

# Analysis of data from microarray experiments, the state of the art in gene network reconstruction \*

Literature thesis by Bas Dutilh<sup>†</sup>  
Theoretical biology and Bioinformatics<sup>‡</sup>  
Utrecht University

October 29, 1999

## Abstract

Since the development of the microarray technique in 1995, there has been an enormous increase in gene expression data from several organisms. Based on the view of gene systems as a logical network of nodes that influence each other's expression levels, scientists dream of being able to reconstruct the precise gene interaction network from the expression data obtained with this large scale arraying technique. Computer science shows that inference of a logical regulatory network is possible solely from sets of expression data, and mathematicians are working on the question how much data is at least necessary for reverse engineering.

Meanwhile, experimental biologists are experiencing problems in the field. The number of experiments that are necessary before attempting network reconstruction is a lot more than is generally possible in "wet" laboratories, so data compression algorithms are applied to reduce the number of nodes considered. This is however an extremely coarse representation of the intricate interconnections that exist between single genes. The resulting network of only a handful of nodes is therefore usually only sufficient to describe the experiments performed, while any possible predicting properties are absent.

In this literature thesis, I attempt to give an update on the state of the art in computerised network reconstruction techniques, and explicitly relate this to actual biological gene networks. I will go into the model formalisms used to describe genetic networks, and explain their specific advantages and disadvantages. Also, a separate chapter will be dedicated to several experimental results obtained in the research of genetic networks, and finally, a short discussion and some hypothesising is added.

---

\*Literature thesis written from July to October 1999. Since a lot of the information concerning this modern method is available via the Internet, I have composed a web-page which may come in handy to the interested reader (<http://www-binf.bio.uu.nl/~dutilh/gene-networks>).

<sup>†</sup><http://come.to/dutilh>, [bedutilh@yahoo.com](mailto:bedutilh@yahoo.com)

<sup>‡</sup>Padualaan 8, 3584 CH, Utrecht, The Netherlands.

# 1 Introduction

Many useful derivations have already been made from the technique of hybridisation of labelled strands of nucleic acid to complementary sequences immobilised on a solid surface. In this technique, developed by Southern (1975), two phases can be discerned, the mobile and the immobile phase. If one of these phases contains known DNA sequences, the other can be identified in terms of the known phase. In practice, a known mobile phase is used for either identification (as in cDNA fingerprinting), or also quantification (as in Southern and northern blot) of unknown nucleic acid sequences using specifically designed probes. A known immobile phase can be used for mapping a labelled query mixture. This last method has recently entered a new era when Schena et al. (1995) developed a technique for quantitative monitoring of gene expression patterns with a complementary DNA (cDNA) microarray.

Microarraying is suitable for studying the expression patterns of large numbers of genes, however, it is a costly technique. Other techniques suitable for analysis of differential gene expression developed recently (such as serial analysis of gene expression (SAGE) and differential display, which gives an indication of the genes that are differentially expressed between two tissues) are lauded since they are for instance cheaper and easier to carry out (Kozian and Kirschbaum, 1999). Nonetheless, despite the price tag, the large scale of research as is possible with the DNA-chip, is the reason that this method has by far made the greatest impact in the past few years, and at this moment the price of such research is also decreasing as an increasing number of research groups and companies get interested (Marshall, 1999).

In this technique, the immobile phase is a collection of single stranded pieces of known (c)DNA attached to a nylon filter, glass slide or silicon chip. The unknown mobile phase is a mixture of labelled copies of mRNA purified from the query tissue: after hybridisation the label is localised on the spots where the unknown phase matches the known phase. The big advantage of this method is that an enormous amount of test sequences can be arrayed in a single experiment. Not rarely this means that the entire genome (DeRisi et al., 1997), or at least a large number of cDNA's (Schena et al., 1995), are attached to a small glass slide (over 1000 clones are possible per square centimetre), making it possible to array this entire set with a minimum of material. The small scale, which can be obtained because the glass chip allows very precise spotting of target DNA, is also the main improvement of this technique over previous experiments using arrayed cDNA libraries, which still made use of filter membranes for attachment of the solid phase (Lennon and Lehrach, 1991).

That the microarray technique is a very promising one, is apparent from the large amount of recent reviews published on this topic (e.g. Gerhold et al., 1999; Lander, 1999; Duggan et al., 1999), and the fact that the National Human Genome Research Institute (NHGRI) has started the microarray project, a collaborative research effort between multiple divisions of the National Institutes of Health (see Internet page).

Though the velocity at which results can be obtained with this technique is a major advantage, there are a few drawbacks: the error of quantitation can reach levels of 30 to 50%, whereas this is 15 to 20% in the traditional methods (e.g. northern blot or quantitative PCR). Some more technical points that deserve attention are summarised by Vingron and Hoheisel (1999); Vingron (1999); the general opinion of the authors is however positive.

Another important point is, that only known genes can be analysed using the DNA-chip method. This limits the experimental potential to a small (but increasing) set of organisms of which the genome has been fully sequenced, or to a small subset of genes, which have already been studied in some detail.

Because of the ease with which gene expression of large numbers of genes can be monitored, a source of equally huge amounts of data has been generated with the development of microarrays. The means to handle this abundance of information is of course the computer, which is used as an instrument at multiple points in the analysis. Several of these will be addressed in this literature thesis: I will briefly address the means by which researchers store and handle their results, and which attempts there are to set up a central database for all microarraying results found. The nature of these projects makes the Internet the ideal place to find information: programs for collecting and digitising the fluorescence data are often commercially available at companies with detailed web-sites, and the stereotyped means to make a central database universally accessible is of course also Internet. Therefore I will refer to the Internet page <http://www-binf.bio.uu.nl/~dutilh/gene-networks> where necessary. On this site you can find links to a number of other microarray-related sites.

The main aim of the thesis however, is to obtain an insight in the state of the art of the methods used to process the obtained data. The ultimate goal researchers may dream of pursuing is, to infer from the data obtained from microarray experiments, the genetic interaction networks that lie at their basis (Somogyi and Kitano, 1999). The idea is, that the genetic induction- and inhibition routes are determined through specific molecular interactions, and can

thus be described in detail by a logical or mathematical model. In chapter 3 I will elaborate on the techniques used for reverse engineering of gene networks from microarray data. It will be explained that though one can collect data on the expression patterns from thousands of genes, it is often rather necessary to have long time-course data, or data from a large number of independent expression experiments to infer an underlying network. This means there is in practice often a shortage rather than an excess of data, and the techniques with which investigators attempt to mine this data by e.g. reducing the number of nodes or genes considered are treated. The features of biological gene networks which are of importance in the preparation of a relevant model are treated in paragraph 3.2, which gives the reader an impression of the complexity of the object system. Consequently, we see that in most of the model formalisms presently used to describe genetic interaction nets, a number of the important properties are left unconsidered. A number of model formalisms are listed and their specific properties are mentioned.

In the “Results” (chapter 4) some notable experimental papers are treated. This is to show the reader a selection of the results that have been reached in the search for gene networks.

## 2 The experimental routines

Two main types of DNA chips can be discerned, the microarray and the DNA-chip. Both are based on the same principle, however the method of addition of the nucleotide stretches to the chip differs. The microarray is closest to the old fashioned spot-blotting array. A set of solutions containing amplified strands of DNA is spotted onto the solid support by a robotically directed precision pipette (printing tip). Many small PCR-reactions can each amplify a separate DNA stretch, from each of which one or several spots are placed on the microarray (e.g. DeRisi et al., 1997; Wang et al., 1999). Detailed information on the preparation of an arrayer (which can print DNA onto chips), using it to prepare DNA-chips, and protocols for microarray experiments can be found in “the Brown Lab’s complete guide to microarraying for the molecular biologist” (see Internet page). Though especially building the arrayer will need a technician’s expertise, the protocols are clearly outlined in the mentioned guide, as long as you keep to them precisely. This includes purchase of a number detailed parts necessary for e.g. the 3-D precision printer and its computerised conduction.

The process of printing target DNA on a glass chip is very comparable to the experimental setup of the original Southern blot. For a comparison and a review on the behaviour of oligonucleotides on chips, cf. Southern et al. (1999). The glass slide is first cleaned and prepared with poly-L-lysine (which is routinely used for promoting cell adhesion to solid substrates) so that the DNA is able to attach. After linking the printed spots of target onto the prepared slide using UV light, the rest of the slide has to be blocked to prevent the probe from attaching to the coating. We are then left with microarrays ready for hybridisation.

In the case of a DNA-chip, the oligonucleotides are synthesised directly onto the chip. The solid surface is prepared such, that there are 3’-OH ends sticking out, to which nucleic acids can be attached in sequence. This can be done either by photo-lithography (Pease et al. (1994); a technique now commercially exploited by Affymetrix; cf. Internet page) or through an electro-chemical process (Livache et al., 1998). Out of many reviews about these techniques, I refer to Ramsay (1998); Marshall and Hodgson (1998).

In most experiments performed so far, researchers have used the microarray, because it is easier to purify mRNA than to synthesise the to-be-tested sequences from scratch nucleic acids. Because the basic principles for both methods are identical, they will not be discerned in the current thesis.

The source of hybridisation probes to be used for annealing to the chip can be very diverse. Optimally, one could use the mRNA purified from a single cell, and amplify it with a nucleic acid mixture containing fluorescent dUTP. This yields probes which can be used for a microarray. The difficulty here obviously lies in the precision with which mRNA can be purified and amplified, and in practice, researchers tend to purify the query mixture from a specific tissue (e.g. Wang et al., 1999) rather than from a single cell. Clearly this will increase the error rate, and it is even possible to make misinterpretations about the co-expression of two genes which are actually only expressed in an exclusive manner (Szallasi, 1999).

### 2.1 Processing the expression data

Once the labelled probe has successfully hybridised to the chip, all that needs to be done is to measure the amount of fluorescent marker present at each location in the chip. This can be done by excitation with a laser beam, and measurement of the amount of light with the label specific wavelength that is emitted. This information is usually immediately digitised for easier computer processing. There are already numerous specialised computer programs available for quantitative analysis of the intensity of the light spots (see Internet page).

Other methods that are used for determining at which sites nucleic acid has hybridised to the immobile phase, are use of radioactively labelled hybridisation probe, and detection of the phosphorus present in the hybridised DNA by e.g. resonance ionization mass spectrometry (Arlinghaus et al., 1997). This can only be done if the immobile phase does not contain phosphorus, which would otherwise interfere with the measurements. For this purpose, Arlinghaus et al. (1997) attach peptide nucleic acids (PNA) to the chip, a DNA analog in which both the phosphate and the deoxyribose are replaced with polyamides.

For the next step, commercially available computer programs can mine the massive gene expression data generated from a microarray experiment. The result of such microarray image analysis tools is a table of expression values for all investigated genes under one or more conditions. In the available programs (see e.g. Ermolaeva et al. (1998) and Internet page), one can select any number of such data tables to compare and explore the database using a variety of visualisation and data-compression methods, such as scatter plots, histograms, neural network clustering, time series analysis, principal components analysis (PCA) and cluster analysis (cf. paragraph 3.1 below).

### 3 Network reconstruction

Microarray experiments can quickly result in expression values for large numbers of genes. Because of the potency of the new measurement system, the expectations of the results have also grown. The goal chip-researchers have in mind is ultimately to decipher the precise connections of the genetic network: for each gene, they will want to know which other genes it influences, and in what way. While in the previous chapter the encountered problem was how to make the huge amounts of data readily accessible and sharable with other laboratories, the problem in this chapter is mainly how to approach the ambitious goal of gene network reconstruction, given the relatively few data points measured.

#### 3.1 Clustering

If the research covers a large amount of genes, currently what researchers do, is to find clusters of genes which have similar expression patterns (e.g. Michaels et al., 1998). This could mean their optimum in a time series experiment coincides (Wahde and Hertz, 1999), or also that their expression behaviour in the different mutants or tissues arrayed is alike (Eisen et al., 1998). Arguments used may be, that co-expressed genes are probable to have related functions, or simply that no better means to analyse the data is available at this moment.

Though in theory it is a big step from simple correlation analysis to gene interaction networks, several papers indicate that the clustering of gene expression data does result in groups of genes that have related functions (for as far as functions can be ascribed). Eisen et al. (1998) form a dendrogram of practically all *Saccharomyces cerevisiae* genes, the expression of which has been arrayed under several conditions. 35% of these genes have been studied in some detail, and the clusters that result from their analysis show large sets of grouped related genes (cf. paragraph 4.2.2). As a clustering algorithm, they revert to the pairwise average linking method (Sokal and Michener, 1958), which calculates a pairwise similarity matrix for all pairs of genes, groups the two most similar in a cluster, and then recalculates the matrix using the average properties of both (all) the genes in the cluster. They mention that the paper is mainly meant to indicate how obvious the grouping of functionally related genes is, even with the classical clustering method used. Additionally Eisen et al. (1998) introduce a method for easier visual estimation of the expression patterns of arrayed genes using colour gradients. They show how the results of clustering are well visible using this method.

Törönen et al. (1999) use a self-organising map (SOM, Kohonen's map) for cluster analysis of gene expression data. In this algorithm, one chooses a geometry of nodes, which are randomly mapped in  $n$ -dimensional space. Then follows an iterative process of adjustment of the coordinates of all the nodes such, that they move towards a randomly chosen data point; the closer they are together, the faster the node moves. Eventually, all the nodes are distributed over the clusters, and a map can be prepared from the initial grid in which the nodes were laid down.

Since Törönen et al. (1999) only had access to part of the data used by Eisen et al. (1998), the networks resulting from their analyses do not contain as many clusters with functionally related genes as the dendrograms in the latter article did. Particularly, Törönen et al. (1999) recognise the problem of expression patterns that are not influenced by the given treatment. If there is only data available from a single treatment (in this case the expression profiles of 6400 yeast genes during diauxic shift, which were available on the Internet, were used), the genes that are hardly influenced by this treatment tend to form an unstructured group in the centre of the SOM. Basically, this means the clustering algorithm has trouble discerning expression patterns that are quite similar. Tamayo et al. (1999), who have also used a SOM to cluster data from microarray experiments, have implemented a variation filter for data points with no significant change across the sample points, to enhance the system's sensitivity and prevent nodes from being attracted to large sets of invariant genes. With Törönen et al. (1999), they come to the conclusion that a SOM reliably clusters functionally equivalent genes, and is especially suitable here, since similar patterns will occur as neighbours in the SOM.

Hartuv et al. (1999) have developed a new clustering algorithm, based on a graph theoretic approach. The motivation for the study was the need to cluster a collection of cDNA's by their oligonucleotide fingerprints, however the results from such a study are very comparable to e.g. time-resolved data obtained from a microarray experiment. Basically, a similarity graph is defined, in which edges are drawn if the pairwise similarity value exceeds a threshold. From this graph, a "cut" is taken by removing a minimum number of edges, and the connectivity of this cut is evaluated. If the subgraph contains more than  $\frac{n}{2}$  edges, it is called highly connected. These steps are repeated, until finally, the total data set is grouped into a number of highly connected subgraphs, the clusters. The authors show that the algorithm runs in low-degree polynomial time, the limiting step being the minimum cut algorithm. The results using this algorithm are good for both simulated and real cDNA fingerprinting data tested; the authors make no statement about their expectations of the algorithm in other clustering tasks.

The above exposition indicates that clusters of genes organised by their expression patterns can give hints as to their function. Nonetheless, all the authors mentioned are naturally cautious in saying more than that, and indicate that suggestions given by the cluster analyses may be worth being looked into further. The techniques used for gene clustering in the microarray area mainly fall back upon classical methods (Eisen et al., 1998; Törönen et al., 1999; Tamayo et al., 1999), and it is probably necessary to scan the developments in related fields for novel approaches (Hartuv et al., 1999).

## **3.2 Properties of regulation networks**

There is a large number of factors that complicate simple mathematical modelling of biological genetic networks. In this paragraph, several of these are listed, and publications that have addressed the problems are shortly reviewed.

### **3.2.1 Stochasticity**

Many complex interactions on the molecular scale have been described (McAdams and Arkin, 1997; Savageau, 1998), which rely on presence of specific factors that can either enhance or inhibit the expression of certain genes. McAdams and Arkin (1997) point out that the time interval between the switching on of the first promoter and its effect on the next promoter can vary widely across otherwise identical cells, as a result of stochastic processes. Mechanisms put forward to explain the stochastic nature are for instance the degradation of gene products, the spatial collision necessary before a reagent can exert its influence, and the reversible reaction equations for e.g. dimerisation and reaction complex formation. For a cell to create less noisy output, it is necessary to produce more frequent transcripts with fewer proteins per transcript, which is related to a higher energy cost. Both McAdams and Arkin (1997) and Szallasi (1999) argue that stochastic models are most likely to yield realistic results.

### **3.2.2 Gene duplication**

Wagner (1994) made a weight matrix model (cf. paragraph 3.3.1) to study the effects of gene duplications on the organisation of the genome. He states that duplication of several genes can alter the equilibrium expression pattern of network genes, while only duplication of the complete genetic network will not. The effect of duplicating  $k$  out of  $n$  genes increases monotonically from zero as  $k$  is increased from 0, or also if  $k$  is decreased from  $n$ , leading to a maximal effect for some intermediate  $k$ , which is what could be expected intuitively.

### **3.2.3 Genomic organisation and network dynamics**

Thieffry and Thomas (1998) have examined how the nature of possible feedback loops (or equilibria) present in genetic networks depends on the properties of these networks. The feedback loops may be denominated “positive” or “negative”, indicating the ultimate influence of one of the nodes in the loop on itself. This property can be easily be calculated by multiplying the signs of the rates of influence of each subsequent genetic interaction. They find that the nature of the feedback loop is a necessary (but not sufficient) condition for the distinctive type of behaviour in the system: negative feedback loops may give stable oscillatory behaviour, while at least one positive regulatory circuit is necessary to generate multi-stationarity.

In their study on the relationship between genomic regulatory element organisation and gene regulatory dynamics, Wolf and Eeckman (1998) showed that dynamical system behaviour, stability of equilibria and their bifurcation potential can be largely determined from regulatory element organisation. Their differential equations model is generalised from Shea and Ackers (1985), who have analysed in detail the gene network responsible for the switch from lysis to lysogeny in the lambda phage. The complex transcription level, promoter control model contains a number of interactions, such as promoter regulated transcription and RNA polymerase binding. It is assumed that the linear model represents prokaryotic gene networks, which are dominated by local promoter control by regulator proteins.

It is shown that a monomer-controlled, one-gene, one-operator site gene regulation system have a single stable equilibrium point independent of the parameters. More complex systems, still with a single gene, but with an arbitrary number of protein binding sites are globally stable, and may bifurcate to multiple equilibrium points. In the systems as investigated by Wolf and Eeckman (1998), there is a simple relationship between the number of operators and the maximum number of stable equilibria: in a monomer-controlled gene-regulation system with  $n$  operator sites, there is

a maximum of  $1 + \frac{n}{2}$  stable equilibria for even  $n$ , while for odd  $n$ , this is  $\frac{n+1}{2}$ . In case of multimer-controlled system, this maximum lies at  $2 + \frac{n}{2}$  for even  $n$ , and at  $\frac{n+3}{2}$  for odd  $n$ .

### 3.3 Models for gene networks

When we want to make an estimate of the genetic network in the organismal system of interest, it is convenient to have a search image in mind. As explained in paragraph 3.2, there are many particularities to be considered, though sometimes simplifications have to be made. For instance, though stochastic models show more realistic dynamics (McAdams and Arkin, 1997), models of gene networks are usually deterministic. The reason for this simplification is the difficulty to infer an underlying network if the expression patterns are the result of a stochastic process. Several model formalisms used for description of gene networks are overviewed below.

#### 3.3.1 Weight matrices

Weight matrices are the most established method for deduction of gene networks. A weight matrix consists of  $n \times n$  weight values, each of which indicates the influence of one specific gene on another. The advantages of modelling regulatory networks with weight matrices are stated by Weaver et al. (1999), who present an algorithm (TReMM: Transcription REgulation Modelled with Matrices) for a classical example of a weight matrix model. The weights  $W_{ij}$  represent the influence of gene  $i$  on gene  $j$ , and the complete input into a certain gene  $j$  is given by the summation over the inputs of all genes  $i$ , multiplied by their weights. The answer to this calculation is fed to a normalisation formula which outputs an expression value between 0 and 1 for gene  $j$  (cf. paragraph 3.3.4).

The  $n \times n$  weights composing the matrix are unknown when initiating a study of the genetic regulation network in a certain biological system. The idea is, that they can be approximated from expression data in the process of network reconstruction. The deduction of the values is usually done by one of several classical learning mechanisms, such as simulated annealing (Spears, 1996), neural networks or genetic algorithms (GA's).

As is clear from the above exposition, a weight matrix model considers the interactions between all combinations of genes, many of which are of course 0. Because it is not known at fore-hand which ones are 0, this means a hard computational task. In a consideration of this problem, Hertz (1998) looked at the set of all the networks consistent with the input/output examples found in experiments: this set decreases as the number of experiments increases. He shows that if the genetic connectivity or in-degree  $k$  is much lower than  $n$ , and we have sufficient experiments (i.e. more than  $\mathcal{O}[k \times \log(n/k)]$ ), "knowledge" is the same as if it was known in advance which connections were 0. Hertz' poster purely investigates in what situations it is possible to infer an underlying network, and gives no further indications as to how this should be done. A few simplifying assumptions are mentioned, that have a big influence on the result: additive regulation, on/off classification of the genes, and knowledge of the network's time steps. The assumption that the state transition pairs are completely uncorrelated is of influence, and will be addressed by Hertz in the future.

Reinitz and Sharp (1995) have made a gene circuit model describing the mechanism of stripe formation of the *Drosophila melanogaster* gene *eve*, which plays a role in the segmentation of the insect. The gene circuit model is a spatially explicit weight matrix system: the interaction parameters were optimised by use of simulated annealing, and the spatial factor was implemented by including a diffusion term for the gene product influence between adjacent cells. Basically the model is comparable to that of Turing (1952), though generalised to include  $n$  reacting species, and specified to be able to model nuclear division, degradation of proteins, and with explicit functions substituted for Turing's general reaction functions  $f$  and  $g$ .

#### 3.3.2 Boolean networks

A Boolean network consists of  $n$  nodes (e.g. representing genes) which can either be repressed or expressed (the node has state 0 or 1, respectively). The dynamics of the network are determined by a list of  $n$  (Boolean) functions which each receive input from  $k$  specified nodes. Every node has its own specific function, which can determine its next state from the current states of all the input nodes.

Compared to gene networks, Boolean networks are of course a coarse simplification. Gene expression is never a case of all-or-nothing, and it is also unlikely that the number of input nodes of each gene is specifically bounded by  $k$ , a necessary condition for proper reconstruction of the gene networks in algorithms that use Boolean network models (Liang et al., 1998; Akutsu et al., 1999). Nevertheless, Boolean networks provide a framework in which genes can

have complex interactions (e.g. XOR connections), and they show behaviour comparable to features of biological gene networks (e.g. global complex behaviour, self-organisation, redundancy; cf. Somogyi and Sniegowski (1996)). These features are likely in natural systems but are often not considered in gene network models using weight matrices (see paragraph 3.3.1). The networks are very suitable for studying the reconstructive capabilities of algorithms, since one defines the network in advance, and all connections are known. Data comparable to that resulting from microarray experiments can be generated, such as time-series data (a series of states of the nodes obtained from input-output calculations) or cellular gene expression patterns (comparable to an attractor situation in which the input configuration of the network equals the output state) can be computed and fed to the algorithm. Though no literature has been found on this so far, it should even be possible to simulate knock-out mutants, and find the attractors in situations that a specific node is, or is not expressed. Since in case of designed Boolean networks, the results of reconstruction can be compared to the true network, it is possible to evaluate the used algorithm, which is of course impossible in natural systems.

The algorithm of Akutsu et al. (1999) can identify a Boolean network of  $n$  nodes from  $\mathcal{O}[\log n]$  state transition pairs, by exhaustively searching all possible Boolean functions until a set that fits all the data is found. (Note, that like Hertz (1998) did, these authors use uncorrelated, random state transition pairs as their input data for inference. As mentioned in paragraph 3.3.1, this does influence the ease with which results can be found, while it is unlikely that such completely independent data pairs result from microarray experiments). The scale of the networks that can be resolved by this algorithm is large, and the authors show that the problem can be solved in polynomial time for a determined maximum number of input nodes (an example with an in-degree  $k = 2$  is shown in the paper). However it would be preferential to have a more efficient algorithm, in which part of the exhaustive search could be bypassed.

The main advantage of the REVerse Engineering ALgorithm (REVEAL) of Liang et al. (1998), is that the in-degree is not fixed at fore-hand. Rather, the minimum  $k$  for each node is determined in the algorithm, so that the minimum effective network (Somogyi and Fuhrman, 1997) is inferred. The algorithm is based on comparison of the Shannon entropies of the input and output data, which can reveal the in-degree. Once the input nodes have been specified, the search for a rule that fits the data is exhaustive like Akutsu et al. (1999).

The methods of inference of both Akutsu et al. (1999) and Liang et al. (1998) are grafted upon Boolean networks: deterministic and discrete data are necessary, two conditions that do not apply to biological gene networks. However a discretion assumption shall always have to be made in case of computer simulations. Both authors mention that the systems can be extended to include more possible states of the nodes, however it is unclear whether it is possible to have a ranking order in these multiple states (which is necessary for discretised gene expression values).

Nonetheless, a framework of Boolean functions is valuable: they allow a wealth of possible gene interactions, which is a good starting point for realistic modelling of gene networks (Yuh et al., 1998). Thieffry and Thomas (1998) have reviewed the use of logical Boolean net-style models and continuous differential models for gene networks, especially focusing on the nature of possible feedback loops (or equilibria) present in the systems (cf. paragraph 3.2.3). They claim that the stereotyped method for conceivable realistic modelling of genetic interaction nets, is the logical model formalism (i.e. composed of Boolean functions).

### 3.3.3 Differential equations

Wahde and Hertz (1999) followed Reinitz and Sharp (1995) in modelling genetic networks as a set of nonlinear differential equations. Thus, they can search for parameters that indicate the rates of change of a certain gene, and the assumption of discrete time steps for the network's next state is unnecessary. In stead, they used a GA to determine the time constants for each of the  $n$  nodes in the system. The other parameters were deduced similarly:  $n \times n$  gene interaction weights, and  $n$  bias terms. Because the use of GA's can generate very different results in alternate runs, the generated parameter values were averaged over several runs of the algorithm; this also gave insight in the error in each parameter. Unfortunately, lack of data restricted Wahde and Hertz (1999) to the use of first order terms in the differential equations, which results in a reconstruction method of the level of simple additive weight matrices. If more data would be available, including higher order terms could make it possible to implement a range of complex interactions, comparable to the gene interactions possible in Boolean network models (cf. paragraph 3.3.2).

There are two more assumptions made by Wahde and Hertz (1999) that deserve attention here. The first is the activation function: once the inputs from all genes multiplied by their weight factors and the gene specific bias term have been processed, this function translates the result to gene a expression level (0 to 1, cf. paragraph 3.3.4).

Secondly, they use the average trajectories of coarse clusters of genes with similar expression patterns as nodes. For these authors, this simplification is inspired by the fact that for reliable determination of the network parameters, a

minimum requirement is that the number of useful data points exceeds the number of parameters. Gene expression data series often count only few measurements, and clustering of genes into sets with similar temporal expression patterns severely reduces amount of parameters to be calculated.

The method of Wahde and Hertz (1999) was used to determine the parameters of both a predefined artificial network and the network behind a set of gene expression data from rats. The evaluation shows that for a situation with 4 nodes, the method is able to obtain an accurate representation of the network parameters. In situations where there is only time-series data available, the accuracy of this prediction decreases, though a rough estimate can still be made.

Chen et al. (1999) also propose a differential equation model for gene expression, and provide a method to construct the model from temporal expression data. They make a number of assumptions, among which again a linear transcription function for each gene and feedback of the gene translation product on the transcription rates. They discern in their model transcription, translation and degradation of RNA and proteins, for which parameters can be found using a technique called Fourier transform for stable systems. This approach is specific for genes with periodic expression, such as are important in the cell cycle, and all genes considered in the model are assumed to show this kind of expression pattern. By realising that, assuming a linear transcription model, realistic genetic networks have to be stable, Chen et al. (1999) manage to greatly reduce the space in which to search for parameters. They derive, that all the real parts of the eigenvalues  $\lambda_j$  are non-positive (otherwise  $q_{ij}(t) \times e^{\lambda_j \times t}$  is an exponential function leading to unlikely growth of a gene product), and the polynomials  $q_{ij}(t)$  (which are functions of  $t$  and elements of the  $2n \times 2n$  transition matrix comprising the set of differential equations) have to be constants. Because Chen et al. (1999) can reduce their search space so much, they can consider more features of gene expression than simpler models (such as that of Wahde and Hertz (1999), above) in their linear transcription model. However the authors do not mention how much influence these model complications have on the results: how important is it that these extra assumptions are taken into account?

To increase the amount of data from which to start network inference, D'Haeseleer et al. (1999) started their analysis with the calculation of a non-linear interpolation curve of the gene expression time series data points. Though they used data from three different experiments, with very alternate time-scales (time intervals between measurements ranged from half an hour to two months), they managed to combine the data to an interpolated time-series using a cubic interpolation on the log of the expression levels (taking the log prevents negative values). The next step was to derive the entries in an interaction matrix as described in the models above: their model of the genetic network was also a set of linear differential equations in which the change of expression of a gene depends on the weighted inputs of all other genes. These interaction weights were approximated using the least squares fit to the interpolated time series. The result of their approach is an accurate fit of the data points, but they do give some comment on the method. First, the mechanism does not minimise the number of gene interactions: each gene is modelled by a weighted sum of all other genes. Secondly, the simple linear and additive modelling of the genetic interactions can only capture the primary linear components of the system; and finally, because the data that these researchers had at their disposal was so non-uniformly spaced, a larger weight was given to the more widely spaced data points. A point that also needs consideration is to what extent the data is over-fitted in this method. The authors mention that an indication may be obtained by constructing a series of similar models by disturbing the input data within the known standard deviation for each measurement, and by using different non-linear interpolation schemes. Comparison of these models could tell us how sensitive the results are with respect to small amounts of noise in the input data.

### 3.3.4 Translation to gene-expression levels

In weight matrix models as well as in models composed of differential equations (both treated above), the direct result of the stepwise calculation is often of an arbitrary magnitude, this being the result of the additive nature by which the input signals from all nodes are combined into the node of interest. To overcome this problem, researchers scale this primary output value to a gene-specific expression level. An important assumption is made here: namely that the maximum expression level per gene is known, and that the calculated primary output value can be mapped to a 0 to 1 fraction of this maximum. Often a sigmoidal function is used for this mapping, such as  $L_p = \frac{1}{1+e^{\alpha \times L_o + \beta}}$  (Weaver et al., 1999). In this function,  $L_p$  and  $L_o$  are the primary output value and the resulting fraction of the maximum expression value, respectively.  $\alpha$  and  $\beta$  are parameters that determine the shape and location of the sigmoid relative to zero, respectively: these parameters can differ per node.

This normalisation also protects the system from the exponential growth of e.g. a pair of mutually inductive genes, which might otherwise grow to unnatural levels.

### 3.4 Coupling promoter analysis with expression patterns

Combining gene transcription regulation data with sequence analysis can give insight in the factors that influence the regulation of co-expressed genes. Br̄azma et al. (1998); Helden et al. (1998); Roth et al. (1998) have looked for sequence motifs in the non-coding DNA upstream of *Saccharomyces cerevisiae* ORF's that all showed a high sensitivity to a specific treatment. In the upstream regions of 3 sets of co induced or -repressed genes, which were found in the analysis of separate treatments of *Saccharomyces cerevisiae*, Roth et al. (1998) found a number of short sequence motifs in common. Some of these motifs could be identified as known transcription regulators, however 40% of the motifs found had not been described previously. The authors mention that this percentage is a likely level of false-positives to be found under the used threshold values: more stringent threshold values about which ORF's are "counted" in the analysis will reduce the number of false-positives, but also increase number of false-negatives. An examination of the sequences for transcription factors known from the literature resulted in identification of a number of false-negatives, one of which could clearly be attributed to a shortcoming of the algorithm (i.e. 7% of the total number of motifs found in the analysis).

Helden et al. (1998), who have a model based on the regulation of nitrogen metabolism in *Saccharomyces cerevisiae*, go into some of the difficulties of computational detection of regulatory sites in eukaryotes: the consensus sequences that are recognised by transcriptional factors are generally much shorter than in prokaryotes, they can be quite variable, and they can be dispersed over very large distances, and can be active in both the upstream, as well as in the downstream direction. In their research, several known transcription regulation sites did not show any signal with a high significance, because of complicating factors: Helden et al. (1998) conclude that when a signal is selected as highly over-represented, it is likely to correspond to a functional regulatory site, although the opposite is not true. This can be compared with a high threshold in the algorithm of Roth et al. (1998), where they prefer false-negatives over false-positives. The method of Helden et al. (1998) differs from that of Roth et al. (1998), in the fact that they use a calibrated table with an estimate for the frequency of each oligonucleotide. This allows them to assess the significance of the number of matching sequences; they perform a straightforward exhaustive search, checking each of the sequence motifs with the calibration table.

Br̄azma et al. (1998) searched for similar sequence patterns in the upstream regions of all (i.e. more than 6000) yeast genes, as well as in the upstream regions of genes with similar expression profiles. The basic idea of their algorithm is again similar to that of both Helden et al. (1998) and Roth et al. (1998) (above), though they may use slightly different rules or parameters. Ultimately, similar analyses should be able to provide insights in the nature of the transcription factors that are responsible for the co-regulation of genes.

## 4 Results

In this section I will discuss a number of experiments that fall in the range of the topics that are dealt with in this thesis. This is to give the reader an impression of what practical results can actually be achieved with the techniques described above.

### 4.1 Molecular classification of cancer

In several fields of research, it is currently investigated what help the microarray technique can offer. The molecular screening of carcinogenic tumours is an important application which is likely to be aided by the method.

Wang et al. (1999) have composed a DNA-chip containing 5766 cDNA's, from several cDNA libraries prepared from tumours in human ovaries. Genes with at least a reproducibly 3-fold expression level difference were assumed to play a role in cancer formation, and compared with results from semi-quantitative PCR, and other experimental investigations. Though the microarray experiment of Wang et al. (1999) identified fewer differentially expressed genes than previously used techniques, the authors mention that this may also be the result of the technique used for preparation of the cDNA library. They conclude that the results from microarray hybridisations are consistent with other experimental approaches, and can be of importance for tumour classification or to delineate the progression of cancers based on their gene expression patterns.

Indeed Golub et al. (1999) describe a generic approach to cancer classification based on gene expression monitoring by DNA microarrays. They illustrate it with an example of distinction of two acute leukemias. The first issue was to determine whether there were genes with expression patterns that correlated strongly with the class distinction to be predicted. This was done by designing an idealised expression pattern, of an imaginary gene which perfectly indicated the different clusters to be found, and using this as a vector for supervised clustering of the set of 6817 human genes, in search of examples which were more than randomly similar to this idealised gene. Then, a collection of known samples was used to create a class predictor, which was capable of assigning a new sample to either of the classes. In case a prediction needs to be made, the genes in this informative subset all cast a weighted vote, depending on its expression level in the new sample and the correlation with the predicted classes. All of the votes are then summarised to determine in which class the sample will fall, and a 0 to 1 prediction strength value is calculated.

Of a collection of 38 acute leukemia samples, an arbitrarily 50-gene predictor set grouped 36 in either of the two acute leukemia classes, 2 samples were uncertain (prediction strength smaller than 0.3). The classifications were consistent with the clinical diagnoses. A new 50-gene predictor set derived from the 38 samples was composed, which then was then allowed to classify a sample set from peripheral blood, instead of bone marrow. Though the samples were from a different tissue, from different patients and from a number of different laboratories, 29 samples out of 34 were assigned correctly, and only 5 were uncertain.

Another problem Golub et al. (1999) addressed, was cancer class discovery. If the two classes of leukemia would not have been known in advance, would it be possible to discern between them just from the gene expression patterns? This entails two issues: discovery of clusters in the set of samples, and determining whether the classes found are meaningful. The initial 6817 tumour genes were filtered: all genes with less than five-fold variation throughout the samples were left unconsidered. To cluster the resulting set of data into two groups, a SOM was used (cf. paragraph 3.1). This gave two clusters, in which respectively 77% and 96% of the two types of leukemia were grouped. Now, for the completely unprejudiced discovery of new classes in gene expression data sets, Golub et al. (1999) proposed to combine the two methods above in an iterative process. First, classes could be predicted by clustering the expression data. Then, on the basis of the obtained data, an  $n$ -gene class predictor should be made, which should assess the samples. Meaningful clusters will result in high prediction strength values. This means the described method is a good way to independent class discovery and subsequent prediction, which was shown again when Golub et al. (1999) attempted subgroup identification of the two leukemia classes examined above: one of the two groups could be split with high probability, while fission of the other group did not lead to subgroups with much difference.

### 4.2 Co-fluctuation of functionally related genes

In a number of articles, it has been shown that the clusters formed by cluster analyses of gene expression data, which results in groups of genes with similar expression patterns (cf. paragraph 3.1), often contain functionally related genes. Unknown genes that also fall into the clusters are hinted for further research. Here, two examples in which

such results were found are treated: co-regulated genes in the development of the nervous system of rats, and genes that are similarly expressed over a variety of situations in the yeast *Saccharomyces cerevisiae*.

#### 4.2.1 Identification of developmental stages of the rat CNS

A relatively simple experiment carried out by Wen et al. (1998) shows how, in the development of the rat central nervous system (CNS), a number of major developmental stages can be discerned. The measured temporal patterns of 112 genes (by quantitative PCR of RNA's isolated from the complete spinal cord of a rat) could be clustered into 4 waves and a group of genes whose expression profile was characterised as constant over the time interval measured. These waves are indicative of 1) an immature proliferative stage (decreasing expression levels), 2) neurogenesis (increase and maintenance of high a level of gene expression), 3) initial excess cell growth, a large portion of which dies in later development (or other genes with a hill-shaped expression pattern), and 4) gliogenesis and the final maturation of the tissue (exponential increase during the whole period measured). Characteristic in these waves is the clear mapping of gene groups with functional relationship to the processes that take place in the periods of their optimum. What was striking in the results was that no oscillatory or U-shaped expression pattern could be found. However it remained undetermined whether this was the result of a bias in the selection of the genes, or is perhaps characteristic of gene expression patterns in the developing spinal cord.

To be complete, it should be mentioned that Michaels et al. (1998) also did various cluster calculations on the same data set, which were of similar tenor.

#### 4.2.2 Functional grouping of essentially all genes from yeast

The research groups in Howard Hughes Medical Institute at Stanford University School of Medicine have already done a lot of gene expression experiments on the budding yeast *Saccharomyces cerevisiae*. The experiments include time courses of the mitotic cell division cycle (Spellman et al., 1998), sporulation (Chu et al., 1998), the diauxic shift (DeRisi et al., 1997), and several unpublished results. Eisen et al. (1998) have combined all these data to form a set of 79 gene expression measurements, which was clustered using pairwise average linking (Sokal and Michener, 1958). Though too much of the article by Eisen et al. (1998) places emphasis on the method used to enhance interpretation by the use of colour gradients, a few biologically interesting results are also mentioned. They find the extent to which gene expression patterns are sufficient to separate genes into functional categories across a relatively small and redundant collection of conditions surprising, and mention that it seems likely that the addition of more and diverse conditions can only enhance these observations. Indeed this is what is likely to be expected: they find a large cluster with little internal structure, composed of inert genes that barely respond to the given treatment (comparable to that found in the middle of the SOM by Tamayo et al. (1999); cf. paragraph 3.1). By increasing the number of experiments performed, more and more genes are given a chance to move out of this cluster. Together with functional relatives, their changed expression pattern will provide a handle for distinction by clustering algorithms. I will go into these expected patterns a bit more in the discussion (paragraph 5.1).

When reviewing yeast gene expression research, the work of Hauser et al. (1998) should be mentioned. The paper describes a thorough investigation of several aspects of microarray research, including the techniques used for RNA isolation, which can severely influence the reproducibility of experiments. Hauser et al. (1998) also mention that for increasing the statistical significance of measurements, re-use of DNA-chips is strongly advised, since preparation of a DNA-chip is often subject to more fluctuations than is desirable.

### 4.3 Logical description of the regulation of a sea urchin gene

Science magazine featured a remarkable research article last year, in which the authors combined previous knowledge and detailed analysis of the *Strongylocentrotus purpuratus* gene *Endo16*, to result in a precise logical description of the influences of the transcription factors of this gene (Yuh et al., 1998). *Endo16* is an endodermal gene that encodes a poly-functional protein, transcription of which is activated early in the development of the organism. The goal of the researchers was to develop a computer program mapping the *cis*-regulatory system of this gene, which could simulate the expression of *Endo16* given the levels of the transcription modules as input. The realisation of the program was a step-wise process: the time-series data of the experiments were tested against model output from the computer, by a procedure of minimising the square of a single free parameter per interaction model (Yuh et al., 1996). For completion of the model system, the logic statements were revised, or new statements were added progressively. If the model

came up with a dilemma, an experiment was designed to decide between alternatives. Ultimately, a logical system was composed, which could account for all the experiments carried out. The system contained both all-or-none effectors and transcription factors that had a gradual effect, proportional to their concentration.

Though this experiment did not make use of DNA-chips, it is a good example of how the goal of specifying the interactions of genes can be achieved. A warning that may be taken from the example, is that to resolve the complete topology of gene connections, a lot of detailed research may be necessary, since even a very small (8-node) network could be very complex. If all gene networks turn out to be as complex as the *cis*-regulatory system of *Endo16*, attempts to reconstruct the networks underlying all the genes arrayed in large microarrays with over 6000 nodes is a hopeless task. The only pathway to a reliable solution is to severely restrict the scale at which the system is studied. A question that arises, is whether the logical net found by (Yuh et al., 1998) is indeed the only solution, and also if the solution(s) might be found using a more general reconstruction approach, given the experiments they performed.

Arnone and Davidson (1997) review some research done on the organisation and function of genomic regulatory systems. They argue that the only way to understanding of the hardwiring of development, lies in direct analysis of *cis*-regulatory systems situated at all levels of the network. In their review, the methods used to infer regulatory circuitry from data collected using the microarraying technique are not at all mentioned; their opinion is that "correlation analysis" is close to worthless in analysis of real regulatory circuitry (E. Davidson, pers. comm.).

## 5 Discussion

### 5.1 Theory versus biology

An important assumption made in many of the theoretical models on which reverse engineering is tried out, is that the data comes in a set of random state transition pairs (Hertz, 1998; Akutsu et al., 1999). In biological experiments however, the results are often rather expression values of the genes at a number of sequential time-steps, which are far from uncorrelated (Eisen et al., 1998; Tamayo et al., 1999). As Tamayo et al. (1999) have mentioned, a lot of genes in the biological system remain uninfluenced by the experimental treatment, and in expression pattern clustering, they are therefore hardly distinguishable. To make this point explicit: if reconstruction of a genetic network is pursued, in which for all genes there are interaction parameters represented, we need to give all genes the opportunity to exert their influence. During an experimental treatment by which a gene is not affected, it will just stay at its equilibrium value: this is a likely scenario, since the vast majority of the interaction values in an  $n \times n$  interaction matrix are expected to be zero (Hertz, 1998). According to these considerations, what may be expected, is that a within the genetic network a number of subnetworks can be discerned, with at most just a few cross-links between them.

In this light it is perhaps also not surprising that clustering of genes by their expression patterns reveals genes that are functionally related. If a certain experiment is performed, the genes specific for the processes involved in that experiment will respond. If the genetic network is, as proposed above, composed of a large number of subnetworks, with only minor interconnections, then it can only be expected that clusters in the genetic expression patterns indicate functionally related genes.

### 5.2 Prospects

An important question in the modeling of genetic networks, is to what extent the net is modular. If we can assume well specifiable subnetworks, the detailed inference of the connections would require much less data points than inference of the complete genomic network. Two interesting findings in this light are the following:

- The clustering of the expression patterns in microarray experiments leads to specific groups of functionally related genes (Eisen et al., 1998). These groups of genes are involved in the processes whose reaction is invoked by the experiments performed in the experiment. The subsets of genes that are used in processes left unaffected by the experiment stay in their equilibrium, and thus form a large and rather inert cluster which is difficult to place (Tamayo et al., 1999).
- Some sequence homologs (genes supposedly originating from a single ancestor) are found scattered over all the expression pattern clusters (Hogeweg, preliminary results). This may mean that if a certain function is needed in different “response groups”, for each group, a special gene is used. Regulation of its expression specifically in the situation in question may be done by e.g. different regulation sites in the *cis*-region of the gene.

Unless the scale and experimental precision of microarraying increases very severely, this technique will never be sufficient as a basis from which to infer complete detailed genetic interaction nets. However, we may be able to build on the picture of gene networks as proposed above, and assume detached subnetworks (which may be identified by e.g. cluster analysis of expression data). Difficulties can be expected, since the basis of the clusters is a shared expression pattern: subnetwork reconstruction may only yield trivial connections. It is important to get a grip on the nature of the interconnections between the subnetworks, and to develop models that may contain the modular view of the genetic network. This could be an aid in network reconstruction, hinting at which interaction parameters need not be considered. Techniques as described in this thesis may be sufficient for network reconstruction within the modules, assuming complete connectivity therein.

As we have seen in chapter 4, microarraying can serve many other useful purposes, when it comes to the identification of gene transcripts or the discovery of coexpressed clusters of genes. In these investigations, the goal of reverse engineering is put aside, and the results of expression studies are used as heuristics in other research targets.

## References

- T. Akutsu, S. Miyano, and S. Kuhara. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. *Pacific Symp. Biocomp.* 99, 4:17–28, 1999.
- H.F. Arlinghaus, M.N. Kwoka, and K.B. Jacobson. Analysis of biosensor chips for identification of nucleic acids. *Anal. Chem.*, 69:3747–3753, 1997.
- M.I. Arnone and E.H. Davidson. The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124:1851–1864, 1997.
- A. Brāzma, I. Jonassen, J. Vilo, and E. Ukkonen. Predicting gene regulatory elements *in silico* on a genomic scale. *Genome Res.*, 8:1202–1215, 1998.
- T. Chen, H.L. He, and G.M. Church. Modeling gene expression with differential equations. *Pacific Symp. Biocomp.* 99, 4:29–40, 1999.
- S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P.O. Brown, and I. Herskowitz. The transcriptional program of sporulation in budding yeast. *Science*, 282:699–705, 1998.
- J.L. DeRisi, V.R. Iyer, and P.O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686, 1997.
- P. D’Haeseleer, X. Wen, S. Fuhrman, and R. Somogyi. Linear modeling of mRNA expression levels during CNS development and injury. *Pacific Symp. Biocomp.* 99, pages 41–52, 1999.
- D.J. Duggan, M. Bittner, Y. Chen, P. Meltzer, and J.M. Trent. Expression profiling using cDNA microarrays. *Nat. Genet. suppl.*, 21:10–14, 1999.
- M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.
- O. Ermolaeva, M. Rastogi, K.D. Pruitt, G.D. Schuler, M.L. Bittner, Y. Chen, R. Simon, P. Meltzer, J.M. Trent, and M.S. Boguski. Data management and analysis for gene expression arrays. *Nat. Genet.*, 20:19–23, 1998.
- D. Gerhold, T. Rushmore, and C.T. Caskey. DNA chips: promising toys have become powerful tools. *Trends Biochem. Sci.*, 24:168–173, 1999.
- T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.
- E. Hartuv, A. Schmitt, J. Lange, S. Meier-Ewert, H. Lehrach, and R. Shamir. An algorithm for clustering cDNAs for gene expression analysis. *RECOMB 99*, pages 188–197, 1999.
- N.C. Hauser, M. Vingron, M. Scheideler, B. Krems, K. Hellmuth, K.-D. Entian, and J.D. Hoheisel. Transcriptional profiling on all open reading frames of *saccharomyces cerevisiae*. *Yeast*, 14:1209–1221, 1998.
- J.v. Helden, B. André, and J. Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, 281:827–842, 1998.
- J. Hertz. Statistical issues in reverse engineering of genetic networks. *Pacific Symp. Biocomp.* 98, 1998. Poster, see <http://www.nordita.dk/hertz/projects.html>.
- D.H. Kozian and B.J. Kirschbaum. Comparative gene-expression analysis. *Trends Biotech.*, 17:73–78, 1999.
- E.S. Lander. Array of hope. *Nat. Genet. suppl.*, 21:3–4, 1999.
- G.G. Lennon and H. Lehrach. Hybridization analyses of arrayed cDNA libraries. *Trends Genet.*, 7:314–317, 1991.

- S. Liang, S. Fuhman, and R. Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symp. Biocomp.* 98, 3:18–29, 1998.
- T. Livache, H. Bazin, P. Caillat, and A. Roget. Electroconducting polymers for the construction of DNA or peptide arrays on silicon chips. *Biosens. Bioelectron.*, 13:629–634, 1998.
- A. Marshall and J. Hodgson. DNA-chips: an array of possibilities. *Nat. Biotech.*, 16:27–31, 1998.
- E. Marshall. Do-it-yourself gene watching. *Science*, 286:444–447, 1999.
- H.H. McAdams and A. Arkin. Stochastic mechanisms in gene expression. *Proc Natl. Acad. Sci. USA*, 94:814–819, 1997.
- G.S. Michaels, D.B. Carr, M. Askenazi, S. Fuhman, X. Wen, and R. Somogyi. Cluster analysis and data visualization of large-scale gene expression data. *Pacific Symp. Biocomp.* 98, 3:42–53, 1998.
- A.C. Pease, D. Solas, E.J. Sullivan, M.T. Cronin, C.P. Holmes, and P.A. Fodor. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci. USA*, 91:5022–5026, 1994.
- G. Ramsay. DNA-chips: state of the art. *Nat. Biotech.*, 16:40–44, 1998.
- J. Reinitz and D.H. Sharp. Mechanism of *eve* stripe formation. *Mech. Dev.*, 49:133–158, 1995.
- F.P. Roth, J.D. Hughes, P.W. Estep, and G.M. Church. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, 16:939–945, 1998.
- M.A. Savageau. Rules for the evolution of gene circuitry. *Pacific Symp. Biocomp.* 98, 3:55–65, 1998.
- M. Schena, D. Shalon, R.W. Davis, and P.O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, 1995.
- M. Shea and G. Ackers. The  $o_R$  control system of bacteriophage lambda. A physical-chemical model for gene-regulation. *J. Molec. Biol.*, 181:211–230, 1985.
- R.R. Sokal and C.D. Michener. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci Bull.*, 38:1409–1438, 1958.
- R. Somogyi and S. Fuhman. Distributivity, a general information theoretic network measure, or why the whole is more than the sum of its parts. *IPCAT 97*, 1997. In press.
- R. Somogyi and H. Kitano. Gene expression and genetic networks. *Pacific Symp. Biocomp.* 99, 4:3–4, 1999.
- R. Somogyi and C.A. Sniegoski. Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation. *Complexity*, 1:45–63, 1996.
- E. Southern. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.*, 98:503–517, 1975.
- E. Southern, K. Mir, and M. Shchepinov. Molecular interactions on microarrays. *Nat. Genet. suppl.*, 21:5–9, 1999.
- W.M. Spears. Simulated annealing for hard satisfiability problems. In D.S. Johnson and M.A. Trick, editors, *Cliques, coloring and satisfiability: second DIMACS implementation challenge*, volume 26 of *DIMACS series in discrete mathematics and theoretical computer science*, pages 533–558, 1996.
- P.T. Spellman, G. Sherlock, M.W. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, 9:3273–3297, 1998.
- Z. Szallasi. Genetic network analysis in light of massively parallel biological data acquisition. *Pacific Symp. Biocomp.* 99, 4:5–16, 1999.

- P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA*, 96:2907–2912, 1999.
- D. Thieffry and R. Thomas. Qualitative analysis of gene networks. *Pacific Symp. Biocomp.* 98, 3:77–88, 1998.
- P. Törönen, M. Kolehmainen, G. Wong, and E. Castrén. Analysis of gene expression data using self-organizing maps. *FEBS L.*, 451:142–146, 1999.
- A.M. Turing. The chemical basis of morphogenesis. *Trans. R. Soc. Lond B*, 237:37–72, 1952. Reprinted in *Bull. Math. Biol.* 52:153–197, 1991.
- M. Vingron. Computational analysis of expression data. In E. Bronberg-Bauer, A. De Beucklaer, U. Kummer, and U. Rost, editors, *Proceedings of workshop on computation of biochemical pathways and genetic networks*, Heidelberg, 1999.
- M. Vingron and J. Hoheisel. Computational aspects of expression data. *J. Mol. Med.*, 77:3–7, 1999.
- A. Wagner. Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization. *Proc. Natl. Acad. Sci. USA*, 91:4387–4391, 1994.
- M. Wahde and J. Hertz. Coarse-grained reversed engineering of genetic regulatory networks. *IPCAT 99*, 1999.
- K. Wang, L. Gan, E. Jeffery, M. Gayle, A.M. Gown, M. Skelly, P.S. Nelson, W.V. Ng, M. Schummer, L. Hood, and J. Mulligan. Monitoring gene expression profile changes in ovarian carcinomas using cDNA microarray. *Gene*, 299:101–108, 1999.
- D.C. Weaver, C.T. Workman, and G.D. Stormo. Modeling regulatory networks with weight matrices. *Pacific Symp. Biocomp.* 99, 4:112–123, 1999.
- X. Wen, S. Fuhrman, G.S. Michaels, D.B. Carr, S. Smith, J.L. Barker, and R. Somogyi. Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci. USA*, 95:334–339, 1998.
- D.M. Wolf and F.H. Eeckman. On the relationship between genomic regulatory element organization and gene regulatory dynamics. *J. Theor. Biol.*, 195:167–186, 1998.
- C.-H. Yuh, H. Bolouri, and E.H. Davidson. Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, 279:1896–1902, 1998.
- C.-H. Yuh, J.G. Moore, and E.H. Davidson. Quantitative functional interrelations within the *cis*-regulatory system of the *s. purpuratus endo-16* gene. *Development*, 122:4045–4056, 1996.