Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica A 364 (2006) 297–314
PHYSICA A
www.elsevier.com/locate/physa

Relationships between the global structure of genetic networks and mRNA levels measured by cDNA microarrays

Lina A. Shehadeh[a], Larry S. Liebovitch[a,b,c,d,*], Viktor K. Jirsa[a,e]

[a]Center for Complex Systems and Brain Sciences, Florida Atlantic University, Boca Raton, FL 33431, USA
[b]Center for Molecular Biology and Biotechnology, Florida Atlantic University, Boca Raton, FL 33431, USA
[c]Department of Psychology, Florida Atlantic University, Boca Raton, FL 33431, USA
[d]Department of Biomedical Sciences, Florida Atlantic University, Boca Raton, FL 33431, USA
[e]Department of Physics, Florida Atlantic University, Boca Raton, FL 33431, USA

Received 15 April 2005; received in revised form 16 August 2005
Available online 10 October 2005

Abstract

Previous studies of genetic interactions have focused on measuring and analyzing the level of mRNA expressed by each gene and then finding statistical correlations between them. Here we focus on assessing the global pattern, not the level of activity of specific genes. We formulate models of different architectures of genetic interactions, including random and scale free patterns with homogeneous, heterogeneous, and symmetric connection topologies. We then compute the probability density function (PDF) of the mRNA levels produced by each model. Thus, we make a dictionary of the mRNA patterns produced by different architectures of genetic interactions. Then we compare these results to the statistics of mRNA levels experimentally measured by cDNA microarrays. We then read our dictionary backwards, starting from the statistical patterns of the mRNA levels, to find the pattern of genetic interactions responsible for the observed experimental data. We find that the data is best represented by a model where different genes have different patterns of input regulation from the other genes but the same patterns of output regulation to the other genes.

Keywords: Microarrays; mRNA; Genetic network; PDF; Connectivity matrix; Small world

1. Introduction

Biological organisms are complex systems in the natural world which contain an internal description of their evolution, development, structure, and function encoded in the DNA sequences of their genes. As in other complex systems, their functions and responses to environmental conditions arise not only from the structural and biochemical properties of their individual components, but also from the collective interactions of the system as a whole. Consequently, a key emerging question in biology is: How do thousands of genes and their

*Corresponding author. Department of Psychology, Center for Complex Systems and Brain Sciences, Florida Atlantic University, Boca Raton, FL 33431, USA. Tel.: +1 561 297 2239; fax: +1 561 297 2223.
E-mail address: liebovitch@ccs.fau.edu (L.S. Liebovitch).
URL: http://www.ccs.fau.edu/~liebovitch.

0378-4371/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2005.08.069

RNA and protein products function in a complicated and orchestrated way in a given living organism? That is: How are the complex genetic networks of various organisms organized to function as robust and dynamical systems?

Genes interact in the following way. The genetic sequence is transcribed into mRNAs which are then translated into proteins by ribosomes in the cytoplasm. These proteins can then bind to DNA and promote or inhibit the expression of other genes. The expression level of a gene is reflected in the amount of its corresponding mRNA present in the cell. Traditional methods in molecular biology generally work on a "one gene in one experiment" basis, which means that the whole picture of gene function is hard to obtain. In the past several years, a new technology, called cDNA microarrays [1,2], has been developed that makes it possible to measure the mRNA levels of thousands of genes simultaneously, which is why this is called high-throughput technology. For example, microarrays have been used to measure the expression profile of some or all the 13,600 genes in the fruitfly *Drosophila Melanogaster* at different points in time under different environmental conditions (for example [3]). A central problem in molecular biology today is to develop methods to analyze this huge load of data so that we can understand the patterns of genetic interactions that tell us about the web of simultaneous, complex connections in biological systems.

Current analyses of patterns of genetic interactions have focused primarily on statistical measures such as correlation or cluster analysis of the mRNA levels expressed by genes [4–11]. These approaches have concentrated on the role of specific genes and on grouping genes into clusters based on the waves of co-expression. A complementary approach, that we present here, is to find a pattern that is consistent with the pattern of interactions between the genes. Each approach has its advantages and limitations. The traditional "bottom-up" approach can identify specific genes, but cannot deal well with the complex interactions between them. Our new "top-down" approach deals with the global pattern of genetic interactions, but cannot deal well with the role of specific genes in that pattern.

Since the pioneering work of Jacob and Monod [12] on the *lac* operon there have been quantitative studies on the interactions of small numbers of genes. Lately, there have been attempts to reconstruct models for gene regulatory networks on a global, genome-wide scale using ideas from genetic algorithms [13], neural networks [14], and Bayesian models [15]. Most recently, researchers have used linear models and singular value decomposition [16,17,7,18–21] to reverse-engineer the architecture of genetic networks. Quantitative models of these genetic regulatory networks have used Boolean networks, dynamical systems, and hybrid approaches. In Boolean models the state of each gene is either on or off [22–32]. These have been useful in understanding the role of feedback loops and multiple stable states of gene expression and have also been used to analyze experimental data from co-regulation studies. The dynamic features of Boolean models are less complete than dynamical system models based on differential equations where the states of the genes can take on a continuous range of values and can also include stochastic fluctuations in the numbers of molecules [33–43]. Hybrid models have also combined both the Boolean and dynamical systems approaches [44–46].

Our models of the global interactions between genes were motivated by the NK model of Kauffman [25], the "small world" network models of Watts and Strogatz, and scale free networks studied by Barabasi et al. [47–49].

In his NK model, Kauffman assumes that there are N genes, each of which regulates K other genes. If $K = 0$, then each gene has no influence on any other gene. Since a mutation in any one gene does not effect any other genes, it will have only a small effect on the progeny. If $K = N$, then each gene influences all the other genes and so a mutation in one gene affects all the genes, and the change is so extreme that the new progeny will die. As K increases from 0 to N, there is a phase transition in mutations from those having too little an effect to mutations having too much of an effect. He argued that evolution will adjust K so that it is at the value of the transition point, so that evolution is efficient. From his models, he estimated that this transition occurs at $K = 2$, that is, each gene should influence about two other genes.

A more complex and realistic model of genetic interactions must take into account the fact that some genes have a wider influence than other genes. But what is the pattern of these interactions? The network of genetic interactions consists of nodes (genes) and links (genetic connections) between them. Many other natural networks also consist of such nodes and links. There have been several studies on such networks such as the pages on the Internet linked by their URLs, the substations of the electrical power grid of the western United States linked by their transmission lines, actors linked by the number of movies that they appeared in together,

neurons linked by their synapses or gap junctions in the nematode worm *C. elegans*, and the substrates in metabolic pathways linked by the number of reaction pathways that separate them. All these examples of networks have a known network topology [47–49]. These networks have structures that are "scale-free" or "small world" or both [52]. Structures in scale free networks extend over a wide range of scales. They can arise when the nodes and their connections in such networks have growth and preferential attachment or rearrangement over time resulting in structure formation on all scales [21,50,51]. Scale-free networks have a hierarchy of connections that is self-similar across different scales or levels of structure. If $g(k)$ nodes have k links, these networks obey a power law distribution $g(k) = Ak^{-a}$. In "small world" networks, relatively few links are needed to go from one node to any other node in the entire network.

Therefore, in developing analogous models of genetic regulatory networks we study models with different distributions $g(k)$, where $g(k)$ genes are regulated by k other genes, just as $g(k)$ nodes have k links to other nodes in these other network models.

2. Approach

There is no general method to compute the global architecture of genetic interactions from their observed mRNA levels. Therefore, the main idea of our work was to formulate different models of genetic interactions, compute the global statistics of the mRNA expression levels of each model, compare those to the same statistical measures from cDNA microarray experimental data, and so deduce the topology of the underlying genetic networks.

In order to accomplish this we first formulate five different types of models of genetic networks. These models include both random and scale free models, with homogeneous, heterogeneous, and symmetric connection topologies referring to the input and output regulatory connections between genes. These different models represent possible biological patterns of genetic regulation. Then we compute the steady state mRNA levels from each model and evaluate the probability density function (PDF), that is the probability that the mRNA levels, x, are between x and $x + \mathrm{d}x$. We form a dictionary whose entries are the PDFs of the mRNA expression levels of the different models. Next, we determine the PDF of 54 sets of mRNA expression levels. We then read our dictionary backwards, starting from the PDFs of the theoretical mRNA levels, to find the model of genetic interactions that best matches the observed statistics of the experimental data.

3. Models of genetic interactions

For a model with N genes and k connections we represent the genetic interactions by an $N \times N$ connection matrix M. If the jth gene regulates the expression of the ith gene, then $M_{ij} \neq 0$, otherwise $M_{ij} = 0$. A gene can also regulate its own expression. However, in our extremely sparse matrices, self-connections, though possible, are very unlikely to exist. We choose the matrix elements of M so that $g(k)$ genes are regulated by k other genes. The structure of the genetic regulatory network between the genes is described by the distribution $g(k)$ that defines how many genes, $g(k)$, interact with k other genes. For example, if every gene regulates exactly 6 other genes, then $g(k) = N$ for $k = 6$, and $g(k) = 0$ for all other k. If an ever smaller number of genes have an ever wider influence on other genes, then $g(k)$ might have the power law form, characteristic of a scale-free network, that $g(k) = Ak^{-a}$. Different forms of the matrix M can produce the same distribution of genetic interactions $g(k)$. The distribution $g(k)$ can refer to the regulatory input from the genes, the regulatory output of the genes, or both. The elements X_i of the column vector $\mathbf{X}$ represent the mRNA expression levels of the N genes.

An $N \times N$ connection matrix M means that we restrict ourselves to linear interactions between the genes. The nonzero elements in M are positive elements if they represent stimulatory genetic regulation, or negative if they represent inhibitory genetic regulation. Connectivity matrices with both stimulatory and inhibitory regulation can cause a broad range of complex behavior, including steady state, periodic, and quasi-periodic dynamical behavior. We are therefore faced with a trade off between studying less complex models that we can successfully analyze and more complex models which would be much more difficult to analyze. Therefore, in these initial studies, we restrict ourselves to connection matrices with stimulatory behavior where $M_{ij} \geqslant 0$ so that the mRNA levels will always reach a steady state with a characteristic PDF that is the object of our

analysis and which can serve as a starting point for more complex models. Furthermore, in this present work, we use a constant strength for the connections in each matrix. We know, on one hand, that this is a simplification of a realistic biological network where the strength of regulatory connections is influenced by cellular as well as environmental processes. We wish to include weak as well as strong regulatory connections in future work. On the other hand, there is an advantage over the Boolean system where the genes are either on or off. The quantitative interactions in our models are more realistic in terms of biochemical networks.

Both our networks and random Boolean networks (RBNs) are first-order networks that depend on pairwise interactions between the nodes. Although our network model is a more quantitative approach than the RBN (due to its larger state space), it is restricted to less types of interactions that are possible in a RBN. While our models are based on the distribution or topology of the genetic network, RBNs are used as explicit simulation schemes for the dynamics of a genetic system. It would be interesting to translate the dynamics into expression patterns and compare them with the expression patterns that we generate from different network topologies. Particularly in Ref. [28], it seems that when limiting the interaction function of RBN to a specific subset of nodes, the dynamic behavior is dominated by the topology of the underlying graph. In this sense, our models can be viewed as an extreme case of RBN, as we exclude all effects of specific interactions and attempt to discuss universal features of the linkage between topology and gene expression pattern.

We also consider two general kinds of matrices:

Markovian matrix: If M_{ij} is the probability that a gene switches from one expression level to another, then this connection matrix can be thought of as a Markovian probability matrix [53] where $0 \leqslant M_{ij} \leqslant 1$ and $\sum_i M_{ij} = 1$ that is, the sum of the elements M_{ij} in each column is equal to 1. The advantage of using these constraints on M is that Markovian matrices have well known analytical properties that can help us in the computation and in understanding how the steady state mRNA levels depend on M. For example, this formulation guarantees that the transition matrix will always have an eigenvalue of 1 and a corresponding stationary eigenvector as explained in Appendices A, B and C. On the other hand, there are also disadvantages of this Markovian formulation. First, since all $M_{ij} \geqslant 0$, it cannot model inhibitory regulation. Second, the constraints on M mean that we cannot independently set (1) the strengths of the interactions between the genes and (2) the distribution of genetic interactions that $g(k)$ genes interact with k other genes. For example, since the sum of the elements in each column must be equal to 1, if there are m_j equal nonzero elements of M_{ij} in column j, then each element $M_{ij} = 1/m_j$. If there are different numbers of nonzero elements in different columns, then $1/m_j$ will be different in each column j and so the input from each gene cannot have the same strength.

Non-Markovian matrix: The inability of the Markovian matrix to represent arbitrary genetic interactions has lead us to also use a non-Markovian formulation for M. There are no constraints on our choice of the elements M_{ij} in this formulation and we choose the elements M_{ij} to represent different distributions of genetic interactions $g(k)$. The advantage of the non-Markovian formulation is that we can develop much more complex models of genetic interactions. The disadvantage is that the mRNA expression levels may grow or decay exponentially as they are evolved in time.

In designing the scale free distributions of the connectivity matrices, $g(k) = Ak^{-a}$, we vary the slope, a, of the distribution from 1 to 6. We derive a smooth equation for the density of nonzero elements and fill up the matrix accordingly. As described in Appendix D, the density equation used to fill up the 1000×1000 matrices, in the desired hierarchical distribution is

$$Y = \left[\frac{A}{1001 - R}\right]^{1/(a-1)}, \tag{1}$$

where Y is the density of nonzero elements per matrix line, R is the distance between the first cell in the matrix, M_{11} and the desired matrix line, and the constant A determines the total number of gene interactions. A line in the matrix refers to a row in the homogeneous model, a column in the heterogeneous model, and a diagonal in the symmetric model.

We formulate five different types of models. We name the models by the pattern of output or input regulation into the genes. Two of the models have random connections between genes. Three of the models have both scale free and "small world" connectivity. Scale free forms, which have a power law scaling of

$g(k) = Ak^{-a}$, are not the same as "small world" networks in which distant nodes can be reached through relatively few connections [52]. However, we will show later that these particular scale free networks in our models do, in fact, have "small world" properties. In four models the pattern of input regulation into the genes differs from the pattern of output regulation of the genes and in one model the patterns of input and output regulation are the same.

Model A1. Random output: The output of each gene regulates six other genes chosen at random. Each gene is regulated by the input of an average of 6 other genes. The distribution of genetic interactions is the single point distribution $g(k) = N$ for $k = 6$ and $g(k) = 0$ for all other k. In each column, the connection matrix M has 6 equal, randomly chosen elements $M_{ij} = \frac{1}{6}$. All the other M_{ij} elements are zero. This is the only model among the five models that has a connection matrix with Markovian properties.

Model A2. Random input: Each gene is regulated by the input from six other genes chosen at random. The output of each gene regulates an average of 6 other genes. The distribution of genetic interactions is the single point distribution, $g(k) = N$ for $k = 6$ and $g(k) = 0$ for all other k. In each row, the connection matrix M has 6 equal, randomly chosen elements $M_{ij} = \frac{1}{6}$. All the other M_{ij} elements are zero.

Model B. Scale free homogeneous input: In this model, the pattern of input regulation into each gene is the same. The distribution of the regulatory output from the genes has the scale free form $g(k) = Ak^{-a}$. Each gene i has a few inputs from the lower order j genes, more inputs from the middle order j genes, and many more inputs from the higher order j genes. The input from each gene j that regulates gene i has $M_{ij} = 1$ and all other $M_{ij} = 0$.

Model C. Scale free heterogeneous input: In this model, the pattern of input regulation into each gene is different. The distribution of the regulatory input into each gene has the scale free form of $g(k) = Ak^{-a}$. As the order of the gene i increases, there are an ever larger number of inputs from j genes. The input from each gene j that regulates gene i has $M_{ij} = 1$ and all other $M_{ij} = 0$.

The homogeneous and heterogeneous models are related to each other. The outputs into the genes in the homogeneous model and the inputs from the genes in the heterogeneous model have the same scale free form of the distribution of genetic interactions $g(k)$. Taking the transpose of M, that is interchanging the rows and columns, reverses the direction of the genetic interactions. Thus M in the homogeneous model is the transpose of M in the heterogeneous model. This suggests we also consider a model with symmetric input and output distributions.

Model D. Scale free symmetric input/output: The distribution of genetic interactions $g(k)$ of both the input into and the output of each gene is the same. The input from each gene j that regulates gene i has $M_{ij} = 1$ and all other $M_{ij} = 0$.

The connection matrices M for these five models are illustrated in Fig. 1 for $N = 1000$ genes. The nonzero elements are black dots (each of which is larger than the resolution of one row or column). The input and output distributions of genetic connections, $g(k)$, are shown in Figs. 2 and 3. We start with an $N = 10$ matrix and then increased N in logarithmic steps to 1000. The results were not sensitive to N. In the models presented here, all the matrices have $N = 1000$, as compared to 13,600 genes in the fruitfly *D. melanogaster* and about 50,000 genes in humans. The number of nonzero elements in each matrix was normalized so that each model had exactly a total of 6000 connections between genes. Therefore, the average number of connections per node in each network $\langle k \rangle = 6$. We chose $\langle k \rangle = 6$ in approximate agreement with the popular notion of "6 degrees of separation" in social networks [54].

4. Evidence of "small world" properties in our models

Two important properties of a "small world" network are a short characteristic path length and a high clustering coefficient [52].

The average path length, ℓ_{path}, of a network is defined as the number of edges in the shortest path between two nodes, averaged over all pairs of nodes [49]. Given the same number of nodes, N, and the average number of connections, $\langle k \rangle$, both a random network and a "small world" network have a similarly small average path length, ℓ_{path}.

The clustering coefficient C_i of node i that has k_i edges which connect it to k_i other nodes, is equal to the ratio between the number E_i of edges that actually exist between these k_i nodes and the total number of all

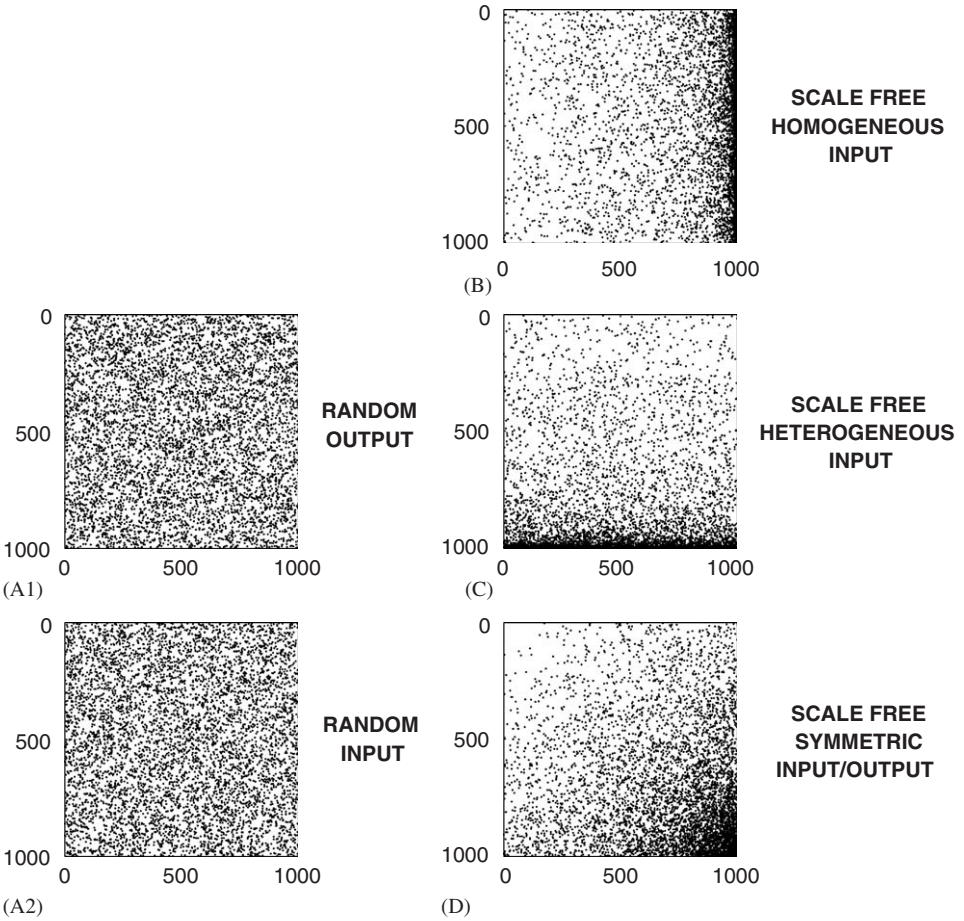

Fig. 1. Models of genetic interaction. Black dots represent the nonzero elements M_{ij} indicating that the jth gene regulates the ith gene. (Each dot is larger than the resolution of one row or column.) The connection matrix is 1000×1000 in each model. (A1) Random output: The output of each gene regulates 6 other genes chosen at random. (A2) Random input: Each gene is regulated by the input from 6 other genes chosen at random. (B) Scale free homogeneous input: The pattern of input connections into each gene is the same. (C) Scale free heterogeneous input: The pattern of input connections into each gene is different. (D) Scale free symmetric input/output: The pattern of inputs and outputs of each gene is the same. For the scale free models here $a = 2$.

possible edges $k_i(k_i - 1)/2$,

$$C_i = \frac{2E_i}{k_i(k_i - 1)}. \tag{2}$$

The clustering coefficient of the whole network is the average of all the C_i's from every node i [49]. Given the same number of nodes, N, and average number of connections, $\langle k \rangle$, a "small world" network has a higher clustering coefficient than a random network.

For each nonrandom model (B, C, and D), we calculate the average path length, ℓ_{path}, between the nodes, and the clustering coefficient, C, of the network and compare these values to those of the random networks that have the same number of nodes, N, and average number of connections, $\langle k \rangle$. We compute ℓ_{path} and C for 10 realizations of each model and average the results.

Figs. 4 and 5 show that the models B and C show a systematic decrease in the characteristic path length, ℓ_{path}, with increasing slope, a. These differences are of one or two links in the network which are considered small. Recent work has shown that the cellular networks of 43 organisms from the three domains have an average path length of 3 [55]. Similarly, the thirteen key metabolites in the bacterium *E. coli* also have an average path length of 3 [56]. The clustering coefficient, C, in the "small world" models is higher than that in

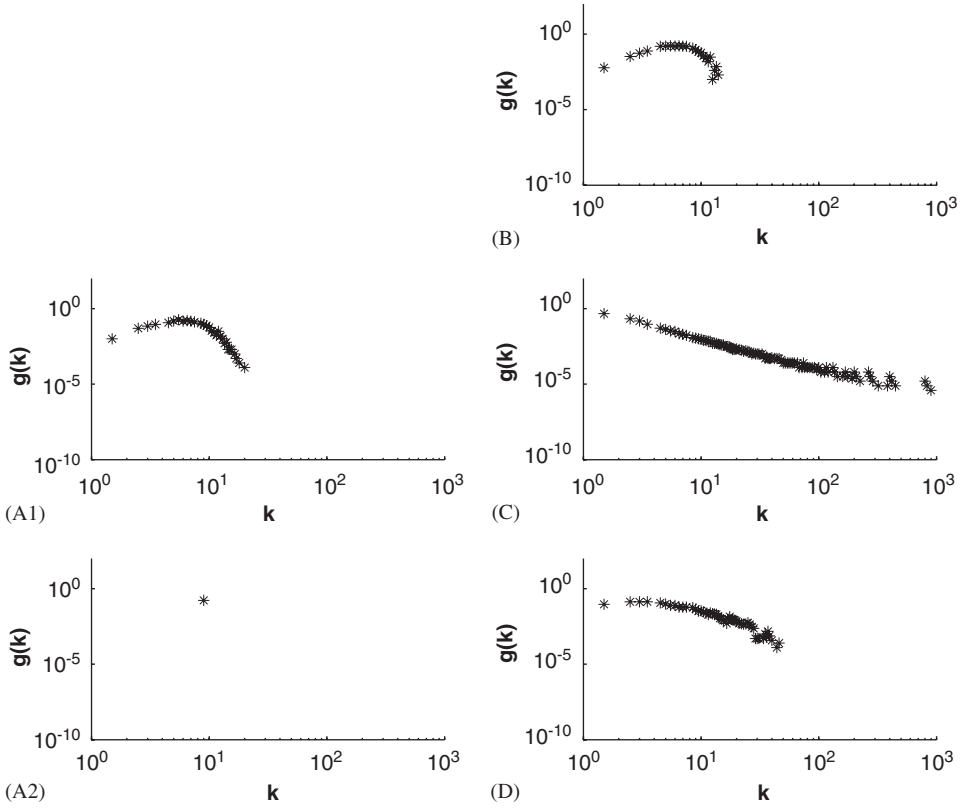

Fig. 2. Degree distribution $g(k)$ of input connections of the networks.

the random models. The clustering coefficient also decreases with increasing a, the scaling exponent of $g(k) = Ak^{-a}$.

The models are constructed with a power law $g(k)$ and thus are scale-free. They also have the properties of "small world" networks, namely short path length, ℓ_{path}, and higher clustering coefficient, C, than random models with the same number of nodes N and average number of connections $\langle k \rangle$. Therefore, they have both scale-free and "small world" properties.

5. Computing the statistics of the mRNA levels of each model

The connection matrix M represents how each gene i is regulated by gene j at any moment in time. The elements X_i of the 1-D column vector represent the mRNA level expressed by each gene. We start with initial values of the elements of X_i. At each subsequent time step the new elements of X_i' are then computed from

$$\mathbf{X}' = M\mathbf{X}. \tag{3}$$

We iterate this equation, evolving the mRNA levels in time, until they reach a steady state. A steady state is reached when $\varepsilon = 5 \times 10^{-32}$, where $\varepsilon = \sum(X' - X)^2$. We do not include noise in our model.

The mRNA values X_i of the model (A1) with the Markovian matrix must reach a steady state. Since $\sum_i M_{ij} = 1$, the connection matrix M has an eigenvalue $\lambda = 1$ with the corresponding stationary eigenvector $\mathbf{X}'$ as explained in Appendix A. Both the iterative simulations and the direct computation of the eigenvector produce the same values of the steady state $\mathbf{X}'$.

The mRNA values $\mathbf{X}$ from the models (A2, B, C and D) having non-Markovian matrices may grow or decay exponentially as they are evolved. Therefore, at each time step, we renormalize the mRNA levels by dividing each element X_i by the norm $(\mathbf{X})$ which maintains the state vector at constant length. The biological interpretation of this procedure is that the elements of $\mathbf{X}$ represent the fraction of mRNA expressed by each

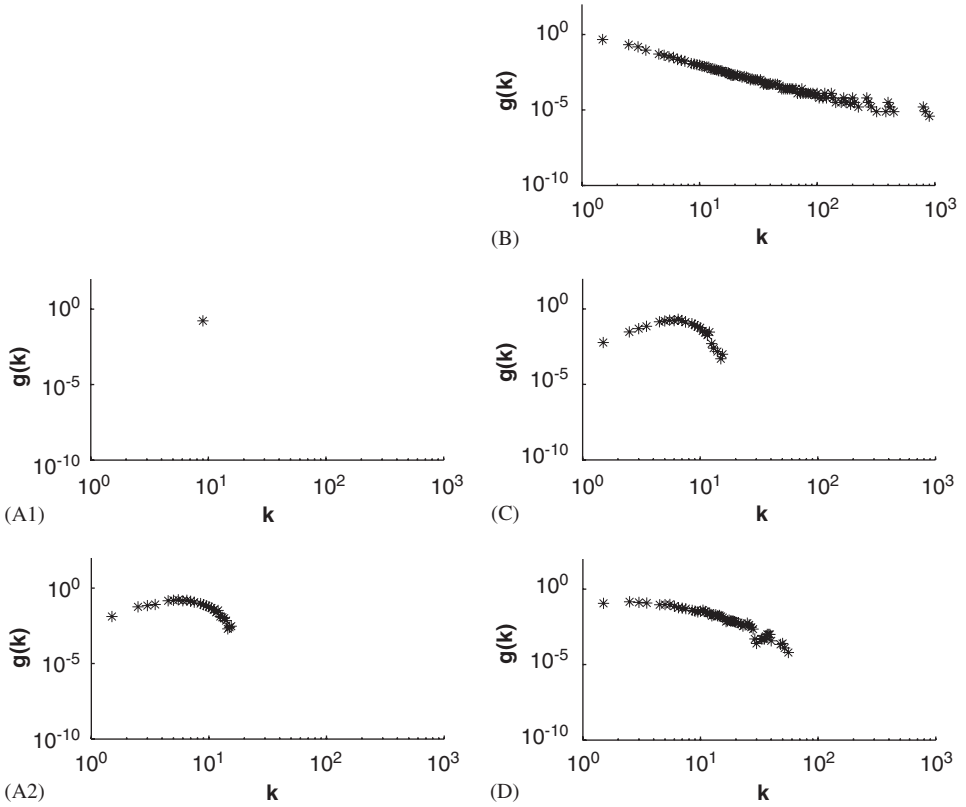

Fig. 3. Degree distribution $g(k)$ of output connections of the networks.

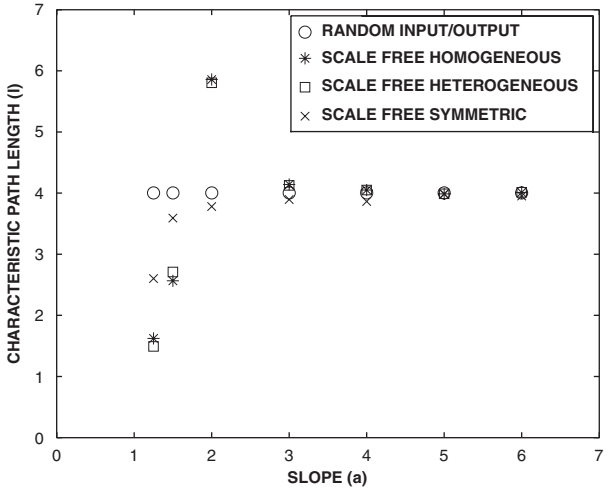

Fig. 4. Characteristic path length, ℓ_{path}, for the random models and as a function of the scaling exponent a for the distribution of connections $g(k) = Ak^{-a}$ for the "scale free" models.

gene and that the total amount of mRNA is confined by biological constraints. That is, given the amount of energy available for a cell, there is a limit on the amount of mRNA molecules that can be synthesized. Some models, such as the random model A1, can be formulated either with a Markovian matrix with nonzero

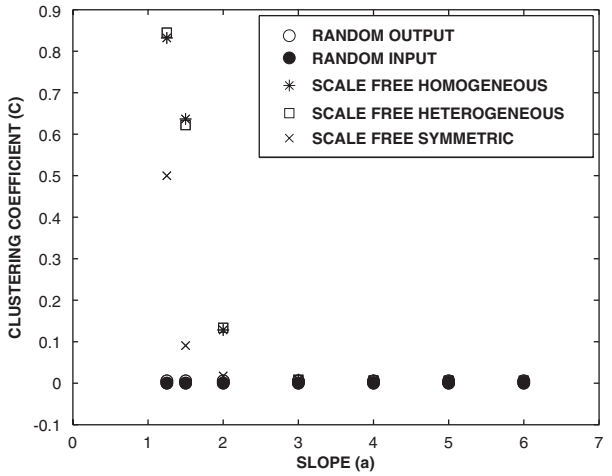

Fig. 5. Clustering coefficient, C, for the random models and as a function of the scaling exponent a for the distribution of connections $g(k) = Ak^{-a}$ for the "scale free" models. The clustering coefficient, C, decreases with increasing a. The clustering coefficient, C, of the "small world" networks is always higher than that of the random networks with equal number of nodes, N, and connections, $\langle k \rangle$.

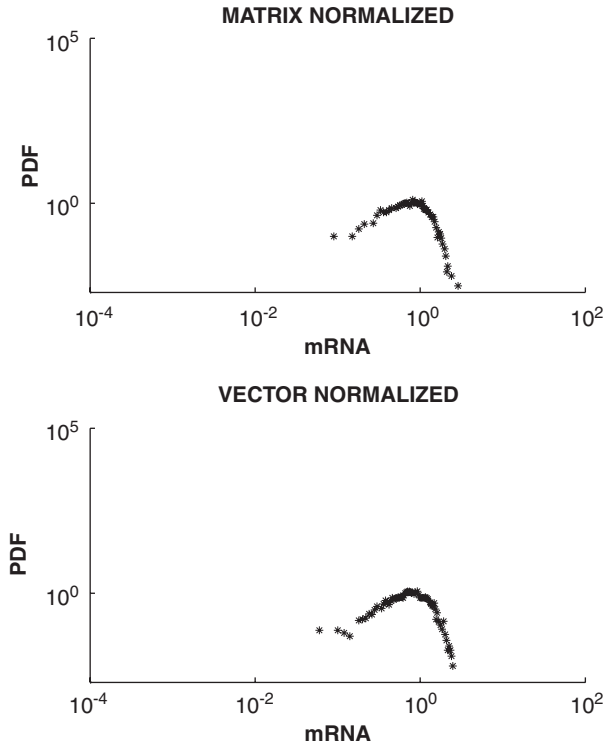

Fig. 6. Normalization test. The distribution of the steady state mRNA levels produced by a Markovian matrix (model A1) or by normalizing the mRNA levels of the equivalent non-Markovian matrix at each step, both produced similar PDFs of the mRNA levels.

elements $M_{ij} = \frac{1}{6}$ or with a non-Markovian matrix with nonzero elements $M_{ij} = 1$. We tested this non-Markovian renormalization procedure by computing the steady state X' both ways. That is, in the first method, each column in the matrix is normalized only at the start and in the second method, the output vector is normalized at each step. Both methods, even though they differ in the time and place the nonlinearity is introduced, produce similar distributions of mRNA levels, as shown in Fig. 6.

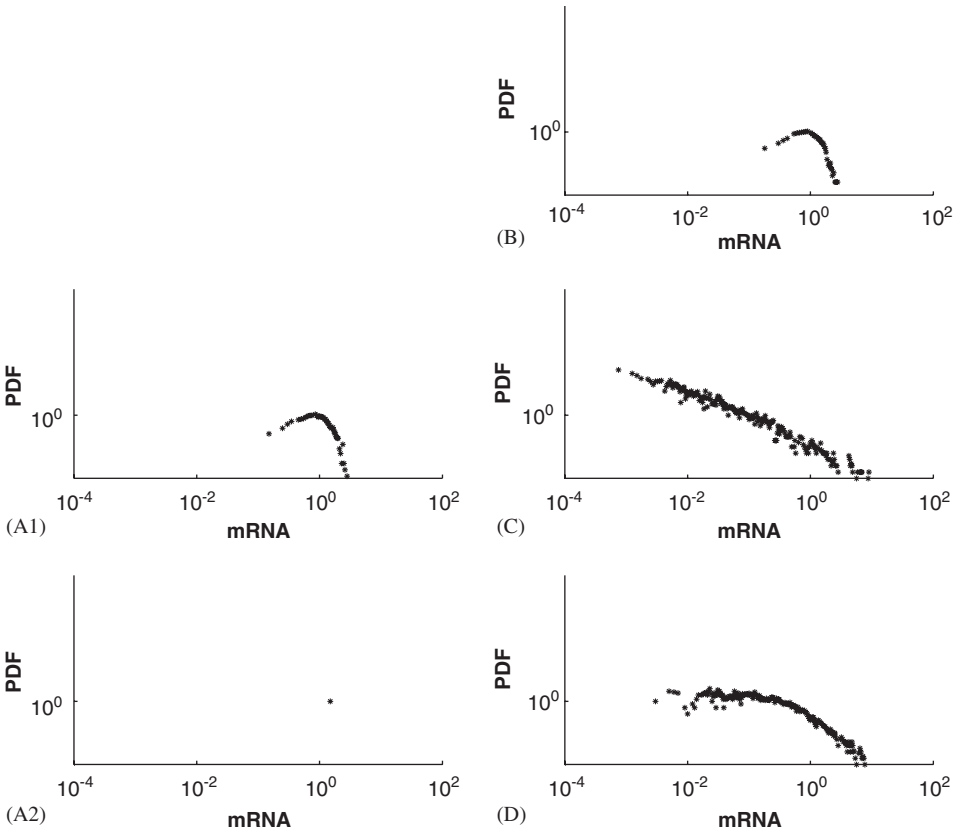

Fig. 7. PDFs of mRNAs levels produced by the models of genetic interaction shown on Fig. 1. Model (A1) is the random output, (A2) is the random input, (B) is the scale free homogeneous input, (C) is the scale free heterogeneous input, and (D) is the scale free symmetric input/output. For the scale free models here $a = 2$.

We then determine the statistical properties of the elements of X_i by evaluating the PDF, that is the probability that any of the mRNA levels, x, is between x and $x + \mathrm{d}x$. We are interested only in the form of the PDF of all the mRNA levels, but not which gene produces what mRNA level. Hence, we study the global pattern of the system, not individual genes. The typical method to determine the PDF is to form a histogram of equal bins. The problem with that method is that the result depends on the size of the bins used. If they are small, they are accurate at small values, but contain too few events at large values. If they are large, they are accurate at large values, but have low resolution at small values. We overcome these limitations by using a multi-histogram method [57], to improve the accuracy of the estimation of the PDF. We compute histograms of different bin sizes, evaluate the PDF from each histogram, and then combine those values to form the completed PDF. We used the PDF to characterize the global, statistical pattern of the mRNA expression levels.

The plots of Log PDF vs. Log x computed from the five different models of the connection matrix M_{ij} are shown in Fig. 7. For the Random input model (A2), the state vector converges to a uniform vector, in which all the components have the same value, as described in Appendix C. This is reflected in the PDF contracting to a single point. For the scale free models, the PDFs depend on the scaling exponent a of $g(k) = Ak^{-a}$.

For $a = 1$, as shown in Fig. 8, as we progress from the homogeneous model (B) through the symmetric model (D) to the heterogeneous model (C), the monotonically decreasing tail of the PDF becomes more prominent. Since the connection matrices B and C are not symmetric, $M^{\mathrm{T}} \neq M$, they both have similar eigenvalues but different eigenvectors. This is evident in the different distributions of mRNA produced by each model. The differences between the PDFs of the three scale free models are interesting. The PDF computed from the symmetric model is intermediate between the PDFs computed from the homogeneous and

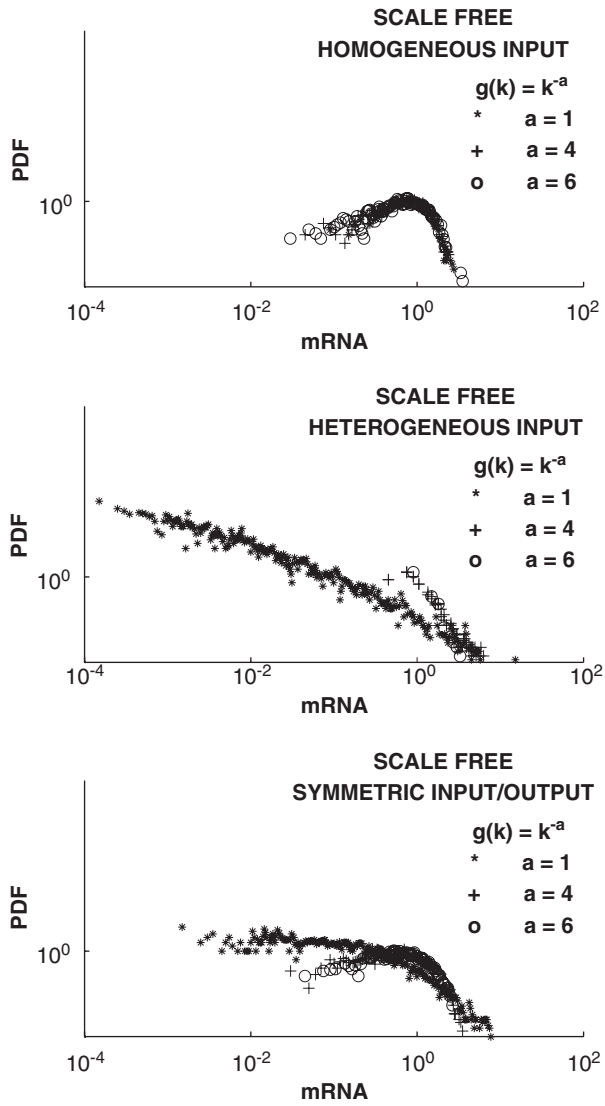

Fig. 8. The dependence of the PDFs of the mRNA levels on the exponent a in the scale free models, $g(k) = Ak^{-a}$.

heterogeneous models which makes sense since the connection matrix M_{ij} of symmetric model is intermediate between the connection matrices of the homogeneous and heterogeneous models. As the exponent, a, in the distribution of genetic interactions $g(k) = Ak^{-a}$ increases, the slope of the tail of the PDF distributions, as measured by least squares method, becomes more sharply peaked as shown in Fig. 9.

In summary, we find that different models of the connection matrix M produce different forms in the PDF of the mRNA expression levels. Thus we have established a direct link between the architecture of the genetic interactions and the global, statistical properties of the mRNA expression levels.

6. The statistics of mRNA levels in experimental data

We evaluate the PDF of the mRNA expression levels from 54 expression sets where cDNA microarray technology was used to measure mRNA expression levels of virtually every gene in the heads of control and *period* null mutant *Drosophila* flies [58]. The gene *period* is one of several genes in *Drosophila* that are necessary

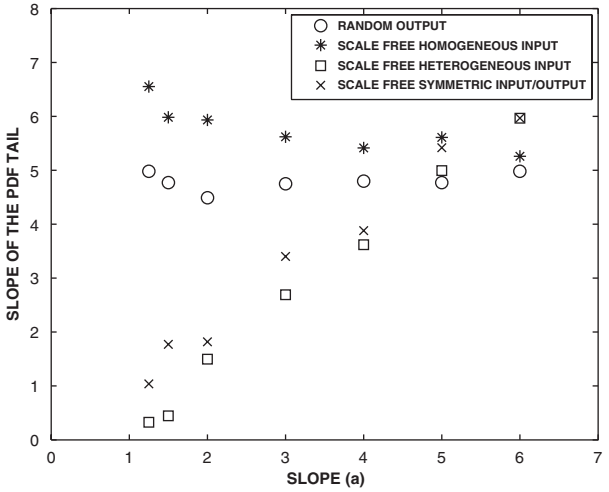

Fig. 9. The slope of the tail of the PDF for each model. For the scale free heterogeneous and symmetric models, the slope of the PDF tail increases with increasing a, the scaling exponent of $g(k) = Ak^{-a}$. For the random models, the slope of the PDF is constant.

for the normal functioning of the circadian time-keeping system. The data sets are distributed among four different experiments. In each experiment, gene expression was measured at 4 h intervals over a 24 h period.

Experiment A. Control dark–dark cycle: Gene expression was measured from control, wild type, flies kept in constant darkness. Two replicate time series were recorded.

Experiment B. Control light–dark cycle: Gene expression was measured from control, wild type, flies kept in light-dark cycles. Three replicate time series were recorded.

Experiment C. period *null mutant dark–dark cycle*: Gene expression was measured from *per* mutant flies kept in constant darkness. Two replicate time series were recorded.

Experiment D. period *null mutant light–dark cycle*: Gene expression was measured from *per* mutant flies kept in light–dark cycles. Two replicate time series were recorded.

The mRNA levels recorded from the four experiments and used in our analysis are absolute expression levels. That is, the expression values are not relative to those in the first time points. The dynamic range of the time series is at least 100. The PDFs of the mRNA expression levels from these 54 mRNA expression sets were quite similar. Representative plots from the last (sixth) time points of Log PDF vs. Log x from each of the four experiments are shown in Fig. 10.

7. Comparing the models and the experimental data

The experimentally measured PDFs have a prominent monotonically decreasing tail that is a line with an approximate slope of 2 on the plots of Log PDF vs. Log x. The experimentally measured PDFs do not resemble the delta function character of the PDFs computed from the random input model of genetic interactions. Considering the remaining four models, the slope of the tail and the width of the distribution of the experimentally measured PDFs are best captured by the scale free heterogeneous input model of genetic interactions where each gene is regulated by a different pattern of inputs from other genes. The slope of the experimental data is closest to models where the distribution of both input and output regulation $g(k) = Ak^{-a}$, where $a = 2$.

By comparing the models to the experimental data, we find that the data is best fit by a model with power structure $g(k) = Ak^{-a}$ of both input and output regulatory connections. This finding helps us understand how a biological system may be controlled by a few master genes that have many links to the other genes. This comparison also allows us to derive quantitative information from the experimental data such as the parameters $\langle k \rangle$, average number of connections, and a, the scaling exponent of the $g(k) = Ak^{-a}$ distribution of the theoretical models. Our results show that genetic networks share structural features, such as scale-free and hierarchical properties, that are common in other networks in nature.

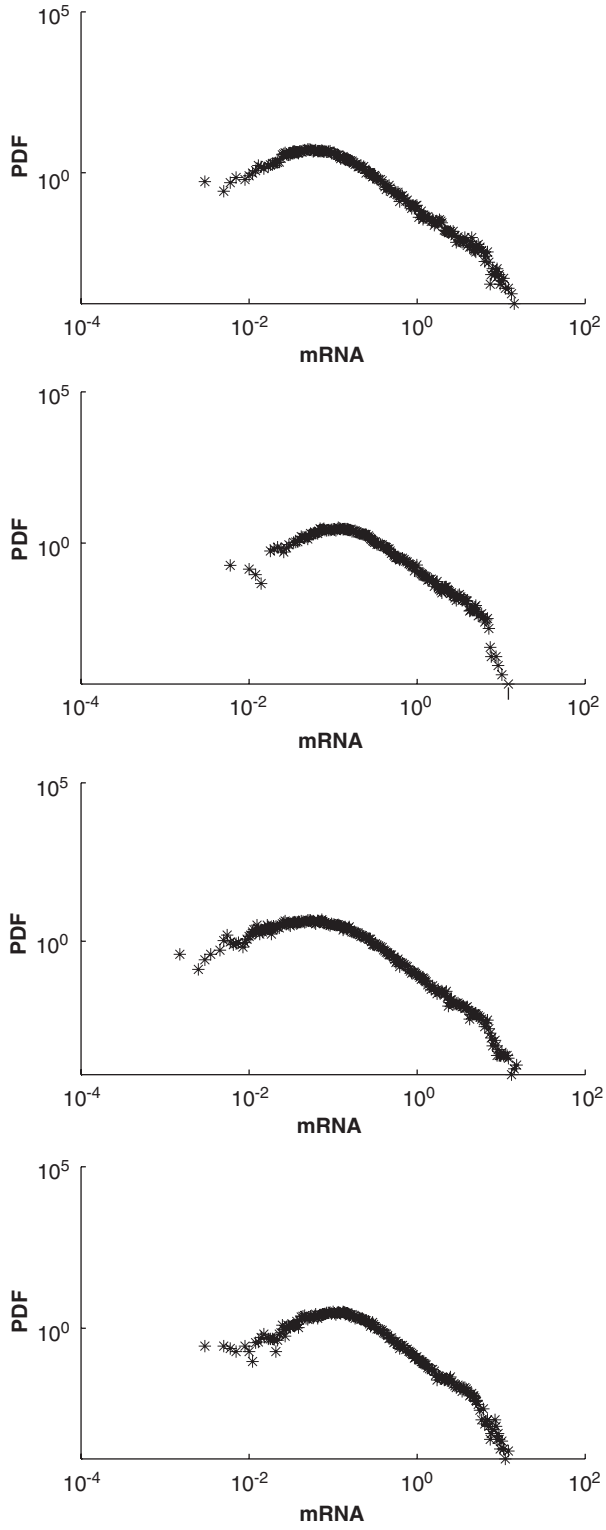

Fig. 10. Typical PDFs of mRNA levels determined from the control Dark–Dark cycle experiment, the control Light–Dark cycle experiment, *per* mutant Dark–Dark cycle experiment, and the *per* mutant Light–Dark cycle experiment. The slope of the tail of the PDF is approximately 2.

8. Conclusion

We are able to formulate some basic models of different types of genetic interactions, compute the statistical properties of the mRNA levels produced by those models, and compare those results to the statistical properties of experimentally measured mRNA levels. By knowing only the statistics of the mRNA levels (PDF) we can approximate which of several different architectures of genetic interaction may be representative of the experimental data. By visual inspection of the experimental PDFs, we conclude that the experimentally measured PDFs are best represented by the PDFs of the scale free symmetric input/output model of genetic interactions where the input and output connections are symmetrical with the value of the scaling exponent $a = 2$ in $g(k) = Ak^{-a}$. If true, this implies that some genes are regulated by the input from only few other genes, while some genes are regulated by the input from a gross average of all genes. Similarly, some genes individually regulate other genes, while many genes collectively regulate a few other genes. In addition, the distribution of input connections is the same as the distribution of output connections. These results are accomplished without knowing which gene is responsible for what mRNA level. This supports the feasibility of our basic idea that we can use the global, statistical information from the observed mRNA levels to infer information about the pattern of genetic interactions.

This work still cannot uniquely determine the topology of a genetic network since it lacks the proof that the PDFs are unique for each model. However, this work suggests that like many other natural networks, genetic networks seem to have scale free topology.

An important conclusion from this work relates to the studies of networks in general. We have found that the PDFs of the activities of nodes at a steady state, which is a measure of the dynamics of the network, have the same functional form as the degree distribution of the network, which is a measure of the structure of a network. This is an important finding that shows that there is a strong interrelationship between the dynamics and the structure of a network.

If true, our results may provide a practical way to screen biological systems to determine which ones are the best candidates for therapeutic intervention. A system where the mRNA expression levels tell us that the genetic interactions depend on a balanced interaction of a very large number of genes, such as models A1 and A2, may be a poor candidate system for basic scientific studies or clinical applications. There are just too many simultaneous interacting genes so that experiments or therapeutic treatments that alter one gene at a time will not be productive. A system that has a scale-free structure or a system in which there is only a limited number of genetic interactions or a strong hierarchy of connections, such as models B, C, and D, may be a good candidate system for basic scientific studies or clinical applications. The small number of simultaneously interacting or controlling genes would make it possible to study the effects of each of these genes separately.

Acknowledgements

We thank Yiing Lin from Washington University Medical School, St. Louis, MI for his continuous help in understanding the Affymetrix data available at (http://circadian.wustl.edu).

Appendix A. Existence of stationary eigenvectors

The condition for stationarity of an eigenvector is that the corresponding eigenvalue of the connection matrix M is $\lambda = 1$, where

$$\lambda \mathbf{X}' = M\mathbf{X}' \tag{A.1}$$

means that $\mathbf{X}' = M\mathbf{X}'$. If the elements of each column in the matrix sum to 1, then the matrix M can be written as,

$$M = \begin{pmatrix} M_{11} & M_{12} & \cdots & M_{1n} \\ M_{21} & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 1 - \sum_{i=1}^{n-1} M_{i1} & \cdots & \cdots & 1 - \sum_{i=1}^{n-1} M_{in} \end{pmatrix}. \tag{A.2}$$

Adding the first $(n-1)$ rows to the last row, the determinant reads

$$\det M = \det \begin{pmatrix} M_{11} & M_{12} & \cdots & M_n \\ M_{21} & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & 1 & 1 \end{pmatrix} \tag{A.3}$$

and then subtracting the first column from all other columns results in

$$\begin{aligned} \det M &= \det \begin{pmatrix} M_{11} & M_{12} - M_{11} & \cdots & M_{1n} - M_{11} \\ M_{21} & M_{22} - M_{21} & \cdots & M_{2n} - M_{21} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 0 & 0 & 0 \end{pmatrix} \\ &= 1 * \det B \end{aligned} \tag{A.4}$$

with

$$B = \begin{pmatrix} M_{12} - M_{11} & \cdots & M_{1n} - M_{11} \\ M_{22} - M_{11} & \cdots & M_{2n} - M_{21} \\ \cdots & \cdots & M_{(n-1)n} - M_{(n-1)1} \end{pmatrix}. \tag{A.5}$$

The determinants of the matrices M and B may be written as the product of the eigenvalues λ_{Mi}, $i = 1, \ldots, n$ and λ_{Bi}, $i = 1, \ldots, n-1$, respectively, that is

$$\lambda_{M1}\lambda_{M2}\cdots\lambda_{Mn} = 1 * \lambda_{B1}\lambda_{B2}\cdots\lambda_{B(n-1)}. \tag{A.6}$$

From this, it follows that

$$\lambda_{M1} = \frac{\lambda_{B1}\lambda_{B2}\cdots\lambda_{B(n-1)}}{\lambda_{M2}\cdots\lambda_{Mn}} = 1 \tag{A.7}$$

because

$$\det B = \det \begin{pmatrix} M_{12} & \cdots & M_{1n} \\ M_{22} & \cdots & M_{2n} \\ \cdots & \cdots & M_{(n-1)n} \end{pmatrix} - \det \begin{pmatrix} M_{11} & \cdots & M_{11} \\ M_{21} & \cdots & M_{21} \\ \cdots & \cdots & M_{(n-1)1} \end{pmatrix}. \tag{A.8}$$

The second determinant is 0 since its columns are identical. From this point on, we will refer to the stationary eigenvalue as $\lambda_0 = 1$.

Hence, if the matrix columns sum to 1, then there is always an eigenvalue $\lambda_0 = 1$ which guarantees the existence of a nontrivial stationary state. The same applies when the matrix rows sum to 1, that is for the transpose M^{T}, since $\det M = \det M^{\mathrm{T}}$.

Appendix B. Convergence to stationary eigenvectors

If an eigenvector is stationary, $\lambda_0 = 1$, all other eigenvectors in the system might or might not converge to it. This convergence will be characterized by the stability of the other eigenvectors and their corresponding eigenvalues λ_i, for $i = 1, 2, \ldots, n$. If $|\lambda_i| > 1$, the system is unstable. If $|\lambda_i| < 1$, the system is stable.

Under the condition in which the elements in the matrix columns or rows sum to one, the numerical simulations in our work have shown that $|\lambda_i| < 1$, $i = 1, \ldots, n$ and there will always be a unique solution where $\lambda_0 = 1$. Therefore, the system will converge to the stationary eigenvector.

Appendix C. Uniform eigenvectors

We study the output column vector $\mathbf{X}'$ of the connection matrix M, where $\mathbf{X}$ is the input vector and norm $(\mathbf{X}) = 1$.

$$\mathbf{X}' = M\mathbf{X}. \tag{C.1}$$

If each row in the sparse matrix M has m equal nonzero elements that sum to 1, and all others are zero, that is $\sum_j^n M_{ij} = 1$, then the model can be written as,

$$\begin{pmatrix} X_1' \\ X_2' \\ \cdots \\ X_n' \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{m} & \cdots & \frac{1}{m} \\ \frac{1}{m} & 0 & \frac{1}{m} & \cdots \\ \cdots & \frac{1}{m} & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \cdots \\ X_n \end{pmatrix}. \tag{C.2}$$

We assume a uniform initial state vector

$$\mathbf{X_u} = \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{pmatrix}. \tag{C.3}$$

Inserting the uniform state vector in (C2), we obtain for each component

$$X_i' = \sum_{n=1}^{m} \frac{1}{m} = 1 \quad \forall i \tag{C.4}$$

and thus the uniform state vector is the unique stationary eigenvector with $\lambda_0 = 1$.

For an arbitrary initial state vector, the following is true: As shown in Appendices A and B, the system will converge to the unique stationary eigenvector, in this case the uniform state vector $\mathbf{X_u}$ with the weight

$$\mathbf{X_u} = \overline{X} \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{pmatrix}, \tag{C.5}$$

where

$$\overline{X} = \sum_i X_i. \tag{C.6}$$

Appendix D. Derivation of the density function

The regulatory influence at each time step of the jth gene on the ith gene is given by the element M_{ij} in the $N \times N$ connection matrix M. If g genes regulate k genes, the distribution of genetic interactions is defined by the function $g(k)$. The scale free networks here have the form $g(k) = Ak^{-a}$, where $1 \leqslant a \leqslant 6$. We need to know how to choose the elements in M to generate the desired $g(k)$.

The choice of A is arbitrary. The distribution of mRNA levels, the elements X_i of the 1-D steady state column vector $\mathbf{X}$ produced by iterating $\mathbf{X}' = M\mathbf{X}$, does not depend on A, since we renormalize $\mathbf{X}$ at each time step. We choose A so there are 6000 nonzero elements in the $N = 1000$ connection matrices.

To produce the desired a we must determine how to choose the density $Y(R)$ of nonzero elements in each line that is at a distance R from the start of the matrix, M_{11}.

A general relationship between $Y(R)$ and $g(k)$ can be found by noting that the number of genes with k connections, $g(k)\,\mathrm{d}k$, over the interval $[k, k+\mathrm{d}k]$ is equal to the number of nonzero elements, $\mathrm{d}R$, with $Y(R)=k$ connections over the corresponding interval $[R, R+\mathrm{d}R]$. To generate the scale free models, if the density has the form

$$Y(R)=\left[\frac{A}{(a-1)(N-R)}\right]^{1/(a-1)} \tag{D.1}$$

then

$$R=N+\frac{A}{1-a}\,k^{1-a} \tag{D.2}$$

and therefore, the distribution is

$$g(k)=\frac{\mathrm{d}R}{\mathrm{d}k}=Ak^{-a}. \tag{D.3}$$

Appendix E. Normalizing the connectivity matrix to *k* connections

Integrating the density equation, we can adjust the parameter A to have K total connections in the connectivity matrix. If

$$Y(R)=\left[\frac{A}{(a-1)(N-R)}\right]^{1/(a-1)} \tag{E.1}$$

then,

$$K=\int\left[\frac{A}{(a-1)(N-R)}\right]^{1/(a-1)}\mathrm{d}R. \tag{E.2}$$

and integrating from $R=1$ to $R=N-1$ we find that for $a\neq 2$,

$$A=(a-1)\left[\frac{K}{(1-(N-1)^{(a-2)/(a-1)})((1-a)/(a-2))}\right]^{a-1}, \tag{E.3}$$

and for $a=2$,

$$A=\frac{K}{\ln(N-1)}. \tag{E.4}$$

References

[1] J.B. Rampal (Ed.), DNA Arrays: Methods and Protocols, Human Press, 2001.
[2] M. Schena (Ed.), DNA Microarrays: A Practical Approach, Oxford University Press, NY, 1999.
[3] Z. Bryant, L. Subrahmanyan, M. Tworoger, L. LaTray, C. Liu, M. Li, G.V.D. Engh, H. Ruohola-Baker, Proc. Natl. Acad. Sci. USA 96 (1999) 5559.
[4] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iver, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, B. Futcher, Mol. Biol. Cell 9 (1998) 3273 (Data online at http://cellcycle-www.stanford.edu/).
[5] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Proc. Natl. Acad. Sci. USA 95 (1998) 14863.
[6] G.S. Michaels, D.B. Carr, M. Askenazi, S. Fuhrman, X. Wen, R. Somogyi, Pac. Symp. Biocomput. 3 (1998) 42.
[7] S. Raychaudhuri, J.M. Stuart, R.B. Altman, Pac. Symp. Biocomput. 5 (2000) 452.
[8] M.P. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares Jr., D. Haussler, Proc. Natl. Acad. Sci. USA 97 (2000) 262.
[9] X. Wen, S. Fuhrman, G.S. Michaels, D.B. Carr, S. Smith, J.L. Barker, R. Somogyi, Proc. Natl. Acad. Sci. USA 95 (1998) 334.
[10] S. Tavazoei, J.D. Hughes, M.J. Campbell, R.J. Cho, G.M. Church, Nat. Genet. 22 (1999) 281.
[11] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, T.R. Golub, Proc. Natl. Acad. Sci. USA 96 (1999) 2907.
[12] F. Jacob, J. Monod, J. Mol. Biol. 3 (1961) 318.

[13] M. Wahde, J. Hertz, Biosystems 55 (2000) 129.
[14] P. D'haeseleer, S. Liang, R. Somogyi, Bioinformatics 16 (2000) 707.
[15] A.J. Hartemink, D.K. Gifford, T.S. Jaakkola, R.A. Young, Pac. Symp. Biocomput. 6 (2001) 422.
[16] P. D'haeseleer, X. Wen, S. Fuhrman, R. Somogyi, Pac. Symp. Biocomput. 4 (1999) 41.
[17] O. Alter, P.O. Brown, D. Botstein, Proc. Natl. Acad. Sci. USA 97 (2000) 10101.
[18] N.S. Holter, A. Maritan, M. Cieplak, N.V. Fedoroff, J.R. Banavar, Proc. Natl. Acad. Sci. USA 98 (2001) 1693.
[19] M.K.S. Yeung, J. Tegner, J.J. Collins, Proc. Natl. Acad. Sci. USA 99 (9) (2002) 6163.
[20] T.G. Dewey, D.J. Galas, Integer Genomics 1 (2001) 269.
[21] A. Bhan, D.J. Galas, T.G. Dewey, Bioinformatics 18 (11) (2002) 1486.
[22] S.A. Kauffman, J. Theor. Biol. 22 (1969) 437.
[23] S.A. Kauffman, J. Theor. Biol. 44 (1974) 167.
[24] S.A. Kauffman, Physica D 10 (1984) 145.
[25] S.A. Kauffman, The Origins of Order, Oxford University Press, New York, 1993.
[26] S.A. Kauffman, C. Peterson, B. Samuelsson, C. Troein, Proc. Natl. Acad. Sci. USA 100 (25) (2003) 14796.
[27] S.A. Kauffman, Phys. Rev. Lett. 93 (4) (2004) 048701-1-4.
[28] S.A. Kauffman, C. Peterson, B. Samuelsson, C. Troein, Proc. Natl. Acad. Sci. USA 101 (2004) 17102.
[29] R. Thomas, R. d'Ari, Biological Feedback, CRC Press, Boca Raton, FL, 1990.
[30] R. Thomas, D. Thieffry, M. Kauffman, Bull. Math. Biol. 57 (1999) 247.
[31] J. Boden, J. Theor. Biol. 188 (1997) 391.
[32] S. Liang, S. Fuhrman, R. Somogyi, Pac. Symp. Biocomput. 3 (1998) 18.
[33] R.J. Bagley, L. Glass, J. Theor. Biol. 183 (1996) 269.
[34] T. Mestl, C. Lemacy, L. Glass, Physica D 98 (1996) 33.
[35] L. Glass, S.A. Kauffman, J. Theor. Biol. 39 (1973) 103.
[36] H. Smith, J. Math. Biol. 25 (1987) 169.
[37] J.C. Leloup, A. Goldberger, J. Biol. Rhythms 13 (1998) 70.
[38] J.M. Mahaffy, D.A. Jorgensen, R.L. van der Heyden, J. Math. Biol. 30 (1992) 669.
[39] W.S. Hlavacek, M.A. Savageau, J. Mol. Biol. 255 (1996) 121.
[40] T.S. Gardner, C.R. Cantor, Nature 403 (2000) 339.
[41] E.H. Davidson, J.P. Rast, P. Oliveri, A. Ransick, C. Calestanni, C. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, O. Otmin, C.T. Brown, C.B. Livi, P.Y. Lee, R. Revilla, A.G. Rust, Z.J. Pan, M.J. Schilistra, P.J.C. Clarke, M.I. Arnone, L. Rowen, R.A. Cameron, D.R. McClay, L. Hood, H. Bolouri, Science 295 (2002) 1669.
[42] P. Smolen, D.A. Baxter, J.H. Byrne, Bull. Math. Biol. 62 (2000) 247.
[43] A. Arkin, J. Ross, H. McAdams, Genetics 149 (1998) 1633.
[44] H. McAdams, L. Shapiro, Science 269 (1995) 650.
[45] M. Kerszberg, J. Changeus, Bioessays 20 (1998) 758.
[46] C.H. Yuh, H. Bolouri, E.H. Davidson, Science 279 (1998) 1896.
[47] A.L. Barabasi, R.A. Albert, Science 286 (1999) 509.
[48] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, A.L. Barabasi, Nature 407 (2000) 651.
[49] D.J. Watts, S.H. Strogatz, Nature 393 (1998) 440.
[50] R.A. Albert, A.L. Barabasi, Phys. Rev. Lett. 85 (2000) 5234.
[51] B.A. Huberman, A.L. Adamic, Nature 401 (1999) 131.
[52] R.A. Albert, A.L. Barabasi, Rev. Modern Phys. 74 (1) (2002) 47.
[53] A. Papoulis, Random Variables, and Stochastic Processes, second ed., McGraw Hill, New York, 1984.
[54] S. Milgram, Psychol. Today 1 (1) (1967) 60.
[55] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, A.L. Barabasi, Nature 407 (2002) 651.
[56] D.A. Fell, A. Wagner, Proceedings from Second Workshop on Computation of Biochemical Pathways and Genetic Networks, Logos Verlag, Berlin, 2001, p. 11.
[57] L.S. Liebovitch, A.T. Todorov, M. Zochowski, D. Scheurle, L. Colgin, M.A. Wood, K.A. Ellenbogen, J.M. Herre, R.C. Bernstein, Phys. Rev. E 59 (3) (1999) 3312.
[58] Y. Lin, M. Han, B. Shimada, L. Wang, T.M. Gibler, A. Amarakone, T.A. Awad, G.D. Stormo, R.N. Van Gelder, P.H. Taghert, Proc. Natl. Acad. Sci. USA 99 (14) (2002) 9562.