

Towards Reverse Engineering of Genetic Regulatory Networks

Zelmina Lubovac^{*}, Björn Olsson
[zelmina,bjorne]@ida.his.se

*Department of Computer Science, University of Skövde,
Box 408, SE-541 28 Skövde, Sweden*

Abstract

The major goal of computational biology is to derive regulatory interactions between genes from large-scale gene expression data and other biological sources. There have been many attempts to reach this goal, but the field needs more research before we can claim that we have reached a complete understanding of reverse engineering of regulatory networks. One of the aspects that have not been considered to a great extent in the development of reverse engineering approaches is combinatorial regulation. Combinatorial regulation can be obtained by the presence of modular architectures in regulation, where multiple binding sites for multiple transcription factors are combined into modular units.

When modelling regulatory networks, genes are often considered as “black boxes”, where gene expression level is an input signal and changed level of expression is the output. We need to shed light on reverse engineering of regulatory networks by modelling the gene “boxes” at a more detailed level of information, e.g., by using regulatory elements as input to gene boxes as a complement to expression levels. Another problem in the context of inferring regulatory networks is the difficulty of validating inferred interactions because it is practically impossible to test and experimentally confirm hundreds to thousands of predicted interactions. Therefore, we need to develop an artificial network to evaluate the developed method for reverse engineering. One of the major research questions that will be proposed in this work is: Can we reverse engineer the *cis*-regulatory logic controlling the network organised by modular units?

This work is aiming to give an overview of possible research directions in this field as well as the chosen direction for the future work where more research is needed. It also gives a theoretical foundation for the reverse engineering problem, where key aspects are reviewed.

Keywords: genetic regulatory networks; artificial network; reverse engineering; modular architecture;

1 Introduction

In computational biology there is a great interest in understanding the regulatory mechanisms that control gene expression. Understanding how genes communicate in the networks of regulatory interactions is a step towards understanding the regulation of gene expression. Knowledge about

^{*} Corresponding author

regulation can in turn help us gain insights into diseases since many diseases are characterised by abnormal gene expression.

The advent of microarray technology, which allows us to measure expression levels of tens of thousands of genes simultaneously, has opened up new opportunities for identifying biologically plausible regulatory interactions between genes. Gene expression data from microarrays is used to construct expression profiles, which is often done by putting together expression levels from different experimental conditions, or time intervals. Similarities and differences between the expression profiles, as well as changes in expression levels, provide important insights into regulatory relationships.

A common assumption in earlier studies is that genes that share similar expression profiles, and thereby act similarly during a time course, are functionally related and may also be involved in the same regulatory mechanism. Therefore, clustering techniques, which constitute **the first generation** of expression analysis techniques, have been used for structuring expression data into co-expressed groups of genes. **The second generation** of techniques relies upon integrating other data sources with expression data. Identification of co-expression is now a prelude for further analysis, like scanning promoter sequences for regulatory elements (Zhu and Zhang, 2000), modelling metabolic pathways (Fellenberg and Mewes, 1999), etc.

Recent research has emphasised reverse engineering of regulatory networks from gene expression data. It relies on the following assumption: Given enough data on gene expression levels, it is possible to infer how genes are regulated. The ultimate long-term goal is to answer the question “what controlling mechanism caused the behaviour that we can observe in gene expression profiles?” However, we need to expand the definition of reverse engineering to encompass other types of biological data, especially when it comes to the more complex examples of gene regulation. This brings us to **the third generation** of analysis techniques which rely on integration of various data sources but also take into consideration more complex aspects of genetic regulation, such as the modular and combinatorial nature of regulatory elements.

In this report, we aim to show how far the research has come in the area of extracting regulatory networks. We will also address the question “What is needed in order to extract biologically plausible networks?” The emphasis will be on the second and third generation of analysis tools, with special focus on the regulation aspects that need further investigation. In section 2, we review some of the key concepts of regulatory networks. Section 3 describes the problem being studied in this work. Section 4 describes the second generation of analysis techniques including information sources that can be useful in the context of deriving regulatory networks. Section 5 describes the third generation of analysis techniques, and finally section 6 gives indications of what is left to do in the future work.

2 Key aspects of regulatory networks

In this chapter, the necessary concepts connected to regulatory networks that constitute the foundation for this work will be introduced.

2.1 Genetic vs. regulatory networks

There is a broad spectrum of different ways to describe genetic networks. This section will start with a description of genetic networks in a broad sense, which will gradually be narrowed down so that more specific aspects of genetic networks can be brought out.

The term **genetic network** can be defined as “an association of many genes which *interact* with each other in cascades or parallel pathways and achieve a specific function or various functions through such interactions” (Noveen et al., 1998, pp 378). Furthermore, Neveen et al. (1998) refer to the interactions between genes in a network as a way of direct or indirect adjustment of genes’ expression or function. The adjustment of gene expression occurs through stimulation or inhibition of transcription, translation or through post-translational modifications. This definition is broad in the sense that it allows regulatory interactions to occur at all levels of control of gene expression.

Wagner (2001) provides a similar definition where the focus is on *gene activity*, which might include changes in gene expression on the mRNA or protein level (methylation state, phosphorylation state, or alternative splicing). According to Wagner (2001), a genetic network consists of a group of genes in which individual genes can change the activity of other genes.

The advent of microarray technologies have led to more specified definitions of genetic networks, which are more suitable to reflect the level of data generated from large-scale microarray experiments, i.e. the transcriptome level. D’haeseleer (2000) used a model of gene networks that represents a higher-level of regulation consisting of regulatory interactions between genes rather than detailed biochemical mechanisms, as described in previous definitions. This definition of gene networks is very narrow compared to the previous definitions since it only allows mRNA levels to reflect regulatory interactions.

A common representation of regulatory networks that reflects D’haeseleer’s definition is shown in figure 1 where connections with arrows represent activation and flat-end connections represent inhibition of gene expression. Connections in the network can be assigned weights that denote the strength of activation/inhibition. In this representation, only the transcriptional state, i.e. the amount of mRNA, is considered to be an activation/inhibition factor. This definition is not complete, since we know that the post-transcriptional level is as significant for the genetic network as the transcriptome level. The implicit assumption that has been used in most gene expression studies is that there is a high correlation between the level of mRNA and the corresponding amount of protein produced in a cell, which is not always correct. Many important interactions occur at post-transcriptional levels of regulation, which can be involved in control of any of the many stages between gene transcription and the translation of corresponding mRNA in the cytoplasm (Latchman, 1998). Such data can not be observed in microarray experiments. A study performed by Hatzimanikatis and Lee (1999) shows that nonlinear stability analysis of continuous models of gene expression proves that describing models of gene networks requires information on both the mRNA and protein level.

In spite of those disadvantages, we should not underestimate the importance of the regulation at the transcriptional level. Hence, while waiting for efficient technology for development of large-scale protein chips, we will restrict this work to the regulatory interactions that occur on the transcriptional level.

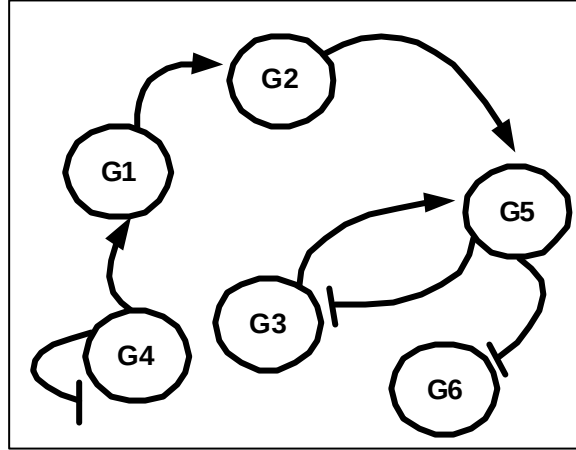


Figure 1: Representation of genetic network.

Even when we narrow down the definition of genetic networks to only consider regulation at the transcriptional level, using only gene expression data is not sufficient to derive regulatory interactions. There is still a need for integration with other information sources to derive regulatory networks. Those sources are reviewed in section 4.

2.2 The role of regulatory elements

A central point of gene expression control is regulation of transcription initiation. This process can be explained as co-operation between *cis*- and *trans*- acting regulatory elements. Simply stated, *trans*-acting transcription factors bind to *cis*-acting regulatory elements in upstream regions of the gene, which either enhance or inhibit transcription. In the next sections, the role of regulatory elements will be explained.

2.2.1 Transcription factors

The presence of transcription factors in regulatory networks is necessary, due to their important role to bind *cis*-regulatory elements, which in turn affects the initiation of transcription and thereby the gene expression level. The second generation of expression analysis tools has been focused on extracting regulatory networks by identifying potential *cis*-regulatory elements, which in turn could be coupled to corresponding transcription factors. Those approaches tend to draw an idealised picture of the problem of identifying regulatory elements. In previous work (Tavazoie et al., 1999; Hughes et al., 2000; Zhang, 1999; Brazma et al., 1998) this problem is described as a matter of finding a set of putatively “co-regulated” genes according to similar shapes of their expression patterns. In their upstream regions they should also share a set of common transcription factor binding sites (TFBs), which in turn can help deriving the connection to transcription factors.

Identifying this connection to transcription factors, which play a major role in regulatory networks, is not as simple as it seems. There are some difficulties that are reviewed in this section.

Regulatory regions that are involved in transcriptional control of eukaryotic genes often contain several TFbs. The hope is that the TFbs point towards transcription factors that play a key role in the transcriptional regulation of the corresponding gene. However, several problems make it difficult to define connections to transcription factors in genetic networks by studying expression data (“grouping genes by expression”) and selecting over-represented oligonucleotides which are treated as putative TFbs. One of these problems, as stated in Pedersen et al. (1999), is that TFbs can bind several different members of a family of transcription factors. The opposite case is also usual, i.e. “there can be great variability in the binding sites of a single factor, and the nature of the allowable variations is not well understood” (Sinha and Tompa, pp 1, 2000). In SCPD (a promoter database of the baker’s yeast) for example, for 200 genes there are 203 binding sites which are in single copy, 69 in two copies, 19 in three copies, 9 in four, 3 in five and 1 in six (Zhu and Zhang, 1999). An example of the complexity of a regular network is the sea urchin *Endo16* gene where the upstream region contains at least 33 TFbs which are organised in modular units (Yuh et al., 1998).

Another difficulty is that the choice of which transcription factor will bind sometimes depends on the availability of transcription factors for binding in a given cell type. The complexity increases even more if we take into consideration that a binding site might never be bound if some sequence elements that must be recognised correctly by the transcriptional mechanism are not present under certain conditions when the gene is transcribed (Pedersen et al., 1999).

Current approaches of analysis of promoter regions of co-expressed genes to find significant regulatory motifs that correspond to putative TFbs are not sufficient to make the pieces of the regulatory network puzzle fit together. Additional sources to locate regulatory regions have been proposed, like phylogenetic footprinting for example (Fickett and Wasserman, 2000).

2.2.2 Binding sites

As mentioned before, transcription factors play a major role in the control of gene transcription. They are proteins which interact with different classes of DNA targets (TFbs) and with each other. For the purpose of this work, only the most usual classes of binding sites will be described here.

The *core promoter elements* are placed immediately upstream of the transcription start site. One of the best characterised core promoter elements is the TATA element which is usually located 25 bp upstream of the transcription start site (Nikolov and Burley, 1997). The TATA box is involved in directing RNA polymerase II to begin transcription and it works most efficiently together with *promoter proximal elements* (Griffiths et al., 1996). Promoter proximal elements can be placed anywhere between 50-200 bp upstream of the start site. The CCAAT box is an example of a proximal element (Griffiths et al., 1996). Those elements typically affect the efficiency of transcription initiation. *Enhancer elements* can be found very far from the transcription initiation site in either direction and they can greatly increase the transcription rate. Similarly, *repressors* can inhibit the transcription rate. Those regulatory sites, which are located on the DNA molecule and bound by *trans*-acting proteins in order to control expression, are also called *cis*-regulatory elements. Hence, the interplay between *trans*- and *cis*-acting elements is an important molecular mechanism which influences the gene expression patterns we observe in e.g. microarray experiments.

The idea of linking genomic sequence and expression data by associating transcription factor binding sites with expression profiles of groups of genes is very well established.

2.3 Coregulation

The term coregulation is very widely used in gene expression studies due to the assumption that coregulated genes should be members of the same regulatory network. A general idea of coregulation is following: Groups of genes which are subjected to regulation by the same transcription factor (“master switch”) or participate in same regulatory pathway are coregulated. There are also many special cases of coregulation that do not follow this definition strictly (see section 2.4). In this chapter, we will highlight different contexts of using the term coregulation.

Some studies imply that genes which share similar patterns of expression (coexpressed genes) during a time course should also be coregulated. Coregulation between a pair of genes is also claimed if genes share a significant number of regulatory elements. However, coexpressed genes which do not share regulatory elements may be indirectly coregulated, i.e. subjected to regulation of, as Yanai et al. (2002) state, a “master controller” which governs more than one regulatory element.

Another common description of coregulation model refers to proximity of the genes at the chromosome. A phenomenon that exemplifies this coregulation model is that a methyl transferase gene is found adjacent to a restriction enzyme gene along the chromosome, in all known occurrences (Yanai et al., 2002). An explanation of this type of coregulation is that both genes need to be switched on at the same time for the defence system to work satisfactorily, which is facilitated by the proximity of those two genes. In prokaryotes, there are sets of genes that are known to be transcribed together and strongly coregulated. These sets of genes that are transcribed together into mRNA to be translated to proteins are known as operons.

A similar “proximity based” mechanism of coregulation in eukaryotes is explained in Kel-Margoulis et al. (2002). They state that specific combinations of transcription factor binding sites located in a close proximity to each other, also called “composite clusters” play a key role in coordinating regulation. This “cluster”- or “modular”- based reasoning, which often relies on synergistic binding of adjacent transcription factors and depends on the order and spacing of the binding sites, is not taken into account by today’s methods for identifying regulatory networks.

McGuire and Church (2000) refer to coregulation in terms of presence of significant regulatory motifs within the set of genes being coregulated (regulon).

It is important to distinguish coregulation from coexpression because coexpression can occur purely by coincidence while coregulation is the consequence of a common molecular mechanism (Werner, 2001).

2.4 Modularity

In this section, a specific type of gene regulation, based on the combinatorial nature of regulation, will be discussed. It is often referred to as the modular organisation of gene regulation due to the fact that several transcription factor binding sites, which are organised in functional units called modules, play a crucial role in gene transcription (Klingenhoff et al., 2000). The traditional

approaches of analysing microarray data do not explicitly address this modular nature of regulation.

Pilpel et al. (2001) discussed that the basic idea behind the modular approach is that transcription factors can be considered as “words” which can be compiled to “sentences” to regulate gene expression. In more detailed, it means that genes can consist of subelements of the DNA sequence, each of which can execute a certain regulatory function. Transcription of a gene can for example depend on which combination of modules is involved in regulation, so that same gene can be expressed in one condition and repressed in another. Each subelement is referred to as a regulatory module and contains multiple target sites for DNA binding factors (Yuh et al., 1998).

This problem has been actualised with the enormous flow of expression data generated from microarrays. All attempts to identify regulatory networks by combining regulatory elements and expression data have revealed more difficult problems, like identifying the synergistic effect of transcription factors, for example. This means identifying a combination of regulatory elements that control the gene expression pattern. Still, there is a lack of methods for reverse engineering of regulatory networks which take into consideration such more complex aspects of regulation. This is probably due to the fact that it takes many years of experimental work to verify *cis*-regulatory systems which are characterised by modular architectures. Examples where we can find this type of regulatory system include the sea urchin *Endo16* system (Yuh et al., 1998) and *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome (Berman et al., 2002).

In spite of the experimental evidence that *cis*-regulatory information is organised into modular units in higher eukaryotes, we are still mostly using traditional computational methods for modelling genes. A gene is often considered as a “black box” where the gene expression level is an input signal and changed level of expression is the output. Figure 2 is an example of modular regulation which shows how the input and output of the “gene box” can vary depending on circumstances of transcription. We need to shed light on reverse engineering of regulatory networks by breaking down information going in to the gene box into more detailed units of information, e.g., by using regulatory elements as input to the gene box as a complement to expression levels.

As mentioned before, we can find a very well studied example of modularity in the sea urchin *Endo16* regulatory system. Two different regulatory effects in *Endo16* are illustrated in figure 2. A general pattern of modularity is described in following hypothetical example. Assume that transcription factor (TF) A alone induces transcription to 60 % of the maximum value. Assume also that TF B does not stimulate expression when it acts individually. On the other hand, when A and B act synergistically, they induce expression at the maximum rate. Hence, their synergistic effect is different from the simple sum of their individual effects. In this case, binding sites for TFs A and B can form a module, which is a functional unit that can regulate expression in different stages and acts autonomously. Now assume that there is a TF C which together with TF B represses transcription. Hence, different combinations result in different effects on gene transcription.

Consider the example illustrated in figure 2 where module A communicates the output of all upstream modules to the core transcription apparatus. In figure 2A, it is shown how modules

CDEF repress the gene. Figure 2B shows how synergism with modules B and G enhance the expression of the gene in later development.

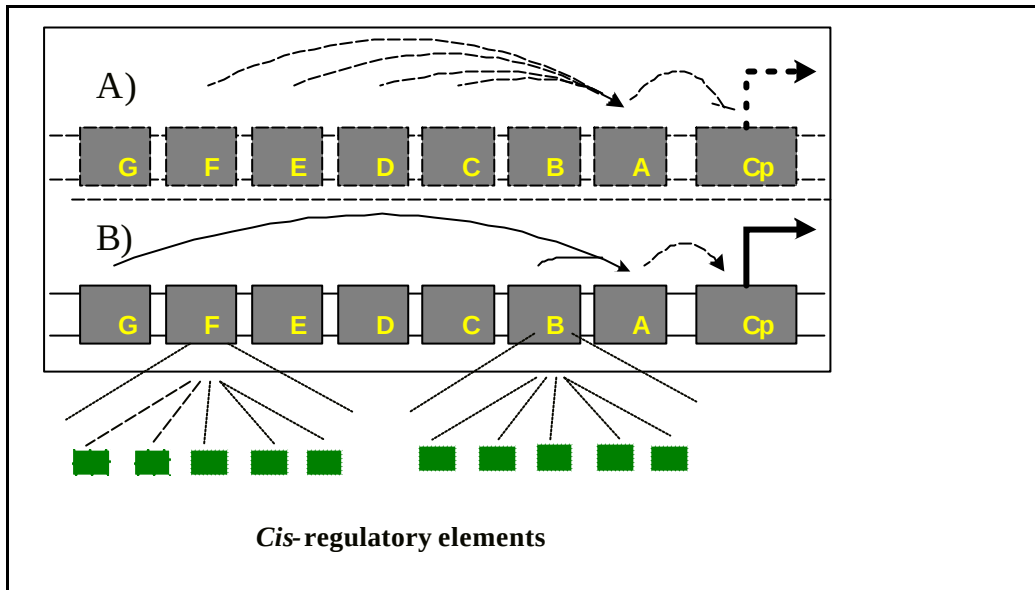


Figure 2: The figure gives an example of modular architecture.

This example shows, as Yuh et al. (1998) refer to, part of genomic *cis*-regulatory logic for sea urchin gene *Endo16*. *Cis*-regulatory logic means, in other words, regulatory programming code specified in *cis*-regulatory DNA sequence, which mediate complex developmental patterns of expression. Example in figure 2 is a part of large regulatory network described in Davidson et al. (2002) where *Endo16* gene is the most closely examined part. It took the last two decades of molecular studies to partly crack this code, or at least provide a wealthy of data that allow researchers to crack the regulatory network that controls the specification of endoderm and mesoderm in sea urchin embryo. It is important to stress that deriving of this *cis*-regulatory network was a result of interplay of multiple approaches. An integration of experimental embryology, *in situ* hybridisation, knock out experiments, comparative sequence analysis has resulted in predictions of *cis*-regulatory binding sites for each gene in the network, which are then tested with classical *cis*-regulatory analysis using artificial reporter constructs (Brown et al., 2002).

2.5 Artificial regulatory network

An important aspect in developing methods for reverse engineering of regulatory networks is testing of the validity of developed methods. How can we decide whether a method has succeeded to reverse engineer the regulatory network from data the source we used? This question was also addressed by Smith et al. (2002). They propose developing a simulated biological system in which all regulatory rules are known, and running various simulations to

generate data from the system. We can then apply various reverse engineering algorithms on simulated data to recover the original underlying network. In this way, the accuracy of the methods for reverse engineering can be tested and, if necessary, modified to obtain a better agreement to the original network. In Smith et al. (2002), a simulator was developed to generate data at multiple levels of organisation: behaviour, neural anatomy, brain electrophysiology, and gene expression of songbirds. The reverse engineering algorithm, called NetworkInference, which uses a Bayesian network to model the system that govern vocal learning in the songbird brain, provided high recovery success on the simulated data.

Gene regulation has also been frequently modelled with Boolean networks. In Boolean networks, the state of each gene can be defined as either *on* or *off*. As discussed before, gene regulation has a combinatorial nature, which can be reduced to wiring diagrams (Somogyi and Sniegowski, 1996). Each element in the wiring diagram uses a logical or Boolean rule to compute its value based on the values of the other elements it is connected to (see figure 3).

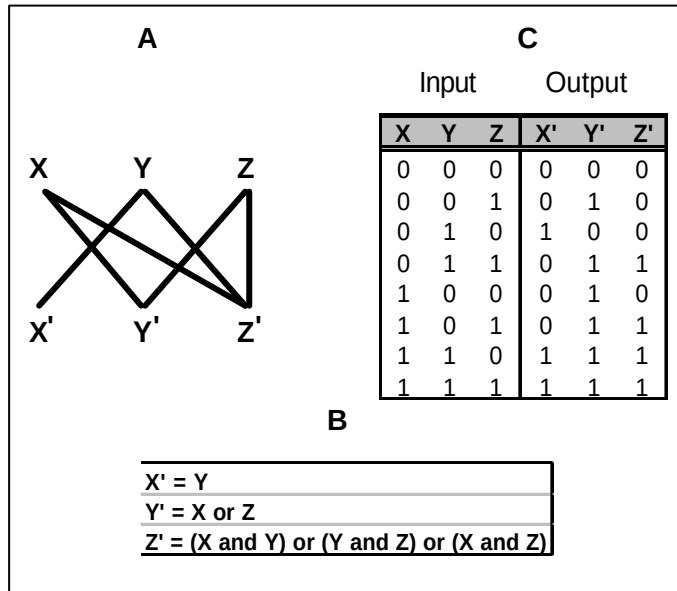


Figure 3: An example Boolean network. A) Wiring diagram, B) Logical rules, C) State transition table, where the input corresponds to state at time t and the output corresponds to state at time $t+1$.

Liang et al. (1998) show an example Boolean network model representing a proximal genetic network, which is also given in figure 3. 3A) shows the wiring diagram, where connections between elements (genes) correspond to regulatory links and the rules in 3B) govern the results of regulatory interactions given a set of input values. 3C) shows the state transition table, where the input corresponds to state at time t and the output corresponds to state at time $t+1$.

Testing the algorithms for deriving network structure from state transition measurements requires knowledge of the original network, which we often do not possess when it comes to biological systems, such as gene regulation. As Liang et al. (1998) state, we can use Boolean networks to simulate some conditions from the “living” regulatory networks, such as size, connectivity etc. and analyse how our reverse engineering algorithms succeed with the background of those varying assumptions. However, if we consider more complex regulation examples, such as regulation of *Endo16*, discussed in section 2.4, where ordering and various combinations of *cis*-regulatory elements affects the regulation, we realise that we need more biologically plausible rules for generating transition states. Those rules can for example be based on data sources that could influence gene regulation. Some of those sources are reviewed in chapter 4.

3 Problem and main research questions

Many biological processes of interest (e.g., diseases, cellular differentiation during development etc.) are controlled by complex interactions over time between hundreds of genes. Furthermore, each gene can be involved in multiple functions. Given the fact that thousands of genes are involved in determining the functional state of an organism, the task of identifying underlying genetic networks is very difficult.

For that reason, large-scale gene expression measurements from microarray experiments have been seen as a possibility to analyse many genes’ behaviour at a transcriptional level simultaneously and hopefully discover interesting gene interactions. The question that many gene expression studies address is: “Given the amount of gene expression data representing genes’ behaviour over a time period, what can we deduce about the underlying regulatory network of interactions that caused that behaviour?” This problem is also called the reverse engineering problem. The goal with reverse engineering is to learn about the structure and the dynamics of the regulatory network only from the measurements/observations of the network, i.e. microarray data. Hence, no detailed information about the structure of the system should be introduced to the reverse engineering algorithm

However, this definition has been criticised due to the fact that microarray data do not provide all information necessary to derive the underlying network. The source of information which is not observable in microarray experiments is the regulation occurring at the post-transcriptional level. Many important regulatory interactions occur at the post-transcriptional level of regulation, which in turn can be involved in control of any of the many stages between gene transcription and the translation of corresponding mRNA (Latchman, 1998). According to a study performed by Hatzimanikatis and Lee (1999), nonlinear stability analysis of continuous models of gene expression proves that describing models of gene networks requires information on both mRNA and protein level. It has been shown earlier in other studies (D’haeseleer, 2000; Szallasi, 2001) that reverse engineering of the complete network is still not possible, due to the lack of gene expression data. Although those datasets usually are very large, they are large along the wrong dimension, i.e. they usually consist of many variables (genes) but few measurements for each variable.

Other examples that stress the need for additional information are regulatory networks that govern modular architecture (see example in section 2.4). To be able to understand the relationships between the modular architecture of regulation and the resulting patterns of gene

expression data, networks need to be fed with information about *cis*-regulatory elements, i.e. the possible ways of combining them and ordering among different modules.

The discussion above brings us to one important aspect for further research, namely forward modelling. It means that we build a regulatory architecture of interacting elements based on, for example, expert-based input (Szallasi, 2001). Such *in silico* regulatory networks can be used to generate as much time-series data as we need so that we can avoid the dimensionality problem. Since we set up all the rules of regulation, it is easy to verify the derived connections against the original ones (see section 2.5).

Another aspect that is worth consideration when developing algorithms for reverse engineering is validity testing of the developed methods. We need to be able to answer following question: How can we measure that our algorithm performed well in deriving regulatory connections? In many biological systems, including regulatory networks, we do not have detailed knowledge of the underlying connections. For example, even if we work with the expression data generated from, what we believe is a “known” regulatory network we can never be sure that all regulatory connections in the network are already found or that the identified connections are correct. Assume that we have identified the connection $A \longrightarrow C$, i.e. gene A activates gene C. We can, later on, discover that this resulted from a long range of interactions, e.g. A represses D and in the absence of D, gene G is transcribed and its product finally activates gene C. Unfortunately, it would be too time consuming to experimentally verify every connection that has been derived by the algorithm. On the other hand, it is hard to claim effectiveness of the algorithm if connections can not be verified. This issue motivates also forward modelling of regulation, where we set up all rules for regulation, which allows us to verify the obtained results.

During the past years, expression analysis methods have mostly been concerned with finding similarities in gene expression profiles, which can help discovering genes that act in a co-regulated manner and may share the same regulation mechanism. That is why statistical methods like clustering have become basic tools in the field of gene expression analysis. The most frequently used clustering techniques in gene expression analysis are hierarchical clustering (Eisen et al., 1998), SOMs (Tamayo et al., 1999) and K-means clustering (Soukas et al., 2000). The problem with this assumption made in most clustering studies is however that regulation can be both convergent (similar expression patterns, different control regions) and divergent (similar control regions, but their joint action is such that the effect on expression is different) (Altman and Raychaudhuri, 2001), which sometimes makes the use of clustering algorithms and expression data insufficient as information source for identifying genes that share common regulatory mechanisms. There is a great interest in combining expression profiles with important knowledge such as gene function, regulatory elements, sequence information, since the clustering approach does not take full advantage of those information sources.

The discussion above gives the background for the following research questions:

- What is an appropriate way to incorporate modular architecture in forward modelling of regulatory networks?
- Is it possible to reverse engineer regulatory networks including modular information from time series of expression data sampled from the model?

- Which information sources are useful to combine with gene expression analysis in order to model biologically plausible regulatory network?
- What is an appropriate way of combining those information sources?

Those research questions need to be answered to be able to meet the main goal, which is defined as follows:

Develop a procedure for identifying putative regulatory networks which encompasses the whole chain from the forward modelling to reverse engineering and validation of regulatory networks by combining gene expression profiles with additional biological information sources.

4 The second generation of analysis techniques

The goal of this project is to develop the analysis method for identifying gene-to-gene relationships with an integrated approach, i.e. integrate different levels of information sources. The approach is intended to serve as a mechanism for deriving interesting hypotheses about relationships between genes that can guide further experimental investigation. The long-term goal is to determine biologically plausible genetic networks of regulation. One of the mechanisms that the approach will be applied on is freezing tolerance in oats. Due to the multigenetic nature of the freezing tolerance, which implies a complex behaviour determined by multiple genes, the problem of identifying interactions between regulatory genes should be approached with bioinformatics methods. In this chapter, section 4.1 describes the basic idea that the approach will be built on and possible information sources that will be integrated. Section 4.2 describes some specific statistical and computational techniques that can be useful in this work, and finally section 4.3 describes the methods to be used to test the significance of the achieved results.

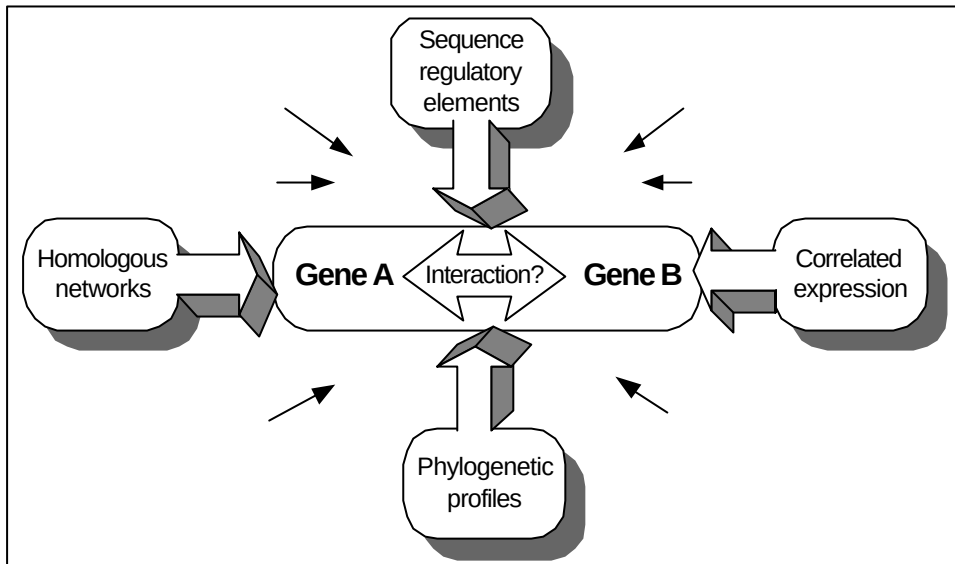


Figure 4: Information sources that could be useful in inferring gene-to-gene relationships.

4.1 Possible information sources

Gerstein and Jansen (2000) discern two parts of expression analysis, namely the “internal” and “external” part. In the internal part, the numerical structure of data is analysed with for instance SOM, hierarchical clustering, PCA etc without taking into consideration other sources of biological information, like protein function or structure. Instead, the relating of expression data to other, external sources of biological information is done in the external part of expression analysis. One of the important steps in this work is to choose external information sources that will be suitable for the intended purpose of the work, i.e. to identify networks of gene interactions. For this purpose, some previously used sources and some proposed sources that could be useful will be reviewed.

4.1.1 Regulatory sequence elements

In previous work, there are some examples of systematic use of external sources to improve deriving of regulatory mechanisms.

In van Helden et al. (2000), *regulatory sequence analysis* has been applied on clusters of co-expressed genes. The question that van Helden et al. (2000) deal with is “is there any underlying mechanism that causes that genes belonging to the same cluster act in a coordinated fashion with respect to transcriptional response?” The answer to this question needs a search for transcription factors that might act simultaneously on the genes belonging to the same cluster. The approach is based on analysing upstream sequences to discover shared DNA motifs, which could correspond to regulatory elements, also called *cis*-acting elements. Potential *cis*-acting elements are then

matched against databases of known transcription factors, which yield putative transcription factor binding sites that could correspond to *cis*-acting elements. This approach seems to be useful in suggesting very likely putative regulatory elements, when small clusters are analysed. However, it is less sensitive when large gene clusters are analysed, since larger clusters (>50 genes) are less likely to be regulated by a single transcription factor (van Helden et al., 2000). This fact can be utilized to build predictive methods for regulatory elements, which control gene expression at the transcriptional level. Hereby, systematic mapping of those elements can be an important step towards determination of the genetic network architecture.

Similar explorations of transcriptional regulatory networks based on clusters of co-expressed genes have been done by Tavazoie et al. (1999), Hughes et al. (2000), Zhang (1999), Brazma et al. (1998) etc. A method called AlignACE has been frequently applied for this purpose. The method returns a series of DNA binding motifs for unspecified transcription factors by searching the regions upstream of the translational start site. Since this method has proven to uncover a set of new regulons (sets of co-regulated genes), and their *cis*-regulatory elements (Tavazoie, 1999), it seems like a suitable source to consider in the integrated approach that will be taken in this work.

Combining regulatory sequence analysis with expression analysis, means, simply stated, that we start from the “big picture” consisting of gene expression profiles that determine the functional state of the cell through innumerable regulatory and other ongoing interactions between genes and gene products. Further, we seek to break down this picture to smaller manageable elements (regulatory elements) that can explain the “big picture”.

4.1.2 Homologous gene networks

The second information source consists of *known regulatory networks* in the species that are related to the organism being studied. We know that many regulatory networks are conserved across the species and we can utilise comparative methods to identify corresponding networks in the studied organism. Noveen et al. (1998) state that genetic networks are under tight evolutionary constraints not to change their circuitry (the way they interact) and function. Since gene networks often act in cooperation with other networks, any changes that imply disruption of the network’s function may lower the fitness of an individual (Noveen et al., 1998). That is why genetic networks usually are stable and have conserved interactions during evolution and speciation. Furthermore Noveen et al. (1998) discuss the concept of *homologous gene networks*, where the members of two networks have similar gene structure and ways of interaction. There are two types of homologous gene networks: *paralogous* and *orthologous* gene networks. Paralogous networks can be found in individuals of the same species, while orthologous networks can be found in individuals of different species. The networks of gene interactions that are conserved across several species are also called “modules”. Van Dassow and Munro (1999) suggest that evolution often proceeds by modifying the relationships between modules (conserved genetic networks), instead of operating on isolated genes. For the purpose of this work, it can be useful to map the conserved interactions across the different organisms to apply this knowledge on the studied mechanism.

The term phylogenetic footprinting (Fickett and Wasserman, 2000), which seems to be closely related to the information source described in this section, has also been suggested as a useful source to be coupled with expression data for unravelling higher order organization in the regulatory language.

4.1.3 Expression profiles

The third information source consists of large *scale gene expression measurements*. Many computational techniques for inferring genetic networks have emerged as a result of the flood of data generated from simultaneous gene expression measurements (often by using microarray technology), where gene expression is quantified as mRNA levels. The expectation is that by comparing complex expression patterns, it will be possible to decipher the regulatory network responsible for the behaviour that can be monitored in patterns. However, it is important to be aware of the limitations of microarray gene expression data. First of all, transcripts that are present at low levels, which are likely to include transcription factors, may not be detected by microarray experiments. Cross-hybridisation is another problem, where a target transcript hybridizes to incorrect probes of sufficient similarity. Lastly, since mRNA levels are used as indicators of gene activity, it is often assumed that there is a significant correlation between the amount of mRNA and the amount of the protein corresponding to mRNA. However, there is evidence that this correlation can vary depending on the type of protein. Microarrays also do not reflect post-transcriptional modifications, which is yet another reason why we should have a combined approach in order to enhance our confidence in hypotheses about gene interactions.

4.1.4 Phylogenetic profiles

The fourth information source concerns *phylogenetic profiles*. Phylogenetic profiles can be constructed as a string with n entries, each one bit, where n denotes the number of genomes. The presence of the homologs to a given gene in the n_{th} genome is marked with an entry of unity, while the absence of homolog is marked with a zero on the n_{th} position (Pellegrini et al., 1999). Genes can then be clustered based on the similarity of their phylogenetic profiles, which according to Pellegrini et al. (1999) results in clusters of genes that should be functionally associated to each other. Links or associations between genes can be derived according to clusters of profiles and criteria for linking profiles and determining similarity can vary from linking genes with exactly the same profiles to linking genes whose profiles differ at for instance one or two entries. There is also conflicting evidence regarding the hypothesis that genes with similar profiles among different genomes exhibit similar function. For instance, DNA polymerases and ribosomal proteins can be found in all genomes, but there is no functional association between them neither are they involved in the same process. Hence, such profiles are less informative for the intended purpose of identifying associations between genes. However, this information could be useful for identifying genes that are necessary and sufficient to maintain cell existence.

4.2 Integrating different sources of information

The main idea with the project is to develop a method that integrates different levels of gene information sources: homology (using prior knowledge about gene-gene association in other organisms and apply the knowledge on the organism being studied), phylogenetic profiles and gene expression, sequence regulatory elements etc., as figure 4 illustrates.

One way of building hypothetical interactions between genes is to start from prior knowledge about genetic interactions and homology matching. Theoretically, this can be done by comparing genes across several genomes and build profiles for each pair of genes that reflect gene-to-gene relationships in different genomes. This information could be used to select the more conserved interactions, i.e. interactions that are represented in several species and also to analyse which

interactions are unique for an organism. Gradually, we can build interactions that include more remote genes by adding genes that are coupled to a certain gene via a third gene. Rules for merging several genes in a network of interactions must be carefully determined.

Another way would be to use only expression profiles and build hypothetical networks by clustering genes with significantly high correlation of expression profiles. The reason for selecting subsets of genes that are highly correlated with each other is that it is often assumed that genes that share the same functional class should be co-ordinately regulated at the level of transcription under the condition tested.

The third information source that figure 4 shows concerns phylogenetic profiles (Pellegrini et al., 1999). Genes can then be clustered based on the similarity of their phylogenetic profiles, which according to Pellegrini et al. (1999) results in clusters of genes that should be functionally associated to each other. This method can be used to predict the function of the unknown protein.

An integrated approach means using both phylogenetic profiles and correlated expression to generate hypotheses about interactions between genes. Hopefully, those profiles will provide a clue about linkage between genes.

How can we know that the inferred interactions are reliable, i.e. biologically plausible? In order to validate the interactions, we can use at least two sources of biological information to infer the interactions and validate against the third source (see figure 5). Hence, the hypothetical networks generated from gene expression and phylogenetic profiles can be validated against “trusted” networks that are generated from the known interactions, also called homologous networks. Likewise, we can do this procedure the other way around, i.e. start from the known interactions (homologous networks) between genes and their phylogenetic profiles that reflect evolutionary history and generate general patterns of gene expression, i.e. how genes are expected to behave on the transcriptional level over a time period.

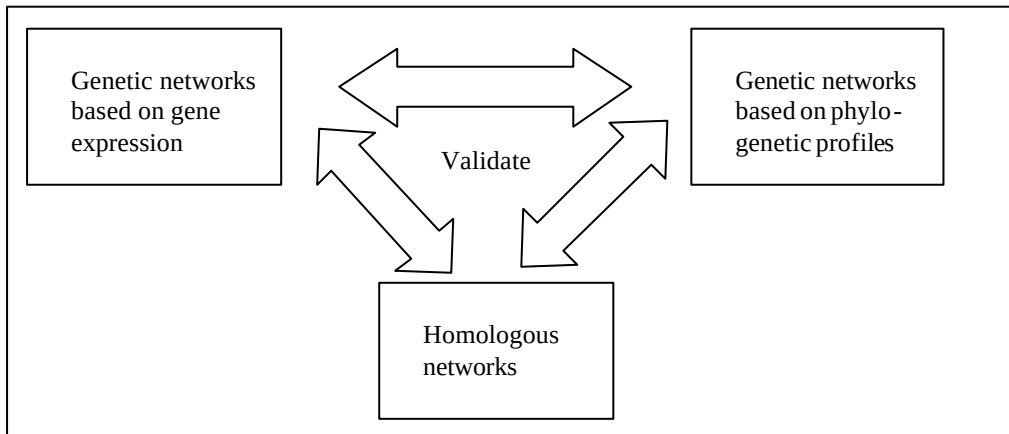


Figure 5: Validation procedure of inferred genetic networks of interactions.

If the agreement between the networks obtained from different sources is significantly high, this method can be used to make predictions about which genes can possibly interact with each other, which in turn can give us a clue about the functional roles of genes.

4.3 Specific methods

This section will describe different methods that concern each of the information sources described in the previous section. The existing integrative approaches for inferring genetic networks will be discussed as well as initial ideas of the integrative approach that will be applied in this work.

4.3.1 Correlation-based approach

There are different ways to express similarity between objects, but most often we use various distance and correlation measures. Observation of correlation between variables has been used in biology for a long time. Simple correlation coefficients (Pearson) reveals linear relationships between genes based on their expression profiles, i.e. it shows if mRNA concentrations co-vary linearly with the time or experimental condition. In D'haeseleer et al. (1998; 1999), Pearson correlation has been applied to static gene expression patterns, and associations between genes have been derived based on significantly correlated expression patterns. Other correlation measures, like Rank correlation for detection of non-linear but tight relationships between variables has also been used for identification of relationships between genes (D'haeseleer et al., 1999). Another similarity measure used in the gene expression analysis is Euclidean distance between patterns of gene expression, which is closely related to positive linear correlation. In terms of gene regulation, significantly high positive or negative correlation between gene A and B can mean the following: gene A regulates gene B, gene B regulates gene A, both genes are regulated by a third gene C (genes are co-regulated), both genes regulate third gene C, or it can happen by accident.

Jackknife correlation

One problem with linear correlation is its sensitivity to outliers. Sometimes, two expression patterns can be totally unrelated in all but one time point and still be highly correlated (Heyer et al., 1999). Removal of the outlier in such a case would show completely insignificant correlation. Such outliers can be the result of experimental error. Heyer et al. (1999) suggest jackknife correlation for making the correlation approach more robust to single outliers and reduce false positives, i.e. pairs of genes with dissimilar curves of expression that are highly scored by the similarity measure. Assume that we have two expression profiles consisting of N data points, where each data point corresponds to an experimental condition c . For $c=1$ to N , the c th observation is deleted, correlation is calculated and c is increased with 1. Correlation is then recalculated with the restored values and observation $c+1$ deleted. As a result of all calculated correlations between a pair of expression profiles, jackknife is the minimal correlation (Heyer et al. 1999). This measure is not only advantageous for reducing the sensitivity to outliers, but also in situations where some genes exhibit similar expression patterns in some condition but not in others. The jackknife correlation process can be implemented systematically or randomly deleting data points for various conditions to find subsets where genes act in a significantly correlated manner with respect to their profiles.

Time-delayed correlation

Another problem regarding analysis of simple correlation of expression patterns is that it is not able to capture time-shifted relationships between genes, such as cascade-like regulation mechanisms, which are very common. Calculating time-delayed correlation between variables for all pairs of different lags can discover such relationships (CBCG, white paper).

The limitation of both time-shifted and linear correlation measures is that neither of those measures is able to capture non-linear relationships between variables.

Linear correlation of dynamic gene expression

There are also different ways of representing expression patterns, i.e. the *static* and *dynamic* way, as introduced in (Reis et al., 2001). Static expression patterns have been used most frequently and refer to simple temporal gene expression patterns ($\text{Exp}_{\text{time1}}, \text{Exp}_{\text{time2}}, \text{Exp}_{\text{time3}}$) that consist of the expression level for each time point. Dynamic gene expression (Reis et al., 2001) refers to the rate of the change in expression level over a time course ($\text{Exp}_{\text{time2}} - \text{Exp}_{\text{time1}}, \text{Exp}_{\text{time3}} - \text{Exp}_{\text{time2}}$). The motivation for using dynamic gene expression is that the dynamics can reveal in which way the expression level of one gene leads to change in the expression level of other gene.

Reis et al. (2001) also applied a linear correlation coefficient as a measure of association of dynamic gene expression profiles. The genetic networks obtained from this method are called relevance networks. Simply stated, a matrix of pairwise gene to gene associations based on linear correlation is calculated and a statistically significant threshold level of association is determined. This level is obtained by permuting the original data set according to the following: the distribution of gene expression values is preserved, but the link between the expression value and the particular time point or experimental condition is broken. After the matrix has been permuted a large number of times and the similarity/association score recalculated each time, the maximum association value obtained from permutations is used to set the minimum threshold value for selecting significant association in the original unpermuted data set. All connections weaker than this threshold are removed, leaving only significantly correlated genes that are interconnected in relevance networks.

4.32 Graph theoretical approach based on genetic perturbations

Wagner (2001) developed a method for reconstructing interactions in gene networks from the effects of genetic perturbations on gene activity. Gene activity is according to Wagner (2001) more than just gene expression, which differs from many other definitions that refer to gene expression, i.e. whether a gene is expressed or not, as mRNA or as protein. Other examples of regulators of gene activity we might consider are phosphorylation and methylation. Wagner (2001) brings an important point concerning using correlation coefficient as described in section 4.3.1. The problem with approaches using correlation coefficient is that correlation only can point to regulatory interactions between genes, and not infer them. Wagner (2001) tries to solve this problem, i.e. resolve the causal structure of a genetic network. For this purpose, large-scale perturbation data has been used, in form of mutations, overexpression, deletion and inhibition of genes or gene products. Wagner (2001) used a directed graph consisting of nodes (representing genes) and directed edges (representing causal regulatory interactions between genes) to represent genetic network. From this graph, two lists have been generated: adjacency list which for each gene G lists all other genes that are directly influenced by gene G and accessibility list

which for each gene G lists all genes that are accessible from gene G , i.e. all genes that are affected by perturbation of gene G . The problem that is addressed is to reconstruct the network structure from its accessibility list for large networks of thousands of genes. Limitation of the approach is that it is only possible to resolve acyclic graphs, i.e. graphs without cycles.

5 The third generation of analysis techniques

5.1 Combining gene expression with regulatory motifs

Common strategies for identifying sequence motifs that are potential transcription factor binding sites have been focused on a “cluster by expression” approach, where clusters with similar expression patterns are identified and their transcription control regions are examined to find shared sequence motifs.

In Chiang et al. (2001), the relationship between expression and sequence data has been evaluated with Genome Mean Expression Profiles (GMEP). This approach assumes that the expression pattern of genes that contain the motif with transcriptional information in their TCR (transcriptional control region) should exhibit non-random behaviour that reflects the activity of the transcription factor. On the contrary, one should not expect the gene expression profiles of the group of the genes to deviate significantly from the entire population if their motifs do not encode transcriptional information.

This hypothesis is tested with GMEP which is calculated for each motif as a mean of all expression profiles of genes containing this motif. The Genome Mean Expression Profile is defined as (Chiang et al. 2001):

$$\text{GMEP}(m)_j = \frac{\sum_{g \in G} w_{mg} * E_{gc}}{\sum_g w_{mg}} \quad (1)$$

where G is a set of genes that contain the motif m , E_{gc} is expression level of gene g in condition c , and w_{mg} is the number of occurrences of motif m in the genome sequence of the gene's nominal transcription control region.

The starting point for computing GMEP is matrix M where each row denotes a gene and each column denotes the experimental condition. Each entry E_{gc} denotes then expression level m of gene g in condition n .

Gene g	Condition c			
	Eg_1c_1	Eg_1c_2	...	Eg_1c_n
	Eg_2c_1	Eg_2c_2	...	Eg_2c_n
	Eg_3c_1	Eg_3c_2	...	Eg_3c_n
	...	...	...	...
	$Eg_m c_1$	$Eg_m c_2$	...	$Eg_m c_n$

Table 1: Matrix M where each cell Eg_{mn} represents the expression level of gene m in condition n .

Motifs in the alphabet $\{A, C, G, T\}$ are counted and motifs have a fixed length L . For each L and matrix M containing expression values (see table 1), a matrix F ($4^L \times c$) is computed where c is the number of experimental conditions. In matrix F each row corresponds to the GMEP of a single motif (see table 2).

Motif m	Condition c			
	GMEP m_1c_1	GMEP m_1c_2	...	GMEP m_1c_n
	GMEP m_2c_1	GMEP m_2c_2	...	GMEP m_2c_n
	GMEP m_3c_1	GMEP m_3c_2	...	GMEP m_3c_n
	GMEP m_4c_1	GMEP m_4c_2	...	GMEP m_4c_n
	GMEP m_5c_1	GMEP m_5c_2	...	...
	...	...	...	...
	GMEP $m_4^L c_1$	GMEP $m_4^L c_2$	...	GMEP $m_4^L c_n$

Table 2: Matrix F where each cell denotes GMEP of motif m in condition c

The question that we may ask is: can GMEP enhance the effect of gene expression profiles so that the clusters based on simultaneous use of GMEP and expression profiles could contain genes that act in a significantly co-regulated manner. This can be tested by clustering the expression profiles with most significant GMEP compared to clustering with the GMEP which is randomly drawn from population.

6 Future work

In this report, we reviewed important aspects of regulatory networks, such as coregulation, modularity etc. with focus on the difficulties that have not been considered to a great extent in today's modelling approaches. We also reviewed information sources that can be used to group genes based on their regulatory properties. Those sources can be useful for developing an artificial regulatory network.

Some problems associated with the term coregulation and approaches for identifying regulatory networks have been highlighted. This review identifies important problems, which need further research. The problems and research question are mentioned earlier. Here, they will be summarised:

- Today's methods for modelling genetic networks are simplified and do not take into account the "modular" context of regulation and synergistic binding of TFs. This needs to be incorporated in future methods. Important approaches that are worth further consideration are phylogenetic footprinting and phylogenetic profiling, which can help identifying TF binding specificities, for example.
- Given the data from different sources that can help identifying regulatory properties and connections between genes (genetic network), what is the best way of using those sources in forward modelling?
- In important issue in developing an artificial model, where all rules are known, is to use information about cis-regulatory elements and their way of clustering to build the modules. This is a prerequisite for understanding the relationship between the modular architecture and expression pattern.
- Validation of the derived interactions is very important, i.e. we need to ask the question of whether we can use two of the mentioned sources to derive regulatory information and validate against the third source. The question "Are those two sources good predictors of the third source?" needs quantitative measures of prediction quality.

References

- Altman, R.B. and Raychaudhuri, S. (2001) Whole-genome expression analysis: challenges beyond clustering, *Current Opinion in Structural Biology*, 11:340-347.
- Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M. and Eisen, M.B. (2002) Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome, *Proceedings of the National Academy of Sciences of the USA*, 99(2): 757-762.
- Brazma, A., Jonassen, I. and Ukkonen, J. (1998) Predicting gene regulatory elements *in silico* on a genomic scale, *Genome Research*, 8:1202-1215.
- Brown, C.T., Rust, A.G., Clarke, P.J., Pan, Z., Schilstra, M.J., De Buysscher, T., Griffin, G., Wold, B.J., Camoron, R.A., Davidson, E.H. and Bolouri, H. (2002) New Computational Approaches for Analysis of *cis*-Regulatory Networks, *Developmental Biology*, 246: 86-102.
- Chiang, D.Y., Brown, P.O. and Eisen, M.B. (2001) Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles, *Bioinformatics*, 17:49-55.
- Davidson, E.H., Rast, J.P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., Otim, O., Brown, C.T., Livi, C.B., Revilla, R., Rust, A.G., Pan, Z.J., Schilstra, M.J., Clarke, P.J.C., Arnonr, M.I., Rowen, L., Cameron, R.A., McClay, D.R., Hood, L. and Bolouri, H. (2002) A Genomic Regulatory Network for Development, *Science*, 295: 1669-1678.

D'haeseleer, P. (2000) *Reconstructing Gene Networks from Large Scale Gene Expression Data*, Ph.D. dissertation, University of New Mexico.

D'haeseleer, P., Liang, S. and Somogyi, R. (1999) Gene Expression Analysis and Genetic Network Modeling., *Pacific Symposium on Biocomputing '99*, Tutorial session on Gene Expression and Genetic Networks, chapters 2 and 5, URL: <http://www.cs.unm.edu/~patrik/networks/PSB99/genetutorial.pdf>.

D'haeseleer, P., Wen, X., Fuhrman, S. and Somogyi, R. (1998) Mining the gene expression matrix: Inferring gene relationships from large scale gene expression data, *Information Processing in Cells and Tissues*, In Paton, R.C., and Holcombe, M. Eds., pp. 203-212, Plenum Publishing.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns, *Proceedings of the National Academy of Sciences of the USA* 95(25):14863-14868.

Fellenberg, M. and Mewes, H.W. (1999) Interpreting clusters of gene expression profiles in terms of metabolic pathways, *Proceedings of the German Conference on Bioinformatics '99*, Poster, URL: <http://www.bioinfo.de/isb/gcb99/poster/index.html>.

Fickett, J.W. and Wasserman, W.W. (2000) Discovery and modelling of transcriptional regulatory regions, *Current Opinion in Biotechnology*, 11: 19-24.

Gerstein, M. and Jansen, R. (2000) The current excitement in bioinformatics – analysis of whole-genome expression data: how does it relate to protein structure and function? *Current opinion in structural biology*, 10:574-584.

Griffiths, A.J.F., Miller, J.H., Suzuki, D.T., Lewontin, R.C., Gelbart, W.M. (1996) *An introduction to genetic analysis*, WH Freeman and Company, New York.

Hatzimanikatis, V. and Lee, K.H. (1999) Dynamical Analysis of Gene Networks Requires Both mRNA and Protein Expression Information, *Metabolic Engineering*, 1:1-7.

Heyer, L.J., Kruglyak, S. and Yooseph, S. (1999) Exploring Expression Data: Identification and Analysis of Coexpressed Genes, *Genome Research*, 9: 1106-1115.

Hughes, J.D., Estep, P.W., Tavazoie, S. and Church, G.M. (2000) Computational Identification of Cis-regulatory Elements Associated with Groups of Functionally Related Genes in *Saccharomyces cerevisiae*, *Journal of Molecular Biology*, 296(5):1205-1214.

Kel-Margoulis, O.V., Ivanova, T.G., Wingender, E. and Kel, A.E. (2002), Automatic annotation of genomic regulatory sequences by searching for composite clusters, *Pacific Symposium on Biocomputing*, 7:187-198.

Klingenhoff, A., Freech, T. and Werner, T. (2000) Regulatory modules shared within gene classes as well as across gene classes can be detected by the same *in silico* approach, *In Silico Biology* (electronic publication), 1:0020, URL: <http://www.bioinfo.de/isb/2000/01/0020/>.

Latchman, D. (1998) *Gene regulation - A eukaryotic perspective*, Stanley Thornes, Cheltenham, UK.

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

McGuire, A.M. and Church, G.M. (2000) Predicting regulons and their cis-regulatory motifs by comparative genomics, *Nucleic Acids Research*, 28(22): 4523-4530.

Nikolov, D.B. and Burley, S.K. (1997) RNA polymerase II transcription initiation: A structural view, *Proceedings of the National Academy of Sciences of the USA*, 94: 15-22.

Noveen, A., Hartenstein, V. and Chuong, C.M. (1998). Gene networks and supernetworks: Evolutionary conserved gene interactions. In Chuong C.M. ed., *Molecular Basis of Epithelial Appendage Morphogenesis*, Landes Bioscience, Austin. pp 371-391.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles, *Proceedings of the National Academy of Sciences of the USA* 99(96): 4285-4288.

Pedersen, A.G., Baldi P., Chauvin Y. and Brunak S. (1999) The Biology of Eukaryotic Promoter Prediction - a Review, *Computers and Chemistry*, 23:191-207.

Pilpel, Y., Sudarsanam, P. and Church, G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements, *Nature Genetics*, 29(2):153-9

Reis, B.Y., Butte, A.S. and Kohane, I.S. (2001) Extracting Knowledge from Dynamics in Gene Expression, *Journal of Biomedical Informatics*, 34: 15-27.

Sinha, S. and Tompa, M. (2000) A Statistical Method for Finding Transcription Factor Binding Sites, In *Proceedings of the Eighth International Conference On Intelligent Systems for Molecular Biology*, AAAI press, 8: 344-354.

Smith, V.A., Jarvis, E.D., Hartemink, A.J. (2002) Evaluating functional network inference using simulation of complex biological systems, *Bioinformatics*, 18: 216-224.

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

Soukas, A., Cohen, P., Socci, N.D., and Friedman, J.M. (2000) Leptin-specific patterns of gene expression in white adipose tissue, *Genes & Development* 14:963-980.

Szallasi, Z. (2001) Tutorial: Genetic network analysis – From the bench to computers and back, *The Second International Conference on Systems Biology (ICSB2001)*, Tutorial session, URL: <http://www.icsb2001.org/SzallasiTutorial.pdf>.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., and 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 USA* 96(6): 2907-2912.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M. (1999) Systematic determination of genetic network architecture, *Nature Genetics*, 22:281-285.

Zhang, M.Q. (1999) Large-Scale Gene Expression Data Analysis: A New Challenge to Computational Biologists, *Genome Research*, 9:681-688.

Zhu, J. and Zhang, M.Q. (2000) Cluster, function and promoter: analysis of yeast expression array, *Pacific Symposium on Biocomputing 2000*, In Altman, R.B., Dunker, A.K, Hunter, L, and Klein, T.E. eds. 5:476-487, World Scientific Publishing.

Zhu J. and Zhang, M.Q. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*, *Bioinformatics*, 15: 607-611.

van Helden, J., Gilbert, D., Wernisch, L., Schroeder, M. and Wodak, S. (2000) Application of Regulatory Sequence Analysis and Metabolic Network Analysis to the Interpretation of Gene Expression Data, *Lecture Notes in Computer Science*, 2066: 155-172.

Von Dassow, G. and Munro, E. (1999) Modularity in animal development and evolution: elements of a conceptual framework for EvoDevo, *Journal of Experimental Zoology*, 285: 307-325.

Wagner, A. (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes, *Bioinformatics*, 15: 776-784.

Wagner, A. (2001) How to reconstruct a large genetic network from n gene perturbations in fewer than n^2 easy steps, *Bioinformatics*, 17: 1183-1197

Werner, T. (2001) The promoter connection, *Nature Genetics*, 29: 105-106.

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

Web-references

CBCG (Center for Bioinformatics and Computational Genomics), Modeling the Cellular Program, Chapter 6, White paper, URL: <http://cbcg.lbl.gov/ssi-csb/Program.html>, as is: 020312.

Yanai, I., Derti, A. and DeLisi, C, Beyond Sequence Similarity, or Sequence Analysis in the Age of the Genome, by, Chapter from "Function Genomics", URL: <http://engpub1.bu.edu/bioinfo/MERL/itai/Chapter.html>, as is: 020401.