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

Natalia G. Berloff [1]

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. 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 [4]. 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. Modeling gene networks using a system of coupled *nonlinear* differential equations 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 [3]. We have developed 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.

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

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

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 $\lambda_i = \lim_{t\to\infty} x_i(t)$ 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 $g(x) = 1/\Big(1 + \exp(-x)\Big)$. and assume $F_i = \sum_{j=1}^{n} a_{ij}x_j + b_i$.

[1]Department of Applied Mathematics and Theoretical Physics
University of Cambridge, Wilberforce Road, Cambridge, CB3 0WA, UK. E-mail: N.G.Berloff@damtp.cam.ac.uk

In [1] 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 modeled by (1). Under the assumption that the transcriptional control in gene networks obeys (1) we devised an algorithms for detecting co-expressed genes based on the similarity of patterns of activation and repression that are represented by the signalling function g. The ith equation of the system (1) has simple analytical solution if either gene i is fully activated ($g \approx 1$) or repressed ($g \approx 0$): $x_i(t) \approx x_i(t_r)\exp(-\tau_i(t-t_r)), \qquad t \geq t_r$, where t_r is the moment of gene suppression (g takes on values close to 0) and $x_i(t) \approx \lambda_i + (x_i(t_a) - \lambda_i)\exp(-\tau_i(t-t_a)), \qquad t \geq t_a$, 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. (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 $q_j^i = 1$ 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 if $\dot{x_i}(t_j) < \min(\dot{x_i}(t_{j-1}), 0), \dot{x_i}(t_j) \leq \dot{x_i}(t_{j+1})$; and 0 otherwise. (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 $q_{j-1}^m = -q_j^l, q_j^m = q_{j-1}^l = 0$.

We applied our algorithm [1]to gene expression data on rat CNS development [2] 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.

The author acknowledges the support from the NIH/NHGRI Grant K25 HG02411-01.

References

[1] Berloff N.G. 2002. Gene expression data analysis: clustering technique using models of transcriptional control of gene networks. Submitted to BioSystems.

[2] 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.

[3] 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.

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