Simulation Studies for the Identification of Genetic Networks from cDNA Array and Regulatory Activity Data

Daniel E. Zak
University of Delaware/
Thomas Jefferson University
Newark DE 19711
zak@che.udel.edu

Francis J. Doyle III
University of Delaware
Chemical Engineering
Department
Newark, DE 19711
doyle@che.udel.edu

Gregory E. Gonye James S. Schwaber
Daniel Baugh Institute
Thomas Jefferson University
1020 Locust Street, Philadelphia,
gregory.gonye@mail.tju.edu
james.schwaber@mail.tju.edu

ABSTRACT

A model genetic regulatory network for the evaluation of genetic regulatory network identification methods is described in the present work. This model is novel because of its mechanistic basis and its relatively large (10 gene) scale. The model is used to evaluate several simple techniques from the literature for identifying genetic regulatory networks from data. Even though simple models derived from cDNA array data alone can fit data very well, it is observed that their applicability for determining network architecture is questionable. A method that combines cDNA array data, regulatory activity data, and genomic sequence data is, of the methods considered, most promising for the identification of network architecture.

1. INTRODUCTION

The prospect of identifying genetic regulatory networks from high throughput data sets such as DNA arrays and proteomics has recently gained considerable interest. While several approaches have been developed to determine networks from data, little has been published on the development of simulation tools that can validate the methods, and what has been published did not have a mechanistic basis or describe realistic networks [11]. It is believed that there is value in being able to compare the results of the network identification methods with the "right answer", and therefore the present work represents the beginning of a larger effort. The model described in this work is unique because it has a mechanistic basis and is of relatively large scale (10 genes).

A description of genetic regulatory networks and methods used to identify them are first given. These are followed by a description of the model system. The results obtained from applying the identification methods to the model system are then shown. Finally conclusions and future directions for this work are discussed.

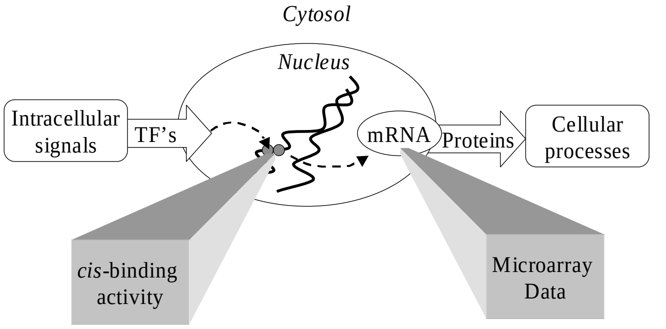
2. GENETIC REGULATORY NETWORKS

Genes function in highly interconnected, hierarchical, and nonlinear chemical networks. Specific phenotypes are often not the result of the expression of single genes but rather the result of interactions of multiple genes, as in the case of some human cancers [8], as well as the past and present intracellular and extracellular environments. New analytical techniques are being developed that allow the quantification of many intracellular factors [5]. These include cDNA array technology [7], which allows the relative transcription levels of thousands of genes to be measured in parallel, and gel electrophoresis and mass spectrometry, which allow levels of hundreds of proteins to be quantified [6]. Also included is "regulatory activity", which in this sense refers to how actively *cis* regulatory sequences are being bound in the nucleus and corresponds roughly to the levels of active transcription factors. CIStem Molecular Corporation (www.cistemcorp.com) is a key provider of this type of data.

The specific form that the genetic regulatory network takes is dependent of the type of data that is used to derive it. A complete genetic regulatory network would be derived from measurements of all relevant cellular components and would be exceptionally complicated. Alternatively, approximate models of genetic regulatory networks can be derived from system-wide measurements of key cellular components. Most work to date related to the derivation of genetic regulatory networks has been using cDNA array data over a time series [3,10,11]. The objective was to determine a relationship between gene expression levels at past and future time points, with the assumption that fundamental information about the genetic regulatory network would be contained within this relationship. While genetic regulatory networks derived from expression data alone may be useful for predicting gene expression levels, their relationship to biological quantities is not clear. One reason is that the interdependencies within these genetic networks may depend on the rate at which the data was collected, due to the existence of different time scales in the actual network. Corresponding changes in gene expression may correlate but are not necessarily causal.

A more structural form of genetic regulatory network may be obtained by coupling expression data with regulatory activity. Combining regulatory activity data with DNA array data gives measures of nuclear input and output, with the genetic regulatory network mapping from one to the other. A scheme of this is shown below in Figure 1.

The approach of using regulatory activity and cDNA array data together holds several advantages over previous efforts that attempted to identify genetic regulatory networks using cDNA array data alone. Since the cDNA array alone approaches attempt to express expression levels at the present time as a function of the expression levels at a previous time, and possibly exogenous inputs, approximately n_m^2 interactions need to be defined for a system of n_m genes. Several groups (reviewed in [11]) have attempted to remedy this by clustering the genes on the basis of similarity of their temporal expression profiles, and then fitting the expression levels of the expression patterns for the clusters or "supergenes."



TF = transcription factor

Figure 1: Nuclear input/output as measured by cDNA array and regulatory activity measurements.

The genetic regulatory network that can be identified from combined cDNA array and regulatory activity data is different from that identified using only cDNA array data. Instead of taking a global, closed loop, black box view of genetic regulation, a more mechanistic local, nuclear, viewpoint is taken. The objective is not to relate relative expression levels across time points, but rather to relate nuclear input to nuclear output. The only interactions that need to be defined are those between the expression levels and the regulatory element activities. There is no longer a need to include exogenous inputs because they are contained in regulatory activity levels, having first been modulated by the intracellular signaling network. Since this approach does take a nuclear view of genetic regulation, however, there is still a need to connect the relative transcriptional levels and the exogenous inputs back to the regulatory element activity to close the loop.

While it may appear that this new method has only the advantage of reducing the number of interactions to be defined from n_m^2 to n_t*n_m, where n_t is the number of *cis* regulatory elements, there are other benefits. In comparison to the methods that used cDNA array data alone, where it was not known a priori which genes would influence which other genes, for the combined approach it is possible to determine which regulatory elements will affect the expression levels of which genes. Through methods such as those of Tavazoie *et al*. [9], genes can be grouped into co-regulated families through analysis of expression profiles across many conditions. For each group, common upstreams regulatory motifs can be found so that it is possible to determine which regulatory elements are present upstream of the coding region for a given gene. If a certain regulatory element is not present upstream of a certain gene, then that is one less interaction that needs to be defined. Given that an average of regulatory elements present for given genes can be on the order of 10, the actual number of parameters that need to be identified is then of order 10*n_m, a significant reduction.

Throughout this paper, it is implied that knowing which *cis* regulatory region is upstream of a certain gene is equivalent to knowing which transcription factor binds upstream of that gene. This information may not always be known, and resort to existing databases, such as Biobase.de (www.biobase.de) that contain this information must be made.

3. METHODS FOR THE IDENTIFICATION OF GENETIC REGULATORY NETWORKS

Determining the network architectures from the cDNA array and regulatory activity data, is nontrivial. The data consists of relative transcription levels and relative activity levels for thousands of genes and tens of *cis* sites over hundreds of time points. Mathematical methods are necessary to determine the interconnections within this data.

The methods evaluated in this study include those that utilize cDNA array data alone as well as combined cDNA array and regulatory activity data. Structurally they are very similar. The three types considered here are linear [3,11], linear – log, and linear + squashing [10]. In the linear approaches, the expression levels of the genes at measurement instance $k + 1$ is a linear combination of either the expression levels and exogenous inputs (for the cDNA array alone approach) or regulatory activities (for the cDNA array plus regulatory activity approach) at measurement instance k. This is expressed mathematically as: $X(k + 1) = A*U(k)$, where $X(k+1)$ is an n_m by 1 vector of expression levels at measurement instance $k + 1$, A is a n_m by n_u matrix, relating the n_m genes to the n_u inputs (expression levels and exogenous inputs or regulatory activities) , and $U(k)$ is an n_u by 1 vector of inputs at measurement instance k. In the linear-log approach, $X_l(k + 1) = A_l*U_l(k)$, where X_l is a vector of the log of the expression levels, and U_l is a vector of the log of the input values. In the linear+squashing approach:

$$x_i(k+1) = \frac{m_i}{1+e^{-\sum_j a_{ij}u_j(k)}} \qquad (1)$$

In (1), m_i is the maximum expression level of gene i and a_{ij} is a coefficient for the interaction between the input u_j and x_i. This can be arranged into:

$$q_i(k+1) \equiv -\ln\left(\frac{m_i}{x_i(k+1)}-1\right) = \sum_j a_{ij}u_j(k) \qquad (2)$$

From (2) it is clear that the system can be expressed as: $Q(k+1) = A_q*U(k)$, where Q is an n_m by 1 vector of the transformed expression levels.

If there are n_p measurements of X and U, where $n_p > n_u + 1$, the linear, linear-log, and linear+squashing systems can be arranged into: $X = A*U$, $X_l = A_l*U_l$, and $Q = A_q*U$, respectively. X, X_l, and Q are now n_m by n_p-1 matrices of X, X_l, and Q in sequential order, and U and U_l are n_u by n_p-1 matrices of U and U_l at the experimental points, respectively, one step behind those in X, X_l, and Q.

The genetic regulatory networks, which are represented by A, A_l, and A_q, can be determined by the simple matrix inversions of U and U_l: $A = X*U^{-1}$, $A_l = X_l*U_l^{-1}$, and $A_q = Q*U^{-1}$. When n_p =

$n_u + 1$, U and U_l will be square, and the inverses are simple matrix inversions. When $n_p > n_u + 1$, the systems are overdetermined and the inverses must be the appropriate pseudoinverse. For the linear system, the Moore-Penrose pseudoinverse will give the least squares solution. For the linear-log and linear+squashing systems, the pseudoinverses should involve nonlinear optimizations for correct weighting of the errors, due to the transformations of the original data that were performed. In this study, however, the Moore-Penrose pseudoinverse was used for all systems when $n_p > n_u + 1$ for convenience.

4. MODEL

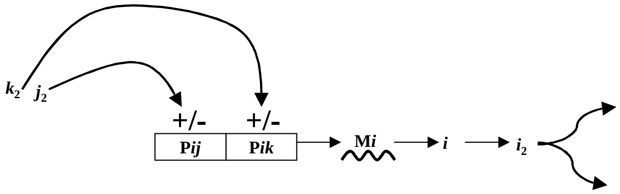


Figure 2: Core process of model. Transcription factor dimers k_2 and j_2 bind to cis regulatory sites P*ik* and P*ij* upstream of gene *i*, respectively. Binding of k_2 and j_2 influences the rate of transcription of mRNA M*i*, which is translated into protein *i*, which dimerizes to bind to other *cis* regulatory sites.

The model genetic network developed and applied in this study is based on the reaction scheme shown above in Figure 2. Binding of transcription factors to *cis* regulatory sites of the gene *i* influences its rate of transcription. Protein *i* is translated from the mRNA M*i*, dimerizes, and then influences the rate of transcription of other genes. In the model system, it is assumed that each transcription factor binds only cis regulatory sequences that are specific to it.

The overall architecture of the model genetic network developed and applied in this study is shown in Figure 3. It contains several structures similar to those in the literature. These include an endogenous oscillator, based on the hysteretic oscillator of Barkai and Leibler [1], a switch mechanism, analogous to the genetic switches of Gardner et al [4] and Cherry and Adler [2] as well as two cascades and a ligand binding mechanism that influences transcription. All transcription factors dimerize before they are active, with the exception of A, which was derived from [1]. Each gene has at least two rates of transcription: with promoter bound and unbound. For genes with multiple promoters, the total rate of transcription is a linear sum of the rates associated with the bound and unbound states of each of the promoters.

Mathematically, the system consists of 97 elementary reactions with 54 species. The quasi-steady state and pseudo-equilibrium assumptions were not made for any reactions. Delays associated with transcription and translation processes were not included. The equations were integrated using the MATLAB ode15s solver (The Mathworks Inc). The parameters in the model, as well as the MATLAB code to run it, are available upon request from the authors.

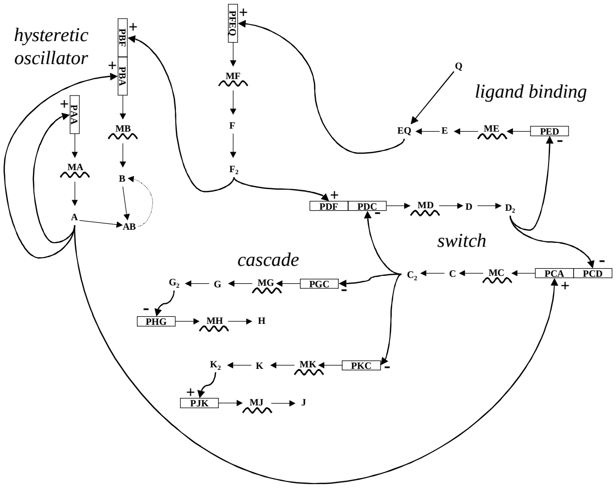


Figure 3: Genetic regulatory network used in this study.* The network consists of a hysteretic oscillator, based on the circadian oscillator of Barkai and Leibler [1]; a switch, analogous to those of Gardener *et al.* [4] and Cherry and Adler [2], a cascade, and a ligand binding mechanism. (* For a larger version, please see Appendix)

The model architecture was chosen so that without ligand input, there would be an endogenous oscillation in production of proteins A and R, as well as with a high levels of transcription factor C, receptor E, and downstream gene H. This could be interpreted as a "dividing state", where the oscillations correspond to a highly abstracted cell cycle. Upon ligand binding, the cell shifts into a state where the oscillations cease, the receptor is down regulated, and transcription factors D and F and downstream protein J are produced in large quantities. This could correspond to "differentiated state", where cell division ceases and a new set of proteins is expressed. Without constant input, however, the cell eventually returns to its dividing state. This presence of the different time scales in the model is believed to make it somewhat representative of a real system, as well as provide a suitable challenge for the network identification methods.

An abstracted genetic regulatory network for this system can be represented simply as the collection of the signs on the interactions between the transcription factors/*cis* regulatory regions and the genes. This is shown in Table 1, where "+" indicates that binding of the transcription factor to the *cis* region upstream of that gene increases its transcription rate, while the "-" indicates that the binding decreases the rate of transcription.

Table1: Genetic regulatory network of model system

	A	CC	DD	FF	GG	KK	EQ
MA	+						
MB	+			+			
MC	+		-				
MD		-		+			
ME			-				
MF							+
MG		-					
MH					-		
MJ						+	
MK		-					

The time course used for analysis is shown in Figure 4. Ligand is injected for 10 minutes at a rate of 100,000 molecules/minute at time 1000 minutes. Number of molecules versus time is shown in Figure 3 for ligand (Q), messenger RNA's (Mi), and activated transcription factors (A, CC, DD, FF, GG, KK, EQ).

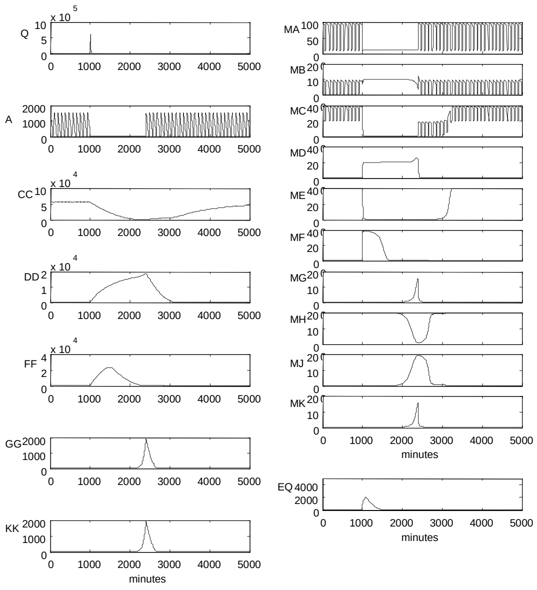


Figure 4: Time courses in mRNA expression levels* (MA, MB, MD, MD, ME, MF. MG. MG, MJ, MK) and active transcription factors (A, CC, DD, FF, GG, KK, EQ) activity in response to a pulse of ligand (Q) at time 1000 minutes. (*For a larger version, please see Appendix)

It becomes apparent how it may be possible to cluster genes according to similarity in expression profile in order to determine *cis* regulatory regions when comparing the time courses of the different genes . This is shown in Figure 5, where genes in each group share a *cis* regulatory region as well as characteristics in the expression profiles. Genes that appear in more than one group are regulated by more than one *cis* regulatory region in common with the others.

These model results should not be over-interpreted, however. The obvious similarities in expression profiles of genes sharing *cis* regulatory regions results primarily from the assumption in the model that simultaneous transcription factor binding to multiple *cis* sites on the same gene have linearly additive effects. This is likely not true in all cases. Also, the resolution of data that would be required to identify the similarities in expression profiles is possibly beyond what would be experimentally feasible. It may be possible to determine which genes share *cis* regulatory regions by perturbing the system with a more complex input than a pulse of ligand, however.

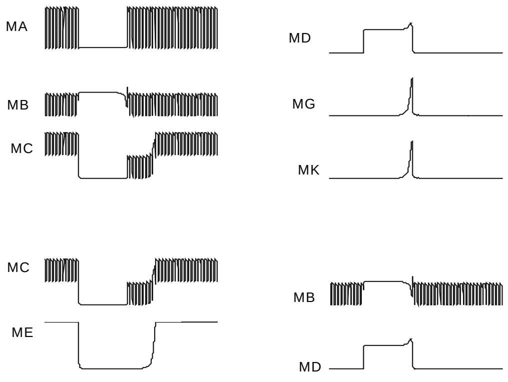


Figure 5: Groups of co-regulated genes from the model

5. RESULTS

In this work, several approaches to identifying genetic regulatory networks were investigated. As described above, the rate of sampling may influence the network that is identified. For this reason, two data sets derived from the data shown in Figure 4 were used. In one, only 12 data points were collected over the 4000 minutes after the injection of ligand. In the other, 120 measurements were taken. The effect of this discretization on the character of the data can be seen below in Figure 6, especially in the time courses of A and MA.

In this section, the results from the genetic regulatory network identification methods that only used cDNA array data are presented first, followed by results from methods that used both cDNA array data and regulatory activity data.

Before proceeding it must be mentioned that it was impossible to distinguish between the genes G and K in any sense. Since they both have the exact same transcription profile, it was impossible to determine if they are regulated by different transcription factors or if they separately regulate other genes through their gene products. To remedy this, G and K, as well as GG and KK were lumped together into a single "super-gene" (G/K) that was negatively regulated by CC and whose dimerized gene product (GG/KK) both negatively regulated H and positively regulated J.

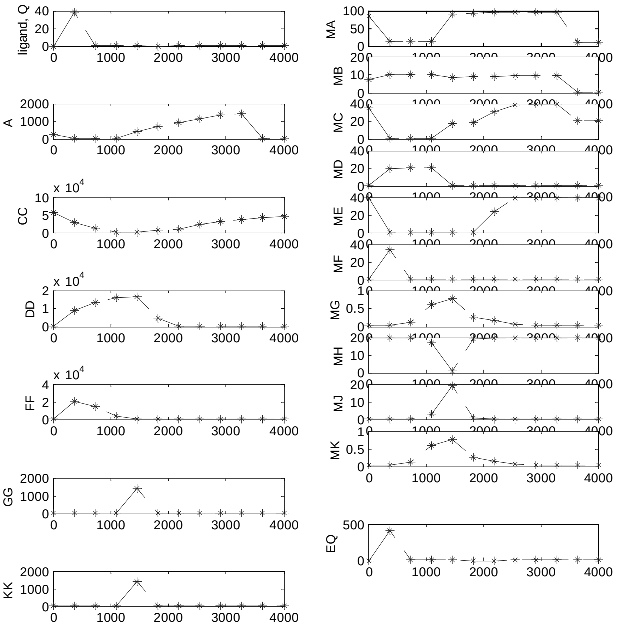


Figure 6: Discretized (12 points) time course after the injection of ligand into model system

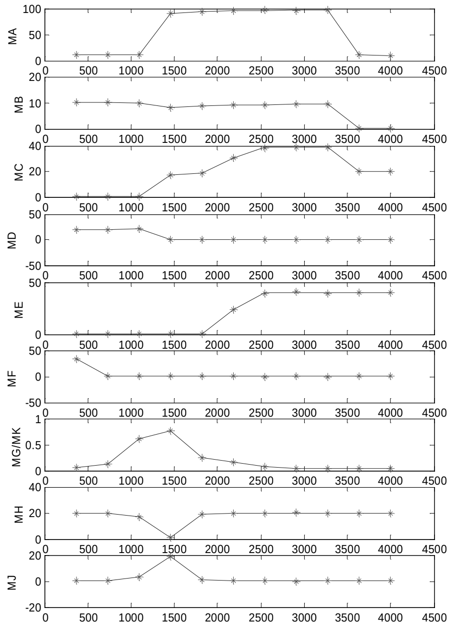


Figure 7: Fit of linear cDNA array model to 12 point data

As can be seen from Figures 7 and 8, the linear identification method that only used cDNA array data, with its 110 parameters, can fit the data very well. Note that these are local fits, where the linear model was used to predict the expression levels for only one step forward in the future. The good fits are not indicative of its ability to identify an underlying structure, however. The other methods were also unable to identify a consistent structure. Shown below in Table 2 are the interdependencies between the genes identified using the linear, linear-log, and linear + squashing methods for the 12 and 120 point data sets only in terms of the sign of the identified coefficient. If the methods identified a consistent structure, it would be expected that the signs of the interactions would be the same. As is seen, there is no consistent structure. The genetic networks identified from cDNA array data alone, for this simple model system, are dependent on both sampling frequency and the method chosen. The linear model therefore seems primarily useful as a data fitting method, rather than a network identification method.

Given that the genetic regulatory network for the model system is already known, it was of interest to see if it is possible to determine it without the prior knowledge of which transcription factor influences the transcription rate of which gene. A study similar to that performed with the identification methods that only used cDNA data, with application of the linear, linear–log, and linear plus squashing approaches, was performed using to both cDNA and regulatory activity data. The resulting interaction matrices that were derived were directly compared to the known network structure (Table 1). The results are shown below in Table 3.

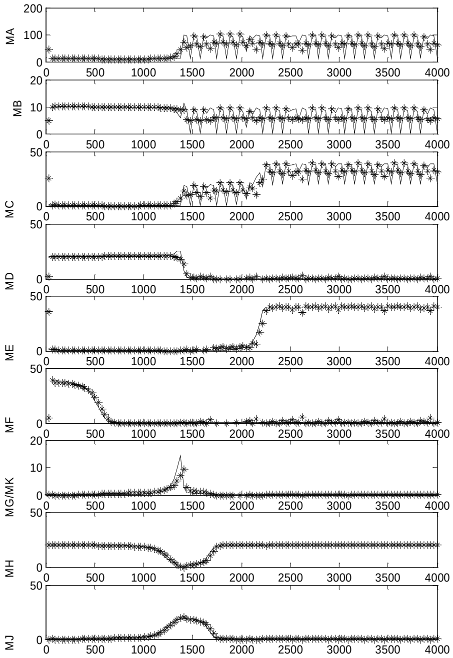


Figure 8: Fit of linear cDNA array model to 120 point data

Table 2: Signs of coefficients of interaction between genes for cDNA array alone, for linear, linear–log, and linear+squashing methods for 12 and 120 point data sets

12 SAMPLES

LINEAR	MA	MB	MC	MD	ME	MF	MG/K	MH	MJ	Q
MA	-	-	+	+	-	-	-	+	+	+
MB	-	-	+	+	-	-	-	+	+	+
MC	-	-	+	+	-	+	-	+	+	-
MD	+	-	-	+	+	+	-	-	-	-
ME	+	+	-	-	+	+	+	-	-	-
MF	+	-	-	+	+	-	-	-	-	+
MG/MK	-	-	+	+	-	-	-	+	+	+
MH	+	+	-	-	+	+	+	+	+	-
MJ	-	-	+	+	-	-	-	+	+	+

120 SAMPLES

	MA	MB	MC	MD	ME	MF	MG/K	MH	MJ	Q
MA	-	+	-	-	+	+	+	+	+	0
MB	-	+	-	-	+	+	+	+	+	0
MC	-	+	-	-	+	+	+	+	+	0
MD	+	-	-	+	+	-	-	-	-	0
ME	-	+	+	-	+	-	+	+	+	0
MF	+	-	-	+	+	+	-	-	-	0
MG/MK	+	+	-	-	+	-	+	-	+	0
MH	-	-	+	+	-	+	-	+	+	0
MJ	+	+	-	-	+	-	+	-	+	0

LINEAR-LOG	MA	MB	MC	MD	ME	MF	MG/K	MH	MJ	Q
MA	-	+	+	+	-	+	-	+	+	-
MB	-	+	+	+	-	+	-	+	+	-
MC	+	-	-	+	+	-	+	-	-	+
MD	-	+	+	-	-	+	-	+	+	-
ME	+	-	-	+	+	-	+	-	-	+
MF	-	+	+	-	-	+	-	+	+	-
MG/MK	-	+	+	+	-	-	-	+	+	-
MH	+	-	+	+	-	-	-	+	+	+
MJ	-	+	+	+	-	-	-	+	+	-

	MA	MB	MC	MD	ME	MF	MG/K	MH	MJ	Q
MA	+	-	-	+	+	+	-	+	-	-
MB	-	-	-	+	+	+	-	+	+	-
MC	+	-	-	-	+	-	+	-	+	+
MD	-	+	-	+	-	+	-	+	+	-
ME	+	-	+	-	+	+	+	+	+	+
MF	-	+	-	-	-	+	-	-	-	-
MG/MK	-	+	-	-	+	+	+	+	+	-
MH	+	-	+	-	-	-	-	+	+	+
MJ	-	+	-	+	+	-	-	-	+	-

LINEAR+SQUASHING	MA	MB	MC	MD	ME	MF	MG/K	MH	MJ	Q
MA	+	+	-	-	+	+	+	-	-	-
MB	+	+	-	-	+	+	+	-	-	-
MC	+	+	-	-	+	+	+	-	-	-
MD	-	+	-	-	+	-	+	+	+	+
ME	-	-	+	+	-	-	-	+	+	+
MF	-	+	+	-	-	+	+	+	+	-
MG/MK	+	+	-	-	+	+	+	-	-	-
MH	-	-	+	+	-	-	-	+	+	+
MJ	+	+	-	-	+	+	+	-	-	-

	MA	MB	MC	MD	ME	MF	MG/K	MH	MJ	Q
MA	-	+	-	-	+	-	+	-	-	0
MB	+	-	+	+	-	-	-	-	-	0
MC	+	+	-	-	+	-	+	-	-	0
MD	-	+	+	-	-	+	+	+	+	0
ME	+	-	+	+	-	-	-	+	+	0
MF	-	+	+	-	-	-	+	+	+	0
MG/MK	-	-	+	+	-	+	-	+	+	0
MH	-	+	+	-	-	-	+	-	+	0
MJ	+	+	-	+	+	+	-	+	-	0

Table 3: Signs of coefficients of interaction between transcription factor and gene expression. Shaded boxes indicate matches with actual network. Percentages of maximum coefficient are shown for methods with best match.

12 SAMPLES

LINEAR	A	CC	DD	FF	GG/KK	EQ
MA	+	-	+	-	-	+
MB	+	+	+	+	-	+
MC	+	0	+	-	-	+
MD	-	+	-	+	+	-
ME	+	+	+	-	-	-
MF	-	+	+	-	+	-
MG/MK	0	0	0	0	-	-
MH	+	+	+	+	+	-
MJ	-	-	+	-	-	+

120 SAMPLES

	A	CC	DD	FF	GG/KK	EQ
MA	91.8	5.1	5.5	-6.1	60.8	-100.0
MB	4.4	0.7	2.0	0.0	-11.6	0.7
MC	26.6	2.4	0.7	-2.4	11.9	-60.1
MD	-7.8	0.3	4.1	0.7	-37.9	22.2
ME	19.5	3.4	-0.3	-2.7	-7.2	-78.5
MF	-3.4	0.3	-0.7	2.0	13.0	69.6
MG/MK	-1.0	0.0	0.7	-0.3	0.3	0.3
MH	22.2	1.4	3.1	0.3	-47.4	-10.6
MJ	-1.0	0.0	1.7	-1.0	29.4	1.7

LINEAR-LOG	A	CC	DD	FF	GG/KK	EQ
MA	+	+	+	+	+	-
MB	+	-	+	+	-	-
MC	+	-	+	-	-	-
MD	+	+	-	+	+	+
ME	+	-	+	-	-	-
MF	+	-	-	-	-	+
MG/MK	+	-	+	+	+	-
MH	+	+	+	-	+	+
MJ	+	-	+	+	+	-

	A	CC	DD	FF	GG/KK	EQ
MA	-17.6	72.6	0.7	-19.1	31.4	-5.1
MB	-46.9	40.4	0.8	8.6	14.7	-3.9
MC	-10.2	50.0	-33.3	-40.6	35.0	2.2
MD	-3.0	-11.4	26.4	53.6	5.5	0.6
ME	1.7	33.9	-52.1	-12.8	-32.9	-9.5
MF	1.7	36.9	-18.0	-22.5	45.0	100.0
MG/MK	-2.1	-73.5	15.1	15.2	15.4	-19.4
MH	0.1	23.6	35.2	0.5	-86.7	-17.0
MJ	-2.5	-62.6	26.6	3.9	87.7	-14.4

LINEAR+SQUASHING	A	CC	DD	FF	GG/KK	EQ
MA	-	+	-	+	+	-
MB	+	-	-	-	-	-
MC	-	+	0	+	+	-
MD	+	0	+	-	-	+
ME	-	0	+	0	+	+
MF	+	-	+	+	-	+
MG/MK	+	0	-	+	+	-
MH	-	-	+	-	-	+
MJ	+	+	-	+	+	-

	A	CC	DD	FF	GG/KK	EQ
MA	0	-	0	+	-	+
MB	+	0	-	-	-	-
MC	+	-	+	+	-	+
MD	+	+	-	-	+	-
ME	+	-	+	+	-	+
MF	+	+	+	-	-	-
MG/MK	+	+	+	+	-	-
MH	-	-	0	-	+	+
MJ	+	+	0	+	-	-

As can be seen in Table 3, the linear and linear–log methods performed about equally well, with the networks determined from the 120 point data set coming closest to the actual network structure. Matches with actual structure are shaded, percentages of maximum coefficient are shown for two best methods. While the linear and linear –log methods correctly identified ¾ of the actual interactions, it must be noted that the methods also identified a large number of false positives. The coefficients for many of the false positives were often large, near 100% of the maximum coefficient for some.

For the final study, it was assumed that the cis regulatory regions upstream of the gene coding regions, and the transcription factors that bound to the regulatory regions, were known. The task remained to determine the parameters in the matrix shown below in Table 4.

Table 4: Parameters to be determined given cDNA array data, regulatory activity data, and knowledge of which cis region is upstream of which coding region

	A	CC	DD	FF	GG	KK	EQ
MA	$a_{a\text{-}a}$						
MB	$a_{a\text{-}b}$			$a_{ff\text{-}b}$			
MC	$a_{a\text{-}c}$		$a_{dd\text{-}c}$				
MD		$a_{cc\text{-}d}$		$a_{ff\text{-}d}$			
ME			$a_{dd\text{-}e}$				
MF							$a_{eq\text{-}f}$
MG		$a_{cc\text{-}g}$					
MH					$a_{gg\text{-}h}$		
MJ						$a_{kk\text{-}j}$	
MK		$a_{cc\text{-}k}$					

The parameters were determined using simple linear regression with the linear, linear–log, and linear+squashing techniques for both data sets. The results are shown in Table 5. The coefficients with the correct sign are in bold.

From Table 5 it can be seen that the linear-log method most successfully identified the signs of the interaction coefficients, incorrectly identifying only one for the 12 sample data set, and correctly identifying all for the 120 point data set.

6. CONCLUSIONS

Through use of the model genetic regulatory network, several conclusions can be drawn regarding the genetic regulatory network identification methods considered. It was observed that for this system, the networks derived from only cDNA array data constitute little more than curve fitting. This is because the structure identified varies considerably with the sampling rate and the method of identification. Similar conclusions can be drawn from the study where cDNA array and regulatory activity data were used without knowledge of which cis regulatory site influences the transcription which gene. While two methods could identify correctly the majority of the signs of the interactions, there were a significant number of false positives. For the case that did use knowledge of which *cis* regulatory site influences the transcription of which gene as well as the combined cDNA array and regulatory activity data, it was possible to determine the correct genetic regulatory network structure. The results were still strongly dependent on the method used, but less dependent on the number of samples used. The linear-log method performed the best while the linear + squashing method performed the worst. Overall the results make a case for the use of as much information as possible for the determination of the regulatory network structure, as well as the use of realistic simulated data to evaluate the network inference methods.

Table 5: Gene-regulatory region interaction parameters determined with cDNA array and regulatory activity data, as well as genomic sequence available. Coefficients with the correct sign are in bold.

	measurements	**12**	**12**	**12**	**120**	**120**	**120**
parameter	**correct sign**	**Linear**	**Lin-Log**	**Lin-Squ**	**Linear**	**Lin-Log**	**Lin-Squ**
a_{a-a}	+	**6.7E-02**	**5.7E-01**	-6.0E-03	**6.4E-02**	**5.1E-01**	-1.8E-03
a_{a-b}	+	**6.6E-03**	**5.3E-01**	-1.1E-03	**5.7E-03**	**1.3E-01**	**6.1E-04**
a_{ff-b}	+	**6.0E-04**	**3.0E-01**	-3.0E-04	**6.0E-04**	**1.1E-01**	-1.3E-04
a_{a-c}	+	**2.7E-02**	**4.1E-01**	-5.9E-03	**2.5E-02**	**4.2E-01**	-2.0E-03
a_{dd-c}	-	5.0E-04	4.7E-02	2.0E-04	0.0E+00	**-8.5E-02**	3.0E-04
a_{cc-d}	-	1.4E-04	**-1.5E-02**	9.0E-05	0.0E+00	**-4.1E-03**	1.5E-04
a_{ff-d}	+	**9.1E-01**	**2.4E-01**	-6.0E-04	**1.2E-03**	**3.3E-01**	-3.0E-04
a_{dd-e}	-	1.5E-04	**-7.2E-02**	3.3E-04	2.3E-05	**-1.7E-01**	3.8E-04
a_{eq-f}	+	**3.9E-04**	**1.4E+00**	**1.3E-02**	**2.6E-02**	**1.2E+00**	-4.1E-03
a_{cc-g}	-	2.6E-06	**-2.4E-01**	6.2E-05	2.1E-06	**-2.4E-01**	1.6E-04
a_{gg-h}	-	1.4E-02	**-1.6E-01**	**-2.2E-03**	2.0E-03	**-2.7E-01**	2.0E-03
a_{kk-j}	+	**7.6E-04**	**5.1E-01**	**2.0E-03**	**1.6E-02**	**8.1E-01**	-4.4E-03
a_{cc-k}	-	2.6E-06	**-2.4E-01**	6.2E-05	2.1E-06	**-2.4E-01**	1.6E-04

7. FUTURE WORK

While the linear–log method used with both the combined cDNA array and regulatory activity data and knowledge of which *cis* regulatory sight influences the transcription of which gene was able to determine the regulatory network structure, significant challenges remain. It needs to be shown whether the methods that failed in this study to produce the correct regulatory network still will fail if a different, more complex, ligand input sequence is used. The effects of noise on the reliability of these methods needs also to be investigated. This will be accomplished by performing stochastic simulations of the model network, to introduce noise that may be fundamental to gene expression. The effect of including potential errors that may result from experimental methods needs also to be considered. The fact that cDNA array and genetic regulatory data is often obtained from populations of cells, rather than single cells, must be investigated. Since different cells may be in different phases of their endogenous oscillations, the effects of treating a population of asynchronous cells with ligand at the same time needs to be considered. The averaging over an asynchronous population that occurs during real experiments may significantly alter the information that may be obtained from the data. These directions for future work will be built upon the model genetic network presented here.

8. ACKNOWLEDGMENTS

DEZ thanks NIH for funding (Training grant NIAAA 5 T32 AA07463-15).

9. REFERENCES

[1] Barkai, N. and Leibler, S. (2000) Circadian clocks limited by noise. *Nature* **403**: 267-268

[2] Cherry, J. L., Adler, F. R. (2000) How to make a biological switch. *Journal of Theoretical Biology* **203**: 117-133.

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

[4] Gardner, T.S., Cantor, C.R. and Collins J. J. (2000) Construction of a Genetic Toggle Switch in *Escherichia coli*. *Nature* **403**: 339-342.

[5] Gombert A.K and Nielsen J. (2000) Mathematical Modelling of Metabolism. *Current Opinion in Biotechnology* **11**: 180-186.

[6] Gygi, S.P., Corthals, G.L., Zhang, Y., Rochon, Y., and Aebersold, R. (2000) Evaluation of Two Dimensional Gel Electrophoresis-Based Proteome Analysis Technology. *Proc.Natl.Acad. Sci. USA*. **97**: 9390-9395.

[7] Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995) Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science* **270**: 467-470.

[8] Szallasi Z., and Liang S. (1998) Modeling the Normal and Neoplastic Cell Cycle with Realistic Boolean Genetic Networks. *Pacific Symposium on Biocomputing* **3**: 66-76.

[9] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, and Church GM. (1999) Systematic determination of genetic network architecture. *Nature Genetics* **22**:281-5.

[10] Weaver, D.C., Workman, C.T., and Stormo, G.D. (1999) Modeling Regulatory Networks with Weight Matrices. *Pacific Symposium on Biocomputing* **4**: 102-111.

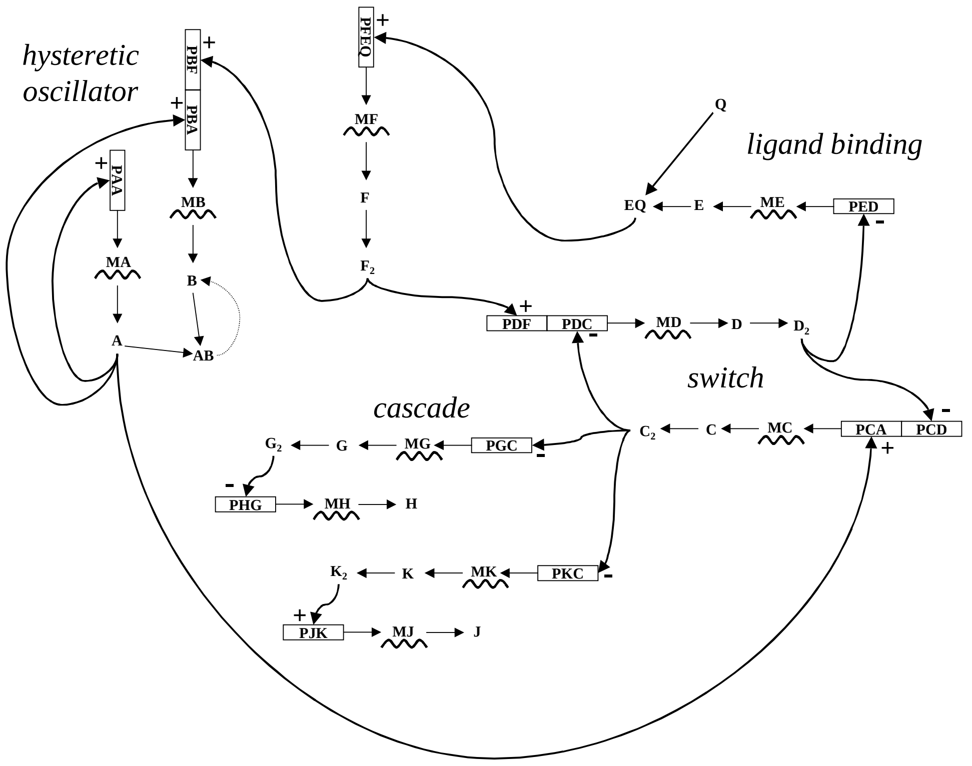
[11] Wessels L. F. A., Van Someren, E. P., Reinders, M. J. T. (2001), A comparison of genetic network models. *Pacific Symposium on Biocomputing* **6**, 508-519.

APPENDIX

A.1 GENETIC NETWORK MODEL

A.2 TIME COURSE IN mRNA EXPRESSION LEVELS AND ACTIVE TRANSCRIPTION FACTORS

A.1 GENETIC NETWORK MODEL



A.2 TIME COURSE IN mRNA EXPRESSION LEVELS AND ACTIVE TRANSCRIPTION FACTORS