Gene expression data analysis: clustering technique using models of transcriptional control of gene networks.

Natalia G. Berloff
Department of Applied Mathematics and Theoretical Physics
University of Cambridge, Wilberforce Road, Cambridge, CB3 0WA, UK
N.G.Berloff@damtp.cam.ac.uk

September 16, 2002

Abstract

An algorithm for clustering gene expression data based on a network of interacting genes is suggested and analysed. Transcriptional controls in gene regulatory networks are modelled by deterministic nonlinear differential equations that take into account nonlinear gene activation and natural degradation of gene product. We suggest a similarity measure between two gene expression profiles based on the underlying biological models.

Keywords: gene expression data, clustering algorithm, transcriptional control, gene regulatory networks.

1 Introduction

As the technologies for DNA expression become more reliable and accessible there is a need for efficient data processing, storing and retrieving information and efficient mathematical analysis of these results. Temporal gene expression patterns are now being obtained for many cell types in response to specific stimuli, or during execution of developmental programs (Wen *et al.*, 1998; Iyer *et al.*, 1999). Such data can help in understanding of how groups of genes control cellular responses to environmental stimuli (Edwards, 1994), and execute stored programs governing the cell cycle. Basset *et al.* (1999) suggested that the greatest intellectual challenges in this area lie in devising ways to extract the full meaning and implications of the data stored in large gene expression libraries. Different statistical and mathematical techniques are being developed and applied to detect the internal structure in the data. Current efforts have focused on identifying underlying patterns in complex data using techniques of clustering points or vectors in multidimensional space, where n points (vectors) in $k-$dimensional space correspond to the quantitative expression level of n genes in k samples. The assumption is that genes with similar expression patterns are likely to be involved in the same regulatory process. But the clustering results can be very different for various measures of similarity that we adopt. The most typically used measures are the Euclidean distance between points or linear correlation coefficient, which is related to the angle between the two $k-$dimensional vectors. Some other distance measures, including rank correlation coefficient and mutual information-based measures, are proposed in D'haeseleer et al. (1998). So far there is no theory how to choose the best similarity measure and there is a vast evidence that different measures produce different clusters.

To achieve a more reliable clustering results, it is necessary to develop a framework for integrating data and gaining insights into the static and dynamic behavior of complex biological systems such as networks of interacting genes. Specific groups of genes may be activated by particular signals and, once activated, regulate a common process and each other's transcription. Such groups are called genetic regulatory systems. Genetic regulatory systems are often activated by signal-transduction pathways in which stimuli lead to second-messenger generation and to activation of transcription factors, often via phosphorylation. Activated transcription factors then bind to DNA sequences known as enhancers and repressors and thereby regulate the transcription of specific nearby genes. Enhancers and repressors affect the transcription of genes for transcription factors, and some transcription factors activate (Jun, myogenin) or repress (Fos) their own transcription. It has become evident (see Smolen *et al.*, 2000 for a review) that nonlinear interactions, positive and negative feedback within signaling pathways, and time delays which may result from mRNA or protein transport, all need to be considered when modeling the operation of genetic regulatory systems.

Two key approximations have been used to model genetic regulatory network: 1) control is exercised at the transcriptional level, and 2) the production of gene product is a continuous process, with the rate determined by the balance of gene activation vs. repression and natural degradation (Rosen, 1968; Smolen *et al.*, 2000).

Modeling gene networks using a system of coupled *nonlinear* differential equations for a purpose of reverse engineering, is a popular way to capture signaling pathways since such models include reasonable (though somewhat simplified) assumptions about the interactions between genes and natural degradation of gene product. But because of their complexity it seemed impractical to use these models for large number of genes directly and it was therefore necessary to construct a coarse-grained description of the system in order to reduce the number of model parameters significantly. Mjolsness *et al.* (1999) used simulated annealing to fit a recurrent neural network with weight decay to four clusters of yeast genes. Wahde and Hertz (1999) used a genetic algorithm to solve reverse engineering problem on four clusters of genes identified in rat CNS development (Wen *et al.*, 1998).

In this paper we shall develop a similarity measure substantiated by the underlying genetic network of a very general form that includes nonlinear effects and natural degradation of gene product. The suggested algorithm can be used to determine simple regulatory signals and to estimate whether the similarity measure is reliable for a given pair of genes. In Section 2 we introduce the nonlinear neural type network that models the transcriptional controls in the system. In Section 3 we consider a simple network of four interacting genes and demonstrate that the linear correlation coefficient (Pearson's r) is a very poor prognosticator of gene interactions. Indeed, in the examples considered the conclusions one would draw by using the linear correlation as measure of similarity are exactly opposite to the actual interactions. We present the details of a new similarity measure based on underlying transcriptional control network in Section 4. We apply this algorithm to rat CNS gene expression data by Wen *et al.* (1998) in Section 5. We conclude in Section 6 with some discussions.

2 Nonlinear model of transcriptional control

Following other authors (Smolen *et al.*, 2000; Tyson and Othmer, 1978; Mjolsness *et al.*, 1991; Reinitz and Sharp, 1995; Wahde and Hertz, 2000) we choose to model gene interactions using a set of coupled nonlinear differential equations.

We shall, therefore, assume that genes regulate one another via the neural network of a general form:

$$\frac{dx_i}{dt} = \lambda_i g\Bigg(F_i(a_{i1}x_1, ..., a_{ij}x_j, ..., a_{in}x_n, b_i) \Bigg) - \tau_i(\mu)x_i, \qquad i = 1, ..., n, \tag{1}$$

where x_i are the expression levels (concentrations of gene product i). $\tau_i(\mu)$ are natural gene product degradation rate functions; in what follows we shall assume that $\tau_i(\mu) = \tau_i$, although vector μ can express hyperbolic or sigmoidal kinetics of gene product degradation. λ_i are the asymptotic maximum expression levels, defined as

$$\lambda_i = \lim_{t \to \infty} x_i(t), \tag{2}$$

when $g \equiv 1$ during the entire time. The matrix $A = [a_{ij}]$ represents the regulatory connection between genes. A positive [negative] value of a_{ij} indicates that the jth gene enhances [represses] the gene i. The parameters b_i correspond to some bias present in the system. The function g is a nonlinear monotonic sigmoidal activation function. In what follows we use

$$g(x) = 1/\Big(1 + \exp(-x)\Big). \tag{3}$$

Different choices of the signaling function produce qualitatively similar results since the correct slope of the signaling function at its inflection point is achieved by rescaling the function F, which represents the mechanism of gene activation (repression).

Under the assumption of cumulative action of the gene products of other genes for activation and repression, we can assume (e.g. Smolen *et al.*, 2000) that

$$F_i = \sum_{j=1}^{n} a_{ij} x_j + b_i. \tag{4}$$

Under the assumption that gene activation (repression) occurs when one of the gene product levels reaches a certain threshold, we can let

$$F_i = \begin{cases} M_i & \text{if } |M_i| > |m_i|, \\ m_i & \text{otherwise,} \end{cases} \tag{5}$$

where $M_i = \max(a_{i1}x_1, ..., a_{ij}x_j, ..., a_{in}x_n)$ and $m_i = \min(a_{i1}x_1, ..., a_{ij}x_j, ..., a_{in}x_n)$. The specific form of function F is not important for the clustering algorithm, but would be crucial for finding genetic regulatory systems via reverse engineering.

So far the model considered is quite general except for the fact that compartmentalization and transport of macromolecules is not included within this simplified model framework. In the analysis and algorithm below we shall also assume that the time interval during which a gene is activated/repressed (the continuous interval on which $0 \ll g(F_i) \ll 1$) is no longer than a typical time interval between gene expression measurements.

3 Linear correlation measure for gene networks

In this section we consider the system (1) with a cumulative action (4) and show that even in a very simple gene regulatory network the linear correlation measure is misleading.

We assume that there are four genes in the system. Gene 3 acts as an enhancer of both Gene 1 and Gene 2 and Gene 4 suppresses the transcription of both Genes 1 and 2. We shall explicitly specify the expressions of control Genes 3 and 4. We assume that Gene 3 expression profile is given by $x_3(t) = \sin t$ and Gene 4 profile is $x_4(t) = \sin((t-1)/2) + 1$ on the interval $[0, 15]$. The expression profiles of Gene 1 and 2 obey (1) with (4), so that

$$\dot{x_1}(t) \;=\; \lambda_1 g\Big(a_{13}x_3(t) + a_{14}x_4(t) + b_1\Big) - \tau_1 x_1(t), \tag{6}$$

$$\dot{x_2}(t) \;=\; \lambda_2 g\Big(a_{23}x_3(t) + a_{24}x_4(t) + b_2\Big) - \tau_2 x_2(t), \tag{7}$$

where the parameters are given in Table 1.

Table 1. The parameters of the gene transcriptional control of (6) -(7). Gene 3 acts as an enhancer of Genes 1 and 2; Gene 4 acts as a repressor of Genes 1 and 2.

i	λ_i	a_{i3}	a_{i4}	b_i	τ_i
1	1	40	-40	-10	0.1
2	4	40	-40	-10	0.9

We integrate system (6)-(7) forward in time with initial conditions $x_1(t=0) = 0.8$ and $x_2(t=0) = 0.1$. Following Mohamad and Gopalsamy (2000) we discretized (1) so that it has the same equilibria as the continuous equation.

We assume the uniform discretization with step h and rewrite (1) with (4) for $t \in [mh, (m+1)h]$ as

$$\frac{d}{dt}\Bigg(x_i(t)e^{\tau_i t}\Bigg) = \lambda_i \tau_i e^{\tau_i t} g\Bigg(\sum_j a_{ij}x_j + b_i\Bigg) \tag{8}$$

Next we integrate (8) on $[mh, t)$ and let $t \to (m+1)h$. As the result we obtain the discrete version of (1)

$$x_i^{m+1} = \phi_i(h)x_i^m + \Big(1 - \phi_i(h)\Big) g\Bigg(\sum_j a_{ij}x_j^m + b_i\Bigg), \tag{9}$$

where $\phi_i(h) = e^{-\tau_i h}$, $x_i^m = x_i(t = mh)$.

Mohamad and Gopalsamy (2000) noted that stability conditions and sufficient conditions for exponential convergence of the solutions of (9) and (1) to equilibrium are independent of h. Such preservation of dynamics is lacking in standard numerical methods. In integration we used $h = 0.1$, $m = \overline{1, M}, M = 150$ and then sampled the resulting solution at $k = 15$ equidistant points, which leads to the discrete time series $x_1(t_j)$ and $x_2(t_j)$, $j = \overline{1, k}$. The resulting time series are shown on Figure 1a. Figure 1b plots the time series for Gene 1 and Gene 2 normalized by their standard deviation:

$$x_i(t_j) \to \frac{x_i(t_j) - \overline{x_i}}{\sigma_i}, \tag{10}$$

where $\sigma_i = \sum_j (x_i(t_j) - \overline{x_i})^2/k$. The linear correlation coefficient r, calculated as

$$r = \frac{\sum_{j=0}^{k} (x_1(t_j) - \overline{x_1})(x_2(t_j) - \overline{x_2})}{\sqrt{\sum_j (x_1(t_j) - \overline{x_1})^2 \sum_j (x_2(t_j) - \overline{x_2})^2}}, \tag{11}$$

yields $r = 0.12$, which would be taken as the indication that the time series are uncorrelated, if the linear correlation were taken as a measure of similarity; whereas according to the transcriptional model they are fully co-regulated (see also the signalling function $g(F_1) = g(F_2)$ plotted in Figure 1b).

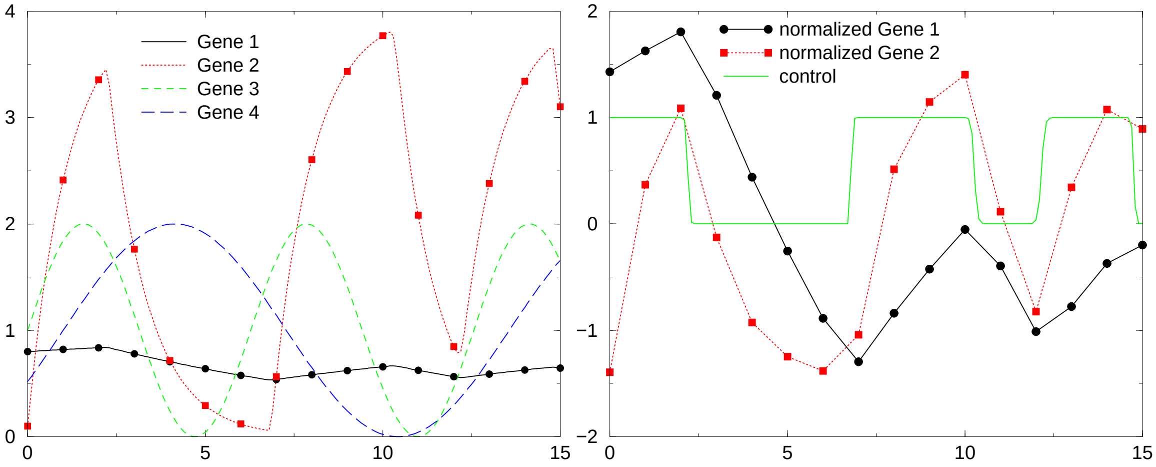
In the next example, Genes 1 and 2 have a very distinct regulation: Gene 3 acts as an enhancer of Gene 1 only; Gene 4 acts as an enhancer of Gene 2 only. Table 2 represents the value of our test parameters in the system (6)-(7).

Table 2. The parameters of the gene transcriptional control in (6)-(7). Gene 3 acts as an enhancer of Gene 1; Gene 4 acts as an enhancer of Gene 2.

i	λ_i	a_{i3}	a_{i4}	b_i	τ_i
1	1	100	0	-10	0.1
2	4	0	100	-10	0.1

The initial values were chosen as $x_1(0) = x_2(0) = 0.1$. The time integration and sampling were performed similarly to the previous example. The linear correlation coefficient between the time series corresponding to Gene 1 and 2 is $r = 0.98$ showing

Figure 1: Plots of the gene expression profiles of Genes 1, 2 3, and 4. The dynamics of gene expression is governed by (6)-(7) with the parameters given in Table 1. The expression levels of Gene 3 and Gene 4 are set to be $\sin t$ and $\sin(t-1)/2+1$ correspondingly. Gene 3 acts as an enhancer of Genes 1 and 2; Gene 4 acts as a repressor of Genes 1 and 2.

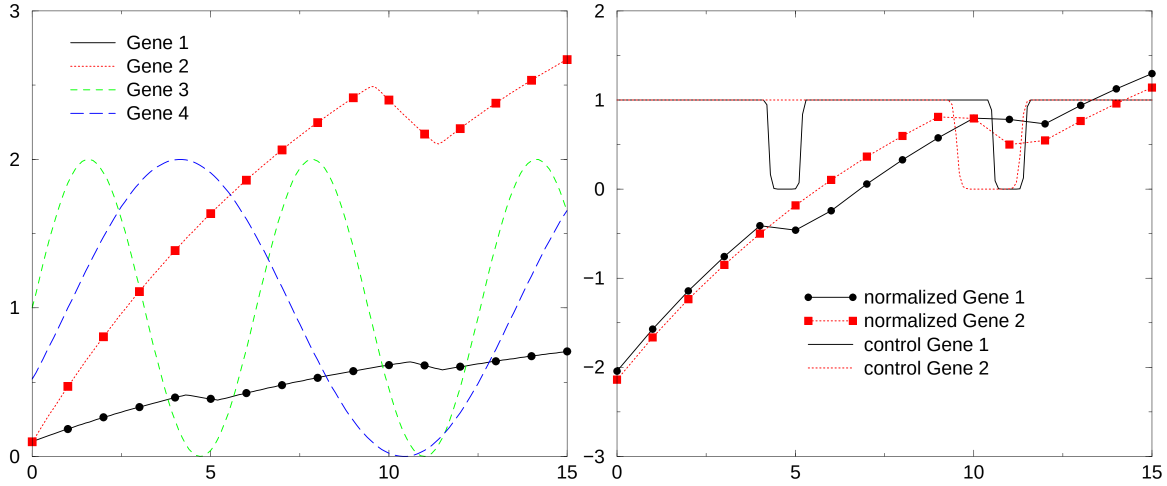


that if the linear correlation were chosen as a similarity measure, Genes 1 and 2 would be decided to be co-expressed. Figure 2 plots the resulting series and the signalling functions $g(F_1)$ and $g(F_2)$. Notice, that there are time intervals where Genes 1 and 2 are oppositely regulated: Gene 1 is "down" while gene 2 is "up" on $[4.3, 5.1]$ and the opposite is true on the interval $[9.5, 10.5]$. In spite of such a different behaviour, these two genes will be clustered together according to a similarity measure that ignores the underlying transcription mechanism.

4 Similarity measure and clustering algorithm based on change in transcriptional control

Under the assumption that the transcriptional control in gene networks obeys (1) we can devise an algorithms for detecting co-expressed genes based on the similarity of patterns of activation and repression that are represented by the signalling function

Figure 2: Plots of the gene expression profiles of Genes 1, 2 3, and 4. The dynamics of gene expression is governed by (6)-(7) with the parameters given in Table 2. The expression levels of Gene 3 and Gene 4 are set to be $\sin t$ and $\sin(t-1)/2+1$ correspondingly. Gene 3 acts as an enhancer of Gene 1; Gene 4 acts as an enhancer of Gene 2.



g. The model (1) offers a natural similarity measure $\mathcal{D}$ between genes m and l

$$\mathcal{D} = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{\dot{x}_m(t_i) + \tau_m x_m(t_i)}{\lambda_m} - \frac{\dot{x}_l(t_i) + \tau_l x_l(t_i)}{\lambda_l}\right)^2. \qquad (12)$$

Close to zero $\mathcal{D}$ indicates that two genes are co-expressed and $\mathcal{D}$ which is close to 1 indicates that two genes are oppositely expressed. Since the values of the natural gene degradation rates τ_i and the maximum expression levels λ_i are not known in advance, it is not feasible to use (12) directly. Instead we suggest to use the properties of the solutions of (1). The ith equation of the system (1) has simple analytical solution if ether gene i is fully activated ($g \approx 1$) or repressed ($g \approx 0$):

$$x_i(t) \approx x_i(t_r)\exp(-\tau_i(t-t_r)), \qquad t \geq t_r, \qquad (13)$$

where t_r is the moment of gene suppression (g takes on values close to 0) and

$$x_i(t) \approx \lambda_i + (x_i(t_a) - \lambda_i)\exp(-\tau_i(t-t_a)), \qquad t \geq t_a, \qquad (14)$$

where t_a is the moment of gene activation (g takes on values close to 1). The solution between adjacent t_r and t_a can be found by smooth matching between $x_i(t_r)$ and

$x_i(t_a)$. Notice that the highest negative rate of change in gene product concentration occurs at $t = t_r$: $\min \dot{x}_i(t) = -\tau_i x_i(t_r)$ and the highest positive rate of change occurs at $t = t_a$: $\max \dot{x}_i(t) = \lambda_i(1 + \tau_i) - x_i(t_a)$. Such maximum values of gene expression rate of change identify the change in control and, therefore, in order to identify the co-regulated genes we should compare *moments of time* when these changes take place. Because of the discrete nature of data, the co-regulated genes can have the highest expression rates shifted by one time interval with respect to one another. The algorithm for determining such co-regulated genes becomes as follows:

1. Calculate the time derivatives for each of the gene expression time series using forward or centered differences, e.g.

$$\dot{x}_i(t_j) \approx (x_i(t_{j+1}) - x_i(t_j))/(t_{j+1} - t_j), \qquad i = \overline{1, k-1}, \tag{15}$$

$$\dot{x}_i(t_k) \approx (x_i(t_k) - x_i(t_{k-1}))/(t_k - t_{k-1}) \tag{16}$$

   Notice that by these definitions, if the forward differences are used for all j except for $j = k$ where the backward difference is used, we will always have $\dot{x}_i(t_k) = \dot{x}_i(t_{k-1})$. This implies that the last approximation for the derivative should be discarded as it carries no additional information. If the central differences are used for all internal nodes with forward difference for the first node and the backward difference for the last node, then the derivatives at all nodes can be used. At the same time it is better to use forward (or backward) differences for internal nodes if data is sparse (as in case of rat CNS development data of Wen *et al.* (1998); see Sections 5 and 6 for further discussions).

2. Represent the expression control experienced by gene i by a vector $\mathbf{q}^i = (q_0^i, ..., q_j^i, ...q_k^i)$ such that

$$q_j^i = \begin{cases} 1 & \text{if } \dot{x}_i(t_j) > \max(\dot{x}_i(t_{j-1}), 0), \dot{x}_i(t_j) \geq \dot{x}_i(t_{j+1}), \\ -1 & \text{if } \dot{x}_i(t_j) < \min(\dot{x}_i(t_{j-1}), 0), \dot{x}_i(t_j) \leq \dot{x}_i(t_{j+1}), \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

   In the definition (17) we set $\dot{x}_i(t_0) = \dot{x}_i(t_{k+1}) = 0$ in order for it to be applicable for the values of q_1^i and q_k^i. The non-strict inequality for the node $\dot{x}_i(t_{j+1})$ takes care of a situation when several nodes are lying on the straight line of the local maximum slope. In this case the definition (17) implies that the moment of activation [repression] takes place at the leftmost point.

3. Two genes m and l are decided to be **co-expressed** if either $q_j^m = q_j^l$ for any $j = \overline{1,k}$, or for any j such that $q_j^m \neq q_j^l$ either $q_{j+1}^m = q_j^l, q_j^m = q_{j+1}^l = 0$ or $q_{j-1}^m = q_j^l, q_j^m = q_{j-1}^l = 0$. Similarly, two genes m and l are decided to be **oppositely expressed** if either $q_j^m = -q_j^l$ for any $j = \overline{1,k}$, or for any j such that $q_j^m \neq -q_j^l$ either $q_{j+1}^m = -q_j^l, q_j^m = q_{j+1}^l = 0$ or $q_{j-1}^m = -q_j^l, q_j^m = q_{j-1}^l = 0$. Table 3 gives the examples of co-expressed Genes 1 and 2, 2 and 4 and oppositely expressed Genes 1 and 3 (also 2 and 3).

Table 3. Examples of q-vectors for Genes 1, 2, 3 and 4. Genes 1 and 2 are co-expressed and oppositely expressed with Gene 3. Gene 4 is co-expressed with Gene 2, but not with Gene 1.

Gene	q_0	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9	q_{10}	q_{11}	q_{12}
1	0	1	0	0	0	0	0	-1	0	0	0	1	0
2	0	1	0	0	0	0	-1	0	0	0	1	0	0
3	0	0	-1	0	0	0	0	1	0	0	0	-1	0
4	0	1	0	0	0	-1	0	0	0	1	0	0	0

Our algorithm correctly recognizes the co-regulation of Genes 1 and 2 in the first example of the previous section and detects the difference in the expression control mechanisms in the second example. It is necessary to emphasize several points. Firstly, as it follows from Table 3, it is feasible to have a situation when gene i is co-expressed with gene j, gene k is co-expressed with gene i, but not with gene j. This drawback also occurs with other similarity measures (e.g. Euclidean norm or linear correlation etc.) that define pairwise "distance" between expression patterns. Such a situation implies that either link between i and j or between i and k is accidental. It is easy to trace such a case with our algorithm. Secondly, additional sources of unpredictability may include external noise or errors in measured data which lead to even less accurate derivative approximations (17). Notice that a small error in data measurements may shift the maximum derivative value by one time interval – this shift is already taken into account by our algorithm. If there is a reason to believe that data is more significantly corrupted, a low-pass filter for data smoothing (e.g. Savitzky-Golay filter (Savitzky and Golay, 1964)) could be applied before clustering is attempted. Recently the singular value decomposition technique has been successfully applied to eliminate noise and experimental artifacts from the gene expression data (Alter *et al.*, 2001)

One of the advantages of the suggested algorithm in comparison with other clustering techniques is that it allows to distinguish between "co-expressed" and "co-regulated" genes. These two terms are often used interchangeably, but there is a significant difference between them. "Co-expressed" genes are activated or suppressed at the same time; "co-regulated" genes have the same transcription factors acting as enhancers or silencers. Two genes can be fully activated during the entire time span of the experiment. We should consider them to be "co-expressed," but there is not enough evidence to judge that they have the same enhancers. On the other hand if two genes were turned on and off simultaneously several times during the experiment, we should assume that they are "co-regulated". Since non-zero entries of $\mathbf{q}$ vector represent the moments of time the genes in the cluster were switched on and off, we can use it as the measure of whether the genes in that cluster are co-regulated. At the same time too many nonzero entries in $\mathbf{q}$ may signal that the data is too erratic (too noisy) to be analysed (the time series representing the gene expression changes the sign of the slope between nodes at almost every time step). Heuristically, we suggest to use the measure $\mathcal{Q} = \sum_i |q_i|$ in a following way: (1) if $\mathcal{Q} \leq 3$, then there is not enough information to judge whether the genes in the corresponding cluster are co-regulated; (2) if $\mathcal{Q}/k > 1/2$, then the data are too corrupted to draw any meaningful comparison. Otherwise, the genes in the same cluster can be assumed to be co-regulated.

5 Rat CNS development data

Wen *et al.* (1998) used reverse transcription-coupled PCR to produce a high-resolution temporal map of fluctuations in mRNA expression of 112 genes during rat CNS development in the cervical spinal cord. Many researches have used these data to elucidate gene interactions (D'haeseleer *et al.*, 1999, 2000; Wahde and Hertz, 2000) by clustering or reverse engineering techniques.

We applied our algorithm to the time series of all 112 genes. The tabulated values of gene expression data consist of 9 time points: embryonic days 11, 13, 15, 18, 21, and post-natal days 0, 7, and 14; and adult day 9. The complete set of genes that are suggested to be co-expressed, oppositely expressed, co-regulated, or too noisy to draw any conclusion are given on the web page

www.damtp.cam.ac.uk/people/ngb23/bio/cluster.html.

Our knowledge of the underlying biological system is insufficient to judge whether the majority of the suggested interactions are meaningful. Figure 3 plots two gene clusters where a co-regulation is suggested ($\mathcal{Q} = 4$). Figure 4 plots two gene expression time series where the opposite regulation is suggested ($\mathcal{Q} = 4$) and Figure 5 shows two co-expressed gene profiles where our analysis suggests that too much noise is present ($\mathcal{Q} = 6$). On these figures each gene expression profile is normalized by (10).

6 Conclusions and discussions

D'haeseleer *et al.* (1998) compared different similarity measures used in clustering algorithms. Linear correlation coefficient (Pearson's r) detects only linear relationships between genes and is a rather poor statistic for deciding whether an observed correlation is statistically significant, or whether one correlation is significantly stronger than another, since there is no universal way of computing the distribution of the expression levels. Non-parametric or rank correlation coefficient allows one to interpret the significance of the observed correlation and generally is more robust than linear

Figure 3: Clusters of co-regulated genes from rat CNS gene expression data (Wen *et al.*, 1998) normalized by (10).

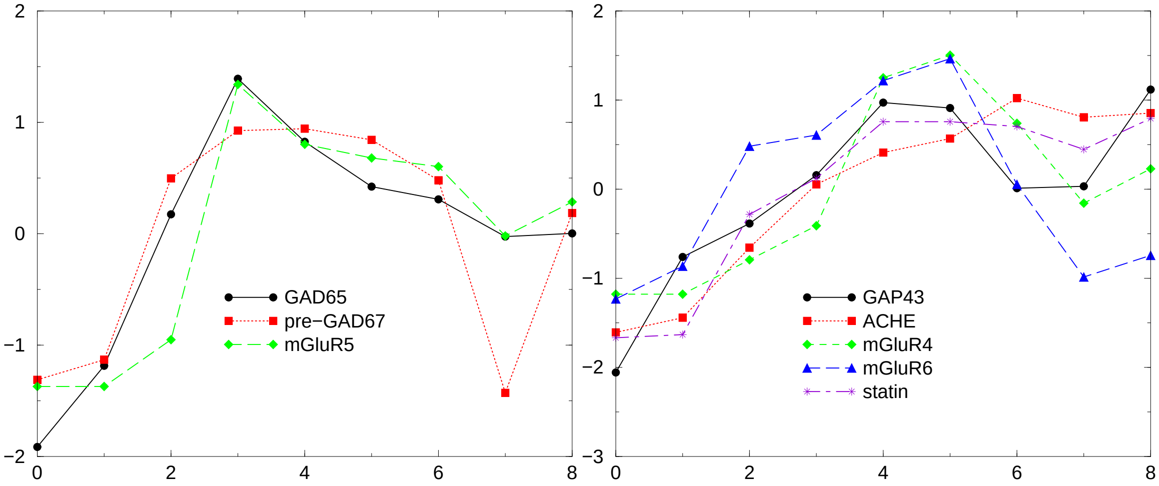


Figure 4: Pair of oppositely expressed genes from rat CNS gene expression data (Wen *et al.*, 1998) normalized by (10).

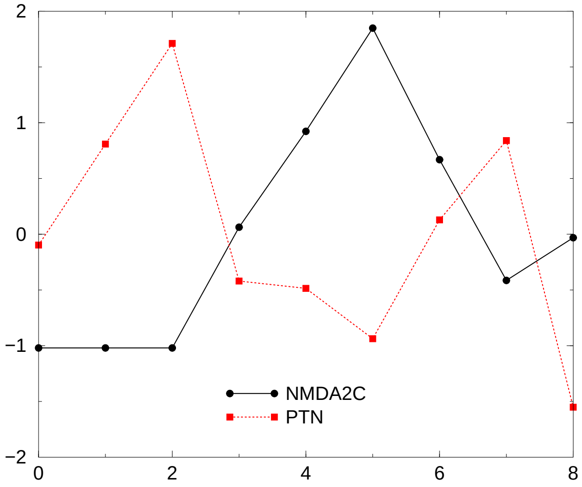
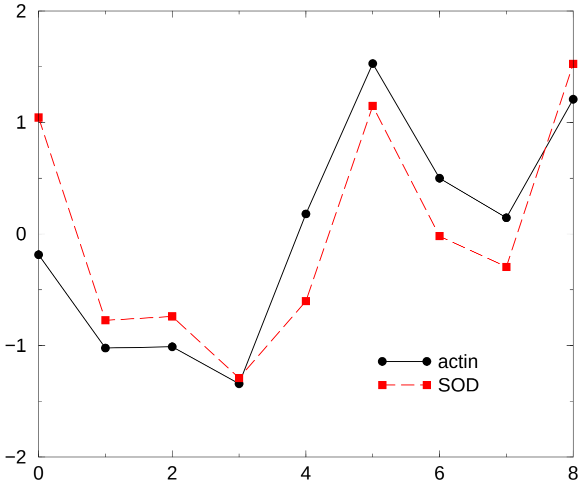


Figure 5: Pair of co-expressed but noisy gene expression profiles from rat CNS gene expression data (Wen *et al.*, 1998) normalized by (10).



correlation coefficient, but it uses only monotonic ordering of the expression levels, which results in information loss. Mutual information based measure uses bins to classify expression patterns, but that makes it inapplicable to continuous time series. So far there is no theory how to choose the "best" similarity measure. Any attempt to introduce such a measure without taking into account the underlying biology of regulatory network will lead to clusters without clear biological significance attached.

We suggested a clustering algorithm based on the network of interacting genes where transcriptional control is modelled by nonlinear differential equations taking into account the signalling network and natural gene product degradation. Our algorithm allows us to filter out the most essential feature of gene interactions that is characterized by the pattern of gene activation/repression. Two genes are assumed to be co-regulated if they are switched on and off simultaneously several times during the experiment.

We demonstrated that the linear correlation measure fails to recognize co-regulated gene patterns or detects similarity in uncorrelated gene expression data in a simple model network of four interacting genes. We applied our algorithm to gene expression data on rat CNS development (Wen *et al.*, 1998) and classified pairs of genes into several categories: co-regulated genes (genes with a similar activation/repression that occurs several times during the experiment, such repeated simultaneous activation/repression can be taken as an indication of similar regulatory mechanisms), co-expressed genes (genes with just a few but similar activations/repressions during the experiment), oppositely expressed genes, and genes with expression levels too noisy to draw any meaningful conclusion. One drawback of our method is the use of approximated derivatives, that are more influenced by the noise in the system than the gene expression data itself. As we already discussed, the method determines some irregularities in the data, so that the noisy data can be discarded. In addition some preliminary low-pass filtering can be used to smooth out the time series before clustering is attempted. There is also some uncertainty which derivative approximation to use: backward, centered or forward difference. For smooth enough data, where measurements of gene expression are taken at small time steps it should not matter which approximation is used. Otherwise, it is better to use forward (or backward) differences, as they represent the slopes between successive data points.

Finally, it may appear that the replacement of the gene expression time series with their **q** vectors, whose coordinates are mostly zeros, leads to a loss of information. We would like to emphasise that while a particular gene i remains fully activated ($g(F_i) \approx 1$) or repressed ($g(F_i) \approx 0$) its time development depends on the inherent features of this particular gene defined by τ_i and λ_i and not on other transcriptional controls or products of gene expression in the system, therefore, in order to detect the similarities in the control mechanisms between different genes, we need to be able to separate them from other elements specific to a dynamics of a particular gene.

Acknowledgments

The author is very grateful to Professor Ken Lange for many useful discussions and acknowledges the support from the NIH/NHGRI Grant K25 HG02411-01. She is also indebted to the program on “Functional genomics” at the Institute of Pure and Applied Mathematics, UCLA that stimulated her interest to these type of problems.

References

1. Alter O., Brown P.O., and Botstein D., 2001 Processing and modeling genome-wide expression data using singular value decomposition. Proceedings of SPIE Vol. 4266, Ed. Bittner *et al.*, 171-186.

2. Basset, D.E., Eisen, M.B., and Boguski, M.S. , 1999. Gene expression informatics–it's all in your time. Nat. Genet. Supp., 21, 51-60.

3. D'haeseleer, P., Wen, X., Fuhrman, S., and Somogyi, R., 1998 in Information processing in Cells and Tissues, Plenum Press, New York.

4. D'haeseleer, P., Liang, S., and Somogyi, R., 2000. Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics 16, 707-726.

5. D'haeseleer, P., Wen, X., Furhrman, S., Somogyi, R., 1999. Linear modeling of mRNA expression levels during CNS development and injury. In: Altman, R.B., Dunker, A.K., Hunter, L., Klein, T.E., Lauderdale, K. (Eds.) Pacific Symposium on Biocomputing '99. World Scientific, Singapore, pp. 41-52.

6. Edwards, D.R., 1994. Cell signaling and the control of gene transcription. Trends Pharmacol. Sci. 15, 239-244.

7. Iyer, V. *et al.*, 1999. The transcriptional program in the response of human fibroblasts to serum. Science 283, 83-87.

8. Mjolsness, E., Mann, T., Castano, R. and Wold, B. 1999. From co-expression to co-regulation: an approach to inferring transcriptional regulation among gene classes from large-scale expression data. Tech. Rep. JPL-ICTR-99-4 Jet Propulsion Laboratory Section 365.

9. Mohamad, S. and Gopalsamy, K., 2000. Dynamics of a class of discrete-time neural networks and their continuous-time counterparts. Math. Comp. Simul. 53, 1-39.

10. Reinitz, J., Sharp, D.H., 1995. Mechanism of eve strip formation. Mech. Dev. 49, 133-158.

11. Rosen, R., 1968, Recent developments in the theory of control and regulation of cellular processes, in International Review of Cytology, G.H. Bourne (Ed.), New York: Academic Press.

12. Savitzky A., and Golay, M.J.E., 1964, Analytical Chemistry, 36, 1627-1639.

13. Smolen, P., Baxter, D.A., and Byrne, J.H., 2000. Modeling transcriptional control in gene networks-Methods, recent results, and future directions. Bull. Math. Bio., 62, 247-292.

14. Tyson, J. and Othmer, G., 1978. The dynamics of feedback control circuits in biochemical pathways. Prog. Theor. Biol. 190, 37-49.

15. Wahde, M. and Hertz J., 2000. Coarse-grained reverse engineering of genetic regulatory networks. BioSystems 55, 129-136.

16. Wen, X., Furhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Barker, J. L., Somogyi, R., 1998. Large-scale temporal gene expression mapping of CNS development. Proc. Natl. Acad. Sci. 95, 334-339.