Towards the Reconstruction of Gene Regulatory Networks

Frank L. Tobin*, Valeriu Damian-Iordache, Larry D. Greller
Mathematical Biology, Bioinformatics
SmithKline Beecham Pharmaceuticals R&D
PO Box 1539
King of Prussia, PA, 19406
*frank_tobin@sbphrd.com

ABSTRACT

Gene expression data in biology is becoming important as the amount and quality of the data rapidly increases. However, the amount generated can be daunting and its direct interpretation is often difficult. The interaction of the genes and the number involved can be large. Is there a dynamical system at play? This paper discusses modeling gene expression data as a computational reconstruction of a dynamical system. The problem is a classic inverse problem – given the data what is the model? A phenomenological model based on a extension of generalized Lotka-Volterra models is developed. One advantage of these models is that they are readily amenable to biological interpretation. The reconstruction is ill-posed and subject to numerical instability problems when is there not enough data of sufficient quality. We will discuss how these problems can affect the results and how we might overcome them. Lastly, we will present some preliminary results and some applications of the reconstructions.

Keywords: genetic regulatory networks, gene expression, ordinary differential equations, Lotka-Volterra

INTRODUCTION

Modern biology is increasingly being driven by its ability to generate large amounts of data that was previously impossible to attain. This revolution started first with sequence data (genomic and EST) and is now starting to encompass the abundances of genes and proteins inside cells. This flood of data has created a paradigm shift. We may have collected the data, but now we need to find the corresponding biological functions. Gene expression falls within this category. We now have the capability to find the expression patterns of thousands of genes simultaneously[1]. For organisms with small genomes, the entire genome can be a placed on a single microarray grid[2, 3].

Large scale gene expression analysis allows an unprecedented opportunity to dissect the function of the genome of cells as we explore the effects of various natural biological conditions – metabolism, growth, morphogenesis, differentiation, development, aging. In addition there are various experiments where we wish to probe cellular function by inducing behavior. In the pharmaceutical industry there is interest in understanding how drugs work and their mechanism of action.

Not only is understanding mechanism important, but so is variation in mechanism. Variation can occur for a variety of reasons – temporal and spatial change[2, 4-7], tissue specific differences[8, 9], development of the organism or the aging of the cell[10, 11]. To understand variation and the resulting behavior of the cell involves deciphering the very complex processes of coordinated expression of genes. By using gene expression studies of thousands of genes through time we can begin delineating the interactions between the genes. This is what we mean by the reconstruction of a gene regulatory network.

THE DYNAMICAL MODEL

We wish to consider the following experiment as the basis upon which the reconstruction will be based:

- All the genes or some carefully selected subset of the genes (presumably one pathway) is chosen and used in the expression experiments
- Their expression levels (absolute abundance or concentration) will be measured at each sampling point
- There will be some continuous, monotonically increasing variable which is the independent, controlling variable for the experiment – time, dose, etc. Samples will be taken at several different values of this variable. For the purposes of this paper, we will always refer to this variable as 'time'.

The matrix of the different gene expression values conventionally has increasing time as the columns and the genes as the rows. The matrix is usually referred to as a 'grid'.

Because we wish to develop completely automated procedures, any modeling approach must scale to tens of thousands of genes for explaining the expression dynamics of full or partial (i.e. pathways) genomes. This requires a robust model that is capable of handling many different situations. Genetic regulatory networks have a long history and many different approaches have been attempted[12-19].

However, in many models, expression is binary; there is only an interaction graph not a dynamical model; or the model is very detailed but only for a limited number of genes and proteins. We propose modeling for explanatory purposes. We don't expect the model to be 'correct' in an absolute sense. We do expect that it is good enough to reproduce the phenomenology of the gene expression patterns that it can serve as a 'clue' generator – a way to steer new experiments to the most important events that are happening.

The reconstruction of the genetic regulatory network is an explanation as a dynamical system of the temporal evolution of the column vectors of the expression grid:

$$G(t)' = \mathrm{F}(G, t;\ T^*) \qquad (1)$$

where G(t) is the column vector of the gene expression values at time t, and T are the interaction parameters that govern the strength of the interactions. The dynamical system F, is unknown, so that we have an inverse problem. We've measured G and t, and hence G', but we do not know F or T. The assumption is being made that this is a locally smooth and continuous dynamical system. This is reasonable since there are biological reasons to believe that if the biological change in the system is 'small' and 'smooth', that concomitant changes in the gene expression will also be 'small' and 'smooth'.

The major problem with this view, is that, by necessity, it is incomplete. The model implies that only genes interact with genes and that genes control the production (i.e. transcription) of new gene expression values. Only the genes are included even though we know that it is the genes and proteins that are coupled and that the proteins are critical to the transcription of the genes. The dynamical system in equation 1 should be expanded to included gene-protein and protein-protein interactions:

$$G(t)' = \mathrm{F}_1(G, P, t;\ T^*) \qquad (2a)$$

$$P(t)' = \mathrm{F}_2(G, P, t;\ T^*) \qquad (2b)$$

where P(t) is the column vector of the protein products corresponding to the genes, G(t). However, this inclusion is impossible, when there are no protein values measured. Since we are excluded from building such a dynamical system, we have chosen to reconstruct the gene-only dynamics (equation 1) as a "phenomenological" model of the interactions. The philosophy behind the reconstruction is that it will provide "clues" to the biology. Such clues will be useful in helping to prioritize and design experiments to pin down the "*true*" interactions.

The inverse problem represented by the gene regulatory network (equation 1) still requires the definition of the dynamical system model. We have chosen extensions to Generalized Lotka-Volterra (GLV) models. GLV models are quite useful for handling situations where there are time varying and interacting populations and where there are resource limitations in growth[20, 21]:

$$G_j' = \sum_k T^1_{j,k} G_k + \sum_k T^2_{j,k} G_j G_k \qquad (3)$$

In this sense the reconstruction of the gene interactions and population modeling are similar:

- the lower bound is zero
- the upper bound is finite
- there are finite resources governing the growth of populations (expression)

The T^1 and T^2 matrices are the interaction strengths of the gene interactions. They can be interpreted roughly as follows:

$T^1_{j,k}$ is the control of gene j by gene k through a serial pathway (e.g. serial tranduction pathway)

$T^2_{j,k}$ is the control of gene j by gene k as a 'regulon' (a gene that activates or inhibits the action of another gene)

Although there is no space in this paper to discuss the properties of equation 3, we note that it can be recast in a traditional Lotka-Volterra form:

$$G_j' = \sum_{k \neq j} T^1_{j,k} G_k + T^1_{j,j} \left\{ 1 + \frac{T^2_{j,j}}{T^1_{j,j}} G_j + \sum_{k \neq j} \frac{T^2_{j,k}}{T^1_{j,j}} G_k \right\} G_j \qquad (4)$$

$$[1] \qquad\qquad [2] \quad [3] \qquad\qquad [4]$$

Terms 2 and 3 can be viewed as the self-limiting logistic growth contributions; term 4 represents the cooperation (i.e. activation), if positive, or competition (i.e. inhibition), if negative of the other gene 'populations'. Similarly, the model (equation 3) can be considered a Ricatti differential equation[22]:

NUMERICAL CONSIDERATIONS

For the purposes of this paper only one example will be provided. A colony of Sacchromyces cerevisiae yeast was fed glucose and the expression of 6153 genes was measured over time[3]. The data for this experiment is shown in Figure 1.

Now that the model is chosen, it is necessary to numerically determine the T^1 and T^2 matrices from the data. This is a classic inverse problem since we know the G(t) values experimentally and can compute the derivatives. While, in principle this is possible, there are two basic problems that must be overcome. The reconstruction can be considered as a linear optimization problem for the determination of the T matrices:

$$\underset{T^1, T^2}{argmin} \left\| G'_j - \sum_k T^1_{j,k} G_k + \sum_k T^2_{j,k} G_j G_k \right\| \qquad (5)$$

The first numerical challenge is to calculate accurate derivatives from what may be noisy, infrequently sampled data.

The second numerical obstacle is that the biological variation in the problem may not be sufficient to distinguish the functional behavior of all the genes. For example, in the diauxic shift experiments in yeast, the yeast colony is supplied a finite amount of glucose. The glucose eventually is consumed and the yeast must changed their metabolism. The problem is that the 'biological variation' is the perturbation of cellular function by 'dosing' with glucose. Not all biology in the cell is exposed – mostly metabolism. However, in this experiment 6153 yeast ORFs (i.e. putative genes) were placed on the grid – the entire genome. Not every gene is going to be affected significantly. We expect to see similar patterns of expression for large numbers of genes. This causes degeneracy problems in the model. Consider the situation of two genes, p and q, that are degenerate. Their contribution to the first term of equation 3 is

$$\begin{aligned} T^1_{j,p} G_p + T^1_{j,q} G_q &= \left\{ T^1_{j,p} + T^1_{j,q} \right\} G_p \\ &= \left\{ T^1_{j,p} + T^1_{j,q} \right\} G_q \end{aligned} \qquad (6)$$

A similar degeneracy occurs for the T^2 matrix. Clearly the determination of the T coefficients is ambiguous in such a situation.

The solution of this problem is to cluster the genes into pattern scaling-invariant degeneracy classes. The reconstruction is then performed not on the genes, but the gene clusters. Such degeneracy can be quite severe. In the diauxic shift case mentioned above, 6153 ORFs can be clustered into approximately 25 clusters. Only 4% of the expression behavior can be used! The degeneracy problem is basically a rank deficiency in a different guise. See Figure 2 for an example of the clustering of the expression patterns of the ORFs.

The third computational impediment arises from limitations with the experiments currently being performed. There are not enough time points sampled to properly determine the system. This has implications for the accuracy of interpolating the expression data and calculating derivatives. More significantly, with only 3 or 5 or 10 time points the system is severely under-determined and the determination of the T matrices is ill-posed. In the diauxic shift example being used, there are 7 time points to determine 1250 coefficients (T^1 and T^2 with 25 gene clusters). We have been exploring the use of regularization techniques to overcome the ill-posedness of the system[23, 24].

The fourth task is to guarantee that the solution is truly a good solution. The computation of the T matrices (equation 5) is a solution only in an average or approximate sense, especially when regularization is used to overcome the ill-posedness. Once the T matrices are computed, then, in principle, it should be possible to integrate the ODEs (equation 3). There may be many T matrices that are acceptable from the optimization/regularization computation, but we have no guarantee that they represent a numerically stable solution of the ODEs. In essence, numerical stability and integrability are additional checks on the quality of the solutions. Given two solutions that are essentially equivalent, stability becomes the deciding criterion. Figure 3 is an example of the reconstruction of the model and its integration. It can be seen that in most cases that the model is doing a good job of describing the data. This is a good example of a reconstruction, but we are still striving to attain this quality for all data sets. We are still exploring techniques for incorporating numerical stability into the optimization, probably as an extension to regularization, in order to make the technique robust for all data sets.

DISCUSSION

In Figure 3 is the reconstruction of the diauxic shift data. In Figure 4 are the corresponding T^1 and T^2 matrices. It can be seen that several gene clusters are strongly regulated by others and that some gene clusters appear to be regulated more than others. This implies that there is specific regulation at work – to the degree that the model is accurate. The next step is to examine the clusters, look at the functional characteristics of the gene components and develop a biological interpretation of the genetic regulatory network. In this example, the interpretation that is developed mirrors the biological knowledge that we already have.

There are several tasks that we have only begun to explore. First, the robustness of the numerical procedures must be improved – regularization and numerical stability. Next we have also started to explore automating the biological interpretation of the reconstructions. This means trying to automate watching the behavior of the system – which gene clusters are strongly coupled – as the dynamics evolve and to find biological 'themes'. Another approach is to do this theoretically by examining the behavior of the dynamical system. So far, the high dimensionality of the system and the arbitrary coefficients has made it difficult to use the analytical and stability properties of Generalized Lotka-Volterra and Matrix Ricatti systems[25-28]. The goal is to try to predict future behaviors of the system as the biological variation is changed. For example, if we have the reconstruction for a cell treated with a drug, we would like to try to predict the behavior for a different drug. Or to predict how the system would behave at much longer times than can be measured experimentally.

It is our belief that the phenomenological reconstruction of genetic regulatory networks is achievable and that the phenomenology is a realistic mirror of the underlying biology. While not a perfect doppelganger, there is sufficient biology being captured that the reconstructions can be used to start computationally exploring the complexities of these systems.

REFERENCES

[1] Lockhart, D.J., Dong, H., Byrne, M.C., et al., Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays, Nature Biotechnology, 1996, **14**:1675-1680.

[2] Chu, S., DeRisi, J., Eisen, M., et al., The Transcriptional Program of Sporulation in Budding Yeast, Science, 1998, **282**:699-705.

[3] DeRisi, J., L., Iyer, V.R., and Brown, P.O., Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale, Science, 1997, **278**:680-686.

[4] McGinnis, W. and Wiechaus, E., Homeobox Genes and Axial Patterning, Cell, 1992, **68**:283-302.

[5] Jackle, H. and Jahn, R., Vesicle Transport: Klarsicht Clears Up the Matter, Current Biology, 1998, **8**:R542-R544.

[6] Ishiura, M., Kutsuna, S., Aoki, S., et al., Expression of Gene Cluster kaiABC as a Circadian Feedback Process in Cyanobacteria, Science, 1998, **281**:1519-1523.

[7] Gilbert, S.F., Developmental Biology, Fifth edition, 1997, Sunderland, MA, Sinauer Associates, Inc.

[8] Galau, G.A., Klein, W.H., Britten, R.J., et al., Significance of Rare mRNA Sequences in Liver, Archives of Biochemistry and Biophysics, 1977, **179**:584-599.

[9] Vasmatzis, G., Essand, M., Brinkmann, U., et al., Discovery of Three Genes Specifically Expressed in Human Prostate by Expressed Sequence Tag Database Analysis, Proceedings of the National Academy of Sciences USA, 1998, **95**:300-304.

[10] Michaelson, J., The Significance of Cell Death, in Apoptosis: The Molecular Basis of Cell Death, Tomei, L.D. and Cope, F.O. editors, 1991, Cold Spring Harbor Laboratory Press, Plainview, NY, p. 31-46.

[11] Tomei, L.D. and Cope, F.O. editors, Apoptosis: The Molecular Basis of Cell Death, Current Communications in Cell & Molecular Biology, Vol. 3, 1991, Cold Spring Harbor Laboratory Press, Plainview, NY..

[12] D'haeseleer, P., Wen, X., Fuhrman, S., et al. Mining the Gene Expression Matrix: Inferring Gene Relations from Large Scale Gene Expression Data. in Proceedings of the International Workshop on Information Processing in Cells and Tissues (IPCAT). 1997. (in press), see http://rsb.info.nih.gov/mol-physiol/homepage.html.

[13] Mestl, T., A Mathematical Framework for Describing and Analysing Gene Regulatory Networks, Journal of Theoretical Biology, 1995, **176**:291-300.

[14] Savageau, M.A., Rules for the Evolution of Gene Circuitry, unknown, 1996.

[15] Othmer, H.G., The Qualitative Dynamics of a Class of Biochemical Control Circuits, Journal of Mathematical Biology, 1976, **3**:53-78.

[16] Mendoza, L. and Alvarez-Buylla, E.R., Dynamics of the Genetic Regulatory Network for Arabidopsis thaliana Flower Morphogeneis, Journal of Theoretical Biology, 1998.

[17] Thieffrey, D. and Thomas, R., Dynamical Behaviour of Biological Regulatory Networks - II. Immunity Control in Bacteriophage Lambda, Bulletin of Mathematical Biology, 1995, **57**:277-297.

[18] Loomis, W.F. and Sternberg, P.W., Genetic Networks, Science, 1995, **269**:649-649.

[19] Mjolness, E., Sharp, D.H., and Reinitz, J., A Connectonist Model of Development, Journal of Theoretical Biology, 1991, **152**:429-453.

[20] Takeuchi, Y., Global Properties of Lotka-Volterra Systems, 1998, Singapore, Wolrd Scientific.

[21] Hernandez-Bernejo, B. and Fairen, V., On the Differential Equations of Species in Competition, Mathematical Biosciences, 1997, **140**:1-32.

[22] Levine, W.S. editor, The Control Handbook, 1995, CRC Press, Boca Raton, Fl..

[23] Neumaier, A., Solving Ill-conditioned and Singular Linear Systems: A Tutorial on Regularization, SIAM Review, 1998, **40**:636-666.

[24] Tikhonov, A.N., Leonov, A.S., and Yagola, A.G., NonLinear Ill-Posed Problems, Vol. 1 and 2, 1998, London, Chapman and Hall.

[25] Brenig, L., Complete Factorisation and Analytic Solutions of Generalized Lotka-Volterra Equations, Physics Letters A, 1988, **133**:378-382.

[26] Cairo, L. and Feix, M.R., Families of Invariants of the Motion for the Lotka-Volterra Equations: The Linear Polynomials Family, Journal of Mathematical Physics, 1992, **33**:2440-2455.

[27] Hirsch, M.W., Systems of Differential Equations that are Competitive or Cooperative. II: Convergence Almost Everywhere, SIAM Journal on Mathematical Analysis, 1985, **16**:423-439.

[28] Smale, S., On the Differential Equations of Species in Competition, Journal of Mathematical Biology, 1976, **3**:5-7.

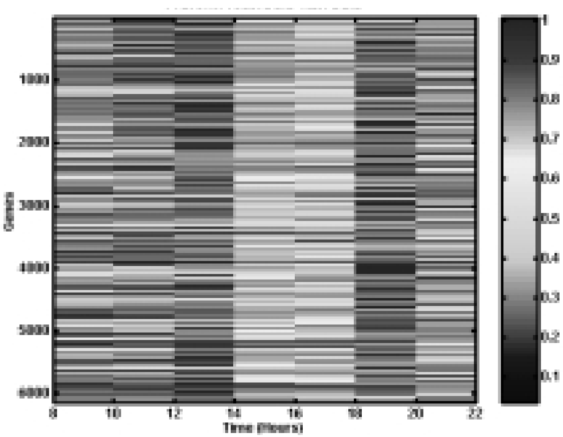

Figure 1

The rows are the 6153 ORFs from the entire genome of Sacchromyces cerevisiae[3]. The yeast have been given glucose and the abscissa is the time into the experiment. The false coloring represents the normalized expression levels of the different genes.

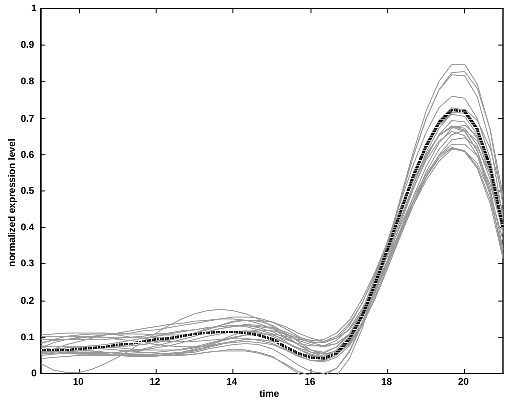

Figure 2

The 6153 yeast ORFs have been scaling clustered into 25 expression clusters where each cluster is scaling invariant. This is an example of one of the clusters.

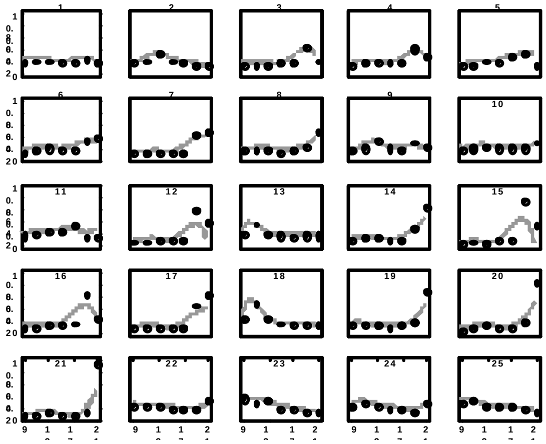

Figure 3

The reconstruction of the genetic regulatory network where each circle represents an experimental data point, the abscissa is time, and the lines are the integrated gene cluster dynamics (equation 3).

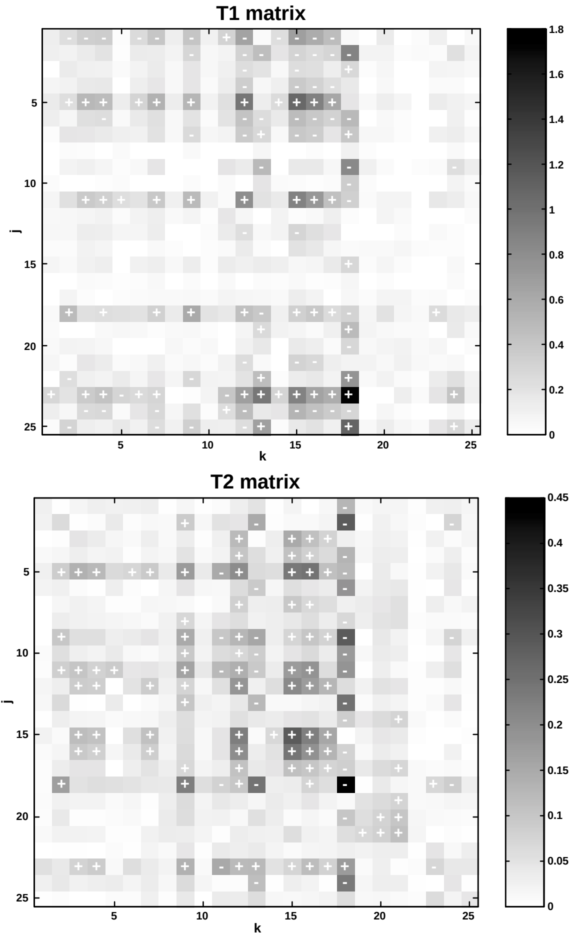

Figure 4

The T^1 and T^2 matrices from the reconstruction of the yeast grid. The y axis is j; the x axis is k. The false coloring indicates the magnitude of the elements as well as their sign. Positive values indicate activation; negative ones inhibition.