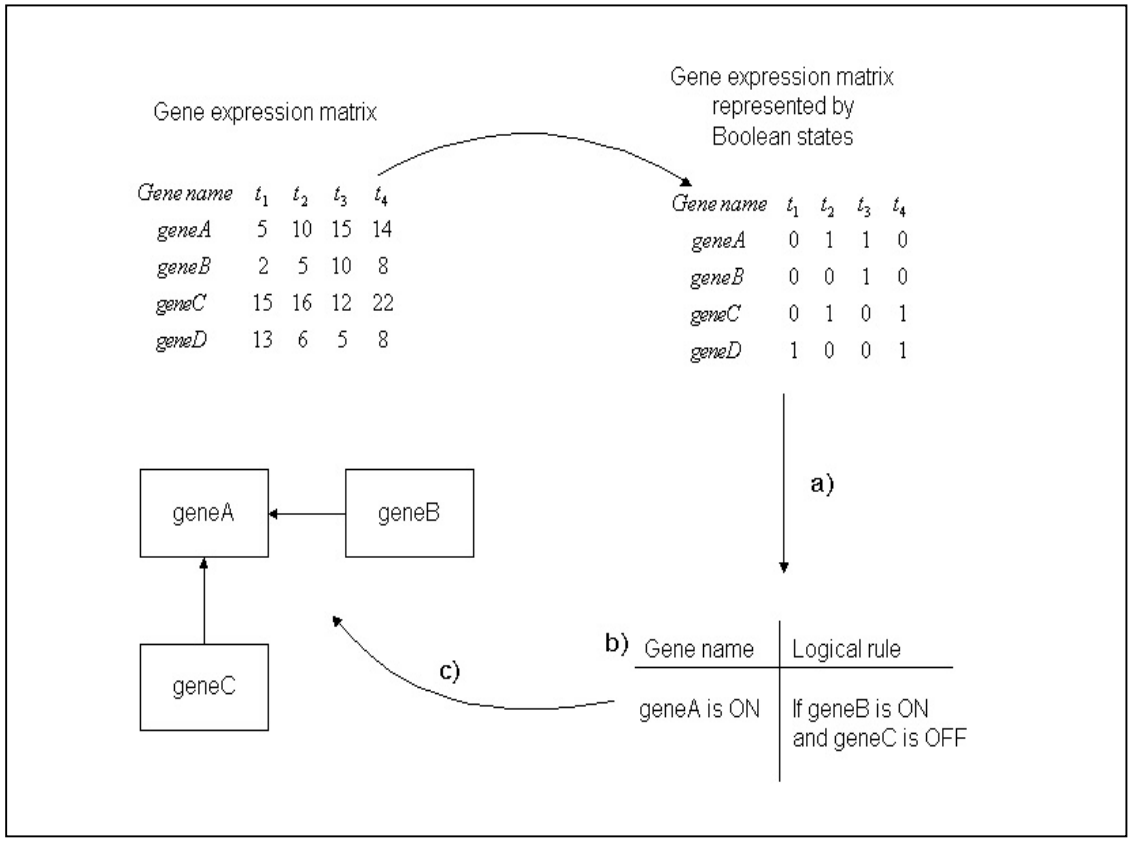


Figure 6. The figure shows the representation of a Boolean network. The boxes and edges represent the genetic regulatory network. The regulatory interactions are a) extracted from the gene expression data, represented by binary characters, b) represented in a matrix with logical rules and c) the genetic regulatory network is derived from the logical rules.

structure of the data, which is a disadvantage (Maki *et al.*, 2000). For example the model has problems in deriving the regulatory interactions if two genes affect each other or if there is a loop structure in the genetic network (Maki *et al.*, 2000). Another disadvantage is that all the genes in the network are assumed to be updated synchronously, which is not the case in real systems (Smolen *et al*., 2000; D'haeseleer *et al.*, 2000). In this model the gene expressions are treated either as completely on or off, which makes assumptions of the regulatory interactions in the genetic network (Weaver *et al.*, 1999). In real systems there are many genes that have intermediate expression levels, where those are also regulatory (Weaver *et al.*, 1999).

Another reflection is that in many methods built on Boolean networks the genetic network is assumed to have a few fixed number of regulatory interactions, often two or three (Weaver *et al.*, 1999; Liang *et al.*, 1998). In real systems it is known that some genes have a lot more interactions than three while others just have a few. This complexity in regulatory interactions is often not taken into consideration in the methods (Weaver *et al*., 1999).

As an example of a proposed method for reverse engineering based on the Boolean network approach is the algorithm REVEAL developed by Liang *et al.* (1998). The algorithm makes use of Shannon entropy and mutual information (also referred to as rate of transmission) to extract the connections, between nodes in the Boolean network from gene expression data. The proposed algorithm was tested on simulated data and not on empirically derived gene expression measurements. The conclusion was that the algorithm performs well when the number of connections between genes is small. It infers the genetic network very quickly for simple networks with only a few interacting nodes, but the computational effort increases with the number of edges.

Ideker *et al.* (2000) developed a method for reverse engineering of a Boolean network through perturbations. The method includes an algorithm which derives one or many possible hypothetical genetic networks from the gene expression data. If more than one hypothetical genetic network is derived, perturbations are used to get additional information about the underlying genetic regulatory network, in order to discriminate among the possible networks that were derived. This is done by a second algorithm, which chooses an additional perturbation among a predefined set of possible perturbations. The method was tested on a number of simulated, hypothetical genetic regulatory networks with inferred gene expression profiles. All of the hypothetical data sets were restricted to not contain any cycles, since these are known to cause instability and oscillations to the Boolean network (Maki *et al.*, 2000). The two algorithms were tested separately. For the first algorithm the evaluation shows that a large percentage of the edges in the derived network are also present in the hypothetical network. A drawback is that as the number of edges in the hypothetical network increases, the percentage of correctly derived edges decreases. As for the second algorithm, a test was performed on networks containing a maximum of two edges from each gene, but with a varying number of genes in the networks. The evaluation shows that as the number of genes increases in the network, so does the number of perturbations required.

The Boolean network approach has also been proposed for forward modeling of genetic networks. Szallasi and Liang (1998) proposed a method, including not only genes as parameters but also parameters such as mRNA levels, the localization of a protein or phosphorylation of a protein. The parameters in the genetic network are modelled as nodes in the Boolean network, and directed edges between the parameters represents the regulatory effect one parameter has on another parameter.

The logical functions, which the Boolean approach is built upon, define the status of a parameter depending on its regulatory inputs. This approach is assumed to produce time series measurements of gene expression levels resembling experimentally measured gene expression levels. It is also proposed that the experimentally measured gene expression levels could be used as an input to the genetic network. The advantage of including additional parameters apart from genes is that the genetic network becomes deterministic, but has the disadvantage of increasing the number of variables in the model (Szallasi and Liang, 1998).

Szallasi and Liang (1998) analysed the proposed method theoretically and their conclusion was that the set of logical rules is the most important factor for avoiding chaotic behaviour, oscillations and biologically unrealistic long cell cycles, which often is the case when modelling genetic networks with the Boolean approach. It was stated that a special subset of logical rules must exist in real biology, which will reduce these side effects. The authors give no clue to how this special subset of logical rules is to be found and the question is if they really exist or if the side effects are inherent in the Boolean approach and therefore cannot be avoided.

3.1.3 Other methods

In this chapter some previous work in reverse engineering and forward modeling will be presented, together with some known advantages and disadvantages.

Differential equations

Differential equations can be used for forward modelling of biochemical systems, such as genetic networks or metabolic pathway networks (D'haeseleer *et al.*, 1999). Here components in the system are modelled as continuous instead of discrete, as in the Boolean network model (D'haeseleer *et al.*, 2000; Chen *et al.*, 1999). Genetic

regulatory systems with continuous behaviour can be more thoroughly analysed with differential equations (Smolen *et. al.*, 2000).

The Boolean network model is favoured because of its ease to model, but differential equations have the advantage of greater physical accuracy (Smolen *et al.*, 2000; Szallasi, 1999). Another advantage is that time delays can be incorporated in the system and those not only have the capacity to model genetic interactions, but can also model other components in the system such as mRNA and protein concentrations, which are important aspects in regulation (Smolen *et. al.*, 2000). Unlike Boolean network models, differential equations can also model negative feedback loops, which have a stabilising effect on the system (D'haeseleer *et al.*, 2000). The disadvantage of differential equations is that they are more computationally intensive than the Boolean network model and are more suited to smaller genetic networks with a few interacting genes (Smolen *et. al.*, 2000). For example, a regulatory network can be modelled by differential equations as

$$\frac{dx_1}{dt} = f(x_n) - k_1 x_1, \qquad \text{(eq 1)}$$

$$\frac{dx_j}{dt} = x_{j-1} - k_j x_j \qquad j = 2, \ldots, n$$

where x_i is a molecule or a gene in the network, $f(x_n)$ is a function that models either activation or repression by increasing x_n, and k_n is the rate constant of the forward or reverse reaction (Smolen *et al.*, 1999; Kyoda *et al.*, 2000).

Hybrid methods

There have also been attempts to develop hybrid models for both reverse engineering and forward modelling. Hybrid models combine two or more approaches in the method, in order to improve the performance.

Maki *et al.* (2000) developed a hybrid model, using a Boolean network model together with an S-system network model for reverse engineering. The Boolean network approach is described in chapter 3.3.2. The S-system model is based on a specific type of differential equation and can handle gene expression data from temporal responses, such as cell development (Maki *et al.*, 2000). Here, it is used for those situations the Boolean network approach cannot handle, for example when there is a loop structure in the network (Maki *et al.*, 2000). The disadvantage of the S-system model is that it requires a large number of parameter estimations. The parameter estimations are done with a genetic algorithm (GA). In this way the strength in each approach is used: the Boolean network model to get a first overall architecture of the genetic network, the S-system for extending the genetic network with those regulatory interactions the Boolean approach cannot handle, and a GA for estimating parameters required in the S-system model. The method was tested on theoretical gene expression data for 30 genes. For this test the hybrid model worked well. The question is how it performs on real gene expression data and larger data sets. The theoretical genetic network in the test contained at maximum two edges from a node, so another remaining question is how it performs with more edges.

Matsuno *et al.* (2000) proposed a Hybrid Petri Net for forward modelling of genetic networks. The Hybrid Petri Net is an extension of Petri Nets. In a Petri Net only discrete factors in the network can be modelled. In the extended Hybrid Petri Net continuous factors can be modelled with differential equations, together with discrete factors. Other factors than genes can also be incorporated, such as the transcription and translation of a gene (Matsuno and Doi, 2000). Including other factors than genes into the model is assumed to generate a more accurate modelling of genetic networks

(Smolen *et. al.*, 2000), plus both discrete and continuous parameters can be incorporated.

Akutsu *et al.* (2000) proposed a hybrid model based on the Boolean network model combined with qualitative reasoning of differential equations, for both reverse engineering and forward modelling of genetic networks. In the method genes are modelled as nodes as in the Boolean approach, but the edges between genes are modelled with qualitative reasoning instead of logical rules as in the Boolean approach. Akutsu *et al.* (2000) developed an algorithm for inferring genetic networks from gene expression data using this approach. The disadvantage of the method is that, in order to perform well, requires many time series data beginning from different sets of initial values from different types of environment or conditions in order to perform well (Akutsu *et al.*, 2000). For this reason this approach is often not applicable, since gene expression data is sparse at the moment (Akutsu *et al.*, 2000).

Weight matrices

Weaver *et al.* (1999) proposed a neural network to model regulatory genetic networks. A neural network is based on a weight matrix, which holds the information about all the regulatory interactions between genes (figure 7). Each gene in the genetic network is represented as a node in the neural network with connections to all the other genes (figure 7). The neural network can be used for forward modelling, to analyse the genetic network model and predict gene expression outputs (Weaver *et al.*, 1999). For each time step t, a gene $r_i(t)$ in the network adds all the input from all other genes in the network, represented by a vector $e_j(t)$, multiplied with the weights from the weight matrix $w_{i,j}$ according to

$$r_i(t) = \sum_j w_{i,j} e_j(t) \qquad \text{(eq 2)}$$

which generates a gene expression output at time step $t+1$, where the level of gene expression is between 0 and 1. The weight matrix is unknown at the start of the modelling, but can be approximated from gene expression data through a learning algorithm for neural networks (Dutilh, 2001). The weight matrix can be approximated through other algorithms, such as a Genetic Algorithm or simulated annealing (Dutilh, 2001). Weaver *et al.* (1999) also proposed that neural networks can be used in reverse engineering, where the weight matrix can be derived from gene expression data, and thereby predict the genetic network, a method was developed for this purpose. The weight matrix method makes assumptions about the genetic network's behaviour. For example, the genetic interactions are assumed to be independent and synchronously regulated, which is not the case in real biological systems (Weaver *et al.*, 1999). In the developed method for deriving the weight matrix from gene expression data the maximal expression of the genes are needed, which makes the assumption that a gene's maximum expression level can be determined empirically (Weaver *et al.*, 1999).

Qualitative analysis

Thieffry and Thomas (1998) proposed a qualitative analysis of regulatory genetic networks, where three matrices can describe the regulatory network. The matrices contain the signs of interactions, the thresholds associated to these interactions and the values of the corresponding logical parameters. In the interaction matrix connections between genes are represented. In the threshold matrix, information on which threshold function is used to a specific connection between two genes is specified. In the third matrix, logical parameters representing the weights of the basal expression, the weights of activation and the weights of combined actions of the genes in the network on a certain gene. A genetic regulatory network is qualitative described by

these three matrices and can be used for forward modelling of the network. The authors state that this approach is especially useful when handling feedback circuits, but also say that the approach depends on the data available and that qualitative data often are lacking, since the thresholds and the weights are often poorly estimated. The authors continue saying the approach can be used as an alternative to differential equations, since it is a useful approach to get a first overview of the dynamical properties of the differential equations and thereby can help refining the model of the genetic regulatory network.

3.2 Comparison measurements for reverse engineering methods

This chapter reviews comparison measurements for reverse engineering methods, developed by Wessels *et al*. (2001) and Ideker *et al*. (2000), chapter 3.2.1 and 3.2.2 respectively. Wessels *et al.* (2001) developed six different measurements of comparison, where the inferential power relates most to the measurement developed by Ideker *et al*. (2000).

3.2.1 Measurement developed for continuous methods

The measurements developed by Wessels *et al.* (2001) were inferential power, prediction power, robustness, consistency, stability and computational cost. The methods compared were restricted to continuous methods (i.e. discrete methods, such as the Boolean network model, were not included).

The *inferential power* measures the capability to accurately estimate the genetic regulatory network (termed the gene regulation matrix in Wessels *et al.* (2001)), which is measured as the similarity between the actual and the derived genetic regulatory network. The *prediction power* measures how well the method

approximates the actual gene expression profile. Gene expression measurements often contain some degree of noise. The *robustness* measures to what degree an accurate gene regulatory network will be derived when there is noise present.

A problem when inferring genetic regulatory networks from gene expression data is that there are a relatively large number of genes compared to the number of measured time points in the gene expression profile. This could result in multiple gene regulatory network candidates from the same gene expression profiles, which is termed inconsistency. A method is *consistent* if it inferres only one genetic regulatory network.

Since concentrations of gene expression products are bounded in the cell, the genetic regulatory network is *stable*. This should therefore also apply to derived gene regulatory networks. If the measurements of the predicted gene expression levels remain bounded over all time, the derived gene regulatory network is said to be stable. The *computational cost* measures how long computation time is needed for the method to derive the gene regulatory network, which a short time is preferred.

The developed measurements by Wessels *et al.* (2001) are good methods for comparing different methods for reverse engineering. The measurements could easily be applied to discrete methods and it would be interesting to make a comparison with other methods, such as different approaches to the Boolean network model, hybrid models and methods based on weight matrices.

As a conclusion by Wessels *et al*. (2001), for this simple test all the methods had low inferential power. This implies that the methods cannot infer the correct genetic regulatory network satisfactorily. Either the methods need to be further developed in order to increase the inferential power or other methods must be used.

3.2.2 Sensitivity and specificity

Ideker *et al.* (2000) developed a method for deriving genetic regulatory interactions from gene expression data using the Boolean network approach and through perturbations (see chapter 2.3.2 for more details). In evaluating the performance of the first algorithm Ideker *et al.* developed two measurements, sensitivity and specificity (figure 7), for this purpose. The definitions of the measurements are:

- *sensitivity*: the percentage of edges in the target network that are also present in the derived network
- *specificity*: the percentage of edges in the derived network that are also present in the target network

The target network is one of the hypothetical networks that were set up for testing the method. High percentage levels on both these measurements are desired. These

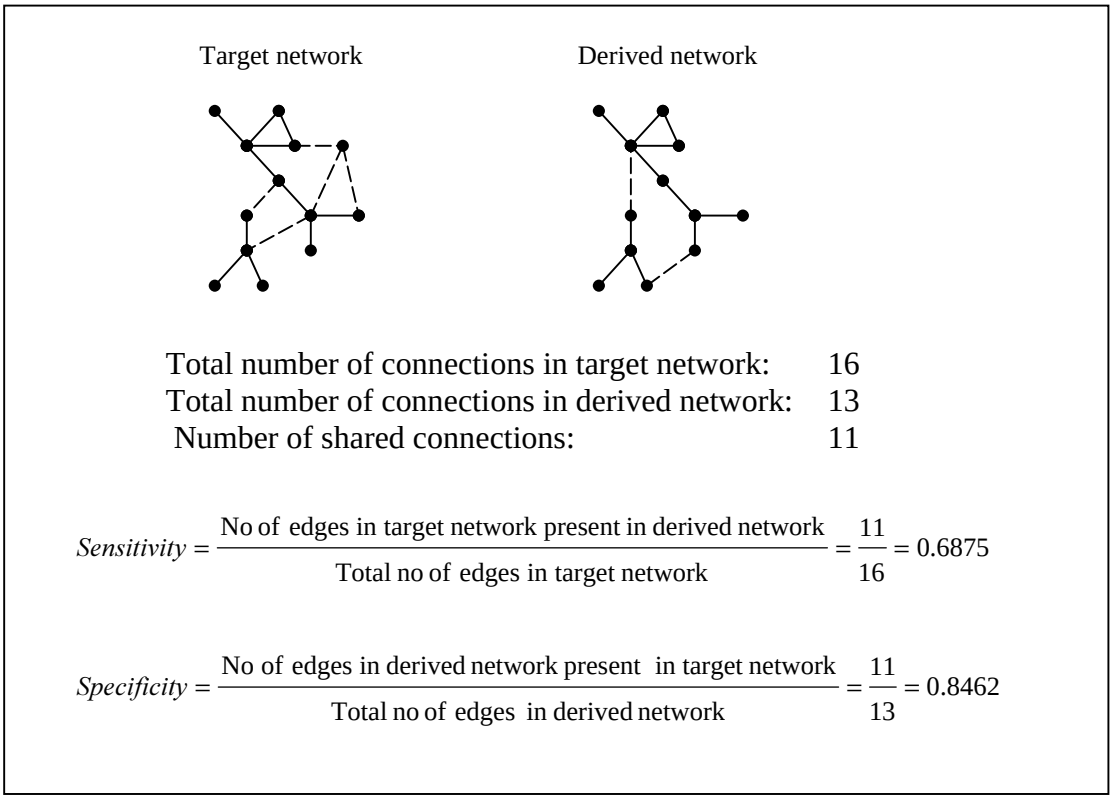


Figure 7. The sensitivity and specificity measurement developed by Ideker *et al.* (2000). In this example the sensitivity is 68.75% and the specificity 84.62%. The solid lines are edges common in the both networks and the dashed lines edges specific to respective network.

measurements are easy to apply to any method developed for reverse engineering and is not restricted to Boolean networks, which is a major advantage. Another advantage is its ease of use, because it requires no complex calculations.

The measurements are similar to the inferential power developed by Wessels *et al.* (2001) as those also measure the similiarity between the target network and the derived network (see also chapter 2.4.1). Those differ in the way that Wessels *et al*. (2001) measures the similarity between two gene regulation matrices, defined as $P_I(W_0, \hat{W}_0) = 0.5(1 + \rho(W_0, \hat{W}_0))$, where W_0 is the target gene regulation matrix, $\hat{W}_0$ is the derived gene regulation matrix and $\rho(W_0, \hat{W}_0)$ is the Pearson correlation.

4 Tested methods

In this chapter the three conceivable methods for deriving the genetic association network will be presented. In chapter 4.1 the correlation measurement approach will be presented and a correlation coefficient will be suggested as a possible method. In chapter 4.2 the Boolean network approach will be presented together with an algorithm based on this approach, which could be used as a method for deriving the network. And in chapter 4.3 the prior knowledge approach will be introduced and a method for based on this approach will be described.

4.1 Correlation measurement approach

In chapter 3.1.1 clustering was presented. It was stated that clustering of gene expression patterns could reveal genes with similar regulation, since genes that are co-regulated should show similar gene expression profiles and the clustering would therefore group these genes together (D'haeseleer *et al.*, 2000; D'haeseleer *et al*., 1999; Michaels *et al.*, 1998; Heyer *et al*., 1999). It is thought that gene expression data reflects the underlying genetic network and it should therefore reflect any type of interaction between two genes. The genetic network does not only consist of regulatory interactions, but also other types of interactions. For example, proteins that form complexes and enzymes interacting with substrates are other types of interactions (Weaver and Hedrick, 1997). If co-regulated genes should show similar gene expression profiles, then genes with other types of interactions should also show similar gene expression.

The distance matrix used to cluster genes with similar expression profiles is usually calculated by statistical methods, such as correlation coefficient measurements

(D'haeseleer *et al.*, 2000; D'haeseleer *et al*., 1999; Michaels *et al.*, 1998; Heyer *et al*., 1999). Genes that are highly correlated end up in the same cluster and if two gene expression profiles are well correlated then it is thought that the two genes either share the same regulatory inputs or that one of the genes regulates the other genes (D'haeseleer *et al.*, 2000; D'haeseleer *et al*., 1999; Michaels *et al.*, 1998; Heyer *et al*., 1999). This could be extended to if two genes are associated with each other, then those two should be well correlated. This statement or hypothesis can be used to derive the connections in the genetic association network, because if two genes are highly correlated then there should be an association between the two genes.

Pearson correlation (Heyer *et al*., 1999) can identify both positive and negative correlation between genes and is easy to implement, which are major advantages. The Pearson correlation coefficient lies between -1 and +1, where -1 means that the two genes have antagonistically expression profiles, +1 means that the genes have identical expression profiles and 0 means that their expression profiles share no similarity (see figure 8 and Heyer *et al.*, 1999).

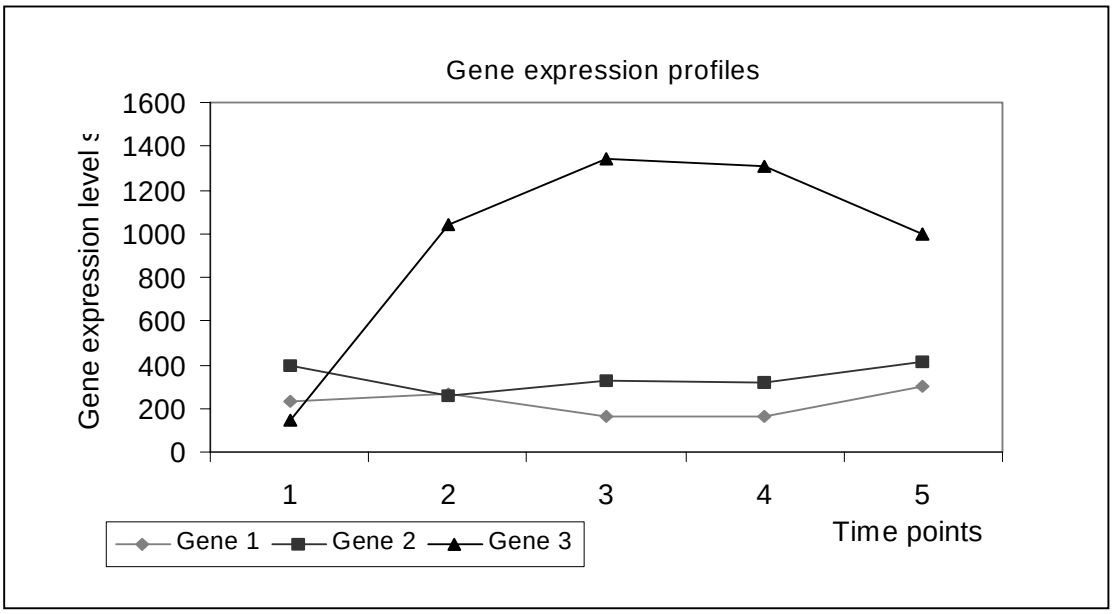


Figure 8. Gene expression profiles for three genes. Gene 1 and gene 2 have correlation 0.34, gene 1 and gene 3 correlation -0.40, and gene 2 and gene 3 correlation -0.56, according to the Pearson correlation measurement.

Pearson correlation coefficient is defined as:

$$C(x,y) = \frac{\frac{1}{n}\sum_i \left[(x_i - \overline{m}_x)(y_i - \overline{m}_y)\right]}{D(x)D(y)} \qquad \text{eq (3)}$$

where n is the number of time points, x_i and y_i are the gene expression levels of x and y at time i, $\overline{m}_x$ and $\overline{m}_y$ are the average expression levels for x and y, and $D(x)$ and $D(y)$ are the standard deviations for x and y, respectively.

4.2 Boolean network approach

This approach has both advantages and disadvantages (see chapter 3.1.2), but has useful qualities that can be used for deriving the genetic association network. In the approach the connections between genes are derived together with logical rules to explain the regulatory interactions. The approach could be used, in a first step, to derive the associations between genes, in the same way as the connections between genes are derived. If the approach derives the associations successfully, then the next step is to infer the logical rules between the associated genes and thereby more specific interactions.

Ideker *et al.* (2000) developed a method for reverse engineering of a Boolean network through perturbations, which was described in chapter 3.3.2. The first algorithm described in the paper, the Predictor, is a possible candidate for deriving the genetic association network, since it is used to derive the Boolean network from gene expression data. The algorithm will be implemented and tested in this study as a conceivable method for deriving the genetic association network. The second algorithm, which is used to propose an additional perturbation experiment, will not be

considered here for two reasons. First, there is no practical possibility to perform the perturbations required to get additional information, which is not just the case in this study but a current common problem. Second, it would be interesting to see how well the algorithm performs on the data available. If the algorithm performs well enough on available data, then there is no need for extra perturbations as was suggested by Ideker *et al.* (2000).

The pseudo code for the method is presented here (see also figure 9) and is based on the Predictor algorithm developed by Ideker *et al*. (2000)

1. Generate gene expression data and order them in a matrix, where row *i* represents the expression profile for gene *i* and column *j* represents the expression level for time point *j*.
2. Translate the gene expression matrix into a matrix with Boolean state symbols.
3. Identify each pair of columns (*j, k*) in the matrix represented by Boolean state symbols for gene *i* where the expression levels differ.
4. For each pair, find the set $S_{j,k}$ of all other genes whose expression levels also differ between two columns (*j, k*).
5. Identify the smallest set of genes S_{min} required to explain the observed differences over all pairs *(j, k)*. This will generate the minimal set covering.
6. Gene *i* will have an association with the genes covered by the minimal set.

The algorithm could produce several possible minimal sets and thereby several possible candidates for the genetic association network. Evaluating all the possible candidate solutions reported by the algorithm could be time consuming and for this reason some heuristics will be implemented, the method will only report the first minimal set it finds and ignore other possible minimal sets.

4.3 Prior knowledge approach

This is a novel approach, proposed in this thesis, where prior knowledge from a known genetic network will be used to derive the genetic association network for a related organism. For simplifying matters in describing the method, following definitions will be used:

- *source organism*: the related organism with a known genetic network, from which prior knowledge will be used
- *target organism*: the organism for which the genetic network will be derived

The known genetic network from the source organism will be mapped to the target

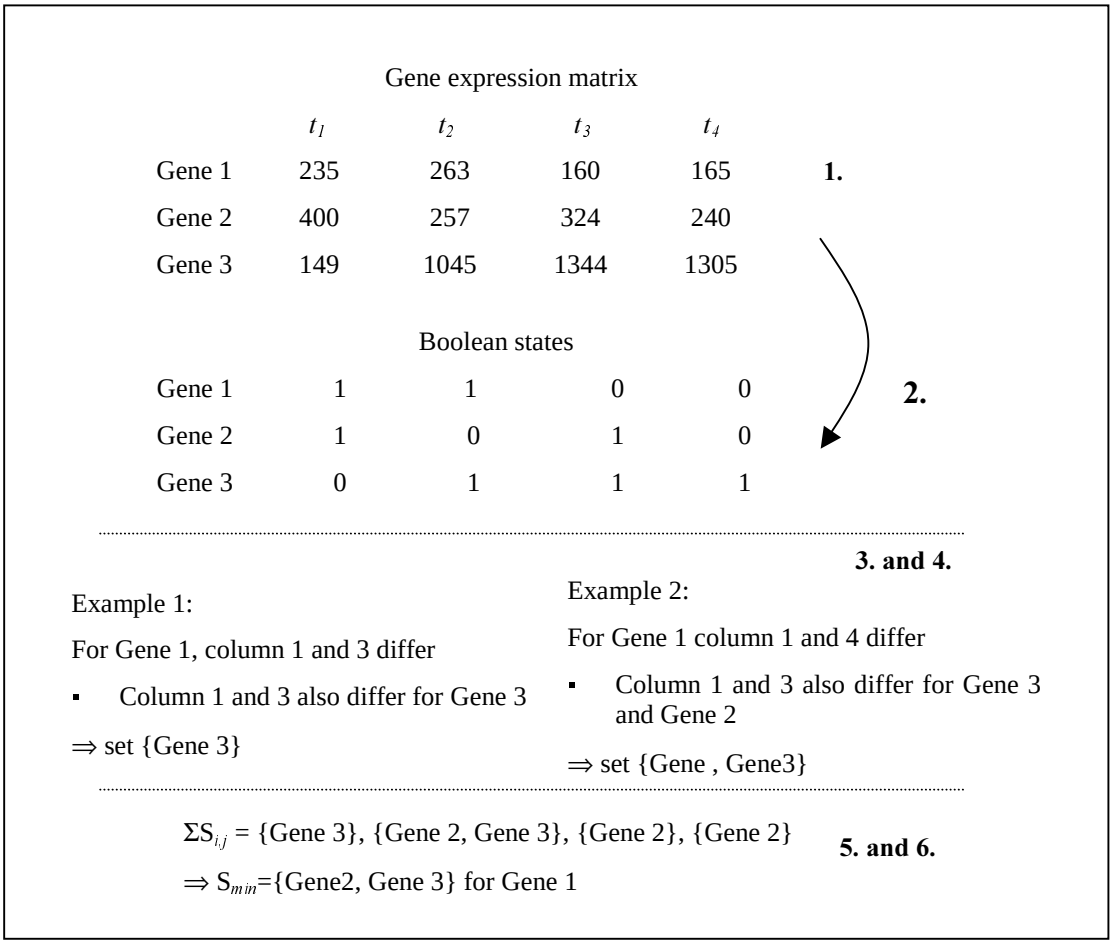


Figure 9. Illustration of the algorithm used to derive the associations between genes, based on Ideker *et al.* (2000). The numbers relate to the numbers in the pseudo code description of the algorithm in the text.

organism by using homolog matching (figure 10). This could be done for two related organisms, because the organisms are related they share a set of homologous genes, i.e. the genes have diverged from a common ancestor (Attwood and Parry-Smith, 1999). This means that the organisms share genes with similar function and sequence and thereby similar interactions (Attwood and Parry-Smith, 1999).

For identifying the homologs, an algorithm that makes pairwise comparisons between all genes from the source organism and each gene from the target organism can be used (see figure 11 a and Attwood and Parry-Smith, 1999). There are some different algorithms that may be used in the homolog finding, for example Smith-Waterman, BLAST or FastA (Attwood and Parry-Smith, 1999). The algorithms differ in some ways, the Smith-Waterman algorithm is for example an exact algorithm, while the BLAST and the FastA algorithm are heuristic. The Smith-Waterman is more preferred since it is an exact algorithm, but has the disadvantage of being computationally costly. For this reason the heuristic algorithms are used instead, especially when examining a large number of sequences. The algorithms perform pairwise comparisons to measure the similarity between two genes; the higher the similarity between the two genes, the higher the probability that the genes are homologs.

Huynen *et al*. (1998) used the Blast algorithm to find homologous genes in their

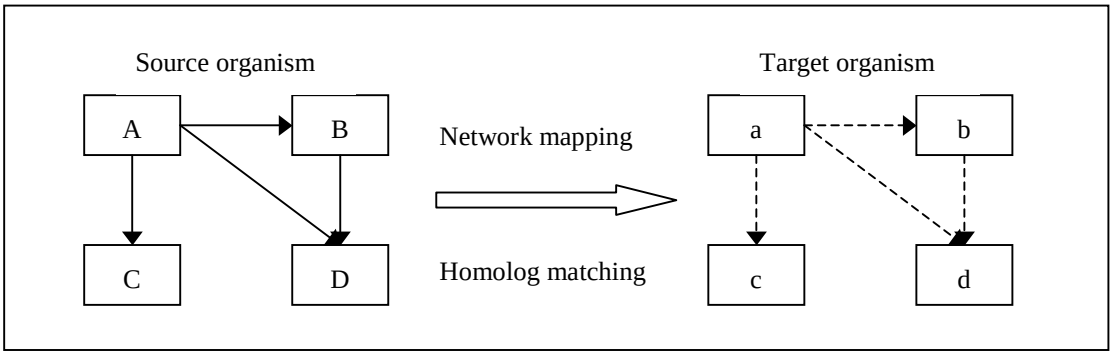


Figure 10. The homologs will be used to map the interactions from the known genetic network to the target organism.

study. They used the *E* value in the algorithm to distinguish genes with significant similarity (see figure 11 b) (Attwood and Parry-Smith, 1999; Huynen *et al*., 1998). The *E* value is used as an indication of homologous genes and for this reason the Blast algorithm (Durbin *et al*., 1997) will be used here. Two genes are homologs if the similarity is significant, which was defined as $E < 0.01$ by Huynen *et al.* (1998).

In pairwise comparison there are usually several genes in the target organism that are significant similar to a gene in the source organism, i.e. more than one gene in the target gene scores an *E* value less than 0.01 for a gene in the source organism. In these cases the gene from the target organism with the lowest *E* value will be chosen, since the higher the similarity between the two genes, the higher the probability that the genes are homologs (figure 11 c).

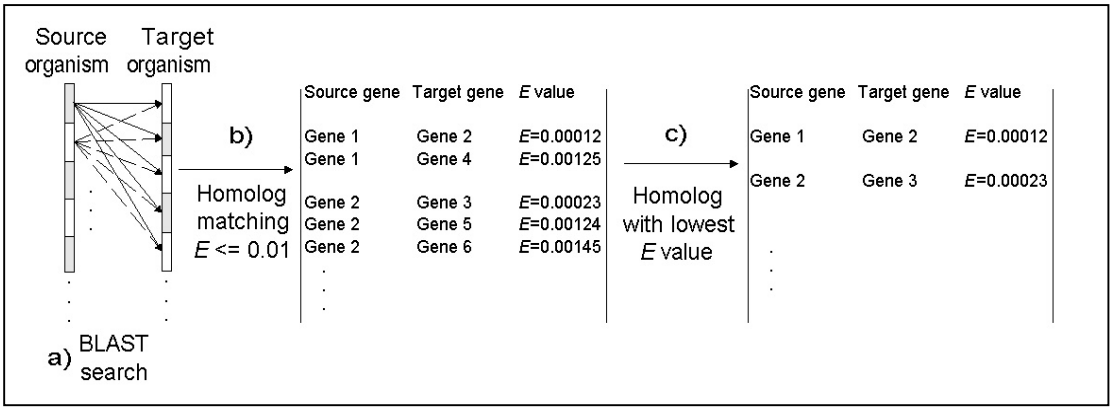


Figure 11. Method for finding the homologs between the source organism and the target organism. In a) the BLAST algorithm makes pairwise comparisons between all genes in the source organism and each gene in the target organism, in b) homologs with an *E* value less than 0.01 are reported and in c) the homolog with lowest *E* value is chosen for the mapping.

5 Evaluation method

In this chapter the testing and evaluation of the methods is described. The methods were tested on trusted data and the data used is presented in chapter 5.1. In chapter 5.2 the experimental part is described, i.e. the deriviation of genetic association networks using the three methods and in the chapter 5.3 the evaluation procedure is described together with the chosen evaluation measurement.

5.1 Testing on trusted data

The methods were tested on trusted data, i.e. verified by literature or experimentally, and not on hypothetical or simulated data as many other methods have been tested on (Liang *et al.*, 1998; Ideker *et al.*, 2000; Szallasi and Liang, 1998; Maki *et al.,* 2000; Akutsu *et al.*, 2000; Weaver *et al.* 1999). This means gene expression data from real experiments were used for the correlation measurement approach and the Boolean network approach. For the prior knowledge approach, sequences for both the source and the target organism were collected and used for the homolog matching. For the network mapping a trusted network for the source organism was used.

The advantage of using hypothetical data is that all information about the genetic network is known and a comparison with the correct answer can be done (Mendes, 1999). But the reason for using trusted data is the lack of methods tested on trusted data and the fact that it reflects the information content received from real experiments. Methods that are tested on hypothetical or simulated data are usually evaluated against a hypothetical genetic network, which is considered to be the correct answer (Liang *et al.*, 1998; Ideker *et al.*, 2000; Szallasi and Liang, 1998; Maki *et al.,* 2000; Akutsu *et al.*, 2000; Weaver *et al.* 1999). When testing on trusted data the

evaluation has to be on trusted data, since this is probably the only way to know the correct answer.

For this study the organism *Saccharomyces cerevisiae* (baker's yeast) was chosen as target organism, i.e. the genetic association network was derived for this organism using the chosen methods. The reason for using baker's yeast is that the whole genome has been sequenced and all the ORFs for the organism are known (Bairoch *et al*., 2000). There is also gene expression data available for the organism (Cho *et al.*, 1998). The gene expression data was used for the correlation measurement approach and the Boolean network approach.

For the prior knowledge approach sequences were needed for the homology search. Protein sequences for *S. cerevisae* were collected from the database Swissprot[1] (Bairoch *et al*., 2000). As source organism *Drosophila melanogaster* (fruit fly) was chosen, because many of the fruit fly's proteins are sequenced and there is a trusted network for about 1000 genes for this organism. Protein sequences for *D. melanogaster* were collected from the database NCBI Batch Entrez[2] (Wheeler *et al*., 2001). About 500 genes in the trusted network had a sequenced protein, which means about half of the genes in the network cannot be used in the homolog matching and the mapping. The trusted network is from the database FlyNets[3] (Sanchez *et al*., 1998). FlyNets contains information of associations between genes in fruit fly during its development. The database stores information for two genes if they have a DNA-protein, mRNA-protein or protein-protein association. The collected protein sequences together with the trusted network from FlyNets were used for the

1. http://www.expasy.ch/sprot/
2. www.ncbi.nlm.nih.gov:80/Entrez/batch.html
3. http://gifts.univ-mrs.fr/FlyNetss/FlyNetss_home_page.html
4. http://www.genome.ad.jp/kegg/regulation.html

homology matching and the network mapping between baker's yeast and fruit fly.

In the evaluation of the methods, a trusted network was used as the correct answer. The trusted network was collected from the database KEGG[4] (Ogata *et al*., 1999). The database contains a variety of information about genes and their activities. One component in KEGG is its pathway database, which consists of graphical diagrams illustrating cellular processes. The pathway database is divided into two catalogues, namely metabolic pathways and regulatory pathways. The trusted network was extracted from the regulatory pathways, illustrating a genetic hybrid network for *S. cerevisiae* during the cell cycle.

5.2 Experiments

Since the conceivable methods were evaluated against a trusted network for the target organism, only these genes had to be considered in the testing and evaluation of the methods. The trusted network from KEGG[4] (Ogata *et al*., 1999) contains 104 genes and in the gene expression data 101 of the 104 genes in the KEGG network were reported (Cho *et al.*, 1998). These 101 genes were used in the correlation measurement approach and the Boolean network approach (appendix 2), but for the prior knowledge approach all the protein sequences for both organisms were used. The reason for why all sequences were used, and not only the proteins that were represented in the KEGG network, is that the fruit fly network only contained the associations for about 500 genes and there was a strong possibility that the network would not overlap the network from KEGG (also see chapter 5.3). This would mean that a lot of the derived associations from the homolog mapping would be excluded in

the evaluation, and it could be of interest to examine all the associations derived from the method.

5.2.1 Correlation measurement approach

Pearson correlation (eq. 3) was calculated between all pairs of the 101 gene expression profiles and the cut off |0.7| was used to distinguish associated genes, i.e. if the correlation between two genes is more than 0.7 or less than -0.7 then the two genes were considered to have an association. The cut off chosen was based on the article by Heyer *et al.* (1997), where a visual inspection of their experiment lead to the assumption that the cut off |0.7| would be appropriate.

5.2.2 Boolean network approach

For the Boolean network approach the algorithm described in chapter 4.2, based on the Predictor by Ideker *et al.* (2000), was implemented. The Boolean state symbols $x'_{i,j}$ were calculated using:

$$x'_{i,j} = \begin{cases} 1 & if \;\; x_{i,j} > \overline{m}_i \\ 0 & otherwise \end{cases} \qquad \text{eq (4)}$$

where $x_{i,j}$ is the expression level for gene i at time point j and $\overline{m}_i$ is the average gene expression level for gene i. If $x_{i,j}$ is greater then the average expression level $\overline{m}_i$ for gene i, then gene i is on ('1'). If $x_{i,j}$ is less then or equal to the average expression level $\overline{m}_i$ for gene i, then it is off ('0'). Each gene i will have an association with the genes covered by the minimal set S_{min} (see chapter 4.3).

In deriving the minimal set S_{min} some heuristic was used - only the first minimal set found was considered. The heuristic was implemented as follows. A matrix m was generated for each gene i represented in the Boolean state matrix (figure 12, a). The

columns in the matrix *m* represent each pair of columns in the Boolean state matrix which differ for gene *i*, and the rows represent all the other genes in the Boolean state matrix which also differ for at least one pair of columns as gene *i* does. For each gene in the matrix *m* a '1' was marked if there was a difference between two columns in the Boolean state matrix for that gene and a '0' otherwise (figure 12, b). This was done for all the genes in the matrix *m*. For each gene in the matrix *m* the row sum was calculated (figure 12, c) and the gene with the highest summation was added to the minimal set covering (figure 12, d). In all the columns in the matrix *m* for which the gene with the highest summation had a '1' was changed to '0' (figure 12, e). For each gene in the matrix *m* the row sum was calculated again and so on. The process was repeated until all the cells in the matrix *m* was marked '0', and thereby the minimal set S_{min} was found.

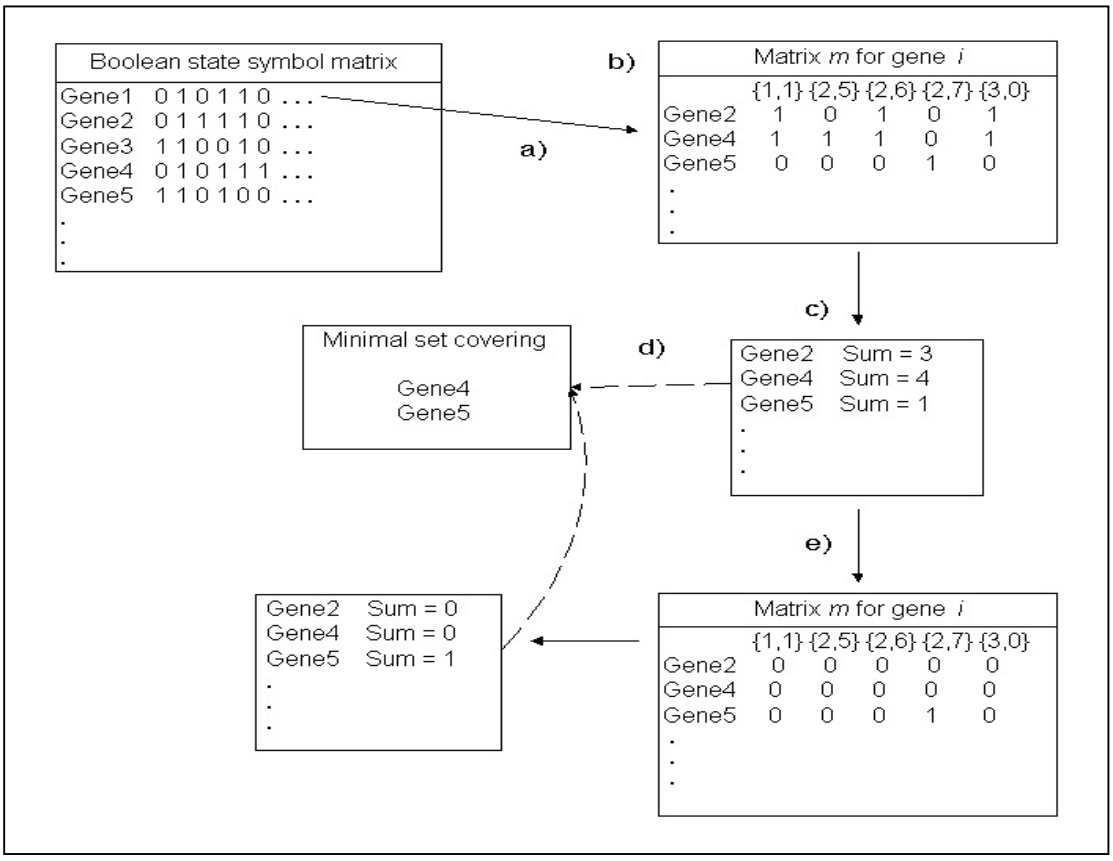


Figure 12. Illustration of the implemented heuristic in the Boolean network approach.

5.2.3 Prior knowledge approach

The homology matching was performed as described in chapter 4.3 between baker's yeast and fruit fly, using the BLAST algorithm and with an *E* value less than 0.01 to identify valid homologs (Durbin *et al*., 1997). If there were several valid homologs reported for a protein, then the homolog with the smallest *E* value was used in the mapping. If two proteins for the target organism received the same valid homolog with the smallest *E* value, then those were excluded from the mapping. This was because there was no way of knowing which protein is the "right" homolog.

The representation of associations, and thereby the mapping, was rather simple. If two genes were associated with each other in the fruit fly network, they were simply paired together. If there was a homologous protein for each of their respective gene in the pair, then the association was mapped to the target network.

5.3 Evaluating results

Paradoxically, using trusted data give problems in evaluating the result from the proposed algorithm. When testing on hypothetical data or simulated data the correct answer is always known. In evaluating methods tested on trusted data this is not the case. Using a trusted network limits the evaluation of the derived genetic association network, because trusted genetic networks are sparse. In this study a trusted genetic hybrid network from the KEGG[4] (Ogata *et al*., 1999) database was collected. The KEGG network contains information of many types of interactions, but because in this study only associations between genes are of interest the network was redrawn as a genetic association network (appendix 1). The redrawn KEGG network was used as the correct answer in the evaluation of the methods.

Since there was a strong possibility that not all the associations derived by the prior knowledge method would be covered by the KEGG network, other evaluation sources had to be taken into consideration. For this reason the mapped associations were also verified against the interactions stored in the Yeast Proteome Database[5] (Costanzo *et al.,* 2000) and the journal database PubMed[6] (Wheeler *et al*., 2001). If two genes in an association were mentioned together in an article then the association was reported, together with the number of articles. This article search was only used as an indicator, to find out if more of the derived associations could be explained than had been done with the two previous evaluations.

For this appraoch the percentage of associations that could be verified by the KEGG network and the Yeast Proteome Database together with the associations reported from the the PubMed search, were used as indications on the performance of the method. The reason for using the measurement only as an indicator is because it is difficult to evaluate the performance of the method, since the entire network for both organisms are not known and therefore may not cover each other properly. For example, only 500 genes of the fruit fly's entire genome was represented in the network, which means that many of associations are missing in the network. The associations in the fruit fly network are not necesserly investigated in baker's yeast and the opposite.

For the correlation measurement approach and the Boolean network approach the sensitivity and specificity measurements developed by Ideker *et al.* (2000), also described in chapter 2.4.2, was used. The evaluations were not totally satisfactory for these two methods either, and a PubMed search was also made, for the same reason as for the prior knowledge approach. An additional evalution measurement was also

5. http://www.proteome.com/databases/index.html

6. http://www4.ncbi.nlm.nih.gov/PubMed/

used, where the distance $\delta(i,j)$ between two genes i and j in an association were reported. This was done by counting the number of edges in the trusted network between the two genes i and j, to indicate how many steps in the trusted network are needed to explain the association.

The overall problem structure for this work is illustrated in figure 13.

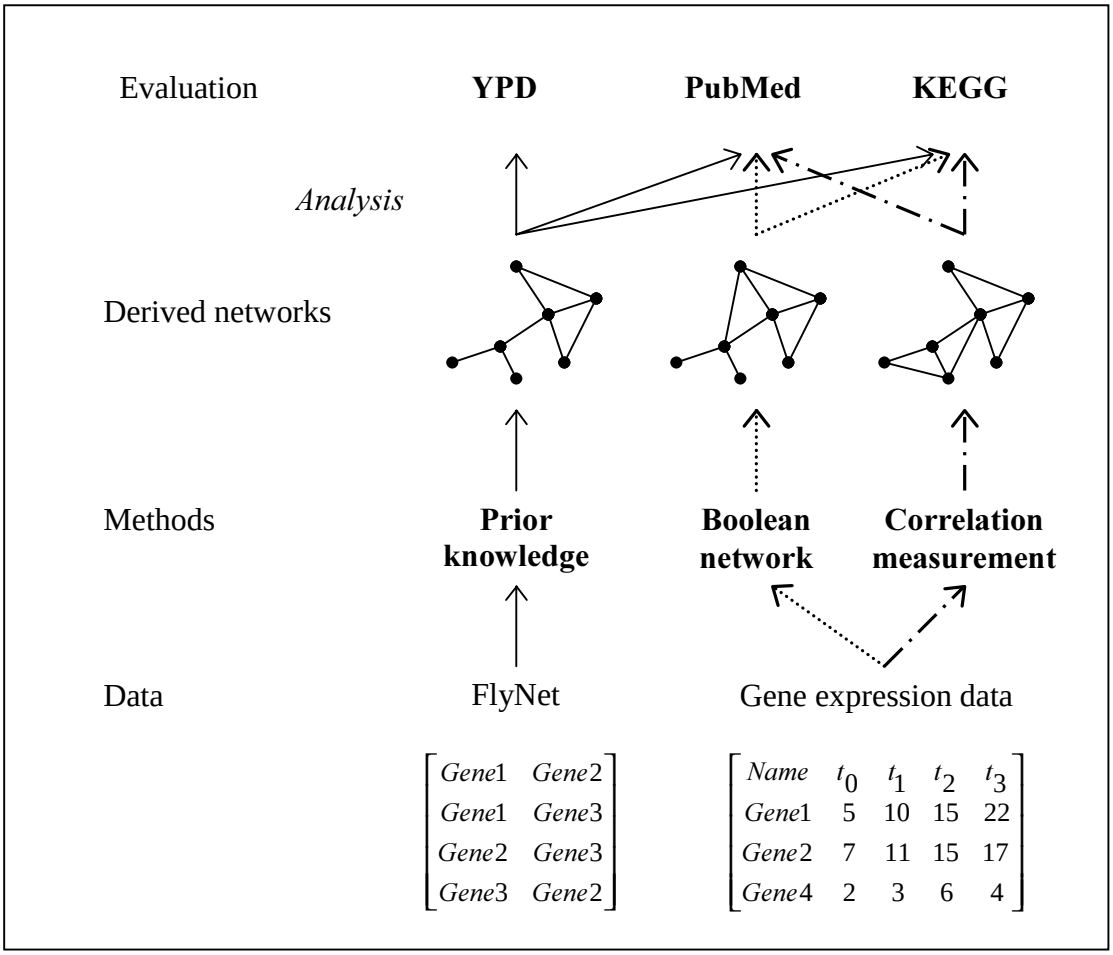


Figure 13. Summary of the problem structure.

6 Results and analysis

The results from the different methods are presented in this chapter together with an analysis of their respective performance; in chapter 6.1 the prior knowledge approach, in chapter 6.2 the correlation measurement appraoch and in 6.3 the Boolean network approach. In chapter 6.4 the correlation measurement and the Boolean network appraoch are compared to analyse if either of the two methods performs better, and also combined for the possibility to increase the overall performance.

6.1 Prior knowledge

In the prior knowledge approach a known genetic association network for *Drosophila melanogaster* was used, which was collected from the database FlyNets[3] (Sanchez *et al.*, 1998). This network contained about a thousand genes, for about half of which (521) there was a sequenced protein. This reduced network was used as prior knowledge in the approach. In the homology matching 164 of the 521 proteins had a valid, unique homolog in *Saccharomyces cerevisiae* (table 1). Valid here means the *E* value was less than 0.01 and the homolog with the smallest *E* value for this protein, and unique means that no other protein had the same valid homolog. For 112 proteins there was no homolog found and for 245 proteins there was a valid homolog, which was not unique (table 1). In the mapping only the proteins with a valid *and* unique homolog was used, which means about 31% of the genes in the network could be used in the mapping.

There were 1206 associations in the reduced network and from these 109 (9%) associations could be mapped to the *S. cerevisiae* network (table 1 and appendix 3 a). One reason for the small number of mapped associations is that in each association

there are two genes that must have a valid, unique homolog. In many associations in the known network none of the genes or only one of the genes in the association had a valid, unique homolog and therefore could not be mapped (table 2). This reduces the number of mapped associations considerably, but the network is still quite large and will give valuable information about the unknown network.

The mapped associations were first verified against the Yeast Proteome Database[5] (YPD) (Costanzo *et al.*, 2000) and the trusted association network for *S. cerevisisae* (appendix 1), originally from the KEGG[3] database (Ogata *et al*., 1999). The trusted network could not verify any of the mapped associations and only 11 of the associations were verified by YPD, which is about 10% of the mapped associations (table 3).

Since many of the mapped associations could not be verified by these two sources a search against the PubMed[6] journal database (Wheeler *et al*., 2001) was done, to indicate if more mapped associations could be verified. If two genes in a mapped association were mentioned together in an article, then it was reported as a possibly verified association. Two genes mentioned together in an article does not prove that those have an association. The genes could be in the same article for other reasons than for some interacting behaviour. Since there was no time to investigate the articles

Number of genes in *D. melanogaster* network	521
Number of associations in D. melanogaster network	1206
Number of genes in *D. melanogaster* network with valid and unique homolog in *S. cerevisiae*	164 (31%)
Number of genes in *D. melanogaster* network with a valid, but not a unique, homolog	112 (22%)
Number of genes in *D. melanogaster* network with no homolog	245 (47%)
Number of associations mapped to *S. cerevisiae* network	109 (9%)

Table 1. Statistics of the network mapping between *D. melanogaster* and *S. cereviasae*.

more closely, the PubMed search should only be considered as an indication of associations that might be verified. The PubMed search resulted in 9 reported associations, which all already were verified by the YPD (table 3). In this case the PubMed search did not generate any new information.

A simple annotation search in the YPD for the cellular role and biochemical function of the genes revealed that in 33 of the associations at least one of the genes had an unknown cellular role, unknown biochemical function or both. This indicates that those genes are not well explored and their interactions are more or less unknown. This may explain for about 30% of the mapped associations why those could not be verified (table 3).

In summary, of the mapped associations 11 could be verified and 33 had a reasonable explanation to why those could not be verified, which is about 40% of the mapped associations. This leaves 65, or 60%, mapped associations that need further investigation. Those should not be discarded immediately as false positives. A more thorough investigation should be done, since those could be true positives as well.

For fruit fly not all the proteins are sequenced, which affects the performance of the approach. As already mentioned, this reduced the trusted genetic network to about half its size. If a gene, or its respective protein as in this case, in the known genetic

D. melanogaster association		*S. cerevisae* homologs (valid and unique)		Mapped?
Gene 1	Gene 2	Homolog 1	Homolog 2	Y/N
Ash1	E(z)	Set2	Set1	Y
Ash1	Scr	Set2	-	N
Ash1	Ubx	Set2	YIL105C	Y
ElB	ElA	-	-	N
Fu	Ci	Ypk2	Mig2	Y

Table 2. Example from the network mapping, where an association can only be mapped if both genes in the association have a valid, unique homolog.

network is not sequenced, it is not possible to find a homolog to this gene in the target organism. This means the associations of the missing gene cannot be mapped to the target network. This, as shown in this study, can have very limiting effects on the performance of the method.

In addition, as also has been shown, many genes in the target organism will not end up with a unique homolog, but will share a homolog with many other genes. Here, these have not been used in the mapping, which have also had a major effect on the performance. Huynen *et al.* (1998) had an additional criterion to distinguish homologs, which was not used in this study. The criterion was that the segment of similarity between two sequences must cover more then 60% of at least one of the sequences in the comparison. This criterion may be used to distinguish additional homologs. For example, if two genes in the target organism receive the same valid homolog and for one of the genes the similarity is less then 60%, then this gene can be discarded, and the other gene can be used in the mapping. For the discarded gene the next homolog in the list may have a similarity that covers more than 60% of the sequence, which means that this gene can also be used in the mapping. This has to be further investigated.

Number of associations mapped to *S. cerevisiae* network	109
Number of associations verified by the known association network	0
Number of associations verified by YPD	11 (10%)
Number of associations reported by PubMed search	9
Number of associations where one or both genes have an unknown cellular role, biochemical function or both	33 (30%)
Percentage of associations that are either verified or have a reasonable explanation for why those could not be verified	44 (40%)

Table 3. Statistics of verification of the mapped associations.

6.2 Correlation coefficient

Pearson correlation was calculated between all pairs of gene expression profiles represented in the known association network for *S. cerevisiae* (appendix 1). The cut off |0.7| was used to distinguish associated genes (appendix 3, b). Pearson correlation generated 203 associations and of these 28 associations was verified by the known association network (table 4). The known association network consists of 372 associations, which gave a sensitivity of 7.5% and a specificity of 13.8%.

To evaluate the Pearson correlation method, these levels of sensitivity and specificity should be compared to what can be expected by chance. A fully connected network consisting of 101 genes would generate (101*100/2) 5050 associations (table 4). In the trusted network there are only 372 associations, which is 7.4% of the fully connected network (372/5050). The Pearson correlation generated 203 associations. If one would generate 203 associations by chance among the 5050 possible associations, one would get about 15 correct associations (0.074*203). This would give a sensitivity of 4.0% and a specificity of 7.4% (table 4). This indicates that the Pearson correlation is slightly better than chance in deriving correct associations.

Since the result from this evaluation showed that many of the associations derived by the correlation measurement could not be verified by the known network, a second

Number of associations in the trusted network		372
Number of associations generated by Pearson correlation		203
Number of correct associations derived by the Pearson correlation		28
Number of correct associations derived by chance	203 * 0.074% =	15
Sensitivity – Pearson correlation	28 / 372 =	7.5%
Sensitivity – chance	15 / 372 =	4.0%
Specificity – Pearson correlation	28 / 203 =	13.8%
Specificity – chance	15 / 203 =	7.4%

Table 4. Pearson correlation result.

evaluation was made by calculating the distance $\delta(i,j)$ between two genes *i* and *j* in a derived association. The $\delta(i,j)$ for the derived associations ranged from one up to ten arcs (table 5).

In the KEGG network there were some interactions that could not be interpreted. For a derived association with a gene that had a non-interpret edge, this evaluation could not be made. This was the case for 18 (9%) of the derived associations (table 5).

There were 28 (14%) associations with $\delta(i,j) = 1$, and for $\delta(i,j) = 2$ there were 26 (13%) associations. The largest group of associations was for $\delta(i,j) = 5$, which was 34 (17%) associations. The rest of the evaluation is presented in table 5. The percentage of arcs was also calculated without the non-interpret edges, which increased the percentage slightly for all groups. This is also presented in table 5 and in addition in figure 14.

The cumulative percentage was also calculated (excluding non-interpret edges), which showed that $\delta(i,j) = 4$ was are needed to cover 50% of the derived associations, and $\delta(i,j) = 6$ was needed to cover 80% of the derived associations (table 5 and figure

	Number of arcs, $\delta(i,j)$											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**?**	**Total**
Asscociations	28	26	24	26	34	19	16	10	1	1	18	203
Percentage #1	14%	13%	12%	13%	17%	9%	8%	5%	0,5%	0,5%	9%	100%
Percentage #2	15%	14%	13%	14%	18%	10%	9%	5%	1%	1%		100%
Cum perc #2	15%	29%	42%	56%	74%	84%	93%	98%	99%	100%		100%

Average	4.1	**Median**	4

Table 5. Statistics for the distance evaluation for the Pearson correlation, $\delta(i,j)$ was counted for every association. The row 'Associations' is the number of associations for each $\delta(i,j)$. '?' mean associations with a non-interpret edge. The row 'Percentage #1' is the percentage of arcs including associations with a non-interpret edge and the row 'Percentage #2' is the percentage of arcs excluding associations with a non-interpret edge. The row 'Cum perc #2' is the cumulative percentage of arcs excluding associations with a non-interpret edge.

14). The average $\delta(i,j)$ is 4.1 and the median is 4.

This evaluation indicates that two genes in a derived association could be rather far away from each in other the network (up to $\delta(i,j) = 10$). The average of 4.1 arcs show that a strong correlation between two gene expression profiles does not imply that the genes have a direct association, i.e. $\delta(i,j) = 1$. In fact, there could be as much as ten arcs between two well-correlated genes. The majority of associations are covered by eight edges and the distribution of the distances is centred on the shorter distances. This means the probability of two genes in an association have a distance of less then $\delta(i,j) = 4$ increases. $\delta(i,j) = 4$ is still a rather long distance between two well-correlated genes, but is much better than a distance of $\delta(i,j) = 10$ and might be tangible.

The distance was also plotted against the Pearson correlation for the associations, which shows the correlations are well spread over the distances (figure 15). This means there is not a higher correlation for those with shorter distance, which would have been preferable. If this had been the case, one could have increased the cut off and got better performance of the method and thereby obtaining a distribution centred

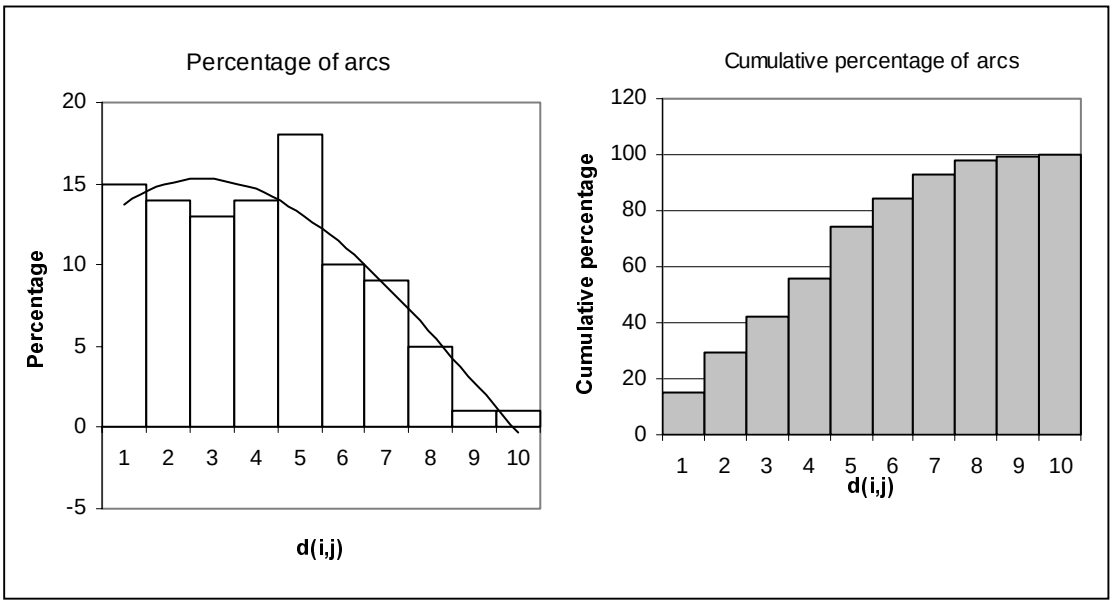


Figure 14. The plot on the left shows percentage of derived associations by Pearson correlation with a given distance $\delta(i,j)$. The plot on the right is the cumulative percentage of derived associations with a distance $\delta(i,j)$.

on shorter distances. And also, if all the true positives had the highest correlation, an increased cut off would exclude more false positives and thereby increase the performance. As it is now, an increased cut off would have the opposite effect. Decreasing the cut off would generate more true positives, but would also generate more false positives and thereby not increase the overall performance.

As was the case with the prior knowledge approach, many of the mapped associations could not be verified by the known association network. For the same reason as in that case a search against the journal database PubMed[6] (Wheeler *et al*., 2001) was done, to indicate if more of the mapped associations could be verified. The PubMed search reported 62 associations, for which 24 was already verified by the known association network (table 6). This means that an additional 38 derived associations might possibly be verified.

If all the associations reported by PubMed could be verified this would mean that both the sensitivity and the specificity increases, but it would affect the specificity much more than the sensitivity (table 6). The specificity would increase to 15% and the sensitivity to 30%. But again, one has to bear in mind that the presence of an article does not prove that there is an association between the two genes. Those could

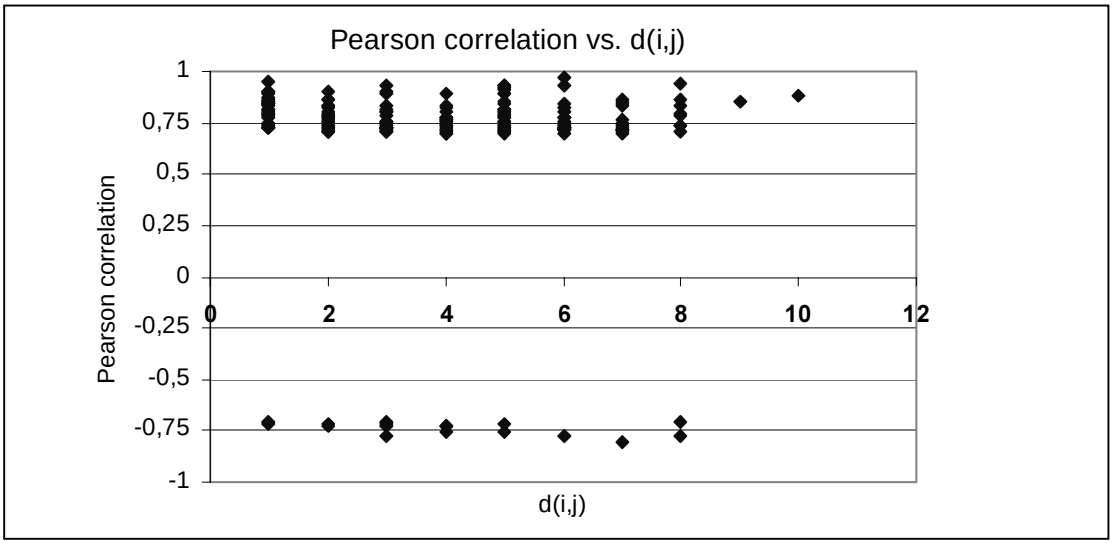


Figure 15. Pearson correlation plotted against the distance $\delta(i,j)$.

be mentioned together for other reasons and even if those were true associations the performance of the method is still rather poor.

6.3 Boolean approach

The associations were derived by the Predictor algorithm described in chapter 4.2, from the gene expression profiles for the genes in the trusted association network, originally from KEGG (Ogata *et al*., 1999). The algorithm derived 323 associations for which 21 could be verified by the known network (table 7 and appendix 3c). The known association network consists of 372 associations, which gave a sensitivity of 5.6% and a specificity of 6.5%.

For the Predictor algorithm the number of correct associations expected by chance was calculated, and turned out to be 23.7. This would give a sensitivity of 6.4% and a specificity of 7.4% (table 7). This indicates that the algorithm is slightly worse than chance and that chance would be a better algorithm in deriving correct associations.

When examining the derived associations it was noticed that one gene, Lte1, was reported very often, nearly a hundred times. It is a fact that a lot of genes are known to be connected to many other genes (Sanchez *et al.*, 1998, Weaver *et al.*, 1997), which could be the case here, but nearly a hundred connections seemed somewhat exaggerated. In the known network, the gene was only connected to one other gene.

Number of associations reported by PubMed		62
Number of associations already verified by the trusted network		24
Additional number of associations reported by PubMed	62-24 =	38
Sensitivity – including PubMed associations	62 / (372+38) =	15%
Specificity – including PubMed associations	62 / 203	31%

Table 6. Statistics of the PubMed search for the associations derived by Pearson correlation.

One possible explanation could be that it was due to the heuristics in the algorithm. The gene Lte1 was listed very early in the file and when choosing the first minimal set found, the genes listed early have a higher probability to be in a derived association. The effect of the heuristics was not examined here, but should be done to get a clearer picture of the effects this issue has.

The effect on the sensitivity and specificity of this gene was examined, in an attempt to investigate if the sensitivity and specificity would increase, and if it did, how much those would increase. The derived associations including the Lte1 gene were excluded, which meant the single true positive for this gene was also excluded. This resulted in 225 derived associations, for which 20 were verified by the known association network (table 7). This result gave a sensitivity of 5.4% and a specificity of 8.9%. This means the sensitivity decreased and got even more worse than chance. The specificity increased to get slightly better than chance. This modification did not increase the algorithm's performance, even if the specificity got better than chance it is still a very poor result.

Number of associations in the trusted network		372
Number of associations generated by the Predictor algorithm		323
Number of correct associations derived by the Predictor algorithm		21
Number of correct associations derived by chance	323 * 7.4% =	23.7
Sensitivity – Predictor algorithm	21 / 372 =	5.6%
Sensitivity – chance	23.7 / 372 =	6.4%
Specificity – Predictor algorithm	21 / 323 =	6.5%
Specificity – chance	15 / 203 =	7.4%
Number of associations without the gene Lte1		225
Number of correct associations derived by the Predictor algorithm, without the associations for the gene Lte1	21 - 1 =	20
Sensitivity – without the gene Lte1	20 / 372 =	5.4%
Specificity – without the gene Lte1	20 / 225 =	8.9%

Table 7. Statistics of the result for the Predictor algorithm.

The distance $\delta(i,j)$ for the associations was also calculated, in the same way as for the associations derived by Pearson correlation (see chapter 6.2). Here, the number of arcs ranged from one up to nine (table 8).

As in the correlation measurement method, there are some derived associations containing a non-interpret edge, for which the evaluation could not be made. This was the case for 36 (11%) of the derived associations (table 8). There were 21 (7%) associations with $\delta(i,j) = 1$ and 17 (5%) associations with $\delta(i,j) = 2$.

Here, the largest group of associations was also $\delta(i,j) = 5$, which was 62 (19%) of the derived associations. The rest of the evaluation is presented in table 8 and the percentage of arcs was also calculated without the non-interpret edges, which increased the percentage slightly for all groups except the first. This is also presented in table 8 and in addition in figure 16.

The cumulative percentage was calculated with non-interpret edges excluded, which showed that $\delta(i,j) = 4$ are needed to cover 50% of the derived associations, and $\delta(i,j) = 6$ arcs are needed to cover 80% of the derived associations (table 5 and figure

	Number of arcs, $\delta(i,j)$										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**?**	**Total**
Asscociations	21	17	60	44	62	42	26	7	8	36	323
Percentage #1	7%	5%	19%	14%	19%	13%	8%	2%	2%	11%	100%
Percentage #2	7%	6%	21%	15%	22%	15%	9%	2%	3%		100%
Cum perc #2	7%	13%	34%	49%	71%	86%	95%	97%	100%		100%

Average	4.5	**Median**	5

Table 8. Statistics for the distance evaluation, $\delta(i,j)$ was counted for every association. The row 'Associations' is the number of associations for each $\delta(i,j)$. '?' mean associations with a non-interpret edge. The row 'Percentage #1' is the percentage of arcs including associations with a non-interpret edge and the row 'Percentage #2' is the percentage of arcs excluding associations with a non-interpret edge. The row 'Cum perc #2' is the cumulative percentage of arcs excluding associations with a non-interpret edge.

16). The average $\delta(i,j)$ is 4.5 and the median is $\delta(i,j)$ is 5 arcs. This method also showed that two genes derived by the method could be rather far apart from each other in the trusted network.

In this case the majority of associations are also covered by eight edges, but the $\delta(i,j)$ values of associations seem to have a normal distribution centred around five arcs. This has the effect that the probability of being two or seven arcs between the two genes in the derived association is the same, which is not satisfactory. Here, one cannot state that it is more likely that there are two arcs rather than seven arcs between the two genes in the derived association, as one might for the Pearson correlation.

In this approach, as for the two other approaches, many of the mapped associations could not be verified by the known association network. A search against the journal database PubMed[6] (Wheeler *et al*., 2001) was done, to indicate if more of the mapped associations could be verified. Here, the PubMed search reported 56 associations, for which the trusted network had already verified 19 (table 9). If all the reported associations from PubMed could be verified this would result in a sensitivity of 14%

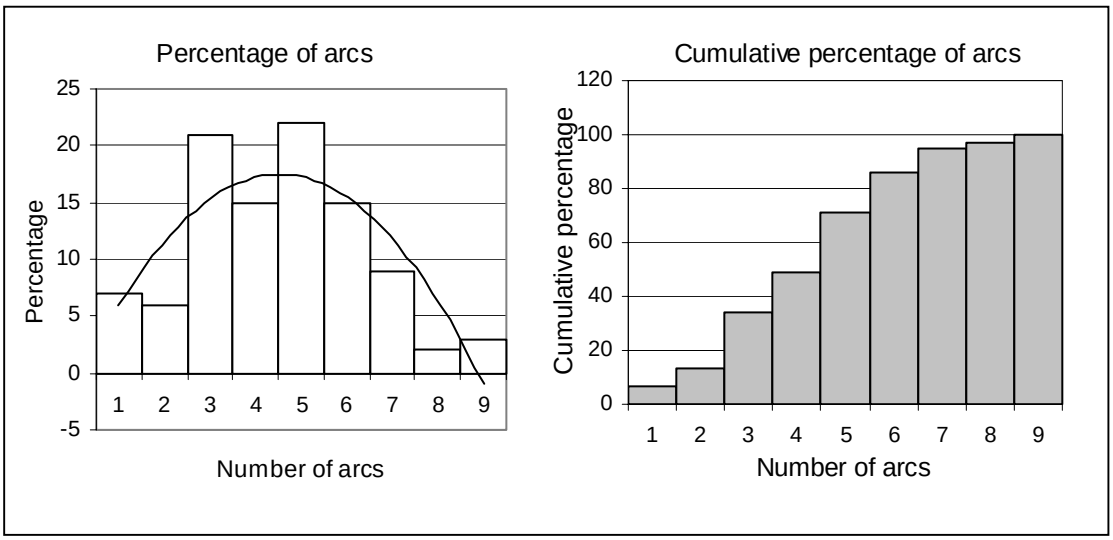


Figure 16. The diagram on the left is percentage of derived associations with a distance $\delta(i,j)$. The diagram on the right is the cumulative percentage of derived associations with a distance $\delta(i,j)$.

and a specificity of 17%. And, as stated in chapter 6.2, this should only be considered as an indication of associations that might possibly be verified. If the PubMed reported associations are all true, it means the algorithm is better than chance at deriving correct associations. However, the result would still be rather poor.

6.4 Correlation vs. Boolean, or combined?

In the analysis of the results from the correlation measurement approach and the Boolean network approach none of them have a satisfactory performance. Pearson correlation derived 28 true associations out of 203 and the Predictor derived 21 out of 323. The PubMed search indicated that the Pearson correlation might have derived up to 62 true associations and the Predictor up to 56. This would increase the sensitivity and specificity to about the double, but still their performance is quite poor. In this evaluation the Pearson correlation performed slightly better than the Predictor.

The distribution of distances seemed for the Pearson correlation to be centred on shorter distances, and decreased nearly logarithmic after four arcs (figure 14). For the Predictor the distribution seemed to be normal distributed, centred around $\delta(i,j) = 5$ (figure 16).

As discussed before, a distribution as in the Pearson correlation case is preferable, since it is more likely to get a shorter distance than a longer distance, as a result of the distribution. The same inference cannot be made from the results for the Predictor.

Number of associations reported by PubMed		56
Number of associations already verified by the known network		19
Additional number of associations reported by PubMed	56 - 19 =	37
Sensitivity – including PubMed associations	56 / (372+37) =	14%
Specificity – including PubMed associations	56 / 323	17%

Table 9. Statistics of the PubMed search.

Here, it is the same probability the distance is two arcs or seven arcs. So again, the Pearson correlation seems to be a better choice, even if none of the methods perform well enough. Pearson correlation being a better choice is also indicated by the average $\delta(i,j)$ and the median of the distances, which is lower for the Pearson correlation.

In the worst scenario there could be a distance of ten arcs for a derived association with Pearson correlation and for the Predictor up to nine arcs. Some of the derived associations could not be evaluated by this measurement, as discussed earlier. Those could possibly show an even longer distance than ten or nine arcs, in a very bad case.

Even if the methods performed poorly, a combination of the two might guide the search for the true genetic network. The Predictor derived 21 true associations and if these associations are the same as the Pearson correlation derived, and the only ones common, then a combination of the two could be used. One could then compare the derived associations from the two methods and from the common associations generate a network. The network would be small, but still a good starting point if all the associations were true positives.

The associations derived from the two methods were compared and the result is presented in figure 17. The evaluation shows that a combination of the two methods is not trustworthy. The two methods had 42 associations in common and of those 9 associations could be verified by the trusted association network, which means 21% of the common associations are true positives. If only the derived associations common to the both methods are used, then the specificity for the combined method would be 21%, which is much better than the specificity for the two methods alone (which for Pearson correlation was 13.8% and for the Predictor 6.5%). But the sensitivity would decrease to 2.4% and this is much worse than chance.

Since the methods have very few common associations, there is a possibility that those derive different parts of the network. The associations derived by the respective methods are illustrated in appendix 4. The Predictor seems to derive associations scattered over the network, with some tendency to complexes. For the Mini-Chromosome Maintenance complex (genes including Mcm2, Mcm3, Mcm6, Cdc54, Cdc46 and Cdc47) there are four derived associations and for the Anaphase Promoting complex (genes including Apc1, Apc2, Apc4, Apc5 and more) there are three derived associations. The Pearson correlation also seems to derive scattered associations scattered, but has centred especially on the Mini-Chromosome Maintenance complex, where all the associations in the complex were derived.

From this examination it seems that the methods could pick up complex associations. But then again, there are other complexes in the network for which none of the methods derived any associations, such as the Origin Recognition complex (including the genes Orc1, Orc2, Orc3 and more) and Condensin (including the genes Scc2, Smc4, Smc2 and more).

For the other associations the methods had in common, those had an average

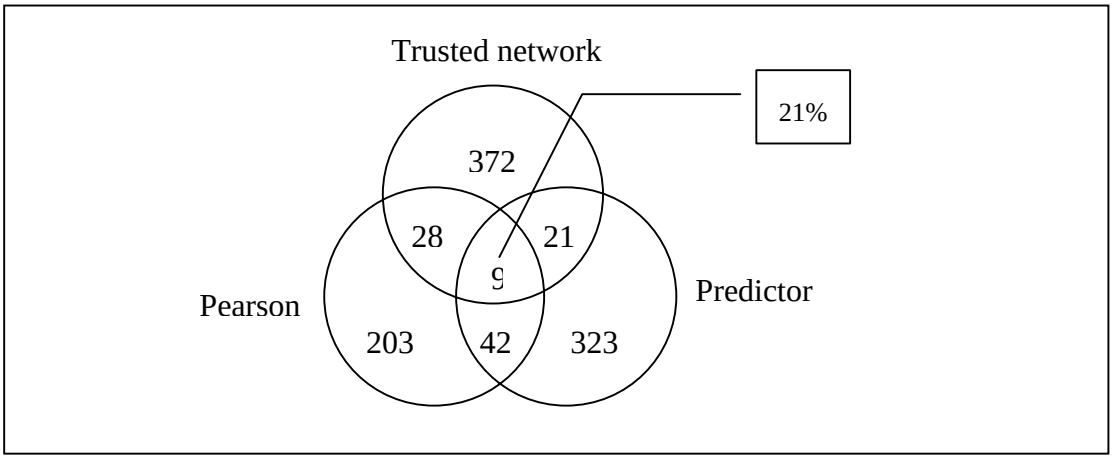


Figure 17. Common associations derived by the Pearson correlation and the Predictor, and the trusted association network from KEGG. Of the 42 common associations by Pearson and Predictor 9 could be verified by the trusted network, which is about 21% of the common associations.

distance of 4.5 and a median of 5 arcs. This is not a satisfactory result either, since one would wish those associations had a maximum distance of two arcs. If this had been the case the common associations could have been useful in deriving the network. One would not get the direct associations, but would know those were closely connected. Since this is not the case here, combining the methods will not give more information of the network and thereby not very useful.

7 Discussion

Regarding the prior knowledge approach, the result was very poor compared to the initial expectation of a much better performance. But there are some acceptable explanations for this result. The two species are rather different from one another. They are related, since they share a common ancestor (as all species do), but one is an insect and the other is a bacterium. All organisms have species-specific genes, due to evolution, and thereby some parts of their genetic networks are species-specific. In the worst case, all the experimentally verified associations in fruit fly may be species-specific and therefore cannot be mapped to baker's yeast.

In many of the associations one or both genes had an unknown cellular role, unknown biochemical function or both, which explained why those associations could not be verified. For the remaining associations, it could be that those associations are still undiscovered and thereby not annotated.

Not all the proteins for fruit fly are sequenced and therefore a lot of homologs might be missed. This means the associations between missing homologs cannot be mapped to the target organism, which reduces the trusted network and thereby the amount of prior knowledge.

Since the genome for baker's yeast is smaller than the genome for fruit fly, many of the proteins in fruit fly will recieve the same homolog. In this study those were not used in the mapping and thereby also reduced the trusted network. As mentioned before, in chapter 3.2.2, the third criterion developed by Huynen *et al*. (1998) could improve this issue. This has to be further investigated in future work.

The many more genes in fruit fly could have an impact on its genetic network - it may have many more interactions than baker's yeast, which are interactions that

baker's yeast does not have. This also is also related to the two species being rather different, as discussed above. It will affect the performance, since those interactions cannot be mapped to the target organism.

So, for future work, two organisms that are more alike, i.e. are more closly related than the two used in this study, would be preferable. The performance of the method depends very much of the prior knowledge, which is an understatement. Therefore, a larger trusted network would probably increase the performance substantially.

For the other two approaches their performance was rather poor at deriving the underlying genetic association network, which makes them not seem like trustable methods for this task. It was stated that the approaches could reveal information of the underlying genetic network (D'haeseleer *et al.*, 2000; D'haeseleer *et al*., 1999; Michaels *et al.*, 1998; Heyer *et al*., 1999; Smolen *et al.*, 2000; Maki *et al.*, 2000; Liang *et al.*, 1998), but it seems very difficult to derive associations from expression data using these two methods.

As shown in the trusted network for baker's yeast, many of the genes in the network do not interact with just one other gene and genes also often have multiple functions (Weaver *et al.*, 1997). Szallasi (1999) discussed the prevailing nature of the genetic network and stated that a network is a stochastic system, where similar cells can follow different gene expression paths between expression states. The complexity of the genetic network reflects this stochastic system and an average expression level for a population of cells can never reveal this system entirely.

The multiple functionality of genes means that the expression level of the gene depends on its current function. If the gene switches between functions during the cell cycle and participates in different pathways, an average expression level of a population of cells during a certain time will never reveal the true interactions for this

gene. For example, in time step one a gene interacts with only one other gene, in time step two the changes that occur in the cell has had the effect that the gene must interact with an additional gene, for the cell to maintain its normal behaviour. This means the expression level increases for this gene, which is reflected in its expression profile. A method for deriving interactions from gene expression levels based on one experiment only may never detect this behaviour of the gene and therefore may never be able to derive the interactions correctly for this gene.

Not all genes have multiple functionality and a natural step would be that the associations for those might be possible to derive. This could have been the case if the methods had been able to distinguish those associations. But since the methods derive so many false positives, this is not possible. It seems that the genes are well correlated for other reasons or simply by coincident. Using correlation as an approach for reverse engineering or an approach based on differences, as the Predictor, may not be sufficient.

Szallasi (1998) also discussed the compartmentalization of the genetic network, and states that a high level of compartmentalization would mean fewer interactions to test and thereby affects the performance of the method. In the trusted network from KEGG (Ogata *et al*., 1999) one could distinguish some compartments in the network, such as complexes. But at the same time the network seems to be well connected, which increases the effort of deriving the underlying genetic network.

As Zhu and Zhang (2000) states as a conclusion of their experiment, clustering gives an overview of the gene expression data, and as Heyer (1999) states the clustering does not give the right answer and therefore should be used as an exploratory tool for identifying candidate solutions, which can be further analysed. The correlation approach could be used as such, since the distance is centred around

$\delta(i,j) = 4$ and the probability for a shorter distance is higher than for a longer distance. This means two genes in an association are within the range of a tangible situation.

Whether the Predictor can be used as an exploratory tool is more difficult to say. The method has an median distance of $\delta(i,j) = 5$, but does not have the same distribution and therefore the same conclusion cannot be made. The method also seems to depend very much on which minimal set is derived, and as was seen one gene was reported nearly a hundred times, which indicates the method is not very trustable. So, the hypothetical test of Ideker *et al.* (2000), indicating that a lot of experiments have to be made in order to derive correct associations, seems also applied for real experimental data.

For future work of the Predictor, it is probably better to not include heuristics in the method. Instead all possible minimal sets should be generated and only the genes that are reported in all minimal sets would be used. The genes that vary among the minimal sets should not be considered or perhaps be further investigated by some other method. This might increase the performance of the method.

In addition, one must test other ways of inferring Boolean state symbols from the expression data (see chapter 4.2). Here, the average expression level for a gene was used, but there are other ways as well. For example, it could be based on the level of the first time point or the average of the first and the last time point. If the steady-states are known for the genes in the study, those should be used preferably. There are many ways of inferring the Boolean state symbols, for which their affect has to be further investigated.

There are also other methods besides the one used in this study for inferring the Boolean network from expression data (see chapter 2.2.2), which are worth to investigate further. Perhaps those are better fitted for this task.

For the correlation approach, it would also be interesting to test other correlation measurements, such as Spearman rank correlation (Sheshkin, 1997). Pearson correlation assumes no ordering of the time points in the data. The different measurements at different time points could be reordered and Pearson correlation would generate the same result. There are other correlation coefficients that are time dependent, and for example calculates how much a change in the level of expression for one gene correlates with a change in a different time point for another gene (Arkin *et al.*, 1997). Considering the time aspect might increase the performance of the approach.

More future work is that the methods need to be tested on other data. The prior knowledge approach, as was discussed above, should be tested for two closer related organisms and for two organisms the interactions are well known. For the other two methods, those should be tested on other expression data and also evaluated on some other trusted network. The associations from PubMed should be further investigated to examine how much it will increase the performance of the methods. The true associations should also be added to the network and the extended network could then be used in a second evaluation, and in addition, on some other expression data.

It would also be interesting to evaluate the methods using the inferential power measurement developed by Wessels *et al*. (2001), where they compared different methods for reverse engineering (see chapter 3.2.1) and compare those methods to the Pearson correlation and the Predictor.

Since there is a trusted network for baker's yeast and there are several available expression data for this organism, it would be interesting to study the expression profiles for two genes in an association and examine what is actually needed to derive the associations correctly.

8 Conclusions

The hypothesis this work was based on was that a correlation measurement, the Boolean network approach or prior knowledge could be used for deriving the genetic association network. The results from testing these approaches show that the hypothesis is still unanswered. It cannot be stated true since the approaches derived many false positives, nor can it be falsified since some true associations were derived. It could be that the methods perform better on other data and that modifications of the methods increase their performances and could also lead to more false negatives being derived.

The results indicate that the correlation measurement is the most promising tool, since it performed the best, but the prior knowledge is probably the one most worth of pursuing. The reason is that the performance of the correlation measurement is too low to be trustable enough for deriving the underlying network, and therefore should be used only as an exploratory tool. The prior knowledge approach on the other hand, still has a great possibility to increase in performance, if a more closely related organism is used as prior knowledge for the target organism, which has to be further investigated.

References

Alberts B., Bray D., Lewis J., Raff M., Roberts K., Watson J. D. (1994). *Molecular biology of the cell.* Garland Publishing, New York.

Akutsu, T., Miyano S., Kuhara S. (1999). Identification of Genetic Networks from a small number of Gene Expression Patterns under the Boolean Network Model. *Pacific Symposium on Biocomputing*, 4:17-28.

Akutsu T., Miyano S., Kuhara S. (2000). Algorithms for inferring qualitative models of biological networks. *Pacific Symposium on Biocomputing*, 5:290-301.

Attwood, T.K., Parry-Smith, D.J. (1999) *Introduction to Bioinformatics.* Addison Wesley Longman, Essex, England.

Bairoch A., Apweiler R. (2000). The SWISS-PROT protein sequence database and ist supplement TrEMBL in 2000. *Nucleic Acids Research*, 28:45-48.

Chen T., He L. H., Church M. G. (1999). Modeling Gene Expression with Differential Equations. *Pacific Symposium on Biocomputing*, 4:29-40.

Cho R. J., Campbell M. J., Winzeler E. A., Steinmetz L., Conway A., Wodicka L, Wolfsberg T. G., Babrielian A. E., Landsman D., Lockhart D. J., Davis R. W. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 21:65-73.

Costanzo M. C., Hogan J. D., Cusick M. E., Davis B. P., Fancher A. M., Hodges P. E., Kondu P., Lengieza C., Lew-Smith J. E., Lingner C., Roberg-Perez K. J., Tillberg M., Brooks J. E., Garrels J. I. (2000). The Yeast Proteome Database (YPD) and *Caenorhabditis elegans* Proteome Database (WormPD): comprehensive resources for

the organization and comparison of model organism protein information. *Nucleic Acids Research*, 28:81-84.

D'Haeseleer P., Wen X., Furhman S., Somogyi R. (1999). Linear Modeling of mRNA Expression Levels During CNS Development and Injury. *Pacific Symposium on Biocomputing*, 4:41-52.

D'haeseleer P., Liang S., Somogyi R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 8:707-726.

Duggan D. J., Bittner M., Chen Y., Meltzer P., Jeffrey M. T. (1999). Expression profiling using cDNA micro arrays. *Science*, 283: 83-87.

Durbin R., Eddy S., Krogh A., Mitchison G. (1997). *Biological Sequence Analysis*. Cambridge University Press.

Dutilh, B. (2001). Analysis of data from micro array experiments, the state of the art in gene network reconstruction. Faculty of Biology, Utrecht University, Netherlands. [http://www-binf.bio.uu.nl/~Dutilh/gene- networks/ thesis.html]

Gerhold D., Rushmore T., Caskey T. C. (1999). DNA chips: promising toys have become powerful tools. *Trends in Biochemical Science*, 24: 168-173.

Huynen M., Dandekar T., Bork P. (1998). Differential genome analysis applied to the species-specific features of Helicobacter Pylori. *FEBS Letters*, 426(1):1-5.

Heyer L. J., Kruglyak S., Yooseph S. (1999). Exploring expression data: Identification and analysis of coexpressed genes. *Genome Research*, 9:1106-1115.

Ideker, T.E., Thorsson, V., Karp, M. R. (2000). Discovery of regulatory interactions through perturbation: inference and experimental design. *Pacific Symposium on Biocomputing*, 5:302-313.

Kyoda M. K., Muraki M., Kitano H. (2000). Construction of a generalized simulator for multi-cellular organisms and its application to SMAD signal transduction. *Pacific Symposium on Biocomputing*, 5:314-325.

Liang S., Fuhrman S., Somogyi R. (1998). REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symposium on Biocomputing*, 3:18-29.

Maki Y., Tominaga D., Okamoto M., Watanabe S., Eguchi Y. (2001). Development of a system for the inference of large scale genetic networks. *Proceedings of Pacific Symposium on Biocomputing*, 6:446-458.

Matsuno H., Doi A., Nagasaki M., Miyano S. (2000). Hybrid Petri Net representation of Gene Regulatory Network. *Pacific Symposium on Biocomputing*, 5:338-349.

Mendes P. (1999). *Metabolic simulation as an aid in understanding gene expression data*. In Bronberg-Bauer E., De Beucklaer A., Kummer U., Rost U., Proceedings of Workshop on computation of Biochemical Pathways and Genetic Networks, Heidelberg, pp 27-33.

Michaels G. S., Carr D. B., Askenazi M., Fuhrman S., Wen X., Somogyi R. (1998). Cluster analysis and data visualization of large-scale gene expression data. *Pacific Symposium on Biocomputing*, 3:42-53.

Ogata H., Goto S., Sato K., Fujibuchi W., Bono H. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27:29-34.

Ramsay G. (1997). DNA chips: state-of-the art. *Nature Biotechnology*, 16:40-44.

Sanchez C., Lachaize C., Janody F., Bellon B., Röder L., Euzenat J., Rechenmann F., Jacq B. (1998). Grasping at molecular interactions and genetic networks in

Drosophila melanogaster using FlyNetss, an Internet database. *Nucleic Acids Research*, 27: 89-94.

Sheshkin D. J. (1997). *Handbook of parametric and non-parametric statistical procedures.* CRC Press.

Smolen P., Baxter A. D., Byrne G. J. (2000). Modeling Transcriptional Control in Gene Networks – Methods, Recent Results and Future Directions. *Bulletin of Mathematical Biology*, 62: 247-292.

Somogyi R., Fuhrman S., Askenazi M., Wuensche A. (1997). The gene expression matrix: towards the extraction of genetic network architectures. *Nonlinear Analysis, Theory, Methods and Applications*, 3:1815-1824.

Somogyi R. (1999). Making sense of gene-expression data. *Pharmainformatics: A Trends Guide*. 17-24.

Szallasi, Z. (1999). Genetic network analysis in light of massively parallel biological data acquisition. *Proceedings of Pacific Symposium on Biocomputing*, 4:5-16

Szallasi Z., Liang S. (1998). Modeling the normal and neoplastic cell cycle with "realistic Boolean genetic networks": their application for understanding carcinogenesis and assessing therapeutic strategies. *Proceedings of Pasific Symposium on Biocomputing*, 3:656-76.

Thieffry D, Thomas R. (1998). Qualitative analysis of gene networks. *Proceedings of the Pacific Symposium on Biocomputing*, 3:77-88.

Weaver R. F., Hedrick P. W. (1997). *Genetics*. Wm C Brown Publisher, USA.

Weaver D. C., Workman C.T., Stormo G.D. (1999). Modeling Regulatory Networks with Weight Matrices. *Proceedings of Pacific Symposium on Biocomputing*, 4:112-123.

Wessels L. F. A., Someren E. P. Van, Reinders M. J. T. (2001). A comparison of genetic network models. *Pacific Symposium on Biocomputing*, 6:508-519.

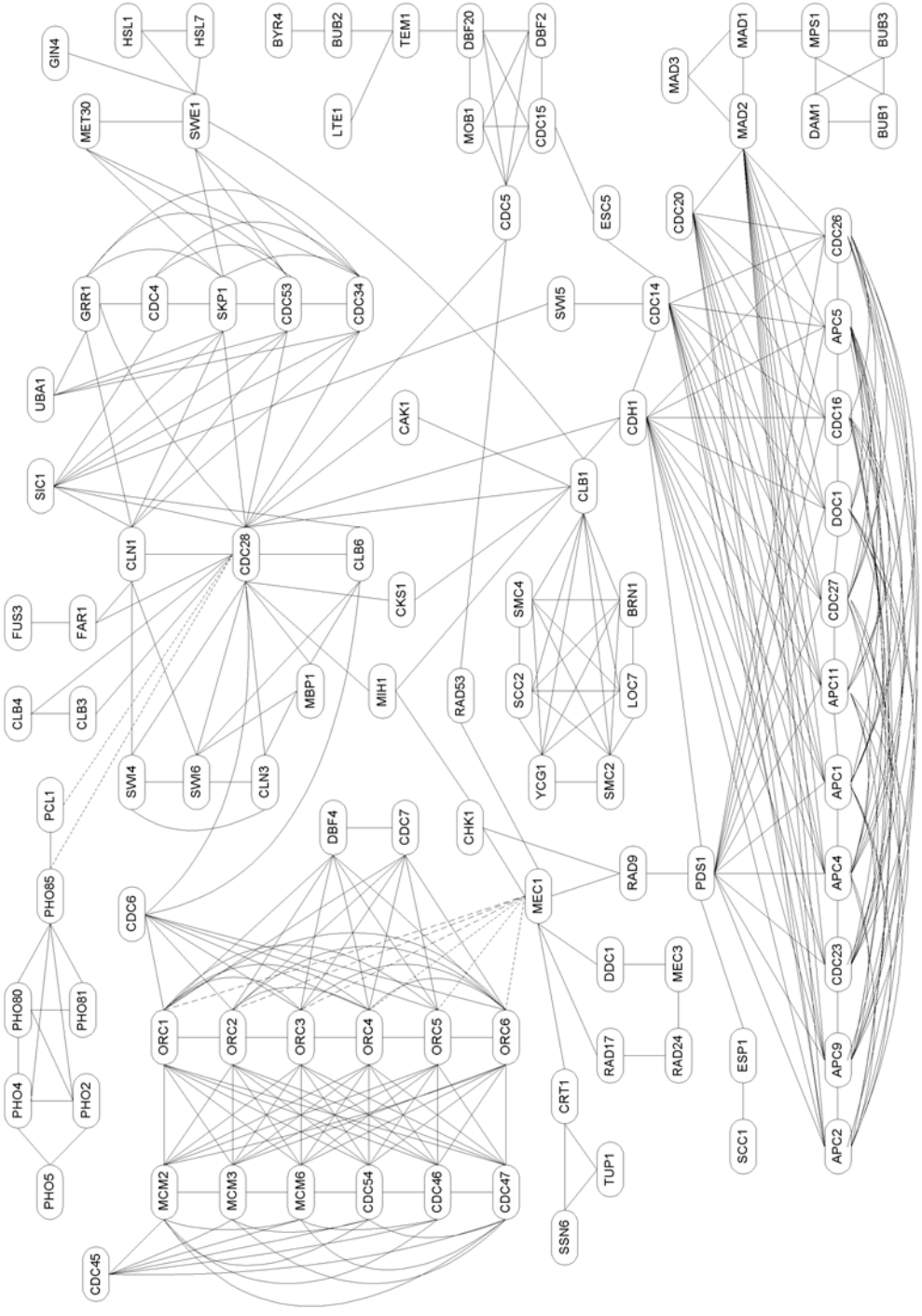
Wheeler D. L., Church D. M., Lash A. E., Leipe D. D., Madden T. L., Pontius J. L., Schuler G. D. (2001). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 29:11-16.

Zhu J., Zhang M. Q. (2000). Cluster, function and promotor: analysis of yeast expression array. *Proceedings of Pacific Symposium on Biocomputing*, 5:476-487.

Appendix

Appendix 1

The redrawn KEGG network (Ogata *et al*., 1999), which was used as the correct answer in the evaluation of the conceivable methods. It is redrawn as a genetic association network.



Appendix 2

The genes in the known association network and for which their respective gene expression profile was used in the correlation measurement approach and the Boolean network approach.

APC1	CDC34	CYC8	MAD3	PDS1	SWI5
APC11	CDC4	DAM1	MBP1	PHO2	SWI6
APC2	CDC45	DBF2	MCM2	PHO4	TEM1
APC4	CDC46	DBF20	MCM3	PHO5	TUP1
APC5	CDC47	DBF4	MCM6	PHO80	UBA1
APC9	CDC5	DDC1	MEC1	PHO81	YCG1
BRN1	CDC53	DOC1	MEC3	PHO85	
BUB1	CDC54	ESC5	MET30	RAD17	
BUB2	CDC6	ESP1	MIH1	RAD24	
BUB3	CDC7	FAR1	MOB1	RAD53	
BYR4	CDH1	FUS3	MPS1	RAD9	
CAK1	CHK1	GIN4	ORC1	SCC1	
CDC14	CKS1	GRR1	ORC2	SCC2	
CDC16	CLB1	HSL1	ORC3	SIC1	
CDC20	CLB3	HSL7	ORC4	SKP1	
CDC23	CLB4	LOC7	ORC5	SMC2	
CDC26	CLB6	LTE1	ORC6	SMC4	
CDC27	CLN1	MAD1	PAS5	SWE1	
CDC28	CLN3	MAD2	PCL1	SWI4	

Appendix 3

A subset of the derived associations from the three methods are presented here.

3 a) Prior knowledge approach, showing a subset of derived associations.

Fly genes		Yeast genes		Occurance in			Comment - unknown	
Gene 1	Gene 2	Gene 1	Gene 2	Kegg (Y/N)	YPD (Y/N)	PubMed	Cellular role	Biochemical function
ash1	E(z)	SET2	SET1	N	Y	3		
ash1	Ubx	SET2	YIL105C	N	N	0		X
bcd	RpII140	YHR217C	RPB2	N	N	0	X	X
brm	ash1	STH1	SET2	N	N	0		
car	g	SLP1	APL4	N	N	0		X
chic	Pi3K92E	CLF1	YIL105C	N	N	0	X	X
fu	ci	YPK2	MIG2	N	N	0		
fu	cos	YPK2	KIP3	N	N	0		
Gl	Dhc64C	NIP100	DYN1	N	Y	1		
gro	Nle	DIP2	YCR072C	N	N	0	X	X
gro	Rpd3	DIP2	RPD3	N	N	0	X	X
gro	arm	DIP2	VAC8	N	N	0	X	X
hop	Stat92E	CMK2	YNK1	N	N	0		
hop	awd	TOS8	YIL105C	N	N	0	X	X
hth	Ubx	FKS1	RPL19B	N	N	0		
kz	Ubx	DHR1	YIL105C	N	N	0	X	X
lace	Dsor1	LCB2	PBS2	N	N	0		
lt	car	VPS41	VPS33	N	Y	0		
lt	dor	VPS41	PEP3	N	Y	1		
lt	g	VPS41	APL4	N	N	0		
Med	CycE	EPL1	CLB5	N	N	0		X
mei-41	grp	ESR1	HSL1	N	N	0		
mle	Sxl	PRP43	NAM8	N	N	0		

3 b) Pearson correlation, showing a subset of derived associations.

Gene 1	Gene 2	Correlation
YAL040C/CLN3	YGR233C/PHO81	0.709689897415221
YAL040C/CLN3	YJL157C/FAR1	0.729220246885089
YAL040C/CLN3	YJR090C/GRR1	0.73616568029001
YAL040C/CLN3	YLR079w/SIC1	0.755205892664614
YAL024C/LTE1	YCR084c/TUP1	0.743358924963828
YAL024C/LTE1	YLR272C/LOC7	0.776260992492998
YAL024C/LTE1	YPL031C/PHO85_ex1	-0.783676679735671
YBL097w/BRN1	YBR133c/HSL7	0.802430976773915
YBL097w/BRN1	YKL022C/CDC16	0.722261066720796
YBL097w/BRN1	YMR055C/BUB2	0.734144391061042
YBL023c/MCM2	YBR202w/CDC47	0.789127261443421
YBL023c/MCM2	YEL032w/MCM3	0.796258380066641
YBL023c/MCM2	YGL201C/MCM6	0.835167436524104
YBL023c/MCM2	YGL116W/CDC20	0.738913012991622
YBL023c/MCM2	YGR092W/DBF2	0.758397122623983
YBL023c/MCM2	YLR274W/CDC46	0.723216260274403
YBL023c/MCM2	YPR019W/CDC54	0.733987280772156
YBL016w/FUS3	YBR060c/RRR1/ORC2	0.707282992616606
YBL016w/FUS3	YDL008w/APC11	-0.724150292741529
YBR060c/RRR1/ORC2	YBR133c/HSL7	0.750786600681088
YBR060c/RRR1/ORC2	YDL008w/APC11	-0.726638398952489
YBR133c/HSL7	YDL056W/MBP1	0.752063226650106
YBR133c/HSL7	YKL022C/CDC16	0.813363519100735
YBR133c/HSL7	YMR055C/BUB2	0.776576512943099
YBR136w/ESR1/MEC1	YBR274w/CHK1	0.722526259730508
YBR136w/ESR1/MEC1	YGR113W/DAM1	-0.802618355782398
YBR160w/CDC28	YDL003W/RHC21/SCC1	0.713022005613655
YBR160w/CDC28	YGR109C/CLB6	0.79993179419015
YBR160w/CDC28	YLR103c/CDC45	0.767583257764419

3 c) Boolean network approach, showing a subset of derived associations.

Gene 1	Gene 2
YAL024C/LTE1	YAL040C/CLN3
YAL024C/LTE1	YBL097w/BRN1
YAL024C/LTE1	YBR093c/PHO5
YAL024C/LTE1	YDL155W/CLB3
YAL040C/CLN3	YBR202w/CDC47
YAL040C/CLN3	YBL084c/CDC27
YAL040C/CLN3	YBR060c/RRR1/ORC2
YAL040C/CLN3	YLR182W/SWI6
YBL016w/FUS3	YAL024C/LTE1
YBL016w/FUS3	YAL040C/CLN3
YBL016w/FUS3	YBR136w/ESR1/MEC1
YBL016w/FUS3	YDR052C/DBF4
YBL023c/MCM2	YAL024C/LTE1
YBL023c/MCM2	YBL084c/CDC27
YBL023c/MCM2	YBR202w/CDC47
YBL023c/MCM2	YPR019W/CDC54
YBL084c/CDC27	YAL024C/LTE1
YBL084c/CDC27	YDL003W/RHC21/SCC1
YBL084c/CDC27	YFL029C/CAK1
YBL084c/CDC27	YGL201C/MCM6
YBL097w/BRN1	YGR109C/CLB6
YBL097w/BRN1	YJL013C/MAD3
YBR060c/RRR1/ORC2	YAL024C/LTE1
YBR060c/RRR1/ORC2	YBL016w/FUS3

Appendix 4

The trusted network for S. cerevisiae. The fat solid lines are derived associations common to Predictor and Pearson correlation. The dashed lines are derived associations for each method, '- - -' lines for Pearson correlation and '- · - ' lines for the Predictor.