Temporal Boolean Network Models of Genetic Networks and Their Inference from Gene Expression Time Series

Adrian Silvescu and Vasant Honavar
Artificial Intelligence Research Laboratory
Department of Computer Science and
Graduate Program in Bioinformatics and Computational Biology
Iowa State University Ames, IA 50011, USA.

Abstract

Identification of genetic signal transduction pathways and genetic regulatory networks from gene expression data is one of key problems in computational molecular biology. Boolean networks [1, 2, 3], offer a *discrete time Boolean* model of gene expression. In this model, each gene can be in one of two states (on or off) at any given time. The expression of a given gene at time $t+1$ can be modeled by a *Boolean* function of the expression of at most k genes at time t, where typically $k << n$, and n is total number of genes under consideration. This paper motivates and introduces a generalization of the Boolean network model to address dependencies among activity of genes that span for more than one unit of time. The resulting model, called the $TBN(n,k,T)$ model, allows the expression of each gene to be controlled by a Boolean function of the expression levels of at most k genes at times in $\{t \cdots t-(T-1)\}$. We present an adaptation of a popular machine learning algorithm for decision tree induction [4] for inference of a $TBN(n,k,T)$ network from gene expression data. Preliminary experiments with synthetic gene expression data generated from known $TBN(n,k,T)$ networks demonstrate the feasibility of this approach.

1 Introduction

The central dogma of modern biology states that the functional state of an organism is determined largely by the pattern of expression of it's genes. Thus, the function of a cell, how well a cell performs its function, and even the determination of a cell's type is controlled by the level at which genes are expressed in the cell. Many biological processes of interest (e.g., cellular differentiation during development, aging, disease) are controlled by complex interactions over time between hundreds of genes. Furthermore, each gene is involved in multiple functions. Therefore, understanding the nature of complex biological processes such as development, cellular differentiation, carcinogenesis, etc., requires determining the spatio-temporal expression patterns of thousands of genes, and, more importantly, seeking out the organizing principles that allow biological processes to function in a coherent manner under different environmental conditions.

A number of emerging high-throughput technologies are revolutionizing the means by which genetic pathways involving hundreds of genes can be studied. Among these the DNA microarray technology allows researchers to determine the mRNA levels of expression at different times during cell life.

These advances in data aquisiton make possible, at least in principle, the inference of models of genetic (regulatory and control) networks from gene expression data. However, the large number of genes involved, complexity of the pathways, and existence of pleiotropic and multigenic interactions [2] make this a challenging task. Less sophisticated methods such as clustering can also be used in order to analyze gene expression data but the information that can be obtained by these methods is restricted to either positive or negative correlations.

A variety of formal models for capturing the interactions and functional dependencies in genetic networks have been proposed in the literature. These include: *electrical circuits* [9], *Boolean networks* [1, 2, 11, 3, 12], *differential equations* [13], *Petri nets* [14, 15], and *Weight matrices* [16]. Each of these approaches has its own strengths and limitations in terms of: faithfulness or accuracy of the model relative to the biological phenomenon being modeled; transparency of the model (or equivalently, its explanatory value); experimental feasibility (in terms of data requirements) of model construction; computational tractability of automated model inference from data.

The focus of this paper is on Boolean Network models of genetic networks wherein each gene can be in one of two states – ON (expressed) or OFF (not expressed). One of their main advantages is their

comprehensibility due to the transparency of the representation. It is easy to extend such network models in order to allow for a discrete set of states for each gene, instead of just two states, or (at the risk of reduced transparency) even to allow states that take on real (and hence an infinite) set of values.

Previous work on inference of Boolean models of genetic networks [3, 11] has focused on data that randomly sample the state transitions typically under the assumption that the state of a gene at a given time step is influenced by the states of a subset of genes in the network at the previous time step. Since many experiments involve obtaining gene expression data by monitoring the expression of genes involved in some biological process (e.g., neural development) over a period of time, the resulting data is in the form of a *time series*. In such a setting, it is of interest to understand how the expression of a gene at some stage in the process is influenced by the expression levels of other genes during the stages of the process preceding it. In order to model the temporal dependencies that span several time steps, we introduce in this paper, a generalization of Boolean Networks called Temporal Boolean networks. This will basically transform the Boolean Networks from a Markov(1) to Markov(T) model where T is the length of the time window during which a gene can influence another gene. We will demonstrate how Temporal Boolean Networks can be efficiently inferred using an adaptation of a greedy decision tree learning algorithm [4].

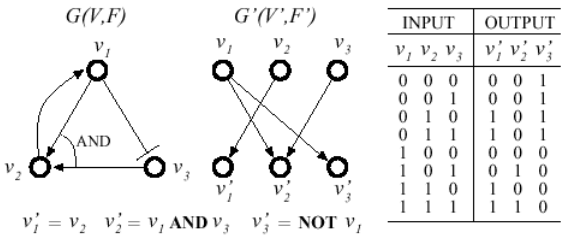
2 Temporal Boolean Networks

2.1 Boolean Networks

Boolean networks, introduced by Kauffman [1] and explored in [2, 3, 11, 12] offer an attractive discrete time, boolean model for gene expression. In this model, each gene can be in one of two states (on or off) at any given time, and the expression of a given gene at time $t+1$ can be modeled by a *Boolean* function of the expression of at most k genes at time t, where typically $k << n$, and n is total number of genes under consideration. We call this family of models $BN(n,k)$ networks. Following Akutsu *et al.* [3] we give the folowing definition:

A Boolean Network $G(V,F)$ consists of a set $V = \{v_1, ..., v_n\}$ of nodes representing genes and a list $F = (f_1, ..., f_n)$ of *Boolean functions*, where a boolean function $f_i(v_{i_1}, ..., v_{i_k})$ with inputs from specified nodes $v_{i_1}, ..., v_{i_k}$ is assigned to each node v_i. For a subset $U \subseteq V$, an *expression pattern* ψ of U is a function from U to $\{0,1\}$. *States* of the boolean Network $G(V,F)$ correspond to expression patterns of V. That is, ψ represent the states of nodes (genes), where each node is assumed to take either 0 (not-expressed) or 1 (expressed) as its state value. Typically, when the usage is clear from the context, we omit ψ. For example, we write $v_i = 1$ for denoting $\psi(v_i) = 1$. In a Boolean network, the expression pattern ψ_{t+1} at time $t+1$ is determined by Boolean functions F from the expression pattern ψ_t. at time t (i.e. $\psi_{t+1}(v_i) = f_i(v_{i_1}, ..., v_{i_k})$).

It is convenient to consider the wiring diagram [2, 11] $G'(V',F')$ of a Boolean network $G(V,F)$ (See Fig. 1). For each node v_i in V, such that $v_{i_1}, ..., v_{i_k}$ are input nodes to v_i in $G(V,F)$. we consider an additional node v'_i and construct an edge from v_{i_j} to v'_i for each $1 \leq j \leq k$. Let $G' = (V',F')$ the network with nodes $v_1, ..., v_n, v'_1, ..., v'_n$ constructed in this way. Then, the expression pattern of the set $\{v'_1, ..., v'_n\}$ is determined by $v'_i = f_i(v_{i_1}, ..., v_{i_k})$. That is, the expression of pattern $\{v_1, ..., v_n\}$ corresponds to one at a time t and the expression pattern of $\{v'_1, ..., v'_n\}$ corresponds to one at time $t+1$. Moreover, it is convenient to consider the expression pattern of $\{v_1, ..., v_n\}$ as the INPUT, and the expression pattern of $\{v'_1, ..., v'_n\}$ as the OUTPUT.



INPUT			OUTPUT		
v_1	v_2	v_3	v_1'	v_2'	v_3'
0	0	0	0	0	1
0	0	1	0	0	1
0	1	0	1	0	1
0	1	1	1	0	1
1	0	0	0	0	0
1	0	1	0	1	0
1	1	0	1	0	0
1	1	1	1	1	0

Figure 1: A Boolean Network (Akutsu *et al.* 1999)

2.2 Motivations for Generalizing the Boolean Network Model

Our proposed generalization of the Boolean network model to a Temporal Boolean Network model is motivated by the following considerations:

- Boolean networks described above are incapable of modeling the existence of latency periods (lasting more than one unit of time) between the expression of a gene and the observation of its effect. For example a gene (say g_4) whose inhibitory effect (say on gene g_5) depends on an inducer (say g_3) first has to bind with this inducer in order to be able to bind to the inhibition site on g_5. Therefore there can be a significant delay between the expression of the inhibitor gene g_4 and its observed effect i.e. the inhibition of the gene g_5.

- Not all variables that can influence the expression level of a gene are necessarily observable. For instance, assume that genes $g_1, g_2, \cdots g_{1000}$ be genes are under study. Suppose that the expression of genes g_1 at time t might turn on a gene g_u at time $t+1$. It is quite possible that g_u is not among the genes that are being monitored in the experiment, or even among the genes that are currently known. Suppose gene g_u being on at time $t+1$ results in a gene g_2 being turned on at time $t+2$. Since the expression of g_u cannot be observed, the Boolean network model described above will be unable to implicate g_1 in the control of g_2. Note that even if all the genes of the cell are monitored in the experiment then the unknown factor g_u may stand for a non-genetic environmental factor. (*Note:* Long temporal delays observed in the our new model might indicate the presence of hidden factors and provide hints for their discovery, therefore contributing to the improvement of the quality of data collected in future experiments).

In what follows, we introduce a generalization of the Boolean network model to address dependencies among the activity of genes that span for more than one unit of time. The resulting model, called $TBN(n,k,T)$, allows the expression of each gene at time $t+1$ to be controlled by a Boolean function of the expression levels of at most k genes at times in $\{t \cdots t-(T-1)\}$.

2.3 Temporal Boolean Networks

In the Temporal Boolean Networks (TBN) model we will allow the state of gene at time $t+1$ to depend on the states of genes at times $t, t-1, ..., t-(T-1)$ instead of only t. The representation of such a network, by analogy with the boolean networks is given in Fig. 2.

The only change from the Boolean Network model shown in Fig. 1. is that the expression level of gene v_2 at time $t+1$ depends on the value of the gene v_1 at time $t-1$ instead of time t. This is represented by a label of -1 label on the edge between v_1 and v_2. By default the edges are assumed to carry a value of 0, which means that the dependency they represent is from time t to time $t+1$. In general, an edge labeled as $-k$ will represent a dependency between the values at time $t-k$ and $t+1$. In general, each edge in the network can have a label $-l$ where l is drawn from the set $\{0, \cdots (T-1)\}$ and will correspond to an edge from level $t-l$ to level $t+1$ in the wiring diagram. The functional dependecy will be represented by a boolean table for each gene, but this time the inputs can be gene expression values at more than one time step in the past.

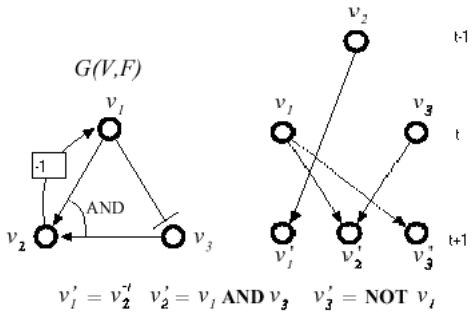


Figure 2: A Temporal Boolean Network

It is also straightforward to extend the Temporal Boolean Networks model in order to allow multiple discrete levels of expression thus yielding Temporal Discrete Networks $TDN(n,k,T,D)$ wherein each gene can be expressed at levels $0, 1, \cdots D-1$.

3 TBN inference from time series

Temporal Boolean Networks model functional dependencies among genes using Boolean functions. Boolean functions have several alternative representations. The simplest (and the most explicit) representation is a truth table shown in Fig. 1. However, functions that correspond to descriptions of natural phenomena typically lend themselves to more compact representations in the Conjunctive Normal Form (CNF), Disjunctive Normal Form (DNF), or in the form of decision trees, decision lists, etc. We have chosen to represent the Boolean functions used to describe Temporal Boolean Networks in the form of Decision Trees.

For example, consider gene g_1 that is induced by a complex formed from the products of two genes g_2 and g_3 and these are the only gene that influence g_1. A decision tree that describes this dependence is given in Figure 3. Similarly, for each gene we can construct such a decision tree. The genes that control gene g_i's expression appear as nodes in its corresponding decision tree.

Since every k-input boolean function can be represented by a decision tree of depth at most k, we can ensures that the resulting representation is expressive enough to describe any member of the Temporal Boolean Network family $TBN(n,k,T)$. We use n decision trees, one for each gene, to collectively describe a Temporal Boolean Network model of a genetic network. The output of each decision tree represents the expression level (0 or 1) of the corresponding gene

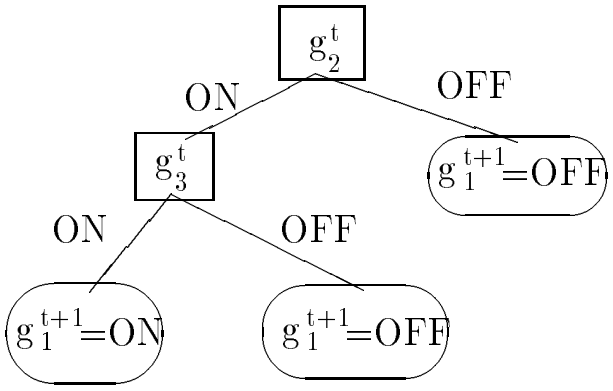


Figure 3: A decision tree that represents the fact that if genes g_2 and g_3 are expressed (ON) leads to the expression of gene g_1 (ON).

based on at most k expression levels over a time window of length at most T.

Several algorithms for inferring decision trees from data (in the form of sample input-output pairs) are available in the machine learning literature. The ID3[4] algorithm and its variants are based on a greedy search through the space of decision trees in order to identify a compact decision tree that adequately models the observed data and (under some reasonable assumptions) has high predictive accuracy on unobserved data. This search for a compact tree, guided by the entropy reduction (or information gain) criterion corresponds to a greedy version of the approach used in REVEAL [11]. Greedy search makes this approach computationally tractable for large genetic networks. A large body of empirical results in the machine learning literature suggest that the decision trees inferred by greedy search compare favorably with those inferred using exhaustive search in terms of predictive accuracy. Hence, we used a greedy search guided by information gain, over the space of depth k decision trees.

The training data for a decision tree modeling the functional dependency of the level the expression of gene g_i consists of observed input output pairs where each input is a nT bit boolean vector encoding the activities of each of the n genes at times $t, t-1, t-2 \cdots t-(T-1)$ and the corresponding output is the observed expression level of g_i at time $t+1$. For a given gene g_i, $m-T$ input–output samples (or training examples) are obtained by sliding a window of length T (where $T < m$) over the rows of a gene expression time series $\mathcal{E}$ (See section 2). Thus, an $m \times n$ gene expression matrix yields $m-T$ training samples for each of the n decision trees. Examples obtained from multiple time series are used for inferring a genetic network.

4 Experimental Results

The experiments described in this section were designed to explore the performance of the proposed approach to genetic network inference on randomly generated temporal boolean networks. In generating a network, we assume that the probability that the expression level of a gene at time $t+1$ depends on the expression levels of genes at time $t-\delta$ is proportional to ζ^{δ} $\forall t$ such that $0 \leq t \leq T-1$ for some choice of ζ where $0 < \zeta \leq 1$. For each network, we generated 20 time series of length 100, by setting the expression levels over a window of length T to random values and recording network's outputs over 100 time steps. Thus, each time series resulted in an $n \times (100+T)$ boolean matrix $\mathcal{E}$. The 20 time series collectively provided 100×20 training examples for each of the n genes. Each time point contains the expression levels for all genes at that moment of time. A decision tree was inferred for each gene.

Then we evaluated the results in terms of the *sensitivity*, *specificity* and *accuracy* of the inferred decision tree DT_i with respect to the corresponding temporal boolean network TBN_i for each gene g_i.

We denote the fact that the decision tree DT_i captures the dependence of the expression level of gene g_i at time $(t+1)$ on the expression level of the gene gene g_j at time $(t-\tau)$ where $\tau \in \{0, \cdots, (T-1)\}$, by writing $(\tau, j) \in DT_i$. Similarly, we denote the fact that the expression level of gene g_i at time $(t+1)$ depends on the expression level of gene g_j at time $(t-\tau)$ if gene g_i were controlled by the temporal boolean network TBN_i by writing $(\tau, j) \in TBN_i$. Let

$$DEP_{DT_i} = \{(\tau, j) | (\tau, j) \in DT_i\}$$

Let

$$DEP_{TBN_i} = \{(\tau, j) | (\tau, j) \in TBN_i\}$$

Then the *sensitivity* of the decision tree DT_i for gene g_i is defined as follows:

$$sensitivity(i) = \frac{|DEP_{TBN_i} \cap DEP_{DT_i}|}{|DEP_{TBN_i}|}$$

The *sensitivity* of the inferred decision tree measures the degree, to which this tree succeeds in capturing the dependency of gene g_i on other genes, with respect to the true Temporal Boolean Network.

The *sensitivity* of the inferred set of decision trees $DT = \{DT_i | i \in \{1, \cdots, n\}\}$ relative to the corresponding TBN is given by:

$$sensitivity(DT, TBN) = \frac{1}{n} \sum_{i=1}^{n} sensitivity(i)$$

The *specificity* of the decision tree inferred for gene g_i is defined as follows:

$$specificity(i) = \frac{|DEP_{TBN_i} \cap DEP_{DT_i}|}{|DEP_{DT_i}|}$$

Specificity of the inferred decision tree measures the degree to which it misleads us regarding the dependency of gene g_i on the various genes in the genetic network.

The specificity of the inferred set of decision trees $DT = \{DT_i | i \in \{1, \cdots, n\}\}$ relative to the corresponding TBN is given by:

$$specificity(DT, TBN) = \frac{1}{n} \sum_{i=1}^{n} specificity(i)$$

The *accuracy* of the decision tree inferred for gene g_i is defined to reflect the degree to which it correctly predicts the expression level of gene g_i (as estimated from a set of gene expression time series data).

Let Λ be a set of gene expression time series used to evaluate the accuracy of a decision tree inferred for gene g_i.

Then we will denote by $accuracy\ (i, \lambda)$ the accuracy of the tree for gene g_i on a time series $\lambda \in \Lambda$. $accuracy(i, \lambda)$ represents the fraction of expression levels of gene g_i from the time series λ that are correctly predicted by the inferred decision tree DT_i, relative to the total size of λ. Thus, if λ is a time series of length 100, and at 80 of the 100 time points, the expression level of gene g_i predicted by DT_i agrees with the corresponding values observed in λ, $accuracy(i, \lambda) = 0.8$. The *accuracy* of DT_i is estimated as follows:

$$accuracy(i) = \frac{1}{|\Lambda|} \sum_{\lambda \in \Lambda} accuracy(i, \lambda)$$

The accuracy of an inferred decision tree DT_i was estimated using an independently generated test set of 20 time series each of length 100 generated from TBN_i. The estimated accuracy of the inferred set of decision trees $DT = \{DT_i | i \in \{1, \cdots, n\}\}$ relative to the corresponding TBN is given by:

$$accuracy(DT, TBN) = \frac{1}{n} \sum_{i=1}^{n} accuracy(i)$$

Each experiment consisted of 10 independent runs of this procedure (each with a new randomly generated network) and the results presented show averages over these runs.

The first experiment presented in Figure 4a) is for a boolean network with $n = 16$ genes, with $T = 3$ and k varying from 2 to 10. Similar results were obtained for networks with 32 genes. The experiment shows that the sensitivity increases as the degree of interaction k increases thus the probability of being able to explain the phenomenon using other genes than the ones that really produce it is decreasing as the complexity of the phenomenon k increases.

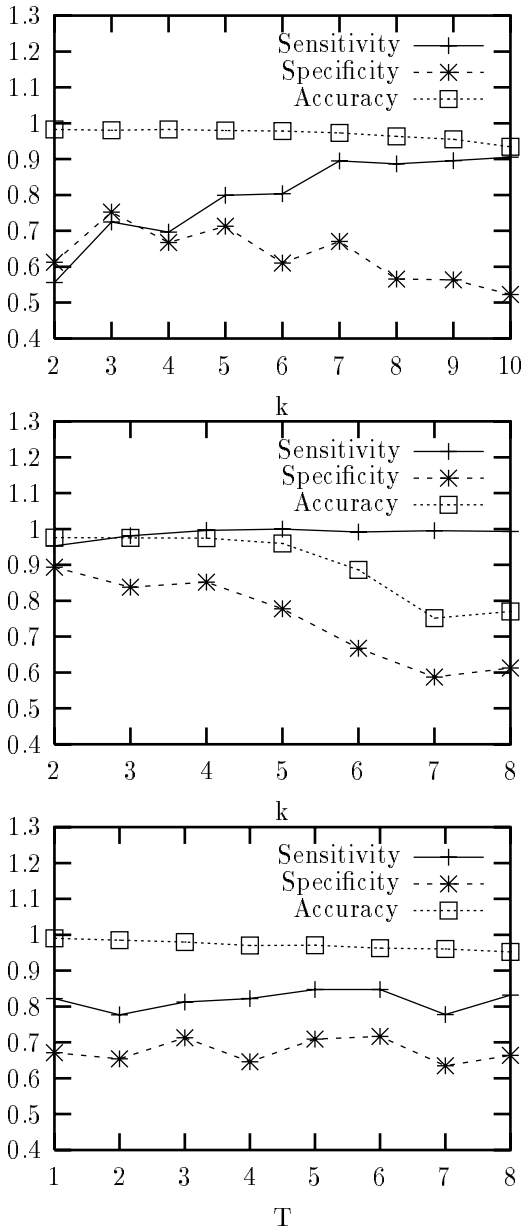


Figure 4: a)boolean net k=2-10, b)4-ary net k=2-8, c) boolean net k=6 T=1-8

The second experiment presented in Figure 4b) is for a 4-ary network with $n = 16$ genes, $T = 3$ and k varying from 2 to 10.(i.e. a Temporal Discrete Network TDN(n=16,k=2-10,T=3,D=4). The goal of this experiment is to observe what is the effect of considering four levels of expression instead of only two. In this case we observe sensitivity and specificity increasing and accuracy decreasing with increase in k.

These experiments show that as we increase the complexity of the model (in this case the number of expression levels) the sensitivity, which represents the probability that we will identify genes that represent the true dependency, increases (compare the Sensitivities in Figure 4a) and Figure 4b)). Further experiments revealed that the reduction in accuracy can be explained by the necessity for additional data as k increases, in the conditions of a 4-ary network were for each gene we have to disambiguate among 4^k 4-ary functions versus 2^k in the 2-ary (Boolean) case.

The third experiment presented in Figure 4c) is for is for a boolean network with $n = 16$ genes, $k = 6$ and T varying from 1 to 8. The goal of this experiment is to determine whether the variation of the time-span T of the dependency has any effect on the performance of the inference algorithm and it shoes that this remains relatively constant as we change T.

5 Summary and Discussion

In this paper we have introduced the Temporal Boolean Networks (and Temporal Discrete Networks) which generalize the Boolean Network model in order to cope with dependencies that span over more than one unit of time. We have shown how the problem of Temporal Boolean Network inference can be reduced to the problem of inferring a set of decision trees. We have demonstrated, through a series of experiments using artificially generated networks, the effectiveness of a simple and fast decision tree learning algorithm for inferring Temporal Boolean Networks and Temporal Discrete Networks from time series data.

Work in progress is aimed at evaluating the effectiveness of this approach for inferring genetic networks from gene expression time series data. Some directions for future work include investigation of variations of the model to accomodate probabilistic Boolean and Discrete valued functional dependencies, and continuous valued expression levels as well as alternative (e.g., event-based and interval-based) representations of time.

Acknowledgments: This research was supported in part by grants from the National Science Foundation (awards 9982341 and 9972653), the Carver Foundation, and Pioneer Hi-Bred Inc. This research has benefited from interactions with Dr. Phil Haydon, Dr. Pat Schnable, Dr. Xun Gu, Dr. Gavin Naylor, and Justin Schonfeld of the Iowa State University Bioinformatics and Computational Biology Program.

References

[1] Kauffman, S.A., The Origins of Order, Self-Organization and Selection in Evolution, (1993).

[2] Somogyi R., et al., *Complexity* 1(6):45-63, (1996).

[3] T. Akutsu, et al., *Pacific Symp. on Biocomp.*, 4:17-28 (1999).

[4] Quinlan J.R., *J of the Op. Research Soc.*, 38:347-352, (1987).

[5] J. L. DeRisi, et al., *Science*, 278:680-686 (1997).

[6] P. T. Spellman, et al., *Mol Bio Cell*, **9**, 3273-3297 (1998).

[7] X. Wen, et al., *Proc Natl Acad Sci U S A*, 95:334-339 (1998).

[8] S. P. Gygi, et al., *Moll. Cell Bio.*, 19:1720-1730 (1999).

[9] McAdams H.H., et al., *Science*, 269:650-656, (1995).

[10] T. Akutsu, et al., *Pacific Symp. on Biocomp.*, 5:290-301 (2000).

[11] S. Liang, et al., *Pacific Symp. on Biocomp.*, 3:18-29 (1998).

[12] T.E. Ideker, et al., *Pacific Symp. on Biocomp.*, 5:302-313 (2000).

[13] T. Chen, et al., *Pacific Symp. on Biocomp.*, 4:29-40 (1999).

[14] Goss P.J.E., et al., *Proc. Natl. Acad. Sci USA*, 95, 6750-6755, (1998).

[15] H. Matsuno, et al., *Pacific Symp. on Biocomp.*, 5:338-349 (2000).

[16] D.C. Weaver, et al., *Pacific Symp. on Biocomp.*, 4:112-123 (1999).

[17] R. Somogyi, et al., *Pacific Symp. on Biocomp.*, 5:288-289 (2000).

[18] Stormo, G., *Pacific Symp. on Biocomp.*, 5:413-414 (2000).

[19] Eisen, M., et al., *Proc. Natl. Acad. Sci. USA*, 95:14863-14868 (1998).

[20] Ben-Dor, A., et al., *J. of Comp. Bio.*, 6:281-297 (1999).

[21] G.S. Michaels, et al., *Pacific Symp. on Biocomp.*, 3:42-53 (1998).