Subject: dynamical systems
Open Problems in the Dynamics of the Expression of Gene Interaction Networks

Larry S. Liebovitch[1] and Vincent Naudot[2],

[1] Florida Atlantic University
Center for Complex Systems and Brain Sciences
Center for Molecular Biology and Biotechnology
Department of Psychology
Boca Raton, FL, 33431 USA

[2] Florida Atlantic University
Department of Mathematical Sciences
Boca Raton, FL, 33431 USA

April 14, 2010

Abstract: Genes influence the expression of each other through a complex, nonlinear, dynamical network of interactions. There are a number of interesting open questions about what kind of information can be determined about the structure and dynamics of this network from limited experimental data.

1 Introduction

DNA (deoxyribonucleic acid) is the molecule that contains the programming code for building, maintaining, and altering living cells. The information in the genes encoded in DNA is transcribed into mRNA (messenger ribonucleic acid) which is then translated to synthesize the proteins that form physical structures and chemical reactions in the cell. But biology is much more complex and adaptive than just such a linear feed-forward chain. The DNA, mRNA, and proteins are all linked together in dynamical interacting feedback networks, see [Liebovitch *et al.*, 2009]. For example, some genes make proteins, called transcription factors, that bind onto DNA and increase or decrease the expression of genes. Some forms of RNA, called micro RNAs, target and destroy mRNA sequences, so that they are never translated into proteins. New technology can simultaneously measure the expression of tens of thousands of genes. Because of the comprehensive amount of data that can now be collected these methods are called ”high throughput” technologies. They include the use of microarray chips that have small pieces of DNA that can bind RNA to identify which genes are being expressed, and real time polymerase chain reaction methods that amplify and then quantitatively assess which RNA is being expressed. Other new methods can also measure DNA-protein and protein-protein interactions, see [MacArthur *et al.*,, 2009]. Therefore, in principle, it should be possible to construct a graph of all these interactions, where the nodes of the graphs are the genes and the edges between them are the net result of all the DNA, RNA, and protein interactions that link those genes together. Understanding the structure and dynamics of such a gene-gene interaction graph would give us great insight into how biology works and how best we could intervene in the network to restore the proper function compromised by disease, see [Liebovitch *et al.*, 2007]. However, the limited information provided by current experimental methods is not sufficient to determine the full structure and dynamics of this gene-gene interaction graph. The open problems presented below ask what is the information about the gene-gene interaction graph that can be determined from the experimental data of gene expression.

To understand the problems described above our approach uses the notion of attractors in dynamical systems. Before stating the problems of this paper, we need to recall several concepts.

2 Attractors for flows and diffeomorphisms

When studying the evolution of a specific system from the theoretical point of view, we need to distinguish between continuous and discrete systems. The former is often the flow of vector field $\mathcal{X}$ defined on a manifold $\mathcal{M}$ and represented by an Ordinary Differential Equation of the form

$$\dot{\mathbf{x}} \quad = \quad F(\mathbf{x})$$

where $\mathbf{x} \in \mathcal{M}$ and F is a smooth function. The flow of $\mathcal{X}$ at the time $t \in \mathbb{R}$

$$\mathcal{X}_t : \ \mathcal{M} \to \mathcal{M}, \ \ \mathbf{x} \mapsto \mathcal{X}_t(\mathbf{x})$$

consists in taking the value of the solution of the above equation with initial $\mathbf{x}$ at the time t.

It is sometimes not appropriate to model the evolution of a system using the approach from ODE and discrete systems are used to model the system. A discret system is represented by a diffeomorphism

$$\Phi : \ \mathcal{M} \to \mathcal{M}, \ \mathbf{x} \mapsto \Phi(\mathbf{x}),$$

that is a one to one differentiable map from which its Jacobian at each point is also invertible. It is well known that a flow is always a diffeomorhpism. However, on the same manifold, a diffeomorhpism is not always the time 1 of a vector field. For instance if one considers an orientation reversing diffeomorphism, such a situation cannot coincide with the time 1 of a vector field on the same manifold. Another difference observed in many models in biology, ecology and finance is when a vector field possesses a homoclinic orbit, that is an orbit which is bi-asymptotic to a singularity. In the case of a flow, if a point is homoclinic, then the whole orbit is also homoclinic. This means that we have a least a curve that is included in the intersection of the stable and unstable manifold. In the case of a diffeomorphism, this generically does not occur and we observe tangle between the stable and the unstable manifold.

For each system, the main interest is with stationary states, i.e., attractors. We say that a set $\mathcal{A}$ is an attractor for a vector field $\mathcal{X}$ if there is an open set $\mathcal{B} \supset \mathcal{A}$ called the basin of $\mathcal{A}$ such that for all $\mathbf{y} \in \mathcal{B}$

$$\lim_{t\to\infty} \mathrm{dist}\Bigg(\mathcal{X}_t(\mathbf{y}), \mathcal{A}\Bigg) = 0,$$

where 'dist' is the Hausdorf distance. When considering a diffeomorphism, we say that $\mathcal{A}$ is a periodic attractor if there exists an integer $p \geq 1$ such that

$$\lim_{n\to\infty} \mathrm{dist}\Bigg(\Phi^{np}(\mathbf{y}), \mathcal{A}\Bigg) = 0,$$

for all $\mathbf{y} \in \mathcal{B}$.

Finding explicit solutions for an Ordinary Differential Equation is very hard in general. However, when concerning the same manifold, analyzing the dynamics of a flow is often easier than analyzing the dynamics of a diffeomorphism. In the next section we present a problem formulated by means of a linear diffeomorphism.

3 Statement of the Problem

Let x be a vector whose elements are the expression levels of the set of genes. At each time step n a function A describes how these genes interact. Thus, $x_{n+1} = A(x_n)$. These time steps are repeated many times, so that from an initial state x_0, the final state

$$\lim_{n\to\infty} A^n(x_0) \quad = \quad x_f \tag{1}$$

Typically, only x_f is known from the experimental data and both x_0 and A are unknown. (Sometimes, there may be some experimental data known about a limited number of the time steps x_n which provides a further extension of the simplest form of this problem.) There is not enough information to determine A from x_f, see discussion below. But, that does not mean that we have no information about A. **The question is: What information can we determine about A from just x_f?** These properties, for example, may include some of the group properties of the function A, or the statistical properties of A. The really hard part of this problem is to determine which are the properties of A

that can be determined from the corresponding properties of x_f. In what follows, we present a discussion to explain why this problem is hard to understand from the mathematical point of view.

3.1 A first attempt

The initial value $x = x_0$ is unknown and so that we may say that it can be chosen randomly. This way we can analyze Eq. (1) for typical values of x_0 and ignore "exceptional cases". For simplicity, we shall assume that the field the matrix is written is the the complex field $\mathbb{C}$. We then identify A with its matrix written in a given basis $(e_1, \ldots, e_n)$ and write A into its Jordan form:

$$A = M^{-1} \cdot D \cdot M,$$

where $D = S[\mathrm{Id} + N]$, S is a semi-simple matrix (i.e., diagonal in the present context), and N is a nilpotent matrix, that is there exists $m \leq n$ such that $N^m \equiv 0$. Moreover, Id is the identical matrix and M is an invertible matrix. Solving Eq. (1) amounts to solving

$$\lim_{n\to\infty} M^{-1} \cdot S^n (\mathrm{Id} + N)^n \cdot M \cdot x_0 = x_f.$$

Put

$$M \cdot x_0 = y_0, \quad y_f = M \cdot x_f.$$

The above equation takes the form

$$\lim_{n\to\infty} S^n (\mathrm{Id} + N)^n \cdot y_0 = y_f. \tag{2}$$

Observe that

$$(\mathrm{Id} + N)^n = \mathrm{Id} + \tilde{N}_n$$

where $\tilde{N}_n$ is another Nilpotent matrix. From this consideration, it is clear that in the statement of Eq. (1) the vector x_f is not arbitrary, but surely

$$x_f \in \mathrm{Im}(S),$$

that is the range of the matrix S. We introduce the set of eigenvalues of S, that is the element of S on the diagonal (counted with their multiplicity)

$$\mathrm{Spect}(S) = \{\lambda_1, \ldots, \lambda_n\}.$$

We can distinguish 4 difference cases.

[a] $|\lambda_j| < 1$: since $|\lambda_i|^n \to 0$ as $n \to \infty$, this implies that the corresponding entry for y_f vanishes

[b] $|\lambda_j| > 1$ since $|\lambda_i|^n \to \infty$ as $n \to \infty$, this implies that the corresponding entry for y_0 vanishes. In the case x_0 (and therefore y_0) is chosen randomly, we must exclude this case. This situation is possible only in the case where we restrict the the initial condition x_0 to a specific domain.

[c] $|\lambda_j| = e^{2i\pi\alpha}$ where $\alpha \in \mathbb{R}\backslash\mathbb{Q}$. In this case since the set

$$\{\lambda_j^m, \mid m \in \mathbb{N}\}$$

is dense in the unit circle, we don't have convergence of the sequences (y_n) and (x_n), therefore this case has to be rejected again.

[d] $|\lambda_j| = e^{2i\pi\alpha}$ where $\alpha \in \mathbb{Q}\backslash\mathbb{N}$. In this case we don't have convergence of the sequences (y_n) and (x_n). However, there exist an integer q such that $\lambda_j^q = 1$ and as we will see in case [e], if we consider A^q instead of A, we may have convergence of the sequences (y_n) and (x_n). In this case, A^n will converge to a period attractor.

[e] $\lambda_j = 1$. In this case we claim that the restriction of A on the corresponding eigenspace is the identity. If not, then the restriction of A on that space takes the form $D(u) = u + N_j(u)$ where $u \in \mathbb{C}^k$, k being the dimension of the eigenspace and N_j is a nilpotent matrix. Therefore for all integer n we have

$$D^n(u) = u + nN_j(u) + \mathcal{P}(N_j)(u)$$

where $\mathcal{P}(N)$ is a polynomial in the endomorphism N i.e.,

$$\mathcal{P}(N_j) = \sum_{\ell=2}^{k-1} \beta_j N_j^\ell.$$

This means that starting by a vector u_0 that we have have no convergence of the sequence (u_n) where $u_{n+1} = D(u_n)$ and since the components of u_n are components of y_n we then conclude that the sequences (y_n) and (x_n) do not converge unless $N \equiv 0$.

From the considerations above and after re-ordering the eigenspaces, we can conclude that the matrix D takes the form

$$D = \begin{pmatrix} \mathrm{Id}_k & 0 \\ 0 & C \end{pmatrix}$$

where Id_k is the identical matrix on $\mathbb{C}^k$ and C is a contraction matrix, i.e., there exists $0 < \alpha < 1$ such that

$$\|C(u)\| \leq \alpha\|u\|.$$

We illustrate the above with the following examples.

3.2 Examples

1) **Example I.** Take the following 2 × 2 matrix

$$D = \begin{pmatrix} 1 & 0 \\ 0 & \alpha \end{pmatrix}$$

where $\alpha = 1$. It is clear that

$$\lim_{n\to\infty} D^n = \left(\begin{array}{cc} 1 & 0 \\ 0 & 0 \end{array} \right).$$

Take now

$$M = \left(\begin{array}{cc} 3 & 7 \\ 2 & 5 \end{array} \right).$$

A straightforward computation show that

$$A^n = M^{-1} \cdot D^n \cdot M = \left(\begin{array}{cc} 15 - 14\alpha & 35 - 35\alpha \\ -6 + 6\alpha & -14 + 15\alpha \end{array} \right) \to \left(\begin{array}{cc} 15 & 35 \\ -6 & -14 \end{array} \right)$$

as n tends to ∞. It is clear from this example that by taking different initial vectors, we end up with different limits. For instance let us consider the following vectors:

$$e_1 = \left(\begin{array}{c} 1 \\ 0 \end{array} \right), \quad e_2 = \left(\begin{array}{c} 0 \\ 1 \end{array} \right)$$

with is example we clearly verify that

$$A^n(e_1) \to \left(\begin{array}{c} 15 \\ -6 \end{array} \right)$$

but

$$A^n(e_2) \to \left(\begin{array}{c} 35 \\ -4 \end{array} \right).$$

Moreover, even if the limit exists, we cannot determine the value of α. Finally the matrix M is unknown.

To make the problem clearer let's start with two additional specific examples.

1) **Example II.** Let A be the m x m adjacency matrix of the gene-gene interaction graph for m genes. Each element of A_{ij}, which is 0 or 1, determines if the gene j influences the gene i. At each time step $x_{n+1} = Ax_n$ and thus the final expression state

$$x_f = \lim_{n\to\infty} A^n(x_0).$$

Given only x_f we cannot determine the elements A_{ij} of the matrix A. We cannot even determine all the eigenvalues or eigenvectors of A. However, perhaps surprisingly, we can determine some important properties of A. For example, x_f is the eigenvector associated with the largest eigenvalue of A. Moreover, [Shehadeh *et al.*, 2006] showed through numerical simulations, for some specific cases, that there is a relationship between the statistics of the elements of A and the statistics of x_f. Let the density of non-zero elements in A be distributed uniformly in the rows of A, and the total number of non-zero elements in each row be proportional to

$$(\frac{1}{m-r})^{[1/(c-1)]},$$

where r is the row number. Then, for the graph defined by A, the in-degree distribution $g(k)$, that is, the number of nodes receiving incoming edges from k nodes, has the power law form

$$g(k) \propto k^{-c}.$$

This leads to a x_f whose probability density function, $pdf(x)$, is a power law, namely,

$$pdf(x) \propto x^{-d}$$

where d is a function of c. These results were compared to the experimental data of mRNA expression as measured by Affymetrix microarray chips. This was done by comparing the probability density function of the mRNA levels on the chips with those computed from the attractors of the matrices with different in-degree and out-degree distributions. Since these pdfs are distributions with long tails, we compared the slopes of those tails, on logarithmic-logarithmic plots, for both the experimental data and the computed attractors.

Comparison with the experimental data suggests that $c \simeq d \simeq 2$, see [Shehadeh *et al.*, 2006] for more details. Thus, the statistical properties of the elements of A define a type of graph with certain a degree distribution, which generates a certain statistical property in x, namely the $pdf(x)$. Again, the hard part of this problem is to define what properties of A can be determined given only the resultant vector x_f. Even for this simplest case, where A is an m x m matrix, this problem is beyond the scope of the usual matrix properties, such as eigenvectors or eigenvalues. There are also computational issues as the total number of interacting genes m is of the order of 40,000. The problem is to define relevant properties (e.g. group, statistical, or other properties) that give us insight into the local or global properties of gene interactions as evidenced only by the relevant properties of the final state of the levels of gene expression. For example, these statistical properties might be the probability density distribution of the values of all the elements of the matrix, or of the elements in a row or column, or the probability density distribution of all the non-zero elements, or of the elements in a row or column, or the higher moments of such distributions in the whole matrix or in parts of the matrix.

2) **Example III.** If there are multiple experimental time series of the expression values as a function of time, then a set of basis functions can be used to identify dynamic functional linkages in the network defined by the adjacency matrix A. For example, singular value decomposition, SVD, has been used to identify limit cycles corresponding to harmonic

oscillators from the mRNA expression data in the cell mitotic cycle in yeast, see [Alter, 2006]. SVD provides a set of orthogonal basis vectors and their eigenvalues that together span the details of the experimental data thereby summarizing the details of the data. They showed that some of these eigenvalues correlated with the activities of molecules previously identified to be the oscillators whose period corresponds to the changes in the cell, the mitotic cycle, as the cell makes copies of its internal constituents and then splits into two daughter cells. These results were shown to be robust to the perturbations and experimental errors in the experiments. The relationship of these basis functions to the other mathematical properties of A and even more so, to the other mathematical dynamical properties on A, is not clear.

4 Experimental Information

For each of the specific problems described in the following section there are three different cases which correspond to different amounts or types of experimental data that may be available about x_f (and also possibly x_n).

1) **Case I.** There is only one experiment, starting from unknown initial levels of gene expression x_0, that converges to a known levels of gene expression in x_f.

2) **Case II.** There are multiple replications of the experiment (presumably each with different unknown initial conditions of the levels of gene expression x_0). Each experiment converges to the same levels of gene expression x_f, or the experiments converge to p different final levels of gene expression $x_f^{\{1\}}, \ldots, x_f^{\{p\}}$. These p attractors may represent some, or all, of the possible attractors of the system.

3) **Case III.** The levels of gene expression x_n do not converge to a final steady state and information is available at multiple time points x_n during the course of the experiments. These time points, may or may not, fully resolve the time behavior of the levels of gene expression. This case extends the simplest form of the problem to broader issues about the dynamics of this system.

5 Theoretical Models of Gene Interaction

For all of the three cases described above, the problem is to determine what properties of A can be determined from what properties of x_f (and also possibly x_n), when we make the following assumptions about the nature of A.

1) **Open Problem I. Determine A where A is the Adjacency Matrix.** This model, often commonly used in systems biology, assumes that genes

influence each other through linear, additive interactions that do not depend on the state of other genes. That is, the contribution of gene j to the expression of gene i is proportional only to x_j and the effect of all the other genes on gene i is the sum of $A_{ij}x_j$.

2) **Open Problem II. Determine A where A is a Nonlinear Function of the Gene Expression of Each Gene.** Biochemical reactions are typically not linear functions and it is more likely that the contribution of each gene j to the expression of gene i is a nonlinear function. A typical such function is that the rate of expression of x_i is proportional to $x_j^h/[1+(x_j/x_{j0})^h]$, where h is the Hill coefficient and x_{j0} is a constant, if gene j stimulates gene i, or $1/[1+(x_j/x_{j0})^h]$ if gene j inhibits gene i. The Hill coefficient is a way of describing the nonlinearity in a chemical reaction. It describes how the amount of the product produced depends on the amount of the reactants in the chemical reaction. For typical biochemical reactions the Hill coefficient is in the range [1,4].

3) **Open Problem III. Determine A where A is a Nonlinear Function of the Gene Expression of Multiple Genes.** Experimental evidence demonstrates that biology is often "context dependent", namely, that the effect produced by the binding of any one molecule A on molecule B is influenced by the presence of a third molecule C. For example, the effect of the binding of transcription factors on the regulatory region of a gene is influenced by the other transcription factors already bound at nearby sites on the DNA, see [Barash *et al.*,, 2003], In this case A is a nonlinear function of two or more of the gene expression levels, x_j, $j = 1, 2, ...q$, $q \geq 2$.

6 Conclusions

The levels of gene expression observed experimentally depend on interactions between many genes executed at the DNA, RNA, and protein levels. There is insufficient information to determine the local and global properties of the gene-gene interaction network A from the final expression x_f of just a few experiments. But, that does not mean that no properties of A can be determined. In fact, important mathematical properties of this network can be determined from that data. For example, [Shehadeh *et al.*, 2006] demonstrated that, for some networks, the statistical properties of the elements of the matrix form of A are related to the statistical properties of x_f. The hard part is to determine, which existing or newly defined properties of these interaction functions A can be determined from the existing or newly defined properties of x_f. If mathematical properties can be found that have important and useful biological relevance, they may give us deep insight into how these networks function and how we could alter them to cure diseases.

References

Liebovitch, LS., Shehadeh. LA., Jirsa, VK., Hütt M-T., Marr, C. [2009] ”Determining the Properties of Gene Regulatory Networks from Expression Data,” *Handbook of Research on Computational Methodologies in Gene Regulatory Networks.* Edited by Das S, Caragea D, Welch S, Hsu WH, (IGI Global, Hershey PA), in press.

MacArthur, BD., Ma'ayan, A., Lemischka, R. [2009] ”Systems biology of stem cell fate and cellular reprogramming,” *Nature Rev Mol Cell Biol* **19**, 672-681.

Liebovitch, LS., Tsinoremas, N., Pandya, A. [2007] ”Developing combinatorial multi-component therapies (CMCT) of drugs that are more specific and have fewer side effects than traditional one drug therapies,” *Nonlinear Biomedical Phys* **1**-11, doi:10.1186/1753-4631-1-11.

Shehadeh, LA., Liebovitch, LS., Jirsa, VK. [2006] ”Relationships between the global structure of genetic networks and mRNA levels measured by cDNA microarrays,” *Physica A* **364**, 297-314, doi:10.1016/j.physa.2005.08.069

Alter, O. [2006] ”Discovery of principles of nature from mathematical modeling of DNA microarray data,” *Proc. Natl. Acad. (USA)***103**, 16063-16064, doi:10.1073/pnas.0607650103.

Barash, Y., Elidan, G., Friedman, N., Kaplan, T. [2003] ”Modeling dependencies in protein-DNA binding sites,” *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology.* Edited by Vingron M, Istrail S, Pevzner P, Waterman M. pp. 28-37, Berlin, http://portal.acm.org/citation.cfm?id=640079