Inferring Genetic Networks from Expression Data with Mutual Information

HS-IDA-MD-02-007

Thorvaldur J. Jochumson

Submitted by Thorvaldur J. Jochumsson to the University of Skövde as a dissertation towards the degree of M.Sc. by examination and dissertation in the Department of Computer Science.

2002-10-28

I certify that all material in this dissertation 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.

Signed: _____

Abstract

Recent methods to infer genetic networks are based on identifying gene interactions by similarities in expression profiles. These methods are founded on the assumption that interacting genes share higher similarities in their expression profiles than non-interacting genes. In this dissertation this assumption is validated when using mutual information as a similarity measure. Three algorithms that calculate mutual information between expression data are developed: 1) a basic approach implemented with the histogram technique; 2) an extension of the basic approach that takes into consideration time delay between expression profiles; 3) an extension of the basic approach that takes into consideration that genes are regulated in a complex manner by multiple genes. In our experiments we compare the mutual information distributions for profiles of interacting and non-interacting genes. The results show that interacting genes do not share higher mutual information in their expression profiles than non-interacting genes, thus contradicting the basic assumption that similarity measures need to fulfil. This indicates that mutual information is not appropriate as similarity measure, which contradicts earlier proposals.

Keywords: Gene Expression, Gene Expression Studies, Gene Expression Analysis, Genetic Networks, Gene Regulation, Similarity Measures, Mutual Information.

Table of Contents

1 Introduction

The field of bioinformatics has changed significantly with a new type of large-scale genomic data called gene expression data. This data reflects the quantitative expression level of genes under different conditions or over a period of time. One important goal of gene expression studies is to reveal information about the underlying mechanism that controls which genes are expressed. This involves studying how gene expression is affected by changes in the internal and the external environment of the cell. Insights gained into this mechanism will provide information about function and regulation of genes, i.e. provide a mapping from the genotype to the phenotype. This is valuable information, e.g. when studying diseases and drug treatments (D´Haeseleer et al, 2000).

The background and motivation for the work is presented in the following sections. In section 1.1 genetic studies are motivated by the central dogma, which states that DNA is the genetic material that controls the synthesis of proteins. In section 1.2 gene expression studies are presented as a method for genetic studies and the domain area of the work is introduced. Finally, in section 1.3 an overview of this thesis is presented.

1.1 Motivation

The central dogma of molecular biology states that the process of gene expression is guided by information from DNA. Transcription and translation are the two main processes, where DNA is transcribed into complementary messenger RNA (mRNA), which is then translated into protein.

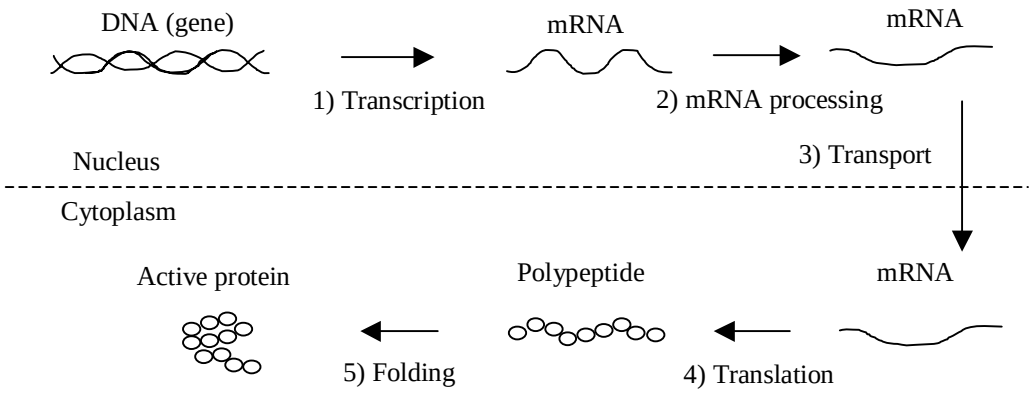

Figure 1. An overview of gene expression in a eukaryotic cell, which explains the connection between DNA (genes) and proteins.

Figure 1 gives an overview of the process of gene expression in eukaryotic cells. The figure identifies five key steps of the process:

1. Transcription: the synthesis of mRNA under the direction of DNA (genes).
2. mRNA processing: important modifications of mRNA, e.g. removal of non-coding regions (introns).
3. Transport: the transport of mRNA from the nucleus to the cytoplasm.
4. Translation: the synthesis of amino acid polypeptides under the direction of mRNA.
5. Folding: determination of the three-dimensional structure of proteins from the order of amino acids.

Each cell in a multi-cellular organism contains in principle an identical copy of the genomic information. Yet, each multi-cellular organism is made of a variety of different cells that perform different activities. Different cellular activities are explained by the various proteins that are produced in the cells, e.g. structural proteins give cells their

shape; enzymes catalyze specific chemical reactions and hormones carry signals between cells. Moreover, as stated by the central dogma, proteins are products of genes, and the activities of a cell are therefore determined by which of its genes are expressed (Weaver & Hedrick, 1997: p.338-345).

Gene regulation is the underlying process that controls which genes are expressed in a cell under different internal end external conditions. Therefore, gene regulation is the fundamental process controlling the activities of cells. When gene expression is not regulated correctly it can lead to serious imbalances in the cell and can cause diseases, e.g. cancer. Thus, the question of how eukaryotic genes are regulated is important both for medical research and for understanding the underlying processes of many biological systems (Campbell et al., 1999).

1.2 Gene Expression Studies

Traditional research approaches in molecular biology have aimed at gathering information about a single gene, protein or reaction, one at a time in isolation from each other. These methods are based on a reductionist approach, i.e. the idea that in order to understand an overall process, one needs to first understand its details (D'haeseleer et al., 2000). This has led to remarkable achievements in studying simple systems. However, this might not be a feasible approach when constructing detailed biochemical models where thousands of genes are involved, such as for the entire yeast cell (Cho et al., 1998). In these situations it might be more appropriate to study the system in a global fashion, i.e. to study the overall behaviour of the system without focusing on

each biochemical reaction. Gene expression studies provide this possibility with methods which measure simultaneously the expression levels of multiple genes. They provide the possibility to study which genes are activated, i.e. which genes are transcribed, at different stages of development, in different tissues or in different conditions of health. Gene expression studies utilise global methods, i.e. they do not only provide information about individual genes and their functionality, but also how genes act together to produce and maintain a functioning organism (D'haeseleer et al., 2000).

The field of gene expression studies can roughly be divided into two sub-fields: 1) gene expression measurements, which are methods for providing reliable gene expression data; 2) gene expression analysis, which are methods and techniques for deriving biological information from gene expression data. These sub-fields will be discussed further in sections 2.2 and 2.3. Gene expression measurements for large genomes provide a tremendous amount of data, which often include several measurements for thousands of genes. This data reflects how genes are regulated under changing conditions in the internal and external environment. The large amount of data is difficult or even impossible to analyse manually, and therefore automatic tools have been developed that support the analysis. These tools can be classified with respect to their goal, e.g. clustering of co-regulated genes and identification of regulatory interactions (D'haeseleer et al., 2000).

The work complements existing methods for deriving regulatory interactions from gene expression data. The common aim of existing methods is to identify genes that share

similarities in their gene expression data. There are several measures that are specially designed for this purpose, called similarity measures. Recent work has shown that traditional similarity measures, e.g. Euclidean distance and Pearson correlation, might not be well suited for predicting interactions between genes (Lindlöf & Olsson, 2002). In the work we consider using mutual information as a similarity measure. We will endeavour to answer the question whether mutual information, or extensions of mutual information, is well suited for deriving regulatory interactions from expression data.

1.3 Overview of this Thesis

Chapter 2, *Background*, presents the background of gene expression. First genetic networks are introduced and defined in biological terms. Then the two sub-fields of gene expression studies, gene expression measurements and analysis, are discussed. Finally, the theoretical background of mutual information is presented.

Chapter 3, *Related Work*, presents previous work that is of particular interest to this study. The work of Liang et al. (1998) is introduced in section 3.1; the work of Butte and Kohane (2000) is presented in section 3.3 and the work of Lindlöf and Olsson (2002) is presented in section 3.3. Finally, these works are compared, and their major similarities and differences are identified and discussed.

Chapter 4, *Problem Definition*, defines the problem that is dealt with in this thesis. Further, the hypothesis is presented and aims and objectives are discussed.

In chapter 5, *Method*, the experiments, carried out to test the hypothesis, are discussed. In the first section a high level design of the experiments is presented, followed by more detailed discussion of the implementation. Further, the presentation of the results and the data used in the experiments are discussed.

In chapter 6, *Results*, the results of the experiments are presented. The chapter is divided into three sections, where each section presents results of different experiments.

In chapter 7, *Discussion*, the results are discussed and compared with results of previous work. The chapter is divided into four sections, where the results of each experiment are discussed in its own section and finally compared with the results of previous work.

In chapter 8, *Conclusions*, the conclusions of the work are identified and finally in chapter 9, *Future Work*, possibilities to take the work further are discussed.

2 Background

In this chapter the background of the thesis is presented. In section 2.1 the mechanisms that control the expression of genetic material are explained. This involves defining the term “genetic network” and introducing how the flow of genetic information is circular, i.e. how the expression of genes is partly controlled by genes. In section 2.2 and 2.3 methods for studying the flow of genetic information are presented. These methods share the common name gene expression studies and involve two sub-fields, i.e. gene expression measurements and gene expression analysis. Finally, in section 2.3 information theory is discussed. This theory has been applied to gene expression analysis and provides the underlying theory for the techniques and methods that are presented in this thesis.

2.1 Genetic Networks

Gene regulation is controlled on a long and short-term basis. Long-term regulations explain how different cell types express various parts of the genome and short-term regulations explain cell responses to signals from the external and internal environment. In all organisms the expression of specific genes is most commonly regulated at the level of transcription. Transcription in eukaryotic cells is performed by enzymes called RNA polymerase. The enzyme binds to a certain location of the DNA sequence, called promoter, and starts the transcription. However, it can not initiate the transcription on its own, i.e. it can not bind directly to the promoter.

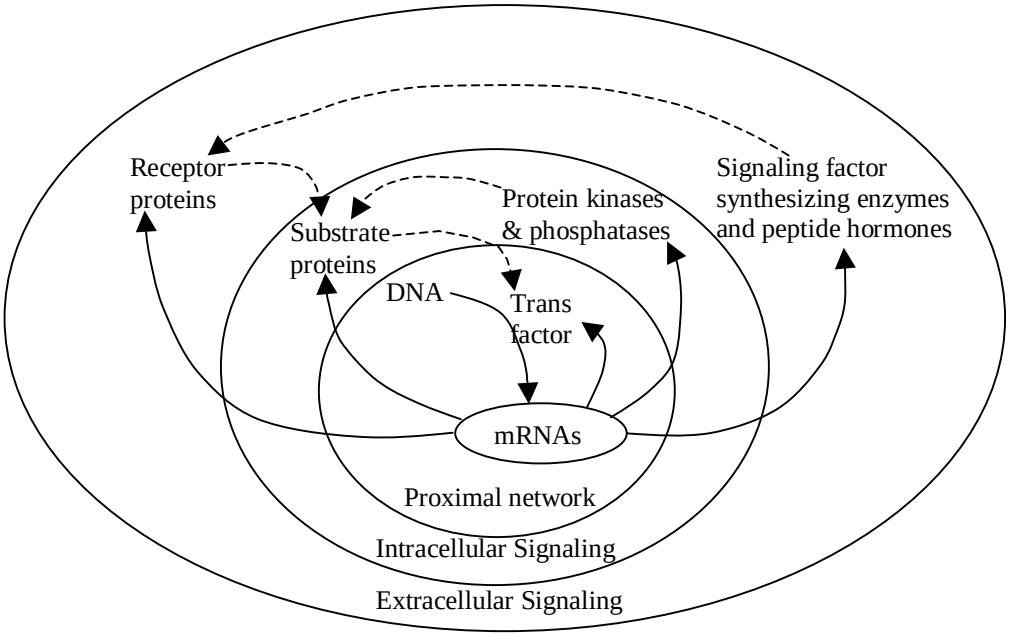

Figure 2. Illustration of a genetic network, showing how the products of genes regulate the expression of other genes, forming signalling loops with different levels of complexity (Somogyi & Sniegoski, 1996).

Transcription factors need to bind to the promoter before the RNA polymerase can do so. Additional transcription factors join the polymerase and form the transcription initiation complex. This initiation complex is the completed assembly of transcription factors and RNA polymerase bound to the promoter. Only when a complete initiation complex has been assembled the polymerase can begin to move along the DNA, producing a complementary strand of RNA. Therefore, the direct control of transcription mainly depends on regulatory proteins, i.e. transcription factors, which bind selectively to DNA and to other proteins (Campell et al., 1999). Since transcription factors are the products of genes it can be said that genes regulate the expression of genes. As suggested by Somogyi and Sniegoski (1996) we will refer to those signalling loops as genetic networks (illustrated in Figure 2). The figure shows how products of expressed genes will interact with a variety of other bio-molecules, which in turn either

directly or indirectly regulate the expression of genes through a complex hierarchy of signalling functions. The regulation can either be positive or negative, i.e. it will either stimulate or depress the expression of the regulated gene. Figure 2 illustrates two kinds of signals, indicated by solid and dotted lines. Solid lines represent information flow from primary sources, i.e. DNA and mRNA, and dotted lines represent information flow from secondary sources. Different kinds of signals produce signalling loops of varying levels of complexity. An example of a simple loop is when the number of transcription factors directly affects the expression of genes that in turn code for other transcription factors. This is shown by the proximal network in Figure 2, i.e. the innermost circle. More complex signalling loops are the result of indirect signalling, which is caused by a cascade of several biochemical activities within the cell. This is illustrated by intercellular signalling networks, the outer circles in Figure 2.

It should be kept in mind that genetic networks, as presented above, are somewhat simplified idealisations of the real mechanism. We have identified two important aspects that need to be considered. Firstly, it takes many transcription factors to form a transcription initiation complex, where the presence of the transcription factors may rely on many different internal or external conditions. Therefore, when searching for regulatory interactions between genes it is biologically correct to consider many factors simultaneously. Secondly, the complexity of eukaryotic cell structure and function provides the possibility of controlling gene expression at additional stages apart from the level of transcription. In fact, gene regulation can occur in every step in Figure 1, but since transcription is considered to be the most important level of gene regulation, other levels are ignored (Campell et al., 1999).

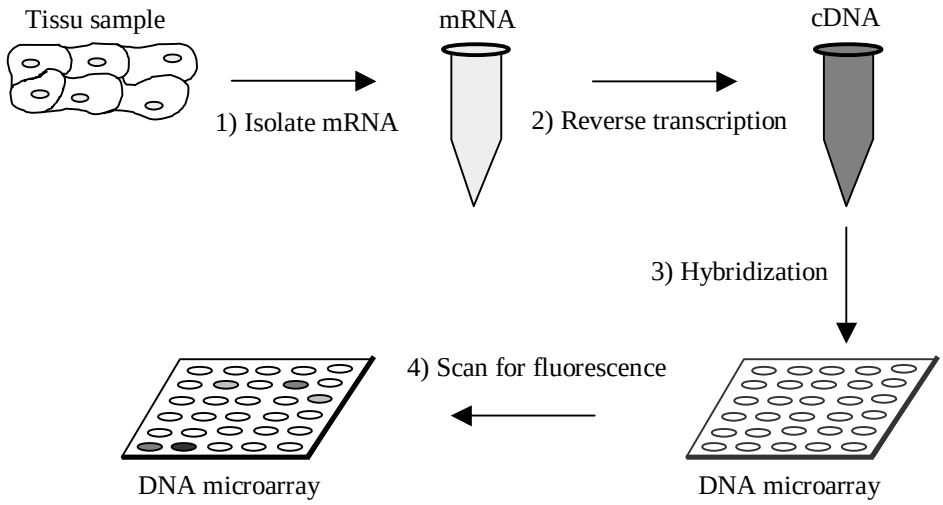

Figure 3. Illustration of the DNA microarray technique, showing how mRNA level is measured from a sample tissue (Campell et al., 1999).

2.2 Gene Expression Measurements

The methods and techniques that are involved in measuring the expression level of genes, constitutes one of the two major sub-fields of gene expression studies. The general strategy is to isolate mRNA in particular cells and use it as a template for making a complementary DNA (cDNA) library by reverse transcription. Finally, the cDNA library is compared with other collections of DNA by hybridization. Several techniques are well known, such as DNA microarrays, automated RT-PCR, 2-D gel electrophoresis and antibody arrays.

Figure 3 illustrates, in more detail, how one of the most frequently used techniques, the DNA microarray, works. Step one in Figure 3 involves isolating mRNA from a tissue sample. Step two involves creating a cDNA library by reverse transcribing the mRNA from step one, using fluorescently labelled nucleotides. Thus, the amount of cDNA in

the library reflects which genes were expressed in the tissue sample.

In step three the cDNA is applied to a DNA microarray. Each spot on the microarray, represented as a circle in Figure 3, contains copies of a short single stranded DNA molecule representing one gene of the organism. There can be thousands of spots on a single microarray, e.g. a microarray with all the genes of yeast contains approximately 6600 spots. After the cDNA has been applied to the microarray it hybridizes to the genes on the array. The final step involves measuring the amount of hybridization for each spot on the array. The amount of hybridization at each spot reflects the expression level of the gene that the spot represents. This is done by scanning each spot on the microarray for fluorescence, where high amount of fluorescence indicate high level of expressions. The final results are values that represent, relative to each other, the amount of expression for a given gene under different situations, which is called a gene expression profile or just expression profile (Campell et al., 1999).

2.3 Gene Expression Analysis

The second major sub-field of gene expression studies involves methods and techniques for extracting information from data provided by gene expression measurements. Gene expression measurements provide a means to measure the output of the underlying genetic network of the cell (Szallasi, 1999). However, biologists are left to face the question of how to make use of large-scale gene expression data. This is certainly not an easy task with the large amount of data produced by typical gene expression measurements, e.g. Cho et al. (1998) provided seventeen gene expression measurements

for the entire genome of the organism *Saccharomyces Cerevisiae* (baker's yeast). It is obvious that there is a need for tools that automate or semi-automate the process of extracting biological information from gene expression data.

Recently, there has been much focus on analysis of gene expression data. One important goal of expression analysis is to extract information about the underlying regulatory network (D'haeseleer et al., 2000). Several automatic methods have been developed for this purpose. These methods share the common name reverse engineering, since they rely on the results of the genetic network, i.e. the expression data, to extract information about it. What these methods have in common is to use changing and unchanging levels of gene expression to identify genes with similar expression profiles. Genes that share high level of similarities in their expression profiles are assumed to be co-regulated, i.e. to be part of the same regulatory process (Szallasi, 1999). Furthermore, it is expected that genes that are part of the same regulatory process share functionality, i.e. the proteins that they code for may be involved in the same biological processes (D'haeseleer et al., 2000). When a similarity measure has been computed between all pair of expression profiles it is possible to apply algorithms that represent characteristics of the genetic network. Traditionally these algorithms infer a genetic network (Ideker et al., 2000; Liang et al., 1998; Maki et al., 2001; Weaver et al., 1999), identify gene interactions (Butte & Kohane, 2000, Lindlöf & Olsson, 2002) or create clusters of genes that share significant similarity in their expression profiles (Brazma & Vilo, 2000; Eisen et al., 1998; Tamayo et al., 1999). However, it should be noted that current methods only predict characteristics of the underlying regulatory network and therefore experimental validation of the results are necessary (Szallasi, 1999).

Several similarity measures have been used to assign similarity between expression profiles. The most frequently used are Euclidean distance and Pearson correlation. Other measures include Euclidean distance between expression profiles and slopes (Wen et al., 1998), the squared Pearson correlation (D'haeseleer et al., 1999), Euclidean distance between pairwise correlations to all other genes (Ewing et al., 1999), the Spearman rank correlation (D'haeseleer et al., 1999), and mutual information (Liang et al., 1998; Butte & Kohane, 2000). In the following sections Euclidean distance, Pearson correlation and mutual information are presented and discussed as similarity measures.

2.3.1 Euclidean Distance

Euclidean distance measures the tendency of two variables to vary together. The formula for Euclidean distance is as follows:

$$ED(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

where x_i and y_i are instances of the variables x and y. If there is a strong correspondence between the two variables, $(x_i\text{-}y_i)^2$ tends to be small and therefore $ED(x, y)$ will be small. However, if there is a weak correspondence between the two variables, $(x_i\text{-}y_i)^2$ tends to be large and therefore $ED(x, y)$ will be large.

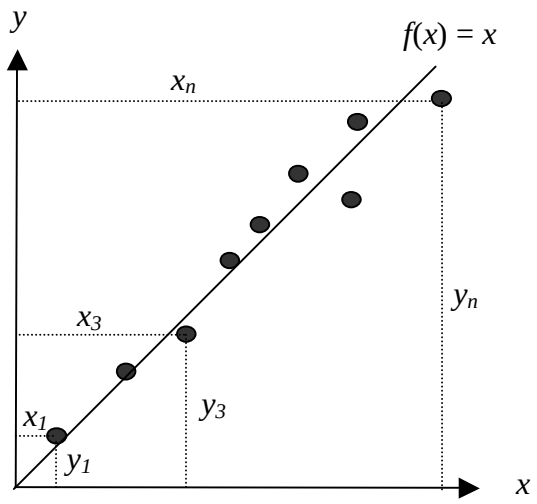

Figure 4. Illustration of Euclidean distance, showing how Euclidean distance measures how close data points tend to lie to the line $f(x) = x$.

This is illustrated in Figure 4, which shows that if the data points lie close to the line $f(x) = x$ then Euclidean distance is close to zero, but as the data points get more scattered around the line, Euclidean distance increases.

There are a few characteristics of Euclidean distance that need to be considered when applying it to gene expression data:

- Biologically it is interesting to identify expression profiles of similar shape while their quantitative expression level is of less importance. This is because the main goal is to detect genes that are controlled by the same regulatory process without concern to their mean expression. However, Euclidean distance is sensitive to the scale of the expression profiles. As is illustrated in Figure 4, Euclidean distance will not detect the relationship between x and y unless every gene expression measurement i provides a similar quantitative expression level for x_i and y_i. Therefore, before Euclidean distance can be used, gene expression

data must be scaled, i.e. all expression profiles must be adapted to the same scale without changing their shape. This is called normalisation and can for example be done with respect to the maximum expression level for each gene, with respect to both minimum and maximum expression levels or with respect to the mean and standard deviation of each expression profile (D´Haeseleer et al, 2000).

- Euclidean distance captures only linear relationships between variables, i.e. variables for which the data points tend to lie reasonably close to the line $f(x) = x$.
- Euclidean distance detects relationships between similar expression profiles and as it does not contrast profiles, only positive regulation is detected.

2.3.2 Pearson Correlation Coefficient

Pearson correlation coefficient measures the strength of a linear relationship between two normally distributed variables (Pearson & Lipman, 1988):

$$PC(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

where x_i and y_i are instances of the variables x and y, and $\bar{x}$ and $\bar{y}$ are the mean values of the instances. Pearson correlation coefficient can have a value between –1 and +1, where +1 indicates perfect positive correlation, 0 indicates absence of correlation and –1 indicates perfect negative correlation. Figure 5 illustrates how Pearson correlation coefficient calculates the correlation between the variables x and y.

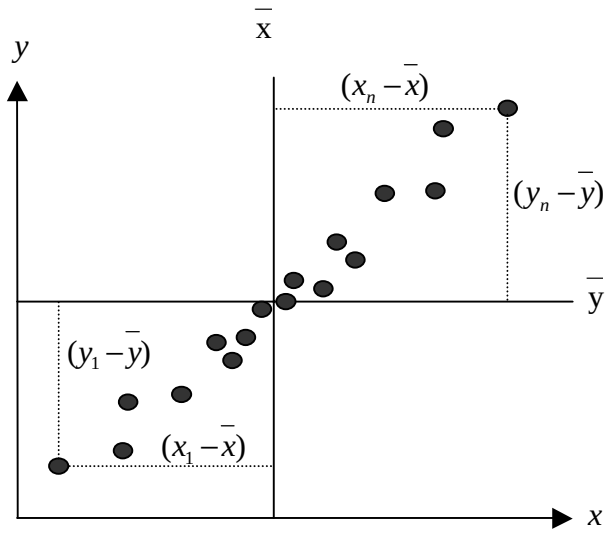

Figure 5. Illustration of how Pearson correlation coefficient is calculated for two variables *x* and *y*. The data points are relatively close to an imaginary straight line and therefore the variables have a strong Pearson correlation.

If the data points, in Figure 5, are concentrated in the bottom left and top right rectangles, then Pearson correlation is positive since the denominator and the numerator of the equation are either both negative or both positive. However, if the data points are concentrated in the top left and bottom right then Pearson correlation is negative, since the denominator and the numerator of the equation do not have the same sign. As a result, if the variables have a strong Pearson correlation the data points tend to lie relatively close to an imaginary straight line and when the variables have a weak Pearson correlation the data points tend to be more scattered.

Some characteristics of Pearson correlation are of great importance:

- In contrast to Euclidean distance, Pearson correlations is not sensitive to scaling and therefore normalisation is not needed.
- Pearson correlation captures positive correlation between similar expression

profiles and negative correlations between contrasting profiles. Therefore Pearson correlation can detect both positive and negative regulation of genes.

- Pearson correlation assumes that the distribution of the data points is a normal distribution. This makes the data points near the centre of Figure 5 of equal importance as data points that are not as close to the mean value of the variables.
- Similar to Euclidean distance, Pearson correlation only captures linear relationships between variables. However, the data points may lie reasonably close to any imaginary straight line, not only $f(x) = x$.

2.3.3 Mutual Information

Mutual information measures the amount of information shared by two variables. It has been used as a similarity measure between expression profiles, based on the assumption that the shared information between expression profiles reflects whether genes are co-regulated or not (Butte & Kohane, 2000; Liang et al., 1998). It is easy to argue in favour of this assumption since shared information between two genes, *x* and *y*, is a measure of information known about the expression of *y* given the expression of *x* and vice versa. Thus, if the genes have high mutual information the expression level of one gene reveals information about the expression level of the other. High mutual information can be explained by co-regulation of genes or regulatory interactions between genes, where the regulation can be either positive or negative. The theory behind mutual information, information theory, is presented in the next chapter (2.4). However, there are a few issues that need to be considered when mutual information is applied to gene expression analysis.

- In contrast to Euclidean distance and Pearson correlation, mutual information does not only capture linear relationships between expression profiles but other relationships as well, e.g. logarithmic relationships.
- Like Pearson correlation, mutual information is not sensitive to normalisation of the data, since different scales of the expression profiles do not affect the calculation.
- Expression profiles that undertake limited changes in their expression level, and are therefore of little interest, do not have high mutual information since they have little information to share.

2.4 Information Theory

When Shannon wrote his original paper on information theory in 1948, he focused on answering two fundamental questions in communication theory:

- What is the ultimate data compression?
- What is the ultimate transmission rate of communication?

However, information theory is more than merely a subset of communication theory and makes a fundamental contribution to statistical physics, computer science, statistical interference and probability and statistics (Cover & Thomas, 1991). Recently information theory has successfully been applied to the field of gene expression studies. In the following chapter important concepts for understanding this theory are introduced. This includes an introduction to quantities called entropy (section 2.4.1 and 2.4.2), and to mutual information (section 2.4.3).

2.4.1 Entropy of Discrete Variables

The concept of information is broad and difficult to define. However, a quantity called entropy has many properties that agree with what a measure of information should be. Entropy is a measure of uncertainty of a variable *x*, i.e. it is a measure of the amount of information required on average to describe the variable *x*. It is defined as (Shannon, 1948):

$$H(x) = -\sum_{k} p(x_k) \log p(x_k)$$

where x_k is a value taken by the variable x and $p(x_k)$ is the probability of that value. The binary logarithm is normally used and therefore the entropy is expressed in bits.

Figure 6 illustrates some of the basic properties of entropy function. It is a concave function that is zero when $p(x) = 0$ or $p(x) = 1$.

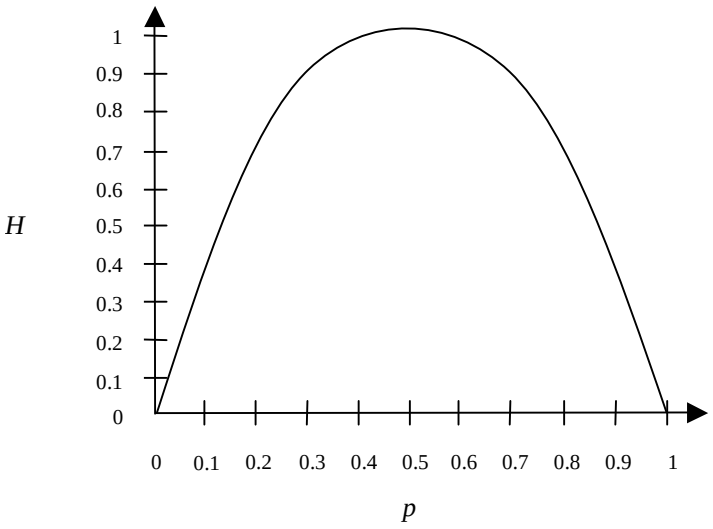

Figure 6. The entropy function of a variable *x*.

This makes sense since when $p(x)$ is 0 or 1 there is no uncertainty of the outcome.

The entropy of a single variable can be extended to the entropy of a pair of variables x and y (Shannon, 1948):

$$H(x, y) = -\sum_{k,l} p(x_k, y_l) \log p(x_k, y_l)$$

where x_k and y_l are possible instances of the variables x and y. In fact, this definition does not include anything new, since (x, y) can be considered to be a single variable z. Conditional entropy is the entropy of a variable x, given another variable y. It is defined as (Cover & Thomas, 1991):

$$H(y \mid x) = \sum_{k} p(x_k) H(y \mid x = x_k)$$

where x_k is a value taken by the variable x and $p(x_k)$ is the probability of x taking that value.

2.4.2 Entropy of Continuous Variables

The entropy of continuous variables is called differential entropy. It is similar to the entropy of discrete variables, yet the probability of a continuous variable x to take specific values is always zero (Durbin et al., 1998). However, the probability of x to take a value in some interval $P(x_0 \leq x \leq x_1)$ can be well-defined and positive. Therefore, entropy for continuous variables is calculated by taking the probability of intervals. Notice that the $P(x_0 \leq x \leq x_1)$ can be written as follows (Durbin et al., 1998):

$$P(x_0 \leq x \leq x_1) = \int_{x_1}^{x_0} f(x)dx$$

where $f(x)$ is a function called probability density, or just density. The probability density function $f(x)$ must satisfy the following conditions (Durbin et al., 1998):

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

An example of a probability density function is illustrated in Figure 7. The area below each interval Δ_i of $f(x)$, is the probability of x taking a value in the interval $[i * \Delta , (i + 1) * \Delta]$. Given the probability density function $f(x)$ the intervals can be minimized by taking the integral of $f(x)$. Therefore differential entropy $h(x)$ can be defined as (Cover & Thomas, 1991):

$$h(x) = -\int f(x) \log f(x)dx$$

However, in order to identify a probability density function for x, the probability distribution of x must be known.

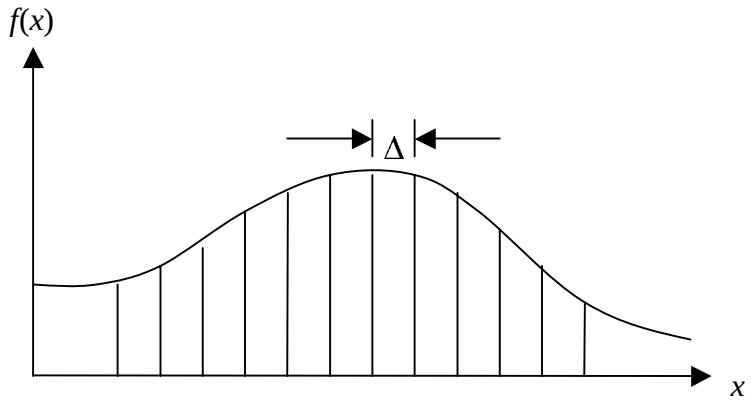

Figure 7. Illustration of probability density function.

Some methods exist for estimating the probability distribution that rely on prior knowledge about the shape of the distribution, i.e. a model is needed. If the hypothesised model is too simple, e.g. a normal distribution, it can only measure the linear relationship between variables. On the other hand, if the model is too complex it might always detect similarity between two variables, even if they are not related (Cover & Thomas, 1991). If there is no information known about the shape of the distribution Cover and Thomas (1991) suggests using the histogram technique. The technique is based on dividing the continuous scale of variables into discrete intervals, i.e. creating a histogram. What this actually does is transforming continuous variables into discrete variables, which enables probability calculations for each interval. When the probability has been calculated the entropy calculation for discrete variables can be used. If the histogram technique is to be used it is important to design the histogram carefully, i.e. carefully choose the number of intervals. If too many intervals are chosen it is likely that all values fall into different intervals and the mutual information will simply be the logarithm of the number of values. On the other hand, if the intervals are too few the power to separate different data points is lost, e.g. if only two intervals are chosen only the linear relationship can be found.

2.4.3 Mutual Information

Another important concept of information theory is mutual information of variables. It is a symmetric measure of the reduction of the uncertainty of a variable, i.e. reductions of its entropy, given another variable. Mutual information between the variables x and y is defined as (Cover & Thomas, 1991):

$$I(x;y) = \sum_{k,l} p(x_k, y_l) \log \frac{p(x_k, y_l)}{p(x_k)\, p(y_l)}$$

where x_k and y_l are values taken by the variables x and y, and $p(x_k)$ and $p(y_l)$ are the probabilities of the values. Further, mutual information has a strong relationship with entropy and can be rewritten as (Cover & Thomas, 1991):

$$I(x;y) = H(x) - H(x \mid y) = H(x) + H(y) - H(x,y)$$

This relationship between $H(x)$, $H(y)$, $H(x,y)$, $H(x|y)$, $H(y|x)$ and $I(x;y)$ is illustrated in Figure 8, which shows that the mutual information $I(x,y)$, is the intersection of the information in x with the information in y.

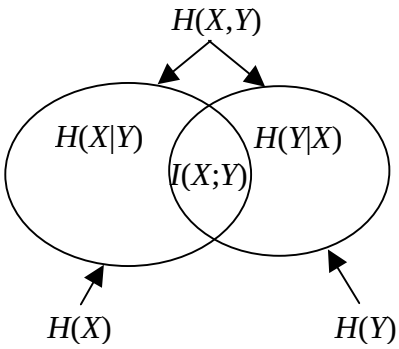

Figure 8. Schematic Illustration of the relationship between entropy and mutual information.

3 Related Work

In this chapter related and previous work of particular interest are presented. Three approaches for extracting relationships between genes are presented. In section 3.1 the work of Liang et al. (1998) to extract regulatory networks from expression data is introduced. Their work is of interest since it indicates that mutual information can be used to infer a model of the underlying regulatory network from gene expression data. In section 3.2 the work of Butte and Kohane (2000) on predicting associations between genes is presented. Their work is particularly interesting since it shows that mutual information can be used for predicting associations between genes. In section 3.3 the work of Lindlöf and Olsson (2002) on predicting interactions between genes is outlined. Their work indicates that traditional similarity measures are not well suited for predicting interactions between genes. Finally, in section 3.4 these works are discussed in comparison to each other.

3.1 Inference of Genetic Network

Liang et al. (1998) presented an algorithm, REVEAL, which infers a model of the underlying genetic network from gene expression data. The model, binary networks, has been shown to capture important information of regulatory networks (Somogyi and Sniegoski, 1996). Binary networks are based on representing genes as elements, and functional links between genes as wiring between elements. In addition, rules are presented that determine the expression of genes given the input signalling from interacting genes. The core of the algorithm is an iterative two step procedure. The

following two steps are iterated for $k = 1$ to n, where n is the number of genes in the experiment:

1. Identify a set of k genes, which we will call x, and a single gene y whose expression is controlled by x. This is calculated by using information theory, where x determines y if $H(y) = I(x; y)$.
2. Construct rules that explicitly state how x regulates y. This requires careful consideration regarding non-redundant and biologically feasible rule inference.

One of the major characteristics of binary networks is that the expression of genes is represented in a binary manner, i.e. either a gene is on (1) or off (0). This idealization simplifies the dynamics in the network and provides a better understanding of the outcomes of complex interactions. However, in reality genes are often expressed at an intermediate level and it might be a dangerous idealization to represent the expression of genes in a discrete manner, i.e. either on or off. Despite this simplification REVEAL is computationally complex. In fact, Liang et al. (1998) infer genetic networks from simulated expression data with less than fifty genes, where each gene has less than four connections to other genes. Yet, normal expression data includes measurements for thousands of genes that typically have four to eight connections (Arnone & Davidson, 1997). Therefore, REVEAL comes with the limitation that it is not practical for extracting information from a large amount of gene expression data.

3.2 Mutual Information Relevance Networks

The work of Butte and Kohane (2000) describes a technique to identify biological association between genes by computing similarity in their expression profiles. This

technique is based on using mutual information as a similarity measure and identifying a threshold value, where similarities above the threshold indicate a biological association between genes. Finally the results are presented in the form of a network, called relevance network, where each gene is represented as a node and associations between genes are represented as wiring between nodes. Butte and Kohane tested the technique on RNA expression data from the organism *S. cerevisiae* (baker's yeast). The results of the experiment indicate that the technique is appropriate to predict biological associations between genes.

Butte and Kohane (2000) calculate mutual information between expression profiles by using the histogram technique with ten intervals. However, they do not discuss the range of the histogram or how the number of intervals is chosen. These issues are of great importance when using the histogram technique, as is discussed in chapter 2.4.3. On the other hand, they use a systematic approach to define the threshold value. It is identified by permuting the expression data and then identifying the highest mutual information between permutated expression profiles. The highest mutual information is hypothesised to be the highest value that could occur by chance and is therefore set as the threshold value.

The aim of the work of Butte and Kohane (2000) is to identify biological relationships between genes, but the term "biological relationship" is not explicitly defined. However, Butte and Kohane (2000) identify four main classes of associations that are found in the experiment. These classes are: identical genes; genes of similar function; genes in the same biological pathway and various types of associations. Butte and

Kohane (2000) show that the majority of associations identified by this technique can be verified by biological literature. However, it is not discussed how many biological associations are expected to exist in the data set and how many of them are identified. It is mentioned that 199 of the 2,467 genes in their data set are identified to have biological associations. Yet, it is estimated that each gene in a multi-cellular organism interacts with four to eight other genes (Arnone & Davidson, 1997), and is involved in ten biological functions (Miklos & Rubin, 1996). Therefore, it is possible that Butte and Kohane (2000) only detect a small fraction of the existing biological interactions.

3.3 Correlation Association Networks

Lindlöf and Olsson (2002) present a technique that derives gene associations from expression data. Similarly to the technique presented by Butte and Kohane (2000) it is based on calculating similarity between expression profiles and identifying a threshold, where expression profiles with similarities above the threshold are predicted to have an association. Three different types of similarity measure are tested, Euclidean distance, Pearson correlation and cross correlations. The technique is tested on experimentally derived expression data from the organisms *S. cerevisiae* and the resulting associations are verified by biological literature. The results are rather discouraging, since a low percentage of the derived associations can be verified by the literature and a low percentage of associations in the literature are detected. This indicates that the correlation measures are not appropriate for the technique or that the technique itself is not well suited for predicting associations between genes.

An interesting aspect of the work of Lindlöf and Olsson (2002) is how the threshold value is defined. It is important to identify a balanced threshold value, which is not too high or low. Lindlöf and Olsson (2002) identify the threshold value for the different correlation measures in an ad hoc manner, i.e. after some initial experiments the threshold was decided. However, the absence of a systematic approach to identify an appropriate threshold value may be dangerous since it can result in an imbalance.

The goal of Lindlöf and Olsson's (2002) work is to detect associations between genes. Associations refer to regulatory interactions between genes, e.g. gene regulation, gene or protein complex, protein-protein interactions and interacting complexes. Despite the fact that the techniques presented by Lindlöf and Olsson (2002) and Butte and Kohane (2000) are highly similar, their aims are not to extract the same kind of information from expression data. The potential effects that this difference might have when the techniques are compared will be discussed further in the next chapter.

Finally, it is interesting to note how Lindlöf and Olsson's (2002) technique is evaluated. In the experiments validated data is used, i.e. experimentally derived expression data, and an association network created from interactions verified by biological literature. The validation is based on applying the expression data to the technique and comparing the derived associations with the associations in the validated network. In contrast to the validation performed by Butte and Kohane (2000), Lindlöf and Olsson (2002) can therefore study the results from two different angles:

1. What percentage of the derived associations are found in the validated network?

2. What percentage of the associations in the validated network are derived?

3.4 Comparison of Related Work

The work of Liang et al. (1998) is in many ways different from the other work that is presented in this chapter. The goal is ambitious, i.e. to extract a complete model of the underlying regulatory network from expression data. This involves extracting the wiring between genes, i.e. to identify interactions between genes, and to define rules that specify how genes are expressed, given an input signal from interacting genes. On the other hand, the work of Butte and Kohane (2000) and Lindlöf and Olsson (2002) is limited to extracting wiring between genes. Furthermore, the different approaches focus on different kinds of information, i.e. the wiring between genes does not indicate the same kind of information. The goal of the work of Liang et al. (1998) and Lindlöf and Olsson (2002) is to identify regulatory interactions while Butte and Kohane (2000) focus on all kinds of biological information.

The techniques presented by Butte and Kohane (2000) and Lindlöf and Olsson (2002) are in many ways similar. Their most important similarity is their goal to derive gene interactions, i.e. biological associations between genes, by computing similarity in expression profiles. However, evaluation of the techniques indicates that the technique presented by Butte and Kohane (2000) is well suited for predicting interactions between genes while the technique presented by Lindlöf and Olsson (2002) is not. This difference can have the following explanations:

- Mutual information is better suited as a similarity measure than correlation.

- The threshold value used by Butte and Kohane (2000) is stricter than the threshold value used by Lindlöf and Olsson (2002). Therefore, fewer but more reliable associations are derived.
- Butte and Kohane (2000) use a wider definition of what associations refer to, i.e. all kinds of biological interactions, while Lindlöf and Olsson (2002) focus more specifically on regulatory interactions. Therefore, the associations identified by Butte and Kohane (2000) might be easier to verify with biological literature.

4 Problem Definition

Recently gene expression analysis has been a popular research area. One of the main focuses has been on identifying regulatory interactions in the underlying genetic network (D'heaseleer et al., 2000). It has been shown that reverse engineering methods are well suited for this purpose when the networks are small with few connections between genes (Ideker et al., 2000; Liang et al., 1998; Maki et al., 2001; Weaver et al., 1999). Their main limitation is computational cost, which increases exponentially with the size of the network. Therefore, these methods are not appropriate for large systems with thousands of genes and many interactions, as real biological systems tend to be (D'heaseleer et al., 2000).

An interesting alternative is to use similarity measures between gene expression profiles to identify regulatory interactions between genes (Lindlöf & Olsson, 2002; Butte & Kohane, 2000). The advantage of these techniques is low computational cost and the ability to derive networks irrespective of the size of the network and the number of connections between genes. However, according to the results of Lindlöf and Olsson (2002) correlation measures are not well suited as a similarity measure. This is because correlations do not discriminate between expression profiles of interacting and non-interacting genes, as illustrated in Figure 9 A. A possible reason for this poor outcome is that correlation measures do not take into consideration the complexity of gene expression (Lindlöf & Olsson, 2002). Therefore there is a need for designing an extended similarity measure that takes this complexity into consideration.

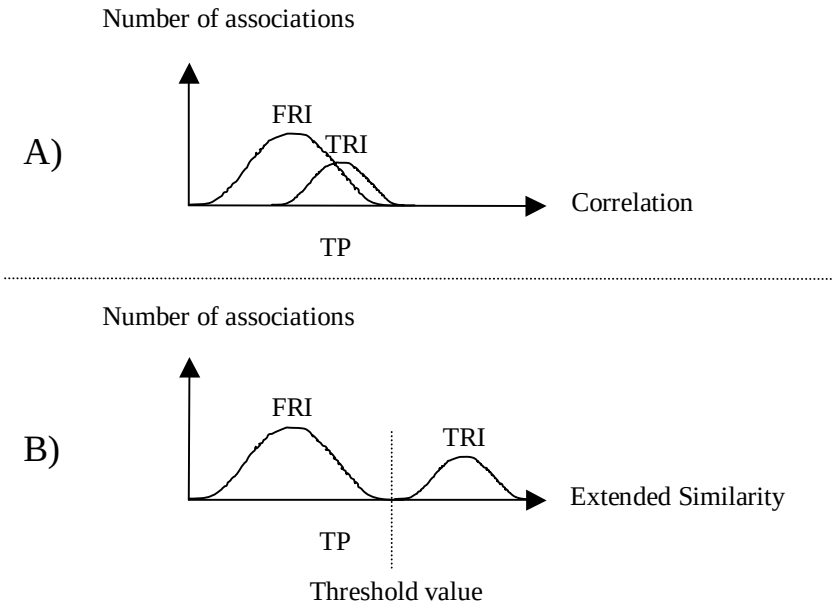

Figure 9. Illustration of how extended similarity (B) is expected to distinguish between true (TRI) and false (FRI) regulatory interactions more efficiently than correlation similarity (A). A: The curves, TRI and FRI, are overlapping and therefore it is not possible to distinguish between them. B: The curves, TRI and FRI, are no longer overlapping and therefore it is possible to identify a threshold value witch discriminates between them.

It is expected that an extended similarity measure assigns higher similarity for expression profiles of interacting genes than expression profiles of non-interacting genes, illustrated in Figure 9 B. In addition, Figure 9 illustrates that with extended similarity measure it might be possible to identify a threshold value, where extended similarity above the threshold identifies interacting pairs of genes and similarity below the threshold identifies non-interacting pairs of genes.

4.1 Extended similarity measure

In order to design a similarity measure that takes into consideration the complexity of

the underlying gene regulation, a few biological characteristics of gene regulation need to be considered. The characteristics that are expected to be of most relevance are presented below.

1) The distribution of gene expression varies for different conditions of the external and internal environment (Somogyi & Sniegoski, 1996). Therefore, it is dangerous to make any assumptions about the shape of the distribution, e.g. assuming a normal distribution.

2) Non-linear relationships between expression profiles can be biologically as interesting as linear relationships, e.g. a logarithmic relationship. Therefore it might be a serious disadvantage to use a similarity measure that only detects linear relationships such as Euclidean distance or Pearson correlation (Chen et al., 1999).

3) Because of varying complexity in intracellular signalling, regulation of genes can be introduced with varying delay, i.e. the state of the genetic network does not affect the expression level of different genes simultaneously. Therefore, it is advantageous to take into account possible time delays when expression profiles are compared (Somogyi & Sniegoski, 1996).

4) Regulation of a gene is typically a result of complex interactions between several genes (Somogyi & Sniegoski, 1996). It is therefore relevant to study complex interactions of more than two genes.

4.2 Hypothesis

The fundamental hypothesis is that profiles of interacting genes share higher level of mutual information than profiles of non-interacting genes. If the fundamental hypothesis holds it is further hypothesised that mutual information is appropriate for predicting regulatory interactions between genes from expression data. It strengthens the hypothesis that mutual information can be designed to take into account the properties of extended similarity measures, discussed in section 4.1. In addition, the results of Butte and Kohane (2000) and Liang et al. (1998) indicate that this hypothesis is plausible.

4.3 Aims and Objectives

The aim is to design an extended similarity measure with mutual information that takes into account the characteristics presented in chapter 4.2. The extended similarity measure will be evaluated and compared with results from previous studies attempting to identify regulatory interactions between genes with similarity measures. The objectives are to:

- Apply mutual information to expression data. This involves studying how mutual information can be applied to expression measurements and how it can be extended to meet the requirements of an extended similarity measure (see section 5.1).

- Design experiments that test the hypothesis. This involves studying what designs are feasible to carry out and to choose the most appropriated design (see section 5.2)
- Choose data for testing. It is important to carefully study the test data since the reliability of the experiment will never be better than the reliability of the data. This involves studying what data is available and determining which to use (see section 5.3).
- Carry out the experiment. This involves applying mutual information to the test data and measuring the performance. Here it is important to have a simple and clear representation of the results (see section 5.4 for detailed discussion about the representation of the results and chapter 6 for the actual results).
- Evaluate the results and make comparison with the result of previous works. This involves analysis and discussion of the performance of the extended similarity measure and in a qualitative manner compare it to the results of previous works (see chapter 7).

5 Method

The content of this chapter is related to the first four objectives of the work (see section 4.3). In section 5.1 we discuss how the first objective was reached, i.e. how mutual information is applied to expression data. In section 5.2 the design of the experiment is discussed, which outlines how the second objective was fulfilled. In section 5.3 the data used in the experiments is discussed, which fulfils objective three. The fourth objective requires a method of presenting the results, which is discussed in section 5.4 and finally, in section 5.5 a detailed description of the procedure of the experiments is given.

5.1 Mutual Information Applied to Expression Data

In this section a description is given of how mutual information is applied to expression data. The description is given on high-level in illustration purpose, for more detailed description and implementation details of the algorithms see Appendix A. In section 5.1.1 a basic approach of applying mutual information to expression data is explained. This approach considers only the first two characteristics outlined in section 4.1, i.e. it does not make any assumption about the shape of the distribution and it can detect non-linear relationships between expression profiles. The basic approach is extended in order to include the last two characteristics presented in chapter 4.1. Firstly, an extension that takes into account time delays between expression profiles is presented in section 5.1.2. Secondly, an extension that takes into account complex interactions among genes is presented in section 5.1.3.

5.1.1 Basic Mutual Information

This section presents an approach to applying mutual information to expression data. It is the most basic approach that is presented in this thesis and is therefore called Basic Mutual Information. Even though it is the most basic approach, it is nevertheless not straightforward to implement (see Appendix A). This is because expression measurements provide data on a continuous scale but mutual information was originally calculated between variables on a discrete scale. In section 2.4.2 two different methods are suggested to calculate entropy between variables with continuous scales, i.e. differential entropy and the histogram technique. Both methods come with limitations and it is important to realize the effects these limitations could have before choosing one method over the other.

In order to use differential entropy the shape of the distribution needs to be provided. However, since the distribution of expression data is not known beforehand, the shape of it has to be predicted, i.e. a model is needed. If the model is too simple then only simple kinds of relationships are detected, e.g. if a normal distribution is used then only linear relationships are detected. On the other hand, if the model is too complex it might always detect similarity between two variables, even if they are not biologically related (Cover & Thomas, 1991). It is therefore important to find balance between the two extremes, which requires a careful study of appropriate models.

When the histogram technique is used the continuous scale is transformed into a discrete scale by breaking it into discrete intervals. However, it is important to choose

the number of intervals with great care. If too many intervals are chosen all pairs of expression profiles will receive high mutual information even if they do not have any biological relationship. On the other hand, if too few intervals are chosen the histogram can only detect simple kinds of relationships, e.g. with two intervals only linear relationships can be detected.

It can be seen that differential entropy and the histogram technique have similar limitations. When differential entropy is used a model for describing the distribution is needed and when the histogram technique is used a number of intervals need to be decided. However, the histogram technique was chosen over differential entropy because of the following two reasons:

1) It is simpler to find the appropriate number of intervals than finding the appropriate model of the distribution. This can be done automatically by trying different numbers of intervals and comparing the results.

2) It is advantageous to use an approach that can be compared with previous works. In sections 3.1 the approach of Liang et al. (1998) to derive binary networks is described, which is theoretically identical to using the histogram technique with two intervals. In section 3.2 the approach of Butte and Kohane (2000) to derive gene networks is described, which uses the histogram technique with ten intervals. If the histogram technique is used it is therefore possible to compare the results of the work with the previous works of Liang et al. (1998) and Butte and Kohane (2000).

5.1.2 Time Delay Extension

In this section an extension of the Basic Mutual Information is presented. The extension is based on taking into account time delay between expression profiles and therefore it is called the Time Delay Extension. Figure 10 illustrates the idea behind the Time Delay Extension.

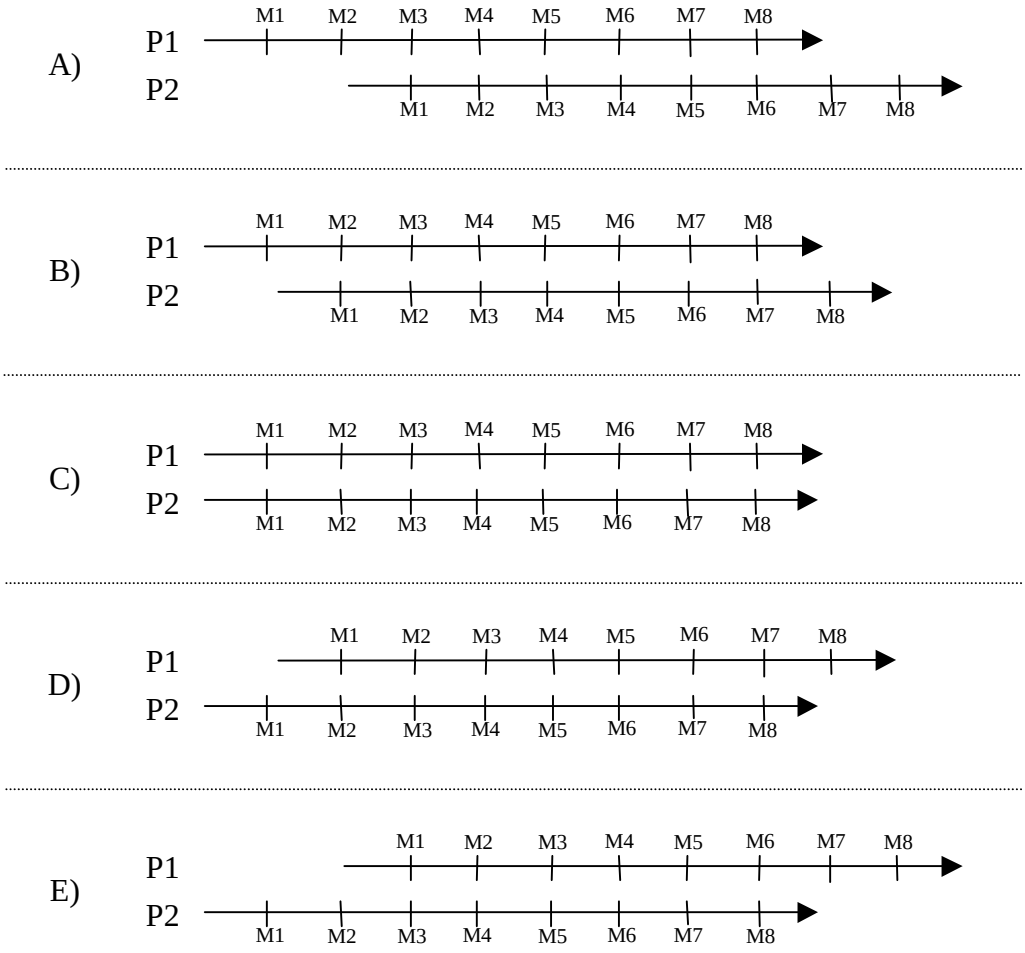

Figure 10. Illustration of how mutual information that takes into account time delays is calculated between profile 1 (P1) and profile 2 (P2). In this illustration, the delay limit is the time between two measurements. Therefore five different mutual information calculations are needed between P1 and P2: A) when P2 is delayed two time steps, B) when P2 is delayed one time step, C) when neither P1 nor P2 are delayed, D) when P1 is delayed one time step and E) when P1 is delayed two time steps. The calculation that gives highest mutual information is then used.

The Time delay Extension is based on several mutual information calculations with a different delay. The delay that results in the highest mutual information is then used. This is based on the assumption that the delay that corresponds with the signalling delay between the genes will give the highest mutual information. The Time Delay Extension has the disadvantage of increasing the possibility of high mutual information that occurs by chance, i.e. between profiles of non-regulating genes. This is because the Time Delay Extension is based on calculating mutual information multiple times between two profiles instead of just once and the possibility of high mutual information by chance is therefore increased. In order to deal with this problem a delay limit is used. This is based on the assumption that the regulatory signals will have a limited time delay. Possible delay limits are measured in intervals, where one interval is the time between measurements in the profiles. In Figure 10 the delay limit is two intervals, i.e. the time between the first measurement (M1) and the third measurement (M3), and therefore five mutual information calculations are needed. A more detailed description of the Time Delay Extension is given in appendix

5.1.3 Complex Extension

In this section the second extension, which we will refer to as the Complex Extension, is outlined. It is based on extending the Basic Mutual Information to include the last characteristic presented in section 4.1. That is, it takes into account that regulation of genes is not always controlled by pairwise interactions between genes, but often by complex interactions between many genes. This is illustrated in Figure 11 where two genes, B and C, together regulate the third gene A.

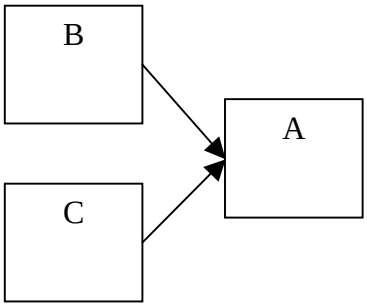

Figure 11. Illustration of how two genes (B and C) can co-regulate a third gene (A).

Figure 12 illustrates the difference between the Basic Mutual Information (Figure 12 A) and the Complex Extension (Figure 12 B) when calculating mutual information between expression profiles A and B. The Complex Extension takes into consideration that the third profile, C, might affect the mutual information between profiles A and B. More explicitly, the complex mutual information between A and B is the mutual information between A and [B,C], where C is the profile in the data set that maximises the mutual information between A and [B,C] . Therefore, the Complex Extension considers that genes B and C might cooperate to regulate gene A. The Complex Extension detects if two genes form a complex that regulates the third gene. However, it does not detect if more than two genes participate in the complex.

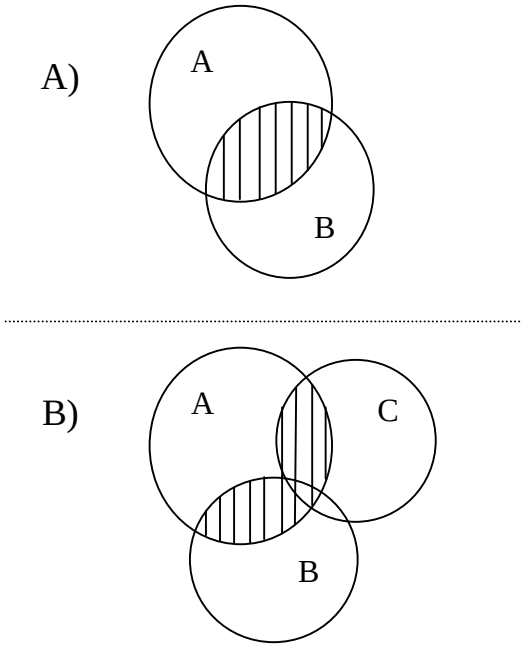

Figure 12. Illustration of the Complex Extension. The circles represent the entropy of given profiles and the shaded parts are the mutual information between the profiles. A) Basic mutual information between A and B, which does not take into consideration how other genes affect their relationship. B) The Complex Extension considers how the third gene C might affect the mutual information between A and B.

It was considered to recursively increase the number of profiles in the complex, which would provide the possibility to find larger complexes, e.g. [B,C,D]. However, this was not done since the complexity of the algorithm increases exponentially with the number of genes in the complex and one of the primary motivations for the work is that mutual information calculations should be feasible for large amount of expression data. On the other hand it is expected that large complexes will be detected implicitly with the Complex Extension. For example, if a complex of three genes B, C and D regulate gene A, it is expected that the mutual information between A and [B,C], A and [B,D] and A and [C,D] will be significant.

5.2 Design of the Experiments

Three experiments were carried out, each testing different a approach to calculate mutual information. The tested approaches are the Basic Mutual Information (see section 5.1.1), the Time Delay Extension (see section 5.1.2) and the Complex Extension (see section 5.1.3). The design of the experiments themselves is the same, which makes it possible to compare the results for the different approaches. Three designs were considered:

1) Simulate expression data with fixed regulatory interactions. The performance of mutual information can be evaluated by experimentally deriving the fixed interactions from the simulated data. This is the approach used by Liang et al. (1998) (see section 3.1).

2) Derive regulatory interactions from real expression data. The performance can be evaluated by comparing the derived interactions to biological literature. This is the approach used by Butte and Kohane (2000) (see section 3.2).

3) Derive interactions from expression data for genes which have known regulatory interactions, i.e. a validated regulatory network. The performance of mutual information can be evaluated by considering if interacting genes share higher mutual information than non-interacting genes.

In this thesis the third design is implemented since it is considered to be the most reliable approach. This is because it does not rely as much on the qualitative judgement of the experimenter, as the first two alternatives do.

The first approach assumes that the simulated data is realistic with respect to the fixed

regulatory interactions. However, we are not aware of any method able to create simulated data that corresponds highly to actual data. Therefore, these methods are based on the qualitative judgement of the experimenter in detecting how a simulated data should be created.

The second approach relies on the qualitative judgement of the experimenter to analyse the results of the experiments. The problem with this approach is that the accuracy of the derived interactions is not known. This means that the experimenter himself needs to identify which derived interactions are valid and which are incorrect with respect to biological literature, and thus, a qualitative analysis of biological literature is needed. For example, there is a great tendency for the experimenter to read the literature in a biased way, in order to make the results look as promising as possible. With this approach it is possible to estimate approximately what percentage of the detected interactions seems to be biologically correct. On the other hand it is not possible to estimate what percentage of existing biological interactions is detected.

The advantage of the last approach is that it does not rely on a qualitative analysis as much as the other two approaches. It is based on comparing the derived interactions with the validated regulatory network and presenting the results in a statistical manner. The validated regulatory network has been created before the experiment is carried out, by a research group that has no connection to the experiment. Therefore, the validated network is created in an objective way, and is not influenced by the experimenter. Obviously the reliability of the regulatory network is important. If the network contains errors it will result in incorrect statistics, and therefore the experiment will never

become more reliable than the data is. Only recently reliable regulatory networks have been available. In section 5.3 the validated network that is used in the work is discussed. A further advantage of this approach is that it provides information about what percentage of the detected interaction are correct as well as what percentage of existing biological interactions is detected.

5.3 Data

In order to carry out the experiments, with the designs presented in section 5.2, expression data and a validated network of regulatory interactions need to be provided. In this chapter the data that was obtained for the experiments is discussed. The choice of data is presented in section 5.3.1 and the pre-processing of the data, carried out in the experiments, is discussed and justified in section 5.3.2. Furthermore, the expression data is permutated. The purpose of the permutation is to remove all biological relationships between all expression profiles. In this way it is possible to estimate the highest mutual information that occurs at random, i.e. between expression profiles that are not sharing any biological relationship. The method used in the experiments to permute the data is presented in section 5.3.3.

5.3.1 Data Used in the Experiments

The data used in the experiments is from the organism *S. cerevisiae* during the cell cycle. The expression data is provided by Cho et al. (1998) and the regulatory network

is constructed from information in the KEGG[1] database (Ogata et al., 1999; Costanzo et al., 2000). This data was chosen because of the following two reasons:

- The expression data and the regulatory network match. This means that the expression data and the regulatory network provide information about the same genes in the same organism under the same biological process.
- The data from the *S. cerevisiae* cell cycle has been used in previous work (Butte & Kohane, 2000; Lindlöf & Olsson, 2002). This makes it easier to analyse the results of the experiment in comparison with previous work.

The regulatory network is constructed from a graphical diagram in the KEGG database. The diagram presents regulatory pathways in the cell cycle of *S. cerevisiae*.

Figure 13 illustrates a regulatory network with two direct interactions, between genes A and C and between genes C and D. It should be observed that the original network is modified to consider transitive interactions. Transitive interactions occur via an intermediate gene, e.g. in Figure 13 genes A and D are considered to interact, even though their interaction is not direct, but through gene C. Transitive interactions are taken into consideration since it is expected that they, as well as direct interactions, will cause high mutual information between expression profiles.

[1] http://www.genome.ad.jp/kegg/regulation.html

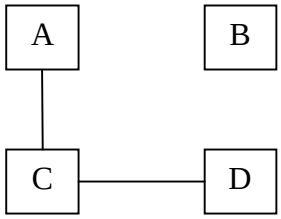

Figure 13. A regulatory network.

5.3.2 Pre-processing of Data

In section 2.3.3 it was discussed that one of the advantages of using mutual information is that pre-processing of the data, such as normalisation, is not needed. However, pre-processing of the data is carried out in the work. The reason for this is that when Basic Mutual Information is implemented, as is described in section 5.1.1, the continuous scale of expression data needs to be divided into discrete intervals. Moreover, as is discussed in chapter 5.1.1, it is important to choose the number of intervals with great care. However, the most appropriate number of intervals changes with the scale of the data. For example it might be appropriate to use ten intervals for expression data with the scale [0,100] and 100 intervals for expression data with scale [0,1000]. Therefore, in order to find a general number of intervals, the expression data needs to be normalised, i.e. the expression profiles are applied to the same scale without changing their shapes. This is done for all expression values (e) in a given profile, with respect to the profile mean value (MV) and standard deviation (SD) where e' is the normalised expression value:

$$e' = \frac{e - MV}{SD}$$

After the normalisation all expression profiles have the same mean value and standard deviation. This has the effect that after the normalisation all profiles have the same probability of receiving high mutual information. This is ideal since profiles that show small changes in their expression are often as important as the profiles that show bigger changes. Profiles that are not showing any changes at all are, however, problematic, as they are not participating in any regulatory interactions that can be detected with the data, and they are therefore of no interest. Thus, it is advantageous to filter out expression profiles that do not show any changes in their expression before the data is normalised. This was done by removing all expression profiles that had entropy equal to zero, i.e. expression profiles that are not showing any changes in their expression. 28% of the profiles in the expression data were removed. This means that interactions that involve removed profiles can not be detected, which is a large fraction of the total number of interactions.

The procedure for pre-processing the data can be summarised as follows:

- Filtering: filter out all expression profiles with entropy of zero.
- Normalisation: normalise with respect to mean value and standard deviation.

5.3.3 Permuting the Expression Data

Given the expression data, the challenge is to remove all relationships between the profiles and simultaneously maintain the distribution of the expression data. It is important to maintain the distribution because it affects the mutual information. Therefore, if the distribution would not be maintained it would not be possible to

compare the distribution of the original and the permutated data.

In the work the permutation involves randomising the order of measurements in each expression profile. More precisely, the position of each measurement is exchanged with the position of another randomly chosen position. Since the relationship between the profiles is temporal, i.e. the order of the measurements is important, the relationship is removed. However, the distribution maintains the same because the measurements themselves are not modified, only their order.

5.4 Methods of Presenting the Results

It is important that the results of the experiments are presented in a suitable manner, which gives as much information as possible and at the same time is easy to understand. Statistical measures called sensitivity and specificity have traditionally been used for this purpose (Ideker et al., 2000). However, in this thesis, a graphical representation is used as well, which presents the results visually. In the following sections these two methods of presenting the results are discussed: the statistical measures sensitivity and specificity are discussed in section 5.4.1 and the graphical representation is discussed in section 5.4.2.

5.4.1 Sensitivity & Specificity

Sensitivity and specificity are statistical measures for presenting the accuracy of an outcome. It requires that the correct outcome be known beforehand and that the actual

outcome can be compared with the correct outcome in a discrete manner. Therefore, when sensitivity and specificity are used to estimate the accuracy of derived regulatory interactions, the correct regulatory interactions need to be known, i.e. a valid regulatory network must be provided. In the context of deriving interactions from expression data, sensitivity answers the question: what percentage of the interactions in the regulatory network were derived? Specificity answers the question: what percentage of derived interactions are correct? Sensitivity and specificity are calculated as follows (Ideker et al., 2000):

- $$Sensitivity = \frac{\text{Number of derived interactions that exist in the regulatory network}}{\text{Total number of interactions in the regulatory network}}$$

- $$Specificity = \frac{\text{Number of derived interactions that exist in the regulatory network}}{\text{Total number of derived interactions}}$$

Sensitivity and specificity are therefore measures that give values between zero and one, where the best outcome is one and the worst is zero. If both measures give the outcome of one, the validated regulatory network has been derived perfectly. More explicitly, sensitivity of one means that all interactions in the validated network were derived and specificity of one means that all derived interactions exist in the validated network.

5.4.2 Graphical Representation

In order to use sensitivity and specificity to represent the results, a threshold value needs to be chosen where expression profiles that receive mutual information above the threshold are considered to interact. The statistics, therefore, depend on the threshold value. However, the aim with this thesis is to determine whether expression profiles of

interacting genes share higher mutual information than expression profiles of non-interacting genes. This would determine whether it is possible to find a threshold value, where mutual information above the threshold is caused by regulatory interactions. This can be done graphically by plotting the mutual information between expression profiles of interacting and non-interacting genes.

The graph is created by counting pairs of profiles that share particular mutual information where mutual information is on the X-axis and the number of profile pairs is on the Y-axis. However, mutual information gives values on a continuous scale, ranging from zero to the greatest entropy of the expression profiles. Since it is not possible to count observations on a continuous scale, the scale is divided into a number of discrete intervals allowing mutual information within each interval to be counted. The actual number of intervals specifies how accurate the graph is. A higher number of intervals increases the accuracy of the graph, but requires an increased amount of data. After trying a different number of intervals, it was decided to choose thirty intervals to represent the results of this thesis.

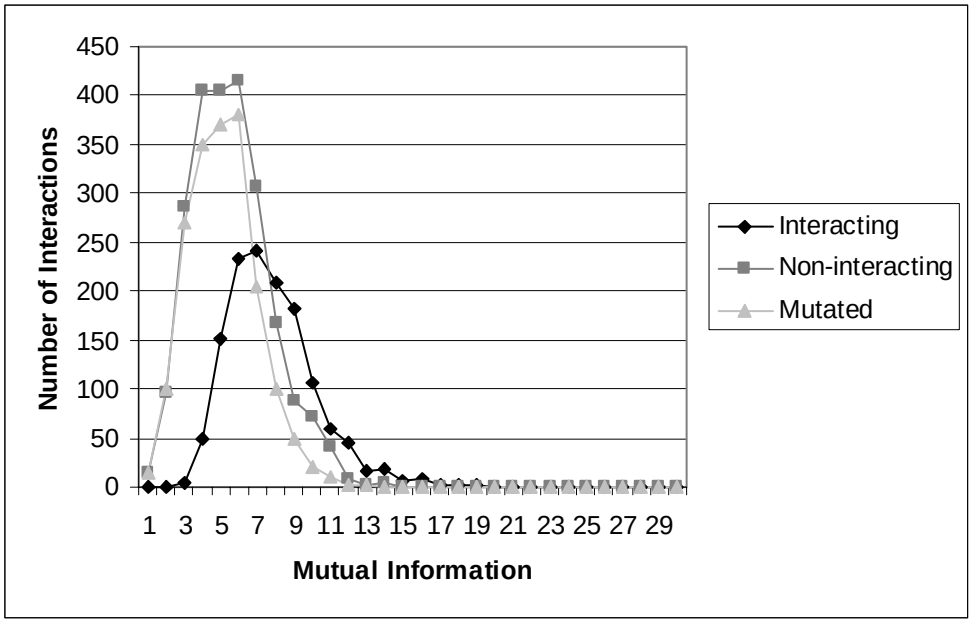

Figure 14. An example mutual information plot. The curves represent the frequency of mutual information observations for interacting, non-interacting and random expression profiles.

An example plot is shown in Figure 14. The black curve (with diamonds) is the distribution of mutual information between profiles of interacting genes and the dark grey curve (with rectangles) is the distribution of mutual information between profiles of non-interacting genes. In this way it is possible to visually observe if the two distributions can be separated by a threshold. Further, it gives a good indication of whether expression profiles of interacting genes share higher mutual information than profiles of non-interacting genes. For example, the mutual information distributions of interacting genes and non-interacting genes in Figure 14 are overlapping and therefore no threshold can separate them perfectly. However, only profiles of interacting genes have mutual information above interval 12. Therefore, if a threshold would be placed above interval 12 the specificity would be maximised, but since only a small amount of interacting profiles share mutual information above interval 12 the sensitivity would be low.

A further advantage with the graphical representation is the ability to plot mutual information between permutated and non-mutated profiles and compare their distributions. The permutated data should not contain any relationship between expression profiles and therefore the mutual information between permutated profiles is random. Further, the highest mutual information between permutated profiles is also the highest mutual information that occurs at random. As suggested by Butte and Kohane (2000) the highest mutual information between permutated expression profiles is an appropriate threshold value. This is because the mutual information above the threshold can not occur at random and therefore it must have a biological explanation.

5.5 The Procedure of the Experiments

In the previous sections important aspects of the experiments have been discussed and justified. However, even if the main ideas behind the experiments have been introduced, they can be implemented in various ways. In this section the procedures of the experiments are presented in more detail. This provides deeper insight into the experiments and makes the work replicable.

Three experiments are carried out, each testing a different approach to calculate mutual information between expression profiles. This is done by testing if expression profiles of regulating genes share higher mutual information than expression profiles of non-regulating genes. In section 5.5.1 the procedure of the first experiment is presented, which tests the performance of the Basic Mutual Information. The procedure of the

second experiment, which tests the Time Delay Extension, is outlined in section 5.5.2, whilst the third experiment, discussed in section 5.5.3, tests the Complex Extension.

5.5.1 Experiment 1: Testing Basic Mutual Information

In the first experiment the Basic Mutual Information implemented as a histogram is tested. The histogram technique is based on changing a continuous scale into discrete intervals. The scale of the data ranges from the smallest expression value to the largest, after the data has been normalised (a description of the normalisation is given in section 5.3.2). However, the most appropriate number of intervals that the histogram should contain is not known beforehand. Therefore, in the experiment different numbers of intervals are tested. The experiment starts by testing mutual information with two intervals. The number of intervals are then increased until further increasing the number of intervals will not give any better results. Finally, by comparing the results it is possible to gain information as to the most appropriated number of intervals. The procedure of the experiment is as follows:

1. Read the expression data and the target network (see section 5.3.1).
2. Filter the expression data by removing all non-changing expression profiles, (see section 5.3.2).
3. Normalise the expression data with respect to the mean value and the standard deviation (see section 5.3.2).
4. Create a permutated copy of the expression data (see section 5.3.3).
5. For *nr_intervals* = 2 while *increasing the interval gives a better results*:

5.1. Create a Basic Mutual Information, where number of intervals = *nr_intervals*.

5.2. Calculate mutual information, with the Basic Mutual Information, between pairs of profiles that are present in the target network, i.e. between interacting profiles.

5.3. Calculate mutual information, with the Basic Mutual Information, between pairs of profiles that are not present in the target network, i.e. between non-interacting genes.

5.4. Calculate mutual information, with the Basic Mutual Information, between the permutated expression profiles.

5.5. Plot the mutual information distributions of regulating, non-regulating and permutated profiles (see section 5.4.2).

5.5.2 Experiment 2: Testing Time Delay Extension

The second experiment tests an extension of the Basic Mutual Information. The extension, called the Time Delay Extension, takes into consideration that a time delay might exist between expression profiles. The extension is based on calculating the mutual information between two profiles with different delays. The correct delay, with respect to the biology, is then expected to result in the highest mutual information. It is advantageous to have a delay limit because in reality regulatory signals have limited time delay and because it decreases the possibility of high mutual information by chance. Therefore, in the second experiment a number of delay limits are tested. The tested delay limits are the time between the first measurement and all other measurements. That is, the most restricted time delay is the time between the first and

the second measurements and the least restricted time delay is the time between the first and the last measurements. The procedure of the experiment is a follows:

1. Carry out steps one to four in Experiment 1.
2. For *delay_limit* = 2 to *number of measurements*:

    2.1. Create a Time Delay Extension (as an extension of Basic Mutual Information with the optimal number of intervals detected in the first experiment), with delay limit = *delay_limit*.

    5.6. Calculate mutual information, with the Time Delay Extension, between pairs of profiles that are present in the target network, i.e. between interacting genes.

    5.7. Calculate mutual information, with the Time Delay Extension, between pairs of profiles that are not present in the target network, i.e. between non- interacting genes.

    5.8. Calculate mutual information, with the Time Delay Extension, between the permutated expression profiles.

    5.9. Plot the distribution of mutual information between regulating, non-regulating and permutated profiles (see section 5.4.2).

5.5.3 Experiment 3: Testing Complex Extension

The third experiment tests the Complex Extension, presented in section 5.1.3. The Complex Extension extends the Basic Mutual Information by taking into consideration how more than two genes interact. When calculating the mutual information between two profiles A and B, it is taken into consideration how the third profile C affects the

calculation. The third profile C is the profile that maximises the mutual information between A and [B,C]. Therefore when calculating mutual information between A and B, mutual information needs to be calculated between A and [B,C] for all profiles C in the data. The procedure for the experiment is as follows:

1. Carry out steps one to four in Experiment 1.
2. Create a Complex Extension (as an extension of Basic Mutual Information with the optimal number of intervals detected in the first experiment).
3. Calculate mutual information, with the Complex Extension, between all pairs of interacting expression profiles, A and B. That is, the mutual information between A and [B,C], where C it the profile that maximizes the mutual information.
4. Calculate mutual information, with the Complex Extension, between all pairs of non-interacting expression profiles, A and B. That is, the mutual information between A and [B,C], where C it the profile that maximizes the mutual information.
5. Calculate mutual information, with the Complex Extension, between the permutated expression profiles.
6. Plot the mutual information distributions of regulating, non-regulating and permutated profiles (see section 5.4.2).

6 Results

In this chapter the results of the experiments are presented. In section 6.1 the results of experiment 1, which tests the Basic mutual Information, are offered. In section 6.2 the results of experiment 2, which tests the Time Delay Extension, are presented and finally the results of experiment 3, which tests the Complex Extension, are given in section 6.3.

6.1 Results for the Basic Mutual Information

In the first experiment Basic Mutual Information is tested with a number of different intervals, as discussed in section 5.5.1. For the sake of simplicity the results for all considered intervals are not discussed here, but the three most important results are reviewed. These are:

- Basic Mutual Information with two intervals. This is theoretically the same approach as was used by Liang et al. (1998).
- Basic Mutual Information with five intervals. This is the number of intervals that was considered most appropriate after initial experiments.
- Basic Mutual Information with ten intervals. This is the number of intervals that Butte and Kohane (2000) used in their study.

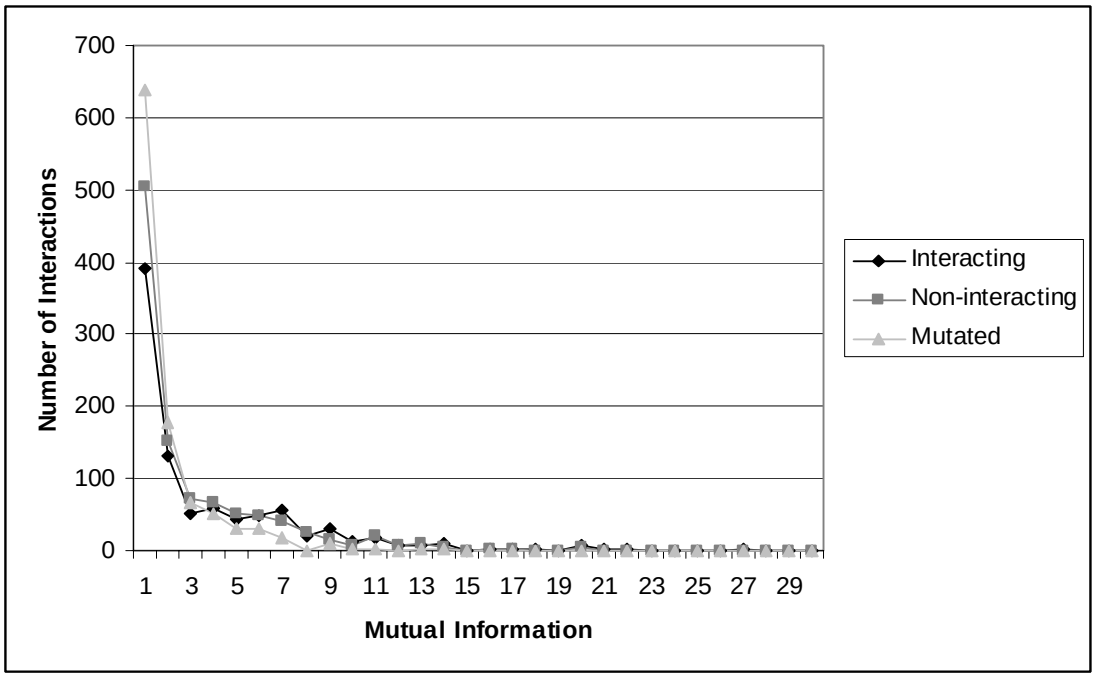

Figure 15. Distribution of the Basic Mutual Information with two intervals.

Figure 15 shows the results for Basic Mutual information with two intervals. As can be seen, the distributions of mutual information between interacting, non-interacting and mutated profiles are highly similar, where the mean mutual information between interacting profiles is 0.11, between non-interacting it is 0.09 and permutated it is 0.05. Firstly, this shows that interacting profiles do not share higher level of Basic Mutual Information than non-interacting profiles when two intervals are used. Secondly, it shows that the Basic Mutual Information between permutated profiles is highly similar to the Basic Mutual Information between non-mutated profiles. This indicates that Basic Mutual Information between non-mutated profiles is as random as the Basic Mutual Information between permutated profiles. In fact, it is not possible to find any threshold value that could give any sensible specificity and sensitivity. Hence, the method suggested by Butte and Kohane (2000), to use a threshold value equal to the highest mutual information between permutated profiles, is not possible to use here.

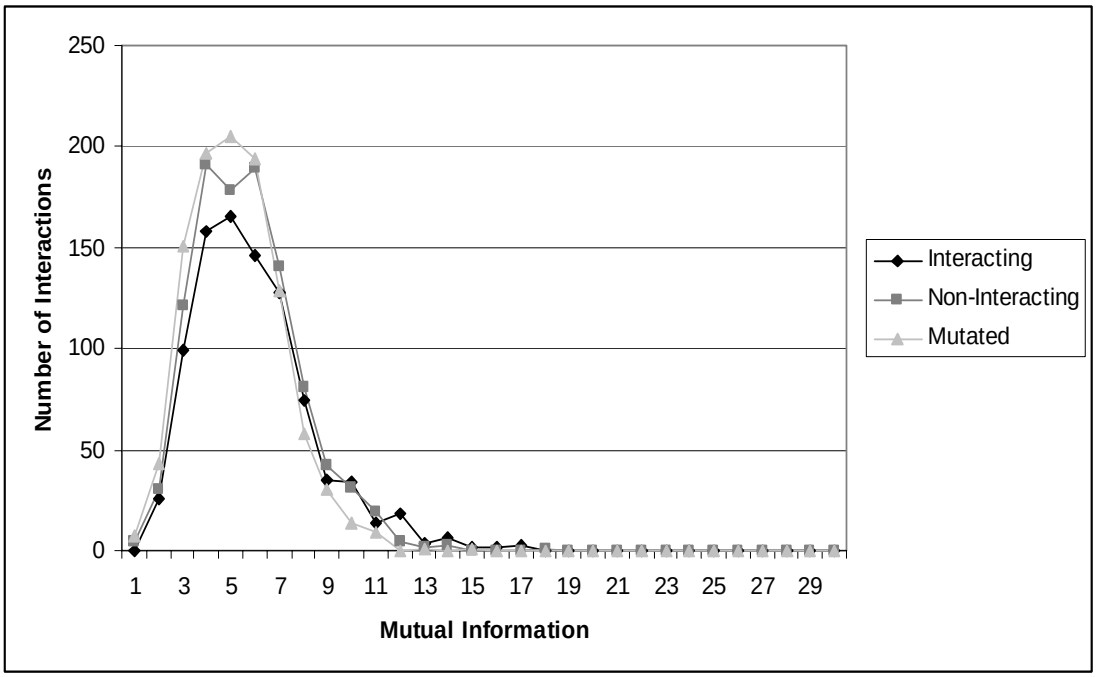

Figure 16. Distribution of the Basic Mutual Information with five intervals.

Figure 16 shows the distribution of Basic Mutual Information with five intervals. It can be seen that the distributions are highly similar, where the mean mutual information between interacting profiles is 0.43, between non-interacting profiles it is 0.40 and between permutated profiles it is 0.34. This means that when five intervals are used, interacting profiles receive in average only 0.03 higher level of Basic Mutual Information than non-interacting profiles. However, the distributions differ considerably at higher level of mutual information, e.g. no permutated profiles share mutual information that falls within a higher interval than 11, while some interacting profiles do. This indicates that some profiles of interacting genes share higher level of mutual information than happens by chance. The most plausible explanation of this is that biological relationships between genes can cause higher mutual information between expression profiles than occurs by chance. Placing a threshold value at the highest mutual information between permutated profiles, as is suggested by Butte and Kohane

(2000), gives specificity of 0.7778 and sensitivity of 0.0025. In other words, only 9 interactions were derived, 7 of them are correct and 2 incorrect. This means that only 7 of 398 associations that exist in the target network were derived.

Figure 17 shows the distributions of Basic Mutual Information when ten intervals are used. From the graph it is evident that using ten intervals does not give any better results than using two intervals. The distributions of interacting, non-interacting and permutated profiles are still highly similar, where the mean mutual information between interacting profiles is 1.08, between non-interacting profiles it is 1.08 and between permutated profiles it is 1.04. This means that interacting profiles do not share higher level of mutual information than non-interacting profiles. Furthermore, unlike the results when using five intervals, Basic Mutual Information of ten intervals does not introduce any difference between the distributions of high mutual information.

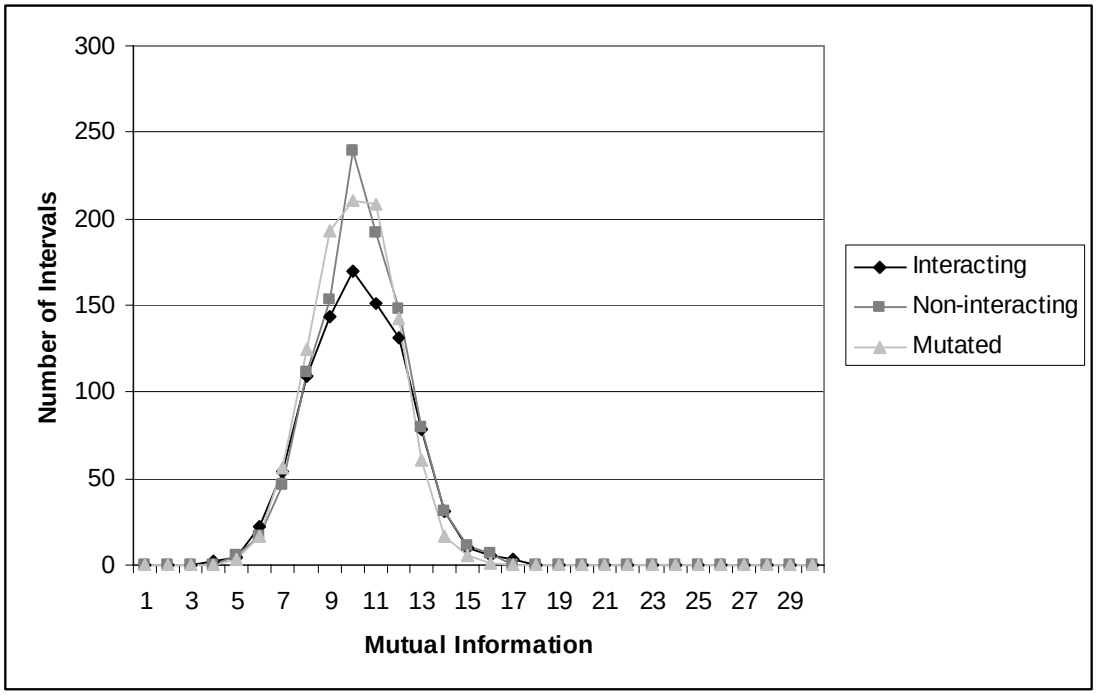

Figure 17. Distribution of the Basic Mutual Information with ten intervals.

This indicates that Basic Mutual Information of ten intervals is not affected by biological relationships between genes. This also means that it is of no significance to place a threshold value at the highest mutual information between permutated profiles since, contrary to the case when using five intervals, it would not lead to any sensible sensitivity and specificity.

6.2 Results for the Time Delay Extension

The purpose of the Time Delay Extension is to take into consideration possible delays between expression profiles. The extension is based on calculating mutual information between expression profiles several times, each time with a different delay. Each calculation is performed with the Basic Mutual Information of five intervals, since the results of the first experiment indicate that it is the preferable number of intervals. After the calculations the highest mutual information is then used. However, with increased number of calculations with different delays, the possibility of receiving high mutual information by chance is increased. Therefore, it is preferable to use a delay limit, i.e. an upper limit of how many possible delays are considered.

Possible time delays are measured in intervals, where one interval is the time between measurements in the profiles. In the experiments for this thesis several different delay limits were considered, but only the results of two of them are discussed here, the delay limits of one and two intervals. These two were found to be the optimal time delay limits and they demonstrate what affect it would have to increase the delay limit, as will be discussed in more details later.

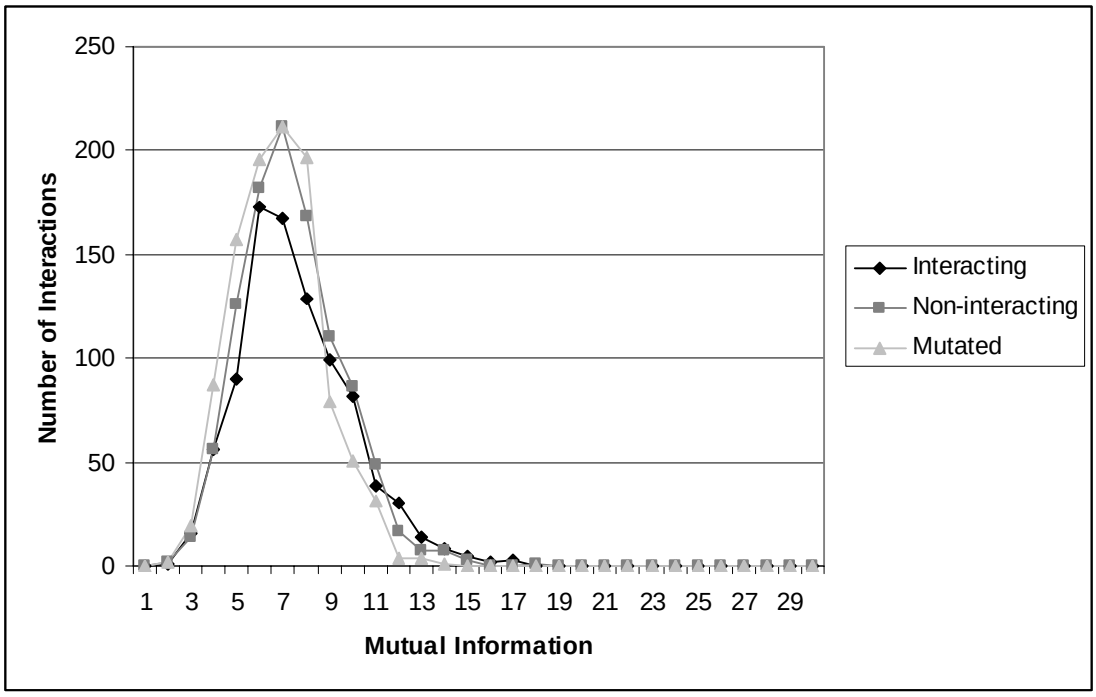

Figure 18. Distribution of mutual information with time delay of one interval.

Figure 18 show the distribution of mutual information for the Time Delay Extension, using delay limit of one interval. It does not show any dramatic difference from the results of Basic Mutual Information of five intervals. The distributions of mutual information between interacting, non-interacting and mutated profiles are still highly similar, where the mean mutual information between interacting profiles is 1.23, between non-interacting profiles it is 1.23 and between permutated profiles it is 1.20. This means that interacting profiles are not sharing higher level of mutual information than non-interacting profiles. However, when it comes to higher level of mutual information the distributions show some differences. This is similar to the results of Basic Mutual Information of five intervals, but this time both interacting and non-interacting profiles share higher mutual information than permutated profiles. This results in a lower specificity, 0.667, and slightly higher sensitivity, 0.013, when a threshold value is placed at the highest mutual information between permutated profiles.

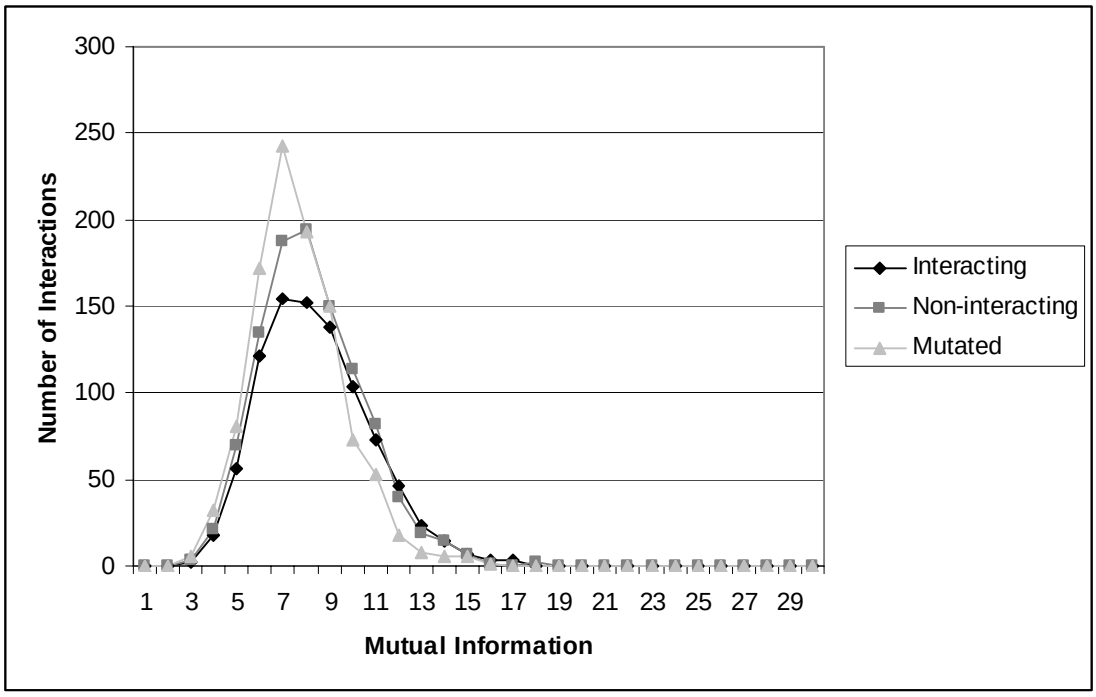

Figure 19. Distribution of mutual information with time delay of two intervals.

Figure 19 show the distributions of mutual information, using the Time Delay extension with delay limit of two intervals. As can be seen, increasing the delay limit does not give any better results. On the contrary, the results are slightly poorer since the non-mutated distributions become more similar to the permutated distributions, where the mean mutual information between interacting profiles is 1.30, between non-interacting profiles it is 1.31 and between permutated profiles it is 1.26. This indicates that increasing the delay limit results in more random mutual information. Further, it can be noted that increasing the delay limit further resulted in distributions of even higher similarity, where the distributions of mutual information between non-mutated and permutated profiles became nearly identical. This also means that it is of no use to attempt to place a threshold value at the highest mutual information between permutated profiles since it would not result in sensible sensitivity and specificity.

6.3 Results for the Complex Extension

The goal with the Complex Extension is to take into consideration that more than two genes can interact. It is based on calculating the mutual information between two profiles, A and B, taken the third profile, C, into consideration, i.e. mutual information is calculated between A and [B,C].

Figure 20 shows the results for the complex extension. Similarly to the results of the other two methods, the distributions are similar. Especially the distributions of mutual information between interacting and non-interacting genes are highly similar. This indicates that interacting profiles do not receive higher mutual information than non-interacting profiles, where the mean mutual information between interacting profiles is 1.31, between non-interacting profiles it is 1.31 and between permutated profiles it is 1.25. It is interesting that the distributions are similar, even for high mutual information. For example, the highest mutual information between permutated profiles is as high as the highest mutual information for non-mutated profiles. This means that it is not possible to place a threshold value at the highest mutual information between permutated profiles and therefore the method suggested by Butte and Kohane (2000) to identify threshold value and calculate sensitivity and specificity can not be used here.

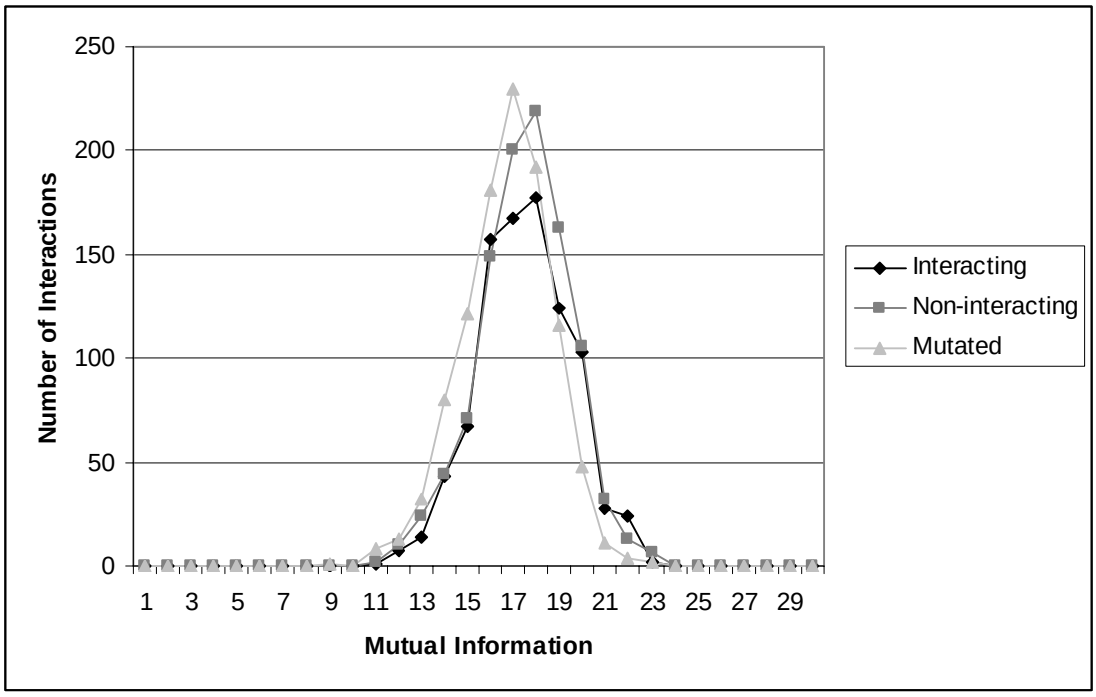

Figure 20. Distributions of mutual information with the Complex Extension.

It can been seen that the overall mutual information received with the Complex Extension is a bit higher than the overall mutual information received with the other two methods, where the mean mutual information is increased approximately 0.1. This is because taking the third gene into consideration increases the chances of high mutual information, i.e. it is a greater possibility that [B,C] share high mutual information with A than just B.

7 Discussion

The fundamental hypothesis of the work states that interacting profiles share higher levels of mutual information than non-interacting profiles. However, the results show that the three approaches to apply mutual information to expression data and which were tested in this thesis, fail to detect higher mutual information between interacting

profiles than between non-interacting profiles. It is nevertheless not possible to disprove the hypothesis, since it is only known that the tested approaches are not successful but it is not known whether more successful approaches exist. However, the results give a reason to believe that that the hypothesis is not true, and that mutual information is not appropriate to predict interactions between genes from expression data.

7.1 Discussion of the Basic Mutual Information

The goal of the first experiment was to test the Basic Mutual Information, which was implemented with the histogram technique. As was expected, the results show that it is important to choose the number of intervals for the histogram with great care. The results presented in section 6.1 show that two intervals are too few and ten intervals are too many. The balanced value of five intervals is considered to give the best results, primarily because it introduced a small difference between the distributions. The difference is however only observable at higher level of mutual information, i.e. there was a considerable amount of interacting profiles that received higher mutual information than occurs between permutated data. This indicates that the interacting profiles share higher mutual information because of their biological relationship. However, it should be noted that it is only a fraction of all interacting profiles that share this high mutual information. In fact only nine interactions are derived, where seven of them were correctly predicted to be regulatory interactions. It should also be noted that the statistical calculations, i.e. sensitivity and specificity, should not be taken too seriously. This is because nine interactions do not provide enough information to extract any reliable statistical information from it.

As was suggested by Lindlöf and Olsson (2002) the non-inspiring results may be caused by the complexity of real data. However, the extensions that were introduced in the work, which take into consideration some important aspects of the complexity of the data, are not giving more promising results. This indicates that there are even more complex aspects in the data that need to be taken into consideration or that other parts of the work need to be questioned. Two parts that can be questioned are: 1) the validated genetic network; 2) the gene expression data.

The validated genetic network: the non-inspiring results might be explained by questioning the validated network that is considered to be the correct outcome. In fact the basic idea of categorising pairs of genes as interacting or non-interacting in a discrete fashion can be questioned. This is because the products of genes will change the internal or external environment of the cell, and therefore lead to changes in the expression of the whole genome. This can be caused by direct interactions of genes, such as by production of transcription factors, or more complex cascade of signals (Somogyi & Sniegoski, 1996). The problem is when cascade of signals is studied all genes in the genome might be interacting in one way or another. For example, in the work we considered transitive interactions (see section 5.3.1), but why not consider double transitive interactions or even more distant relationship, which would make all genes in the genome interacting? Therefore, the idea of predicting pairs of genes as interacting of non-interacting in a discrete fashion might be too simple in order to reveal information about gene regulation, as complex as it is.

The expression data: in order to explain the results it is even possible to question if the expression data is rich enough of information to extract regulatory interactions from it. The fact that 28% of all profiles do not contain any information and are therefore removed before the normalisation indicate the opposite (see section 5.3.2). After that the foundation for finding large amount of interactions has been removed. Further, if that many profiles do not contain any information, how many profiles don not contain the expected information?

7.2 Discussion of the Time Delay Extension

The goal of the second experiment is to test the Time Delay Extension. It is based on calculating mutual information between the profiles with different delays. In order to avoid high mutual information by chance the extension was implemented with time delay limits. In section 6.2 the results for the Time Delay Extension with delay limit of one and two intervals are presented. The results for the Time Delay Extension with delay limit of one are not an improvement of the results for the Basic Mutual Information. The results are in fact highly similar, not showing any differences between the distributions of mutual information except for higher mutual information. Increasing the delay limit to two intervals does not either give any better results. On the contrary the distributions become more similar even for high mutual information. Some additional experiments with increased delay limits were carried out, which resulted in progressively more similar distributions. This indicates that increasing the delay limit increases the possibility of high mutual information between profiles by chance, e.g. between permutated profiles. This means that the distributions become more and more

identical, until there is no difference between them. In other words, the results show that the Time Delay Extension is not an improvement of the Basic Mutual Information. The statistical information extracted from this experiment should be questioned for the same reason as we questioned the statistical results from first experiment. This is because placing a threshold at the highest mutual information between permutated profiles identifies only 16 interactions, which is not sufficient information to extract any reliable statistics from it. The aim of providing statistical results is therefore not to provide comparison between the three mutual information approaches in order to identify the most successful one, but rather to give the reader an indication of how well mutual information would work in practise. The non-inspiring results of the Time Delay extension can be discussed in the same way as the results of the Basic Mutual Information (see section 7.1). However, why does the Time Delay Extension not improve the results of the Basic Mutual Information as was expected? We have identified two main reasons for this: 1) rough alignment; 2) reliability of the data.

Rough alignment: the Time Delay Extension is based on aligning the profiles in order to maximise their mutual information. However, due to how roughly the profiles are aligned the optimal alignment might always be missed. That is, the delay limit is always increased one interval at a time, which is the time between two measurements (see section 5.1.2). However, it might be that in order to find the optimal alignment the delay can not be increase that much each time, which causes the Time Delay Extension to jump over the optimal alignment.

Reliability of the data: as was discussed in the previous section the reliability of both the

validated genetic network and the expression data can be questioned. It is obvious that no extension of the Basic Mutual Information will be successful if the data is not reliable, no matter how it is designed.

7.3 Discussion of the Complex Extension

The goal of the third experiment is to test the Complex Extension. The extension is based on Basic Mutual Information that is extended to take into consideration how mutual information between profiles is affected by other profiles. However, the results for the Complex Extension indicate that it is not a successful extension of the Basic Mutual Information, i.e. the distributions of mutual information for interacting and non-interacting profiles are highly similar. This means that it is not possible to say that interacting profiles share higher mutual information than non-interacting. Furthermore, the distributions of mutual information for interacting and non-interacting profiles are highly similar to the distribution of mutual information for permutated profiles. This indicates that mutual information between non-mutated profiles is random, and not controlled by biological relationships between the expression profiles.

The fact that the Complex Extension does not improve the performance of the Basic Mutual Information can be explained with questioning the reliability of the data, as is done for the Time Delay Extension in previous section. That is, no extension of the Basic Mutual Information will be successful if the data is not reliable (see previous section). However, their exist another possible explanation, which is more related to the Complex Extension it self. The Complex Extension is designed to detect more complex

relationship between three genes, not only pairwise interactions between genes. That is, it is designed to detect the situation when a gene is regulated by a complex of two genes. However, the method is not explicitly designed to detect the situation when the complex contains more than two genes, as frequently occurs in reality. The motivation for this limitation is to maintain low computational complexity, which would increase exponentially with the size of the complex (see section 5.1.3). What is not known is how serious impact this limitation has on the performance of the Complex Extension.

7.4 Comparison with Previous Work

It is of interest how these results compare with results of previous work. Liang et al. (1998) and Butte and Kohane (2000) presented much more positive results, but the results of Lindlöf and Olsson (2002) are similar to the results identified here.

Liang et al. (1998) managed to recreate fixed regulatory networks from simulated expression data. Their algorithm, described in section 3.1 is based on mutual information between binary expression profiles, i.e. genes are either on or off. This is theoretically identical to Basic Mutual Information with two intervals. However, our results do not indicate that Basic Mutual Information with two intervals can be used for this purpose. The different conclusions can be explained by different methods used by Liang et al. (1998) and the ones used in the work, i.e. Liang et al. (1998) used simulated expression data but in the work we used real expression data. Therefore, a possible explanation is that simulated expression data is not comparable to real expression data, e.g. might not take into consideration the complexity of real expression data.

It is suggested by Butte and Kohane (2000) that all profiles that share higher mutual information than happens between permutated profiles must have some biological relationship. They show that this method works but they only predict small amount of all the information that exists in the data. Similarly, Basic Mutual Information with five intervals shows that some interacting profiles share higher mutual information than happens between permutated profiles. This indicates that it is possible to predict some interactions, but these interactions are just a fraction of all interactions that exist in the data. We find it important to point this out because it was left out in the discussion of Butte and Kohane (2000).

The results of Lindlöf and Olsson (2002) are similar to the results of the work, or reflect a slightly poorer outcome of the hypothesis. Differently from the other related work, Lindlöf and Olsson (2002) used correlation between expression profiles, not mutual information, to predict regulatory interactions between genes. The method used by Lindlöf and Olsson (2002) has much in common with the method used in the work and this can explain why the results presented by Lindlöf and Olsson (2002) are similar to the results of the work. In both methods gene interactions were predicted from real expression data, and what is more important, both methods are based on verifying the derived interactions against a validated regulatory network. In fact, the slightly more positive results of the work are because the validated network is extended to take into consideration transitive interactions, and not because mutual information is better suited as a similarity measure than correlation measures. This was realized when the experiments of the work were repeated, only this time without transitive interactions,

and the results turned out to be highly similar to the results of Lindlöf and Olsson (2002). The fact that taking transitive interactions into consideration gives better results indicates that mutual information can, to a certain level, detect transitive interactions between genes.

8 Conclusions

It is the conclusion of the work that none of the approaches to apply mutual information to expression data can successfully discriminate between interacting and non-interacting profiles, which contradicts the fundamental hypothesis. Therefore, none of the approaches are suited to predict regulatory interactions from expression data, which gives a reason to believe that mutual information in general is not appropriate for this purpose.

Another interesting conclusion is that the extensions of Basic Mutual Information, presented in the thesis, that take into consideration the time delay between expression profiles (the Time Delay Extension) as well as complex regulations of genes (the Complex Extension) are not successful.

9 Future Work

This was supposed to be a comprehensive work that casts light on how and if mutual information can be used to infer genetic networks from expression data. But, as often is the case, the work identifies at least as many new questions as it gives answers. This chapter concerns how the work can be continued and identifies new interesting research questions related to the work. According to the results of the work, current methods that predict regulatory interactions from mutual information between expression profiles are not reliable. We identify two ways, with different focus, to take the work further. *Firstly*, by focusing on the data and, *secondly*, by focusing on how mutual information is applied to expression data.

Data: as is discussed in section 7.1 it is necessary to question the data that the experiments are based on. It is discussed that both the gene expression data as well as the validated genetic network might be unreliable. Future work with focus on the expression data might, therefore, involve examining whether the expression data contains enough information to derive a complete regulatory network from it. Another interesting question is if there exists more reliable expression data. At the time this is written a new source of expression data exist at NCBI where it is possible to access recent expression data that might potentially be more accurate than the data used in the work (Edgar et al., 2002). Future work with focus on the genetic network might issue alternative ways to represent genetic information, i.e. using a representation that specifies how the genes interact or representing interactions between genes on a continuous scale, not on a discrete scale where genes are either interacting or not.

Mutual information applied to expression data: we have shown that the three approaches Basic Mutual Information, Time Delay Extension and Complex Extension are not successful. However, this does not prove that there exists no successful approach to apply mutual information to expression data. It is an interesting future work to examine further if it is possible to develop more successful approaches. This might involve:

- Study if there are some other important properties of expression data that need to be considered when applying mutual information to it. If any new properties can be identified, then develop extensions of the Basic Mutual Information that take these properties into consideration.
- Study the possibility of capturing more information from the expression profiles, e.g. is it possible to change discrete measurements to continuous graphs?
- Further develop the Time Delay Extension and Complex Extension to deal with their limitations (see section 7.2 and 7.3).
- Examine the possibility to integrate the approaches presented in this thesis. If the methods are sensitive for different information in the expression data, and therefore identify different interactions, it would be interesting to integrate them.

10 References

Arnone, M. I. & Davidson, E. H. (1997) The hardwiring of development: organization and function of genomic regulatory systems. *Development*, *124*, 1851-1864.

Brazma, A. & Vilo, J. (2000) Gene expression data analysis. *FEBS Letters*, *480*, 17-24.

Butte, A. J. & Kohane, I. S. (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing*, *5*, 418-429.

Campbell, N. A., Reece, J. B. & Mitchell, L. G. (1999) *Biology*. Addison Wesley Longman.

Chen, Y., Bittner, M. L. & Dougherty, E. R. (1999) Issues associated with microarray data analysis and integration. *Nature Genetics*, *22*, 213-215.

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.

Cover, T. M. & Thomas, J. A. (1991) *Elements of Information Theory*. New York: John Wiley and Sons.

D'haeseleer, P., Liang, S. & Somogyi, R. (2000) Genetic Network Inference: From Co-Expression Clustering to Reverse Engineering. *Bioinformatics, 16,* 707-726.

D'haeseleer, P., Fuhrman, S., Somogyi, R. & Wen, X. (1999) Linear modelling of mRNA expression levels during CNS development and injury. *Pacific Symposium on Biocomputing, 4,* 41-52.

Durbin, R., Eddy, S., Krogh, A. & Mitchinson, G. (1998) *Biological sequence analysis: Probabilistic models of proteins and nucleic acids.* Cambridge: Cambridge University Press.

Edgar, R., Domrachev, M. & Lash, A. E (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research, 30,* 207-

210.

Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998) Cluster Analysis and Display of Genome-Wide Expression Patterns. *Proceedings of the National Academy of Sciences of the United States of America, 95,* 14863-14868.

Ewing, R. M., Kahla, A. B., Poirot, O., Lopez, F., Audic, S. & Claverie, J. M. (1999) Large-scale statistical analysis of rice ESTs reveal correlated patterns of gene expression. *Genome Research, 9,* 950-951.

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.

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.

Lindlöf, A. & Olsson, B. (2002) Could correlation-based methods be used to derive genetic association networks? In Caulfield, H.J., Chen, S-H., Cheng, H-D., Duro, R., Honavar, V., Kerre, E.E., Lu, M., Grana Romay, M., Shih, T.K., Ventura, D., Wang, P.P., and Yang, Y., *Proceedings of The Sixth*

Joint Conference on Information Sciences, 1237-1242. USA: Association for Intelligent Machinery.

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

Miklos, G. L. & Rubin, G. M. (1996) The role of the genome project in determining gene function: insights from model organisms. *Cell*, *86*, 521-529.

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

Pearson, W. R. & Lipman, D. J. (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, *85*, 2444-2448.

Shannon, C. E. (1948) A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*, 376-423 & 623-656.

Somogyi, R. & Sniegoski, C.A. (1996) Modeling the Complexity of Genetic Networks: Understanding Multigenic and Pleiotropic Regulation. *Complexity*, *1*, 45-63.

Szallasi, Z (1999) Genetic Network Analysis in Light of Massively Parallel Biological Data Acquisition. *Pacific Symposium on Biocomputing*, *4*, 5-16.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. & Golub, T. R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America*, *96*, 2907-2912.

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

Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Barker, J. L. & Somogyi, R. (1998) Large-scale temporal gene expression mapping of central nervous system development. *Proceedings of the National Academy of Sciences of the United States of America*, *95*, 334-339.

Appendix A: Description of Algorithms

In this appendix a detailed description of the algorithms is given. The description provides, in pseudo-code, implementation details of how mutual information is applied to the expression data. C++ implementation can be obtained by request to b98thojo@student.his.se.

Each algorithm gets the input:

- *NR_INTERVALS*: holds the number of intervals that the histogram is divided into
- *NR_MEASUREMENTS*: holds the number of measurements in each profile
- *X*, *Y*: gene expression profiles to calculate mutual information between (implemented as arrays of expression measurements)

and utilizes the procedure construct_frequency_matrix, which calculates the frequency of measurements that falls within each interval for a given profile.

Algorithm: Basic Mutual Information (BMI)

Initialization:

fx[*NR_INTERVALS*] = construct_frequency_matrix (*X*)

fy[*NR_INTERVALS*] = construct_frequency_matrix (*Y*)

fxy[*NR_INTERVALS*][*NR_INTERVALS*] = construct_frequency_matrix (*X*, *Y*)

mi = 0

Recursion (for *it_x* = 0 to *NR_INTERVALS*)

p_x = *fx*[*it_x*] / *NR_MEASUREMENTS*

Recursion (for *it_y* = 0 to *NR_INTERVALS*)

p_y = fy[*it_y*] / *NR_MEASUREMENTS*

p_xy = fxy[*it_x*][*it_y*] / *NR_MEASUREMENTS*

mi = *mi* – *p_xy* * log2(*p_xy* / *p_x* * *p_y*)

End Recursion

End Recursion

Termination: Return *mi*

Algorithm: Time Delay Extension (TDE)

An extra input is given to the Time Delay Extension, *DELAY_LIMIT*, that holds the maximum delay between profiles *x* and *y*.

Initialization:

max_mi = BMI(*X*, *Y*, *NR_MEASUREMENTS*, *NR_INTERVALS*)

Recursion (*delay* = 1 to *DELAY_LIMIT*)

mi1 = BMI(*X* + *delay*, *Y*, *NR_MEASUREMENTS* – *delay*, *NR_INTERVALS*)

mi2 = BMI(*X*, *Y* + *delay*, *NR_MEASUREMENTS* – *delay*, *NR_INTERVALS*)

Condition (*max_mi* < max(*mi1*, *mi2*))

max_mi = max(*mi1*, *mi2*)

End Condition

End Recursion

Termination: Return *max_mi*

Algorithm: Complex Extension (CE)

A pair of extra inputs are given to the Complex Extension, *PROFILES* that is a array of all expression profiles and *NR_PROFILES* that is the number of expression profiles in *PROFILES*.

Initialization:

fx[*NR_INTERVALS*] = construct_frequency_matrix (*X*)

fy[*NR_INTERVALS*] = construct_frequency_matrix (*Y*)

max_mi = 0

Recursion(*z* = 0 to *NR_PROFILES)*

fz[*NR_INTERVALS*] = construct_frequency_matrix (*PROFILES*[*z*])

fxyz[*NR_INTERVALS*, *NR_INTERVALS*, *NR_INTERVALS*]

=construct_frequency_matrix (*X*, *Y*, *PROFILES*[*z*])

mi = 0

Recursion (*it_x* = 0 to *NR_INTERVALS*)

p_x = *fx*[*it_x*] / *NR_MEASUREMENTS*

Recursion (*it_y* = 0 to *NR_INTERVALS*)

p_y = fy[*it_y*] / *NR_MEASUREMENTS*

Recursion (*it_z* = 0 to *NR_INTERVALS*)

p_z = fz[*it_z*] / *NR_MEASUREMENTS*

p_xyz = fxyz[*it_x*][*it_y*][*it_z*] / *NR_MEASUREMENTS*

mi = *mi* – *p_xyz* * log2(*p_xyz* / *p_x* * *p_y* * *p_z*)

End Recursion

End Recursion

End Recursion

Condition (*max_mi* < *mi*)

max_mi = *mi*

End Condition

End Recursion

Termination: Return *max_mi*