Acknowledgments

This work is supported by grants from the Lipper Foundation and DARPA-ONR Ultra Scale Computing Program.

References

1. McAdams H.H. and Shapiro L. (1995) Circuit Simulation of Genetic Networks. *Science* Vol.269, 4, 650-6, 1995.
2. Somogyi R. and Sniegoski C.A. (1996) Modeling the Complexity of Genetic Networks: Understanding Multigenic and Pleiotropic Regulation. *Complexity* page 45, 1996.
3. Liang S., Fuhrman S. and Somogyi R. (1998) REVEAL, A General Reverse Engineering Algorithm for Inference of Genetic Network Architectures. *Pacific Symposium on Biocomputing* 3:18-29, 1998.
4. Akutsu,T., Kuhara S., Maruyama O. and Miyano S. (1998) Identification of Gene Regulatory Networks by Strategic Gene Disruptions and Gene Overexpressions. *SIAM-ACM Symposium of Discrete Algorithms* 1998.
5. Thieffry D. and Thmos R. (1998) Qualitative Analysis of Gene Networks. *Pacific Symposium on Biocomputing* 3:77-88, 1998.
6. Chen T. et al.(1998) Identifying Gene Regulatory Networks from Experimental Data. *Submitted.*
7. Michaels G.S., Carr D.B., Askenazi M., Fuhrman S., Wen X., and Somogyi R. (1998) Cluster Analysis and Data Visualization of Large-Scale Gene Expression Data. *Pacific Symposium on Biocomputing* 3:42-53, 1998.
8. Wen X., Fuhrman S., Michaels G.S., Carr D.B., Smith S., Barker J.L., and Somogyi R. (1998) Large-Scale Temporal Gene Expression Mapping of CNS Development. *Proc Natl Acad Sci USA*, 95:334-339, 1998.
9. Gary M.R. and Johnson D.S. (1979) Computers and Intractability. *Published by W.H. Freeman and Company*, page 246, 1979.
10. Cho R.J., Campbell M.J., Winzeler E.A., Steinmetz L., Conway A., Wodicka L, Wolfsberg T.G., Gabrielian A.E., Landsman D., Lockhart D., and Davis R.W. (1998) A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle. *Molecular Cell*, Vol.2, 65-73, July 1998.

Theorem 4 *The solutions to Model 4 are of the following form:*

$$\begin{pmatrix} \mathbf{r} \\ \mathbf{p} \end{pmatrix} = \mathbf{Q}(t)e^{\lambda t}$$

where $\boldsymbol{\lambda}$ *are eigenvalues of* S*, and* $\mathbf{Q}(t)$ *is a matrix whose elements are polynomials on* t*.*

This theorem is in the same style as the other theorems we have proved, but apparently weaker: the constraints of the degree of $\mathbf{Q}(t)$ do not hold. All the interesting questions regarding stability of Model 4 can be answered through the studies on the set S. We will leave this discussion to the future.

Limitations of the models and the approaches

Like many other models, the Linear Transcription Model (Model 1) does not consider time delays in transcription and translation. This assumption greatly reduces the complexity of the problem. Although we make an effort to incorporate time delays, no interesting conclusion can be drawn. The most significant limitation comes from ignorance of other regulators such as metabolites. It is known that many genes and other factors directly or indirectly affect the pathway that feeds back to transcription. However, the Linear Transcription Model clearly captures more features of gene expression than other models to our knowledge.

The approach of the Fourier Transform for Stable Systems makes an assumption that gene expressions are periodic in cell cycles. This assumption does not hold for some genes, and cell cycle length may vary too. The other approach of MWSLE assumes the number of regulators of a gene is a small constant, but the actual number may be much larger than expected and the solution may be intractable computationally.

Conclusion and Future Work

In conclusion, we proposed a Linear Transcription Model for gene expression, and discussed two algorithms to construct the model from experimental data. Our future work will apply our methods to real experimental data, and continue to investigate a combined method to reconstruct these models, solutions to the Time Delay Model, and quantitative analysis of experimental designs.

The construction of Protein Model is similar to the MWSLE of Model 1. Because V, L, and U are all diagonal matrices, $(-LVL^{-1} - U)$ and $-LVL^{-1}U$ are also diagonal. Also LC is sparse because C is sparse. Thus

Theorem 3 *Model 3 can be constructed by solving MWSLE on time-series sampling of protein concentrations.*

Time-Delay Model

The real gene expression mechanism has time delays in transcription and translation. Let n-dimensional vectors $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_n)$ be delays for transcription and translation respectively, and $t = (t, t, \ldots, t)$ be the global time clock. Also, r and s are constants. Thus we define our model as

Model 4 *Gene expression with time delay can be modeled as*

$$\begin{aligned} \frac{d\mathbf{r}}{dt} &= \mathrm{C}\mathbf{p}(t-\alpha) - \mathrm{V}\mathbf{r}(t) \\ \frac{d\mathbf{p}}{dt} &= \mathrm{L}\mathbf{r}(t-\beta) - \mathrm{U}\mathbf{p}(t) \end{aligned}$$

Now we take the Fourier transform. We obtain:

$$\begin{aligned} i\xi\hat{\mathbf{r}}(\xi) &= Ce^{-i\xi\alpha}\hat{\mathbf{p}}(\xi) - V\hat{\mathbf{r}}(\xi) \\ i\xi\hat{\mathbf{p}}(\xi) &= Le^{-i\xi\beta}\hat{\mathbf{r}}(\xi) - U\hat{\mathbf{p}}(\xi) \end{aligned} \tag{12}$$

where ξ is the frequency, and $\hat{\mathbf{r}}$ and $\hat{\mathbf{p}}$ are Fourier transforms of $\mathbf{r}$ and $\mathbf{p}$. We simplify these two equations, obtaining

$$(i\xi\mathbf{I}_n + V)\hat{\mathbf{r}}(\xi) = Ce^{-i\xi\alpha}\hat{\mathbf{p}}(\xi) \qquad (i\xi\mathbf{I}_n + U)\hat{\mathbf{p}}(\xi) = Le^{-i\xi\beta}\hat{\mathbf{r}}(\xi)$$

where $\mathbf{I}_n$ is a $n \times n$ identity matrix. We combine these two equations

$$(i\xi\mathbf{I}_n + V)\hat{\mathbf{r}}(\xi) = Ce^{-i\xi\alpha}(i\xi\mathbf{I}_n + U)^{-1}Le^{-i\xi\beta}\hat{\mathbf{r}}(\xi)$$

Therefore $\hat{\mathbf{r}}(\xi)$ is a vector-valued distribution supported on the solutions to the following equation:

$$S = \{\xi \in \mathcal{C} \mid \det(i\xi\mathbf{I}_n + V - Ce^{-i\xi\alpha}(i\xi\mathbf{I}_n + U)^{-1}Le^{-i\xi\beta}) = 0\}$$

Therefore, we obtain the following theorem.

$$e^{-Ut}\int e^{Ut}L\mathbf{r}dt = C^{-1}\frac{d\mathbf{r}}{dt} + C^{-1}V\mathbf{r} \qquad (CC^{-1}C = C)$$

where C^{-1} is a (general) inverse of C. We differentiate this equation,

$$\begin{aligned}\frac{d^2\mathbf{r}}{dt^2} &= C(-U)e^{-Ut}\int e^{Ut}L\mathbf{r}dt + Ce^{-Ut}e^{Ut}L\mathbf{r} - V\frac{d\mathbf{r}}{dt} \\ &= C(-U)(C^{-1}\frac{d\mathbf{r}}{dt} + C^{-1}V\mathbf{r}) + CL\mathbf{r} - V\frac{d\mathbf{r}}{dt} \\ &= (-CUC^{-1} - V)\frac{d\mathbf{r}}{dt} + (-CUC^{-1}V + CL)\mathbf{r}\end{aligned}$$

Then we define our second model:

Model 2 *Gene expression can be partially modeled by the following dynamic system of mRNA concentrations.*

$$\frac{d^2\mathbf{r}}{dt^2} = (-\mathrm{CUC}^{-1} - \mathrm{V})\frac{d\mathbf{r}}{dt} + (-\mathrm{CUC}^{-1}\mathrm{V} + \mathrm{CL})\mathbf{r}$$

There exists one general inverse C^{-1} *that matches the real situation.*

The degeneracy of the transcription matrix C indicates that $\mathbf{r}$ cannot be self-determined. Any solution to $\mathbf{r}$ depends on the initial value of $\mathbf{p}$. This is consistent with our understanding that proteins (and other subsumed feedbacks) are major operators in transcription and translation, and thus determine the fate of gene expression. mRNA concentrations alone, handled in this manner at least, are not sufficient to model the whole system of gene expression.

Protein Model

Similarly to the argument in the RNA Model, we can eliminate $\mathbf{r}$ in Model 1, leaving a dynamic system of $\mathbf{p}$. The final equation is

$$\frac{d^2\mathbf{p}}{dt^2} = (-LVL^{-1} - U)\frac{d\mathbf{p}}{dt} + (-LVL^{-1}U + LC)\mathbf{p} \qquad (11)$$

Here, L is a non-degenerate diagonal matrix and its inverse L^{-1} exists. We define a model of proteins:

Model 3 *Gene expression can be modeled by the following dynamic system of protein concentrations.*

$$\frac{d^2\mathbf{p}}{dt^2} = (-\mathrm{LVL}^{-1} - \mathrm{U})\frac{d\mathbf{p}}{dt} + (-\mathrm{LVL}^{-1}\mathrm{U} + \mathrm{LC})\mathbf{p}$$

where v_{ii}, c_{i1}, ..., c_{in} are to be determined. The equations are underdetermined because $k < n$. There is no unique solution. However, by the argument we made that C is sparse, this problem belongs to the following category [9]:

Problem 1 MWSLE *(Minimum Weight Solutions to Linear Equations): Given k pairs $(\mathbf{a}_1, b_1), (\mathbf{a}_2, b_2), ..., (\mathbf{a}_k, b_k)$, where $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_k$ are n-tuple of reals and $b_1, b_2, ..., b_k$ are reals, is there a n-tuple $\mathbf{y}$ such that $\mathbf{y}$ has at most h non-zero entries and such that $\mathbf{a}_i \cdot \mathbf{y} = b_i$ for all i?*

Solving Equations 7-10 involves solving MWSLE: the unknowns v_{ii}, c_{i1}, ..., c_{in} corresponding to $\mathbf{y}$, and one equation corresponding to $\mathbf{a}_i \cdot \mathbf{y} = b_i$. Unfortunately, MWSLE is NP-complete, and therefore does not guarantee a polynomial time solution. However, there is a hope:

Lemma 1 *If h is a constant,* MWSLE *can be solved in $O(kn^h)$ time.*

Proof: There are at most n^h combinatorial choices of $\mathbf{y}$ which have at most h non-zero entries. For each choice of $\mathbf{y}$, it is over-determined because there are k linear equations but only h variables ($h < k$). The over-determined linear equations can be solved by using *least-square analysis*, which takes $O(k)$ time. Thus it takes $O(kn^h)$ to solve MWSLE. ∎

Without loss of generality, let h be the number of non-zero variables in c_{i1}, ..., c_{in}, which is to say that gene i has at most h regulators. We can apply Lemma 1 into Equations (6)-(9) and obtain the following theorem:

Theorem 2 *Model 1 can be constructed in $O(n^{h+1})$ time.*

The additional "+1" comes from solving n genes. The solution to $d\mathbf{p}/dt = L\mathbf{r} - U\mathbf{p}$ is straightforward, because L and U are diagonal matrices. Thus, Theorem 2 holds.

Extended Models and Solutions

RNA Model

Various recent techniques have focused on profiling mRNA concentrations. A straightforward method for studying the dynamic system in Model 1 is to eliminate the variables $\mathbf{p}$ and leave $\mathbf{r}$ as the only variables. We substitute $\mathbf{p} = e^{-Ut}\mathbf{p}_1$, and from the second equation, we obtain

$$L\mathbf{r} - U\mathbf{p} = \frac{d\mathbf{p}}{dt} = -Ue^{-Ut}\mathbf{p}_1 + e^{-Ut}\frac{d\mathbf{p}_1}{dt} = -U\mathbf{p} + e^{-Ut}\frac{d\mathbf{p}_1}{dt}$$

Therefore, we have $\frac{d\mathbf{p}_1}{dt} = e^{Ut}L\mathbf{r}$, and $\mathbf{p} = e^{-Ut}\int e^{Ut}L\mathbf{r}dt$. Substituting $\mathbf{p}$ into $\frac{d\mathbf{r}}{dt} = C\mathbf{p} - V\mathbf{r}$, we obtain

$$\frac{d\mathbf{r}}{dt} = Ce^{-Ut}\int e^{Ut}L\mathbf{r}dt - V\mathbf{r}$$

matrix $Q = \{q_{ij}\}$, so Equation 3 can be simplified as

$$\mathbf{x}(t) = Qe^{t\lambda} \tag{4}$$

We observe that at every cell cycle, many genes repeat their expression patterns. Although cell cycle lengths may vary even when the environmental conditions are held constant, the transcription analysis of the yeast mitotic cell cycle[10] revealed many similar expression patterns between two consecutive cell cycles. If the cell cycle period is τ, the *Fourier Series* of $\mathbf{x}(t)$ should have $1/\tau, 2/\tau, 3/\tau, ...$ as the frequencies, and every gene has a solution

$$x(t) = \sum_{j=-\infty}^{+\infty} a_j e^{i\frac{jt}{T}} \qquad (a_{-j} = a_j) \tag{5}$$

Equation 4 and 5 are equivalent: each eigenvalue corresponds to a Fourier frequency, and $a_{-j} = a_j$ eliminates non-real terms. We approximate $x(t)$ by the largest k periods, and thus

$$x(t) \approx x'(t) = \sum_{j=-k}^{k} a_j e^{i\frac{jt}{T}} \tag{6}$$

Applying $\mathbf{x}(t_1), ..., \mathbf{x}(t_k)$ into Equation 6, can solve variables $a_1, a_2, ..., a_k$, and thus $x'(t)$ can be uniquely determined. From $x'(t)$, we can approximate the matrix M in Model 1.

Minimum Weight Solutions to Linear Equations

Biologically, the transcription matrix C in Model 1 represents gene regulatory networks: $c_{ij} \neq 0$ indicates gene j is a regulator for the transcription of gene i, and $c_{ij} = 0$ indicates gene j is not a regulator for gene i. C is considered sparse because of many indications that the number of regulators for a gene is small[2], mostly less than 10. In other words, each row of C has only a few nonzero elements.

We apply $\mathbf{r}(t_0), ..., \mathbf{r}(t_k), \mathbf{p}(t_0), ..., \mathbf{p}(t_k)$ into $d\mathbf{r}/dt = C\mathbf{p} - V\mathbf{r}$, and approximate $d\mathbf{r}/dt$ by $\triangle\mathbf{r}/\triangle t$:

$$\frac{r_i(t_1) - r_i(t_0)}{t_1 - t_0} = c_{i1}p_1(t_1) + ... + c_{in}p_n(t_1) - v_{ii}r_i(t_1) \tag{7}$$

$$\frac{r_i(t_2) - r_i(t_1)}{t_2 - t_1} = c_{i1}p_1(t_2) + ... + c_{in}p_n(t_2) - v_{ii}r_i(t_2) \tag{8}$$

$$\cdots \tag{9}$$

$$\frac{r_i(t_k) - r_i(t_{k-1})}{t_k - t_{k-1}} = c_{i1}p_1(t_k) + ... + c_{in}p_n(t_k) - v_{ii}r_i(t_k) \tag{10}$$

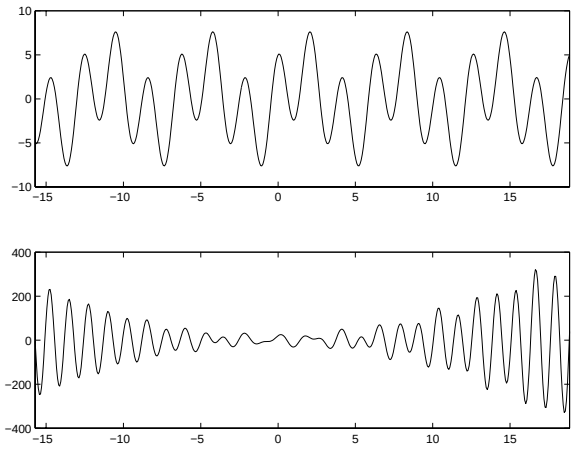

Figure 2: A stable system (top) and a semistable system (bottom).

Reconstructing Models from Temporal Data

Unfortunately, matrix M has yet to be determined because its sub-matrices are mostly unknown. In this section, we will discuss how to determine M from temporal experimental data. We will assume that we obtain a set of time-series samples of $\mathbf{x}(t_0), \mathbf{x}(t_1), ..., \mathbf{x}(t_k)$, where $\mathbf{x}$ includes both mRNA and protein concentrations.

Fourier Transform for Stable Systems

We refine the dynamic system $\mathbf{x}(t) = \mathbf{Q}(t)e^{t\lambda}$ in Theorem 1 to obey real biological meanings. The system is **unstable** if there exists a positive eigenvalue of $\boldsymbol{\lambda}$, because the term $q_{ij}(t)e^{\lambda_j t}$ is an exponential function if λ_j has a positive value. The system is **semistable** if all the real parts of the eigenvalues of $\boldsymbol{\lambda}$ are non-positive. As depicted at the bottom of Figure 2, a semistable system has a polynomial growth rate because of its polynomial term $q_{ij}(t)$. The system is **stable** if it is semistable and all the polynomials $q_{ij}(t)$ are constants. The top curve in Figure 2 shows a stable system.

The gene expression system has to be a *stable* system since an exponential or a polynomial growth rate of a gene or a protein is unlikely to happen. It implies that $q_{ij}(t)$ is actually a constant, denoted as q_{ij} for convenience. Let

where $C = \frac{df(\mathbf{p})}{d\mathbf{p}}|_{\mathbf{p}_0}$ and $\mathbf{s} = f(\mathbf{p}_0) - \frac{df(\mathbf{p})}{d\mathbf{p}}|_{\mathbf{p}_0}\mathbf{p}_0$. Therefore, we may study Equation 1 (near $\mathbf{p}_0$):

$$\frac{d\mathbf{r}}{dt} = C\mathbf{p} - V\mathbf{r} + \mathbf{s} \qquad \frac{d\mathbf{p}}{dt} = L\mathbf{r} - U\mathbf{p}$$

To eliminate $\mathbf{s}$ by variable substitution, we apply $\mathbf{r} = \mathbf{r} + \mathbf{r}_s$ and $\mathbf{p} = \mathbf{p} + \mathbf{p}_s$ into Equation 1 to calculate what constants $\mathbf{r}_s$ and $\mathbf{p}_s$ suffice to eliminate $\mathbf{s}$ and obtain

$$\frac{d\mathbf{r}}{dt} = C\mathbf{p} - V\mathbf{r} + (C\mathbf{p}_s - V\mathbf{r}_s) + \mathbf{s} \qquad \frac{d\mathbf{p}}{dt} = L\mathbf{r} - U\mathbf{p} + (L\mathbf{r}_s - U\mathbf{p}_s)$$

where $\mathbf{r}_s$ and $\mathbf{p}_s$ can be determined by the following equation:

$$\begin{pmatrix} -V & C \\ L & -U \end{pmatrix} \begin{pmatrix} \mathbf{r}_s \\ \mathbf{p}_s \end{pmatrix} = \begin{pmatrix} -\mathbf{s} \\ 0 \end{pmatrix}$$

Because both V and U, the degradation rates, are nonsingular diagonal matrices, we can assume the equation has a unique solution. Therefore it suffices to consider the following dynamic system even if $f(\mathbf{p})$ is nonlinear.

$$\frac{d\mathbf{r}}{dt} = C\mathbf{p} - V\mathbf{r} \qquad \frac{d\mathbf{p}}{dt} = L\mathbf{r} - U\mathbf{p} \tag{2}$$

We can define the Linear Transcription Model as

Model 1 *Let $\mathbf{x} = (\mathbf{r}, \mathbf{p})^T$ be variables for mRNAs and proteins, M be a $2n \times 2n$ transition matrix, and gene expression can be modeled by the following dynamic system:*

$$\frac{d\mathbf{x}}{dt} = M\mathbf{x} \qquad where \qquad M = \begin{pmatrix} -V & C \\ L & -U \end{pmatrix}$$

Solution to Linear Transcription Model

Assume M has $2n$ eigenvalues $\boldsymbol{\lambda} = (\lambda_1\lambda_2...\lambda_{2n})^T$. It is well-known that the dynamic system in Model 1 has the following solution:

Theorem 1 *The solution to Model 1 is of the form*

$$\mathbf{x}(t) = \mathbf{Q}(t)e^{t\lambda} \tag{3}$$

where $\mathbf{Q}(t) = \{q_{ij}(t)\}$ satisfies

$$\sum_{j=1}^{2n} \deg(q_{ij}(t)) + 1 \leq 2n \qquad for\ i = 1, 2, ..., 2n$$

$\mathbf{Q}(t)$ is a $2n \times 2n$ matrix whose elements are polynomial functions of t, and deg() returns the degree of a polynomial function.

molecule degrades randomly, and its components are recycled in the cell. One important feedback missing here is from metabolites to the transcription, which also plays a key role in signaling. Then, Figure 1 can be modeled as a nonlinear dynamic system:

$$\frac{d\mathbf{r}}{dt} = f(\mathbf{p}) - V\mathbf{r} \qquad \frac{d\mathbf{p}}{dt} = L\mathbf{r} - U\mathbf{p} \tag{1}$$

where the variables are functions of time t and defined as follows:

n	The number of genes in the genome;
$\mathbf{r}$	mRNA concentrations, n-dimensional vector-valued functions of t;
$\mathbf{p}$	Protein concentrations, n-dimensional vector-valued functions of t;
$f(\mathbf{p})$	Transcription functions, n-dimensional vector polynomials on $\mathbf{p}$;
L	Translational constants, $n \times n$ non-degenerate diagonal matrix;
V	Degradation rates of mRNAs; $n \times n$ non-degenerate diagonal matrix;
U	Degradation rates of Proteins, $n \times n$ non-degenerate diagonal matrix;

The change in mRNA concentrations ($d\mathbf{r}/dt$) equals the transcription ($f(\mathbf{p})$) minus the degradation ($V\mathbf{r}$), and similarly, the change in protein concentrations ($d\mathbf{p}/dt$) equals the translation ($L\mathbf{r}$) minus the degradation ($U\mathbf{p}$). Here, L, V and U are non-degenerate diagonal matrices, because we assume both the translation rates and the degradation rates are constants for each species. Also, we consider zero time delay in transcription and translation, and leave the time delay case to a later section.

Linear Transcription Model

First we assume the transcription functions, $f(\mathbf{p})$, to be linear functions of $\mathbf{p}$, $f(\mathbf{p}) = C\mathbf{p}$. For example, a combined effect of activators and inhibitors in transcription can be described by a linear function in the form of $w_a[activators] - w_i[inhibitors]$, where w_a and w_i are contributions of the activators and the inhibitors to the gene regulation. Otherwise, we can still make the assumption from the following argument.

We let $\mathbf{p}_0$ be the value of $\mathbf{p}$ at time zero, and take the first-order Taylor approximation:

$$\begin{aligned} f(\mathbf{p}) &= f(\mathbf{p}_0) + \frac{df(\mathbf{p})}{d\mathbf{p}}|_{\mathbf{p}_0}(\mathbf{p} - \mathbf{p}_0) \\ &= C\mathbf{p} + \mathbf{s} \end{aligned}$$

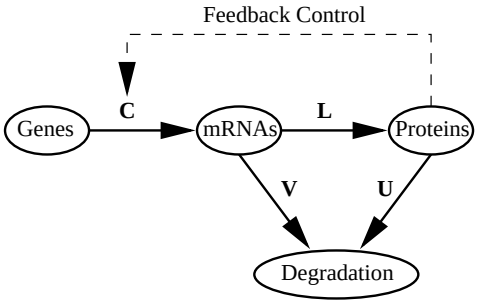

Figure 1: Simplified dynamic system of gene regulation emphasizing feedback on transcription.

mRNAs and proteins: Minimum Weight Solutions to Linear Equations and Fourier Transform for Stable Systems.

- We discuss three extended models: RNA Model, Protein Model and Time Delay Model, among which the Protein Model parameters can be reconstructed through a set of temporal samples of protein expression levels.
- Our results suggest that it is possible to determine most of the gene regulation in the genome level from a minor set of accurate temporal data.

Dynamic System for Gene Expression

The transcription of a gene begins with transcription elements, mostly proteins and RNAs, binding to regulatory sites on DNA. The frequency of this binding affects the level of expression. Experiments have verified that a stronger binding site will increase the effect of a protein on transcription rate. On the other hand, since the DNA sequence is unchanged, the transcription is mostly determined by the amounts of transcription proteins. In translation, proteins are synthesized at ribosomes. An mRNA can be translated into one or multiple copies of corresponding proteins, which can further change the transcription of other genes. A feedback network of genes, mRNAs and proteins is shown in Figure 1.

In Figure 1, we ignore other feedback such as mRNAs to genes, since we subsume such effects in the protein feedback indicated. We assume the translation mechanism is relatively stable (at least for a short time), so the feedback from proteins to mRNAs has no effect. Each mRNA and protein

It is widely believed that gene expression data contains rich information that could discover the higher-order structures of an organism and even interpret its behavior. Conceivably within a few years, a large amount of expression data will be produced regularly as the cost of such experiments diminishes. Biologists are expecting powerful computational tools to extract functional information from the data. Critical effort is being made recently to build models to analyze them.

One of the most studied models is the Boolean Network, where a gene has one of only two states (ON and OFF), and the state is determined by a boolean function of the states of some other genes. Somogyi and Sniegoski [2] showed that boolean networks have features similar to those in biological systems, such as global complex behavior, self-organization, stability, redundancy, and periodicity. Liang *et al.* [3] implemented a reverse-engineering algorithm to infer gene regulations and boolean functions by computing the mutual information between a gene and its candidate regulatory genes. Akutsu *et al.* [4] gave an algorithmic analysis of the problem of identifying boolean networks from data obtained by multiple gene disruption and gene over-expressions in regard to the number of experiments and the complexity of experiments.

In addition to the boolean networks, other models are also studied. Thieffry and Thomas [5] discussed a generalized logical model and a feedback-loop analysis. They suggested that a logical approach can be used to get a first overview of a differential model and thus help to build and refine the model. McAdams and Shapiro [1] proposed a nice hybrid model that integrates a conventional biochemical kinetic model within the framework of a circuit simulation. However, it is not clear how to determine model parameters from experimental data. Gene expression data can also be analyzed directly by statistical and optimization methods. Michaels *et al.* [7] measured gene expression temporally and applied statistical clustering methods to reveal the correlations between patterns of gene expression and phenotypic changes. Chen *et al.* [6] transferred experimental data into a gene regulation graph and imposed optimization constraints to infer the true regulation by eliminating the errors in the graph.

In practice, the determination of the networks has to (1) derive regulatory functions from a small set of data samples; (2) scale up to the genome level; and (3) take into account the time delay in transcription and translation. In this paper, we propose a linear differential equation model for gene expression and two algorithms to solve the differential equations. Potentially, our methods answer the practical questions in (1) and (2), and we also make an effort to incorporate (3). In summary,

- We propose a Linear Transcription Model for gene expression, as well as two algorithms to construct the model from a set of temporal samples of

Pacific Symposium on Biocomputing 4:29-40 (1999)

MODELING GENE EXPRESSION WITH DIFFERENTIAL EQUATIONS [a]

TING CHEN
Department of Genetics, Harvard Medical School
Room 407, 77 Avenue Louis Pasteur, Boston, MA 02115 USA
tchen@salt2.med.harvard.edu

HONGYU L. HE
Department of Mathematics, Massachusetts Institute of Technology
Room 2-487, Cambridge, MA 02139 USA
hongyu@math.mit.edu

GEORGE M. CHURCH
Department of Genetics, Harvard Medical School
200 Longwood Avenue, Boston, MA 02115 USA
church@salt2.med.harvard.edu

We propose a differential equation model for gene expression and provide two methods to construct the model from a set of temporal data. We model both transcription and translation by kinetic equations with feedback loops from translation products to transcription. Degradation of proteins and mRNAs is also incorporated. We study two methods to construct the model from experimental data: Minimum Weight Solutions to Linear Equations (MWSLE), which determines the regulation by solving under-determined linear equations, and Fourier Transform for Stable Systems (FTSS), which refines the model with cell cycle constraints. The results suggest that a minor set of temporal data may be sufficient to construct the model at the genome level. We also give a comprehensive discussion of other extended models: the RNA Model, the Protein Model, and the Time Delay Model.

Introduction

The progress of genome sequencing and gene recognition has been quite significant in the last few years. However, the gap between a complete genome sequence and a functional understanding of an organism is still huge. Many questions about gene functions, expression mechanisms, and global integration of individual mechanisms remain open. Due to the recent success of bioengineering techniques, a series of large-scale analysis tools have been developed to discover the functional organization of cells. DNA arrays and Mass spectrometry have emerged as powerful techniques that are capable of profiling RNA and protein expression at a whole-genome level.

[a] Published in 1999 Pacific Symposium of Biocomputing