18

Genetic Regulatory Networks and Co-Regulation of Genes: A Dynamic Model Based Approach

Ashoka D. Polpitiya[1], J. Perren Cobb[1], and Bijoy K. Ghosh[2]

[1] Cellular Injury and Adaptation Laboratory, Department of Surgery, Washington University, St.Louis, MO 63110
ashoka@cia.wustl.edu

[2] Center for BioCybernetics and Intelligent Systems, Department of Electrical and Systems Engineering, Washington University, St.Louis, MO 63130

Summary. This paper describes a novel approach to obtain co-regulated genes and the underlying genetic regulatory network. Time evolution of gene expression levels is described using a discrete time classical state space dynamical system. The state transition matrix of the model can be used to obtain the adjacency matrix of the regulatory network. Gene expression values over time can also be written as a linear combination of the eigenmodes of the state transition matrix. Based on the relative participation of the eigenmodes, co-regulation is discussed.

Keywords: genes, microarrays, regulatory networks, dynamic models, reverse engineering, eigenvalues, co-regulation, Karhunen-Loève decomposition

18.1 Introduction

The recent advances in microarray techniques [8],[11] have enabled the measurement of expression values of many or all of an organism's genes. This has resulted in a large abundance of data. But the computational techniques available to handle this large-scale data are only a handful. Some challenges in analyzing this data are noisy measurements, stochastic nature of the data and the sheer number of genes involved [9]. There are some other inherent problems with the data itself too. Low time resolution and fewer replicates being the most significant accounting to the costs involved. Until one achieves the "lab-on-a-chip" technology, the techniques should be fine tuned for the available sparse and under-determined data in order to reveal some knowledge about the underlying regulatory interactions. The question of inferring the causal connectivity of the genes, i.e. genetic regulatory networks, using these expression profiles has been addressed by number of authors [6],[14],[4],[7]. But little attention is given to explore the characteristics of the models and the resulting networks.

A.D. Polpitiya et al.: *Genetic Regulatory Networks and Co-Regulation of Genes: A Dynamic Model Based Approach*, LNCIS **321**, 291–304 (2005)
www.springerlink.com

In this report, we first use Karhunen-Loève decomposition to explore the temporal behavior of the sparse time series data from microarrays. Then, we will discuss genetic regulatory networks emerging from the time series data. The idea behind the method is the use of a state-space model ([6], [10]). An algorithm to obtain the model parameters is outlined based on the minimum-norm least square method ([2]). Finally, a novel approach to look at genes that are "co-regulated" is being discussed based on the characteristics of the state-space model.

There are several papers where a simple interpolation scheme is used to obtain a continues representation of time series expression data (see [1], [4]). Here, a piecewise cubic interpolation was done on the data to obtain a finely interpolated time series. The underlying interpolating function $F(t)$, which is C^1 (i.e. continuous in the first derivative) was obtained for each interval of time $t_k < t < t_{k+1}$ and the slopes at t_k were chosen such that $F(t)$ preserves the shape of the data and respects monotonicity. This means that, on intervals where the data are monotonic, so is $F(t)$; at points where the data has a local extremum, so does $F(t)$ (see Figure 18.1).

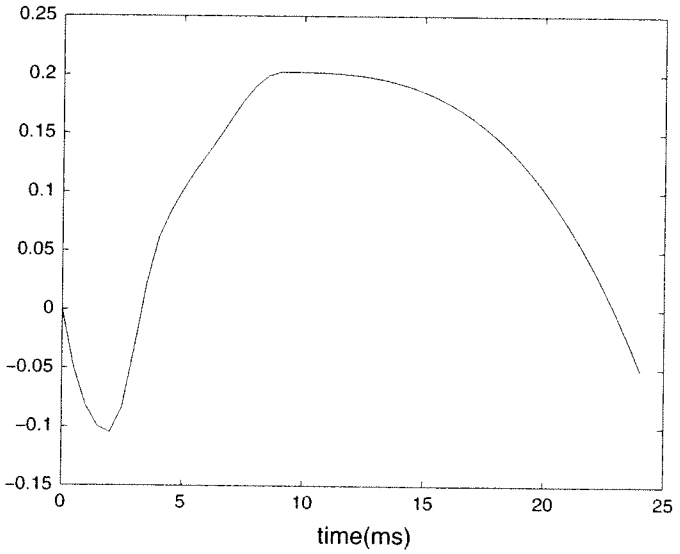

Fig. 18.1. Piecewise cubic interpolated data.

18.2 Temporal Patterns in Gene Expression Data

In order to visualize the temporal variation, the data was projected onto a smaller dimensional space using a series expansion method, the Karhunen-Loève decomposition (KLD) [13].

18.2.1 Series Representation of Gene Expression Data

Gene expression data over time is treated as sample functions of a random process. Let $\mathbf{x}^j(k)$ denote the $N \times 1$ dimensional expression vector at time point k, for the volunteer j. Therefore at kth time point, the expression vector can be viewed as

$$\mathbf{x}^j(k) = [x_1^j(k)\ x_2^j(k)\ \ldots\ x_N^j(k)]^T,$$

where $x_i^j(k)$ is the expression value of the ith gene at time k for volunteer j. The idea is to find a global basis for $\mathbb{R}^N$: ψ_1, ψ_2, ..., ψ_N and to expand $\mathbf{x}^j(k)$ as

$$\mathbf{x}^j(k) = \lim_{N\to\infty} \sum_{i=1}^{N} \alpha_i^j(k)\psi_i$$

where

$$\alpha_i^j(k) = \langle \mathbf{x}^j(k), \psi_i \rangle$$

and $\langle \cdot, \cdot \rangle$ stands for the standard inner product notation.

The orthonormal basis for KLD is selected as the eigenvectors of the average correlation matrix $C_1 \in \mathbb{R}^{N\times N}$, where

$$C_1 = \frac{1}{N_T M} \sum_{k=1}^{N_T} \sum_{j=1}^{M} (\mathbf{x}^j(k))(\mathbf{x}^j(k))^T, \tag{18.1}$$

and N_T is the total number of time samples. The matrix C_1 is symmetric and positive semidefinite, so its eigenvalues are all real and non-negative and the corresponding eigenvectors are orthonormal and forms the global basis for $\mathbb{R}^N$. The eigenvectors corresponding to the largest p eigenvalues of C_1 are called the principal eigenvectors (modes) and the pth order successive reconstruction $\hat{\mathbf{x}}^j(k)$ of the expression value $\mathbf{x}^j(k)$ is given by

$$\hat{\mathbf{x}}^j(k) = \sum_{i=1}^{p} \alpha_i^j(k)\psi_i.$$

It is observed that the first three principal modes capture 97% of the energy content in the expression values (see Figure 18.2). An interesting feature that one can clearly see from Figure 18.2 is that the volunteers treated with the toxic drug have a cyclic response over the 24hrs which correlates well with the fact that they all recover and become normal in 24hrs, and the ones who were given only Saline do not show much variation over time.

18.2.2 Classifying Time Profiles : 2nd KLD

The coefficients $\alpha_i^j(k)$ are uncorrelated with respect to i. The vector function

$$[\alpha_1^j(k)\ \alpha_2^j(k)\ \ldots\ \alpha_p^j(k)]$$

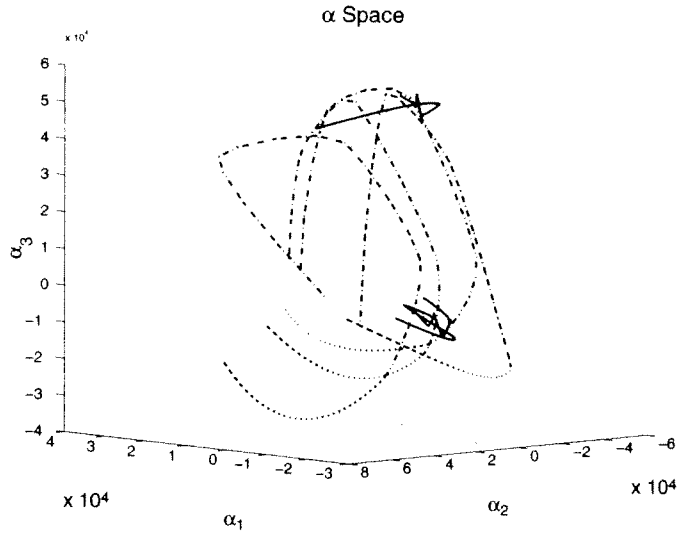

Fig. 18.2. "α-Space": α coefficients for three principal modes of seven human volunteer gene expression data for 5150 genes. Dotted lines correspond to four volunteers treated with a toxic drug (the dosage was administered such that they recover in 24 hrs) and the thick lines correspond to the three volunteers who only had Saline.

can be viewed as a sample function of a vector random process and a 2nd KLD can be performed. Let

$$\zeta_i^j = [\alpha_i^j(1)\ \alpha_i^j(2)\ \ldots\ \alpha_i^j(N_T)], \qquad \xi^j = [\zeta_1^j\ \zeta_2^j\ \ldots\ \zeta_p^j],$$

and it captures the p^{th} order α-space representation of j^{th} data set (i.e., j^{th} volunteer). The covariance matrix $C_2 \in \mathbb{R}^{(N_T \times p)\times(N_T \times p)}$ as in (18.1), will be

$$C_2 = \frac{1}{M}\sum_{j=1}^{M}(\boldsymbol{\xi}^j)^T(\boldsymbol{\xi}^j). \tag{18.2}$$

The q^{th} order successive reconstruction $\hat{\xi}^j$ of the j^{th} vector ξ^j is given by

$$\hat{\xi}^j = \sum_{i=1}^{q}\beta_i^j\varphi_i$$

where φ_i, $i = 1, 2, \ldots, q$ are the eigenvectors corresponding to the largest q eigenvalues of the matrix C_2 and the coefficients β_i^j are obtained by

$$\beta_i^j = \left\langle \xi^j, \varphi_i \right\rangle.$$

Figure 18.3 shows the plot of the "β-space" with the first three principal modes.

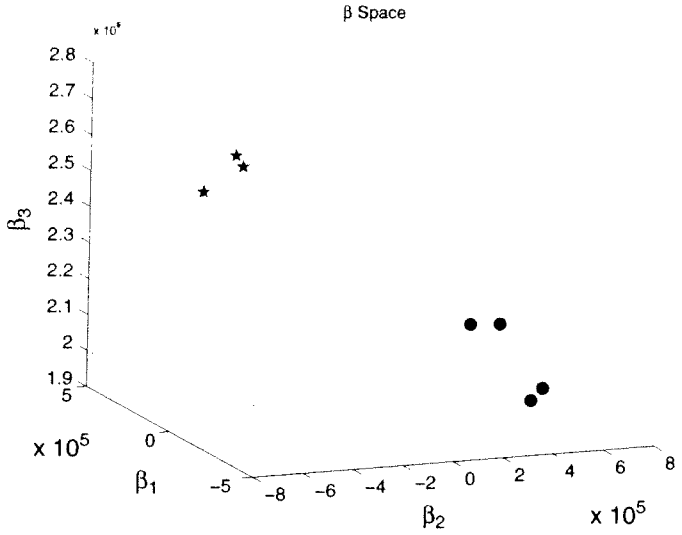

Fig. 18.3. "β-Space": β coefficients for three principal modes. Circles correspond to toxin treated volunteers and the stars correspond to Saline treated.

18.3 Dynamic Model of the Genetic Network

The idea of a dynamic model of the genetic network is that the expression value from one time point determines the expression value seen at the next time point. A complete realistic model should take into account various issues biologically relevant such as, for example, the inherent nonlinearities of biochemical reactions, internal and external noise, etc. But given the nature of the sparse time data available from microarrays, a simple linear time invariant (LTI) discrete time model treating the biological system as a simple state machine, would be a good starting point.

The general form of a discrete time LTI system is as follows:

$$\begin{aligned} \mathbf{x}(k+1) &= A\mathbf{x}(k) + B\mathbf{u}(k) \\ \mathbf{y}(k) &= C\mathbf{x}(k) + D\mathbf{u}(k) \end{aligned}$$

where A, B, C and D are maps between suitable spaces (i.e. matrices of suitable dimensions). u is the *input*, $\mathbf{x}$ is the state variable in an $N-$dimensional state space and finally, y is the *output* of the system. Here we consider the case with no input and $C = Id$. Hence the system becomes

$$\begin{aligned} \mathbf{x}(k+1) &= A\mathbf{x}(k) \\ \mathbf{y}(k) &= \mathbf{x}(k). \end{aligned} \tag{18.3}$$

Figure 18.4 shows a typical network discussed in this paper. Here a_{ij}'s are the elements of the A matrix, i.e., $A = \{a_{ij}\}$. A positive element means that

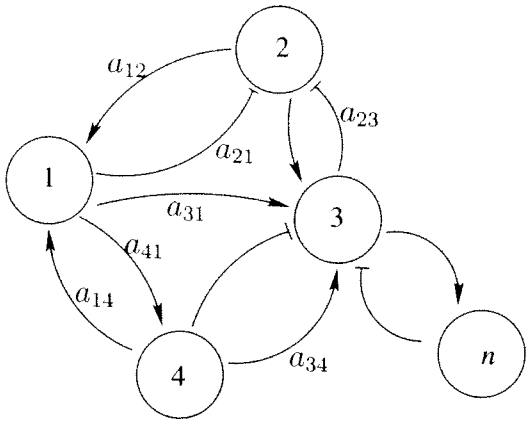

Fig. 18.4. Topology of the genetic regulatory network. Lines with arrow marks indicate positive reinforcing and lines with crossbars indicate negative reinforcing.

the ith gene being positively reinforced by the jth gene expression value at a previous time and vice-versa. In other words, one can write

$$x_i(k+1) = \sum_{j=1}^{N} a_{ij} x_j(k). \tag{18.4}$$

18.4 Reverse Engineering Algorithm for Genetic Networks

Problem 1 (Reverse engineering problem of genetic networks). For N genes, a noisy and sparse mRNA abundance data $\mathbf{y}(k)$ with "sufficiently" large k, is obtained for a known perturbation. Find the matrix A (in turn, find the adjacency matrix of the genetic regulatory network) described in Equation (18.3), such that it fits the data "best".

Fitting a model with $N \times N$ elements with $N \times M$, where $(N >> M)$, data points leads to a highly underdetermined system, i.e. many possible models that fit to the data almost perfectly, exist. However, one can find a optimum solution by imposing an additional constraint on smoothness. This will exclude models that may result erratic behaviors in between data points.

Problem can be formulated in the sense of a minimum norm least square solution[3]. Let

$$X_{k+1} = [\mathbf{x}(2)\ \mathbf{x}(3)\ \ldots\ \mathbf{x}(k+1)]$$
$$X_k = [\mathbf{x}(1)\ \mathbf{x}(2)\ \ldots\ \mathbf{x}(k)].$$

[3]Let's omit the superscript j which denotes each volunteer.

Then the problem becomes one of finding A where

$$X_{k+1} = AX_k$$

while minimizing the L_2 norm of A.

The solution to this problem is given by the well known *Moore-Penrose Pseudoinverse* (see [2]). Thus, A is obtained using

$$A = X_{k+1}{X_k}^{\dagger} \tag{18.5}$$

where

$${X_k}^{\dagger} = \left({X_k}^T X_k\right)^{-1} {X_k}^T \tag{18.6}$$

and $(\cdot)^{\dagger}$ denotes the Moore-Penrose Pseudoinverse. In other words, if

$$B = \sum_{i=1}^{r=rank(B)} \sigma_i u_i v_i^T \quad U = [u_1, \ldots, u_m],\ V = [v_1, \ldots, v_n] \tag{18.7}$$

be the SVD of $B \in \mathbb{R}^{m \times n}$ $(m \geq n)$. Then the pseudoinverse of B is defined as

$$B^{\dagger} = V \Sigma^{\dagger} U^T,$$

where

$$\Sigma^{\dagger} = diag(\sigma_1^{-1}, \ldots, \sigma_r^{-1}, 0, \ldots, 0) \in \mathbb{R}^{n \times m}.$$

In order to obtain the regulatory interactions due to the toxic drug, we consider the average gene expression profiles over time of all the volunteers

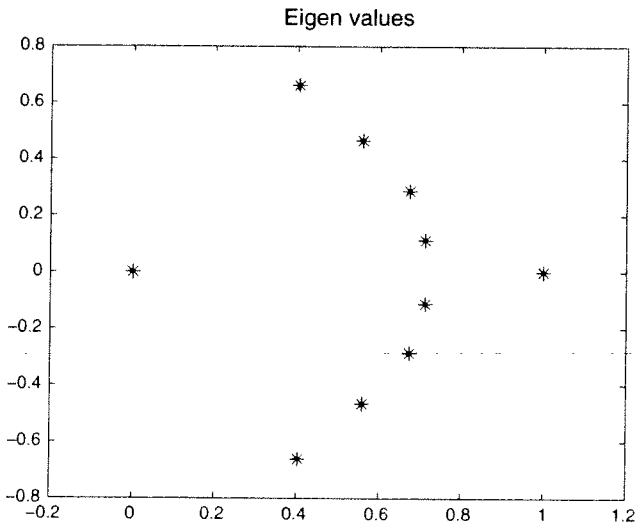

Fig. 18.5. Eigenvalues of the state transition matrix.

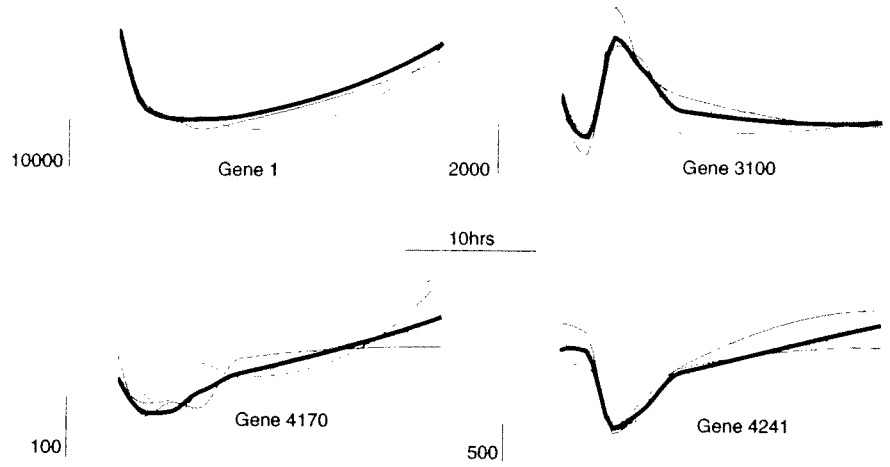

Fig. 18.6. Average gene expression profiles (centroids) for randomly selected four genes. Thick line is the centroid plot and the thin (color) lines are for 4 toxin treated volunteers. Axes are as usual expression value vs. time.

that were treated with the toxin. Figure 18.6 shows few randomly chosen gene profiles and their centroid plots. Figure 18.5 shows the eigenvalues of the resulting A matrix. This will enable one to find the adjacency matrix whose elements are '1', if the corresponding interaction is a positive reinforcement and '-1', for a negative reinforcement. One such network of 220 genes obtained for certain cutoff value and plotted using Cytoscape software[4] [12] is shown in Figure 18.7. The genes shown in light blue are the ones adjacent to NFKB1 (shown in darker green). Genes shown in light green are the ones which are common with the NFKB1 pathway found in Ingenuity©[5] shown in figure 18.8. This shows a significant overlap of gene interactions between what is reported in the literature and our proposed method.

18.5 Eigenmodes and Co-Regulation

Figure 18.5 shows that the state transition matrix of the model described in equation (18.3) has 10 distinct eigenvalues. In fact, the eigenvalue 1 repeats 4 times. Therefore the *Jordan canonical form* (JCF) [3] J of the matrix A has the following form:

$$J = \boldsymbol{diag}(J_0, J_1, \ldots, J_4, J_5) \tag{18.8}$$

[4] http://www.cytoscape.org
[5] http://www.ingenuity.com

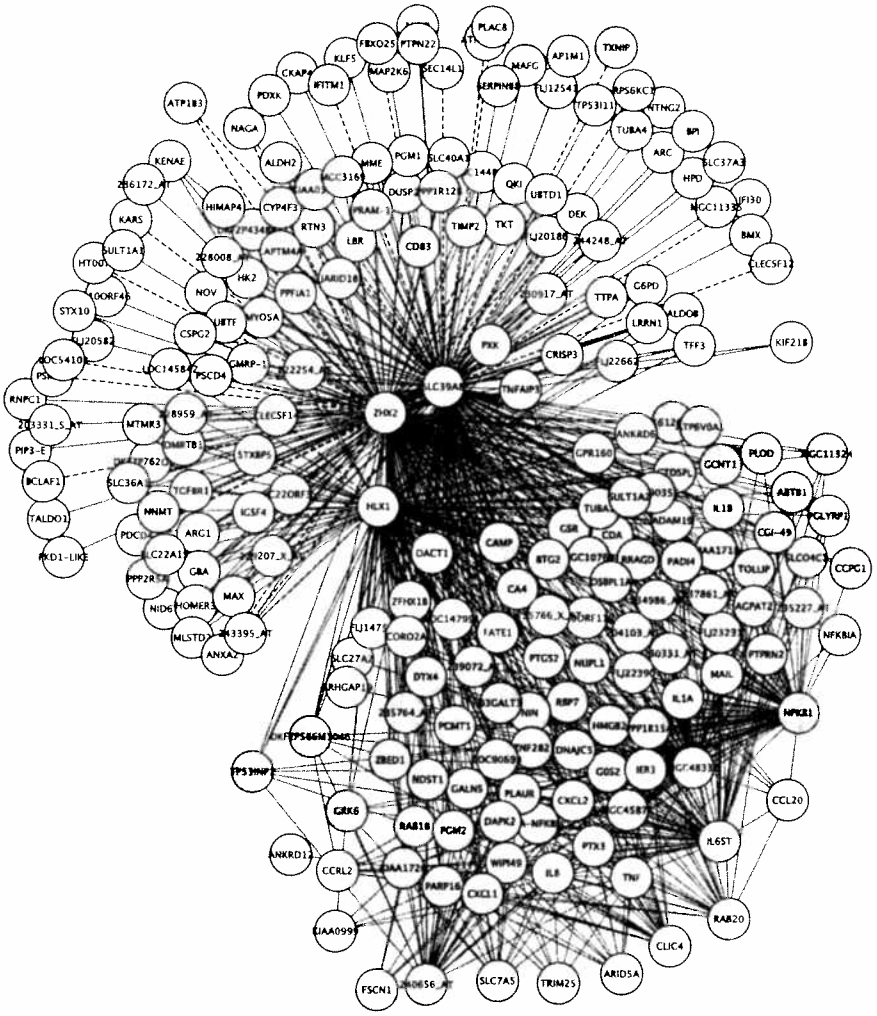

Fig. 18.7. Network of genes obtained from the state transition matrix. Threshold is chosen to be 60% of the maximum coefficient. Each label corresponds to a gene name. A thick edge corresponds to a positively reinforcing causal connectivity and a dotted edge corresponds to one of a negative effect. Shown in green color are the genes that are common with the ones found in the NFKB1 pathway found in Ingenuity© (see Figure 18.8). Plot created using Cytoscape.

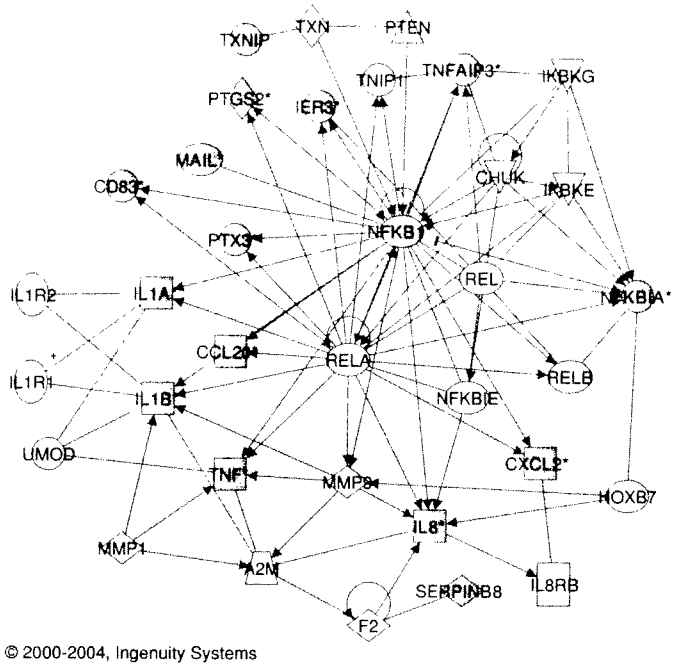

Fig. 18.8. Network obtained using Ingenuity© that shows the NFKB1 pathway. The top functions of this pathway includes immune response, tissue morphology and inflammatory disease.

where $J_0 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$, $J_i = \begin{bmatrix} \sigma_1 & \omega_1 \\ -\omega_1 & \sigma_1 \end{bmatrix}$, $i = 1..., 4$, $\begin{bmatrix} 0 & 1 & \dots & 0 \\ 0 & 0 & 1 & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 \end{bmatrix}$, and $\lambda_i = \sigma_i \pm \omega_i$, $i = 1, \dots, 4$ are the complex eigenvalues. Matrix A can be put in JCF using the similarity transformation

$$J = T^{-1}AT.$$

Now, the solution of the system in equation (18.3) is

$$\mathbf{x}(k) = A^k \mathbf{x}(0)$$

and $A^k = TJ^kT^{-1}$. Due to the block structure, the k-th power of J is obtained by taking the k-th power of each individual Jordan cell. In particular, for a real eigenvalue λ_i,

$$J_i^k = (\lambda_i Id + \delta_{i+1,j})^k = \sum_{n=0}^{k} \binom{k}{n} \lambda_i^{k-n} \delta_{i+n}^j. \tag{18.9}$$

Let $T = [T_0\ T_1\ \ldots\ T_5]$ where T_i are the columns of T associated with ith Jordan block J_i, i.e. $AT_i = T_iJ_i$. Now, with the change of coordinates $\mathbf{x} = T\tilde{\mathbf{x}}$, one can put the system in equation (18.3) into the form $\tilde{\mathbf{x}}(k+1) = J\tilde{\mathbf{x}}(k)$ and further decompose into independent '*Jordan block systems*' $\tilde{\mathbf{x}}_i(k+1) = J_i\tilde{\mathbf{x}}_i(k)$. Then

$$\mathbf{x}(k) = A^k\mathbf{x}(0) = TJ^k\tilde{\mathbf{x}}(0) = \sum_{i=1}^{n} T_iJ_i^k(S_i^T\mathbf{x}(0)) \tag{18.10}$$

where

$$T^{-1} = \begin{bmatrix} S_1^T \\ \vdots \\ S_n \end{bmatrix}.$$

Therefore we can identify the *generalized eigenmodes* of the system from J^k and according to equation (18.10), all solutions $\mathbf{x}(k)$ are linear combinations of these (generalized) modes.

From equations (18.8) and (18.9), we can list all the 12 modes corresponding to eigenvalues $1, \lambda_i = r_ie^{\pm j\theta_i}$, $(i = 1, \ldots, 4)$ as

$$\begin{array}{l} 1, \\ k \text{ terms}, \\ k^2 \text{ terms}, \\ k^3 \text{ terms}, \\ r_i^k\cos(k\theta_i),\ i = 1, \ldots, 4 \\ r_i^k\sin(k\theta_i),\ i = 1, \ldots, 4. \end{array} \tag{18.11}$$

These modes can be treated as a set of basis functions (say, ϕ_j) and the gene expression profiles $x_i(k)$ can be written as,

$$x_i(k) = \sum_{j=1}^{12} \gamma_j\phi_j(k)$$

where γ_j are the corresponding coefficients. Figure 18.9 shows the coefficients of each of the 12 modes for six randomly chosen genes. Based on their relative participation one can cluster the genes. Such a clustering using the K-means algorithm on the coefficients is shown in Figure 18.10 and the corresponding gene expression profiles are shown in Figure 18.11.

18.6 Concluding Remarks

In this paper, we discussed in detail, various analysis approaches one could utilize in analyzing time-series microarray data. Methods are mostly geared toward a systems approach rather than one of a purely statistical in nature. A few interesting observations and ideas will be outlined here.

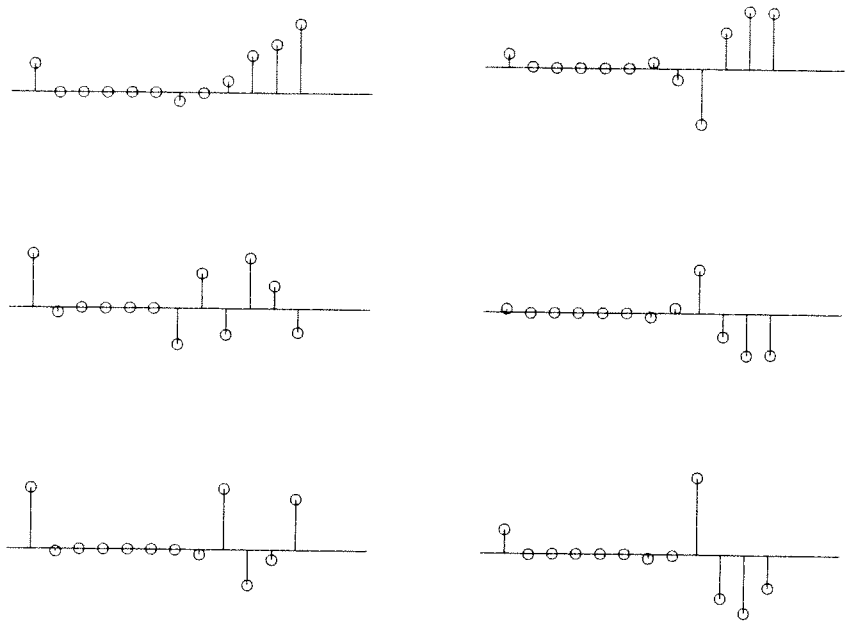

Fig. 18.9. Coefficients showing the relative participation of the 12 eigenmodes for six randomly selected genes. Horizontal axis shows the mode number and the vertical lines are proportional in height to the coefficient value.

The α-space trajectories shown in Figure 18.2 for toxin treated subjects, resemble to 'Homoclinic orbits'[5] found in chaos theory and nonlinear dynamics. Homoclinic orbits occur when the stable manifold and the unstable manifold of a dynamical system intersects. This analogy well explains the idea that the volunteers drift on to an unstable manifold when they get sick and transfers back on to the stable manifold when they become healthy.

The gene regulatory network shown in Figure 18.7 indeed reveals some interesting features. NFKB1 pathway overlaps with the one reported in Ingenuity© well. The top functions of this pathway includes immune response, tissue morphology and inflammatory disease which are all related to the specific toxin response in human volunteers. There are few ESTs (Expressed Sequence Tags) that may worth giving attention.

References

1. J. Aach and G. Church. Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 17(6):495–508, Jun 2001.

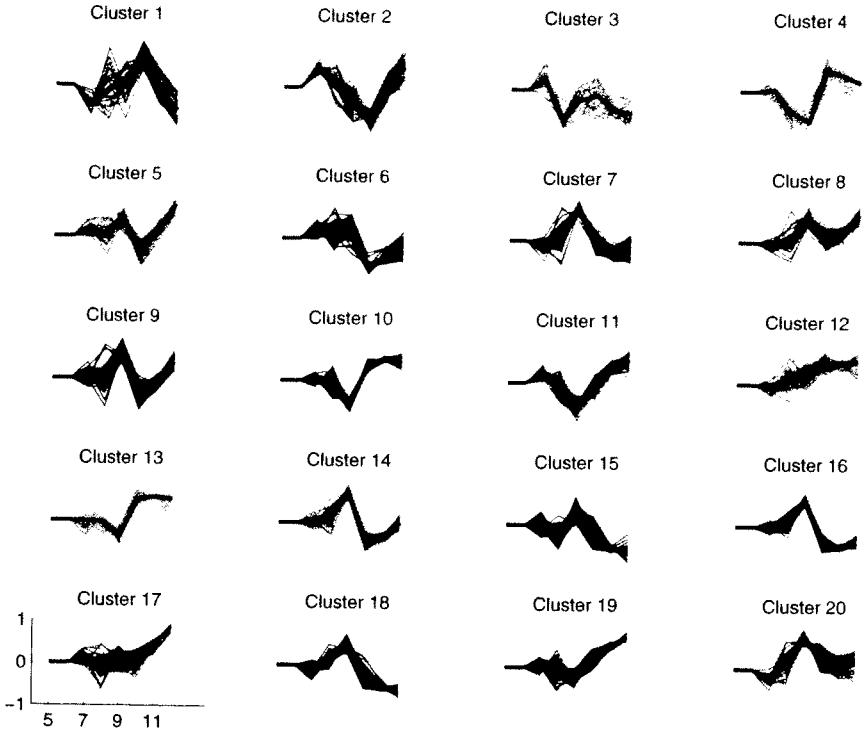

Fig. 18.10. Clusters of genes based on the relative eigenmode participation (only the modes corresponding to complex eigenvalues). Shown here are the coefficients corresponding to gene expression profiles. Horizontal axis shows the mode number (8 oscillatory modes correspond to indices 5-12) and the vertical axis shows the coefficient value.

2. A. Ben-Israel and T. N. E. Greville. *Generalized Inverses: Theory and Applications.* Wiley-Interscience [John Wiley & Sons], New York, 1974. (reprinted by Robert E. Krieger Publishing Co. Inc., Huntington, NY, 1980.).
3. C. T. Chen. *Linear Systems Theory and Design.* Oxford University Press, Oxford, 3 edition, 1999.
4. P. D'Haeseleer, X. Wen, S. Fuhrman, and R. Somogyi. Linear modeling of mrna expression levels during cns development and injury. *Pac Symp Biocomput,* pages 41–52, 1999.
5. J. Guckenheimer and P. Holmes. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields.* Number 42 in Applied Mathematical Sciences. Springer-Verlag, New York, NY, 1983.
6. N. S. Holter, A. Maritan, M. Cieplak, N. V. Fedoroff, and J. R. Banavar. Dynamic modeling of gene expression data. *Proc Natl Acad Sci U S A,* 98(4):1693–1698, Feb 2001.
7. S. Liang, S. Fuhrman, and R. Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac Symp Biocomput.* pages 18–29, 1998.

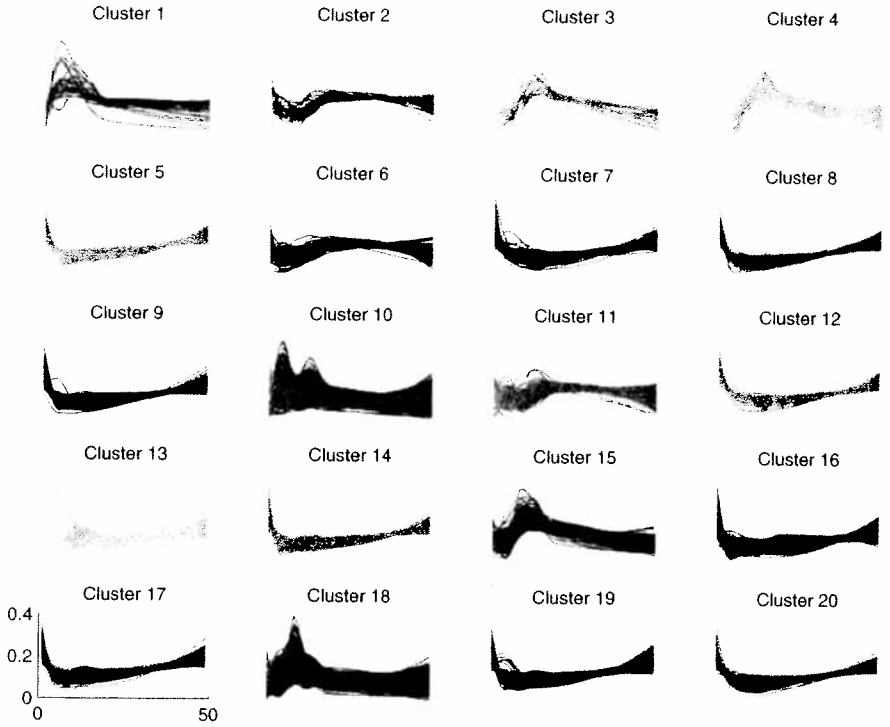

Fig. 18.11. Gene clusters corresponding to the eigenmode clusters shown in Figure 18.10. Axes are expression value vs. time.

8. A. C. Pease, D. Solas, E. J. Sullivan, M. T. Cronin, C. P. Holmes, and S. P. Fodor. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci U S A*, 91(11):5022–5026, May 1994.
9. T. J. Perkins, M. Hallett, and L. Glass. Inferring models of gene expression dynamics. *J Theor Biol*, 230(3):289–299, Oct 2004.
10. C. Rangel, J. Angus, Z. Ghahramani, M. Lioumi, E. Sotheran, A. Gaiba, D. L. Wild, and F. Falciani. Modeling T-cell activation using gene expression profiling and state-space models. *Bioinformatics*, 20(9):1361–1372, Jun 2004.
11. M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, Oct 1995.
12. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504, Nov 2003.
13. H. Van-Trees. *Detection, Estimation, and Modulation Theory. Part I.* John Wiley & Sons, 1968.
14. D. C. Weaver, C. T. Workman, and G. D. Stormo. Modeling regulatory networks with weight matrices. *Pac Symp Biocomput*, pages 112–123, 1999.