BIOINFORMATICS

Vol. 18 no. 11 2002
Pages 1486–1493

A duplication growth model of gene expression networks

*Ashish Bhan, David J. Galas and T. Gregory Dewey**

Keck Graduate Institute of Applied Life Sciences, 535 Watson Drive, Claremont, CA 91711, USA

Received on December 13, 2001; revised on February 8, 2002; April 29, 2002; accepted on May 3, 2002

ABSTRACT

Motivation: There has been considerable interest in developing computational techniques for inferring genetic regulatory networks from whole-genome expression profiles. When expression time series data sets are available, dynamic models can, in principle, be used to infer correlative relationships between gene expression levels, which may be causal. However, because of the range of detectable expression levels and the current quality of the data, the predictive nature of such inferred, quantitative models is questionable. Network models derived from simple rate laws offer an intermediate level analysis, going beyond simple statistical analysis, but falling short of a fully quantitative description. This work shows how such network models can be constructed and describes the global properties of the networks derived from such a model. These global properties are statistically robust and provide insights into the design of the underlying network.

Results: Several whole-genome expression time series data sets from yeast microarray experiments were analyzed using a Markov-modeling method (Dewey and Galas, *Func. Integr. Genomics*, **1**, 269–278, 2001) to infer an approximation to the underlying genetic network. We found that the global statistical properties of all the resulting networks are similar. The overall structure of these biological networks is distinctly different from that of other recently studied networks such as the Internet or social networks. These biological networks show hierarchical, hub-like structures that have some properties similar to a class of graphs known as small world graphs. Small world networks exhibit local cliquishness while exhibiting strong global connectivity. In addition to the small world properties, the biological networks show a power law or scale free distribution of connectivities. An inverse power law, $N(k) \sim k^{-3/2}$, for the number of vertices (genes) with k connections was observed for three different data sets from yeast. We propose network growth models based on gene duplication events. Simulations of these models yield networks with the same combination of global graphical properties that we inferred from the expression data.

Contact: Ashish_Bhan@kgi.edu; David_Galas@kgi.edu; Greg_Dewey@kgi.edu

Supplementary Information: http://www.kgi.edu/html/noncore/faculty/dewey/bioinf.pdf

*To whom correspondence should be addressed.

INTRODUCTION

The transcriptome, the mRNA expression levels of all genes in an organism, can now be explored with DNA microarray technology. Time series expression profiles provide a rich source of biological information and allow the dynamics of gene expression to be modeled. For the computational biologist attempting to interpret and model genome-wide phenomena, this wealth of information provides both significant opportunities and challenges. In principle, we can use functional genomic data as the basis for creating systems models that detail all the specific interactions of each component of the system. Alternatively, we may use the data to attempt to deduce general design features that are independent of the specific details of the system. Both approaches are fraught with difficulties, not only because of the complexity of the systems, but also because the underlying quality of the current data severely limits detailed quantitative modeling. Short of such quantitative descriptions, but better than qualitative phenomenological models, are analyses of the global properties of these systems.

Recent analyses of network properties of protein–protein interactions and of metabolic maps have provided some insight into the structure of these networks (Uetz *et al.*, 2000; Jeong *et al.*, 2000; Barabási and Albert, 1999). In the present work, we describe global analyses based on our method (Dewey and Galas, 2001) for generating gene networks from time series of expression profiles. The statistical properties of networks derived from DNA microarray data reveal unique features characteristic of both scale-free and 'small world' networks. From these features, we can infer some design characteristics of the underlying networks, and speculate about their origins. This is done by simulating network growth with a gene duplication model.

There has been considerable recent interest in the network structure of a diverse range of systems, including the Internet, communities of actors, scholarly citations, metabolic networks and ecological systems, among others (Jeong *et al.*, 2000; Barabási and Albert, 1999; Albert and Barabási, 2001; Strogatz, 2001; Amaral *et al.*, 2000). Three main categories of networks have been used to model these various systems. They are: random networks (Cohen, 1988; Kauffman, 1969), small world networks (Strogatz, 2001; Watts, 1999) and growing random networks (GRNs; Jeong *et al.*, 2000; Barabási and Albert, 1999; Krapivsky *et al.*, 2000; Dorogovtsev and Mendes, 2001). Random graphs have been extensively studied and are constructed by randomly connecting a set of nodes. Small world graphs are generated from a regular starting lattice. Edges in this lattice are then randomly 'rewired' to remote nodes. This provides strong local structure as well as global connectivity. Graphs can also be constructed from non-equilibrium growth models that start with a seed graph and add nodes and connections according to some prescribed set of preferences. Often a 'rich get richer' set of preferences are used, where the newly added nodes are preferentially connected to nodes of high connectivity.

Often the choice of model is dictated by the specific graphical property under investigation. For instance, small world models were originally motivated by the observation of networks that have high clustering coefficients and short mean path lengths. The cluster coefficient characterizes the extent to which vertices adjacent to any vertex are adjacent to each other. In social networks it is the degree that a persons acquaintances are acquainted to each other. The cluster coefficient is calculated by averaging over all vertices, the fraction of vertices adjacent to a given vertex that are adjacent to each other. The cluster coefficient varies from 0 to 1 with 1 indicating that all the neighboring nodes are connected to one another. The characteristic path length is found by determining the number of edges on the shortest path connecting any two vertices and averaging this number over all pairs of vertices. GRNs, on the other hand, were developed to explain the scale-free distribution of node connectivities, a property that the original small world models do not have. Scale-free distributions have no characteristic length scale and follow power law behavior. Random graphs show an interesting phase transition in the 'connectedness' of the graph, but do not show small world or scale-free behavior. When matching a given model to a natural network phenomenon, it is important to examine a range of graphical parameters for full discrimination between potential models.

In this work, we show how network models of gene expression can be obtained from a dynamic model of whole genome expression. The resulting networks are analyzed by determining three global graph properties—the average path length, the clustering coefficient and the connectivity scaling exponent. As will be seen, no existing model can account for the combined properties of the gene expression networks. To explain these results, we propose a new network growth model based on gene duplication events. Computer simulations indicate that this model can adequately describe the gene expression graph parameters.

METHODS AND IMPLEMENTATION

Networks from dynamic models of gene expression

There have been a number of recent attempts to analyze time series data for whole genome expression profiles (Dewey and Galas, 2001; Holter *et al.*, 2001; Heyer *et al.*, 1999; DeRisi *et al.*, 1997; Spellman *et al.*, 1998)). Interestingly, this does not require the complexity of detailed non-linear models of gene expression, but needs only simple, linear models (Dewey and Galas, 2001; Holter *et al.*, 2001)). These previous studies have focused on the cell-cycle and diauxic shift data in the yeast *Saccharomyces cerevisiae* (DeRisi *et al.*, 1997; Spellman *et al.*, 1998). In both cases, the system is prepared in a given physiological state at the initial time point and changes in gene expression levels are measured as it moves to a new state. These experiments have some similarity to traditional perturbation–relaxation experiments in physics and chemistry. Given this analogy, it is perhaps not surprising that the time dependence of the expression profiles can be well represented by simple linear response models.

Our previous analysis of expression time series is based on a simple dynamical model (Dewey and Galas, 2001) that includes both linear and non-linear kinetic terms. This model is briefly summarized here. The time dependence of the system is represented by the rate law given below:

$$A(t) = \Lambda_1 A(t-1) + A(t-1)A^{\mathrm{T}}(t-1)\Lambda_2 \quad (1)$$

where $A(t)$ is a matrix of the gene expression profiles at different points in time. $A(t) = (\hat{a}(2), \ldots, \hat{a}(t))$ where $\hat{a}(t)$ is a vector representing the expression levels of all genes in the genome at time, $t = i$. The ratio values from the public domain data sets were used (DeRisi *et al.*, 1997; Spellman *et al.*, 1998), rather than the log ratios, as these values are proportional to the mRNA concentration and are consistent with a first-order chemical kinetic model. The matrix $A(t-1)$ is a time-lagged matrix given by: $A(t-1) = (\hat{a}(1), \ldots, \hat{a}(t-1))$. The first term in Equation 1 represents a simple linear response and the elements of the Λ_1 matrix λ_{ij}, give the influence of the expression level of the jth gene on the production of the ith gene. The second term, $A(t-1)A^{\mathrm{T}}(t-1)$, in Equation 1, the gene covariance at a previous time, introduces non-linearity into the model.

The two matrices Λ_1 and Λ_2 generated by this data analysis are components of the weighted connectivity

matrix of a graph of interactions between gene expression levels (Dewey and Galas, 2001). We simplify the analysis by using a sparse, binary matrix representation of the adjacency matrix. This is achieved by applying a threshold to the entries in the transition matrix. The absolute values of the matrix elements are set equal to 1 if they are above a certain threshold, ϵ, and are set equal to 0 below this threshold. For high values of the threshold, the resulting matrix will be a sparse adjacency matrix. It is a digraph (non-symmetric matrix) showing the connectivity of the biological network. We do not differentiate here between positive and negative values for members in the transition matrix, as we are only interested in the underlying connectivity. Because the non-linear transition matrix is not of the same dimensionality as the linear term, these matrix elements cannot be directly compared. From a chemical kinetic perspective, the Λ_1 matrix is proportional to a matrix of first order rate constants and Λ_2 is proportional to second order rate constants. A second order rate constant can be converted into a pseudo-first order rate constant by multiplying by the appropriate concentration. We perform the equivalent operation here to compare Λ_1 and Λ_2. Therefore, we use the pseudo-first order matrix defined by: $\Lambda_2^* = \Lambda_2 A^{\mathrm{T}}(t-1)$ which is of the same dimensionality as Λ_1.

The networks derived in this work represent the phenomenological influence of one gene expression level on another. Thus, when the level of expression of gene i influences the expression of gene j, then an arrow is drawn from node i to node j. This network is strictly speaking not a genetic regulatory network. Rather it is a network derived from a kinetic model that shows the influence of one expression level on another. The network obtained from Λ_1 gives the linear response or passive elements of the system. Networks obtained from $\Lambda_2^* = \Lambda_2 A^{\mathrm{T}}(t-1)$ represents a very specific form of a non-linear response within this model and are the active elements of the system.

Gene duplication model

Gene duplication is a mechanism for network growth that is particular to biological systems and has strong implications for their evolution. This work explores specific duplication models to simulate the graph properties of the networks constructed from experimental data. Figure 1 illustrates how a duplication event can affect a network. Duplication results in the creation of a new node that has inherited all the connectivity of the parent node, as would be true of a duplicated gene (including its cis regulatory elements). This results in an increase by one of the number of vertices with the degree of the parent. It also results in an increase of one in the degree of each of the neighbors. In a 'pure' duplication model, this is the only event that occurs. This kind of growth model by itself has some interesting properties but it does not support a scale free distribution of connectivities. We have, therefore, examined a number of 'mixed' models that include gene duplication plus a second event. Features of two such models are illustrated in Figure 1. The 'partial duplication' model (Figure 1b) consists of duplication plus random removal of edges from the daughter node. A second model, 'duplication plus preferential re-wiring' (Figure 1c) involves duplication followed by random rewiring of one of the edges in the network. In our preferential rewiring model, the new node that the edge is rewired to is chosen at random according to the same preference function in the previous GRN models (Jeong *et al.*, 2000; Barabási and Albert, 1999)) i.e. the probability of connecting the edge to a node is proportional to the fraction of edges in the network that are incident at that node. These mixed models have formal similarity to a previous model used to describe the effect of gene duplication on protein–protein interaction networks (Wagner, 2001). In this previous work, network growth was not explicitly treated. Recently, a network growth model that yields scale-free networks has been described that involves gene duplication events (Rzhetsky and Gomez, 2001). This is a specific model involving domain shuffling and is distinctly different from the ones presented in this work. In all of these models, gene duplication is followed by a second event that breaks the parent–daughter symmetry inherent in a pure gene duplication model. This results in a broader range of node connectivities.

To assess the properties of the gene duplication models, we simulated network growth based on these processes. In these simulations, we start with a small initial, seed network. Two different seeds were considered: a random network seed and a network seed with a high clustering coefficient. The influence of the seed reveals those graph parameters that are influenced by initial conditions and those that are due to the dynamics of the growth process. Starting with the seed graph, the network is grown in a probabilistic manner, following the simple set of dynamic rules illustrated in Figure 1. A node from the entire network is chosen at random to be duplicated. In the partial duplication models, edges are then removed at random from the new daughter node. On average, half of the edges are removed. In the mixed model with preferential rewiring, both duplication and rewiring are treated as random processes, each occurring with a probability of one half. This parameter could also be varied, but we found that this condition was sufficient to create satisfactory models. The growth process then proceeds through a random sequence of duplication and rewiring events.

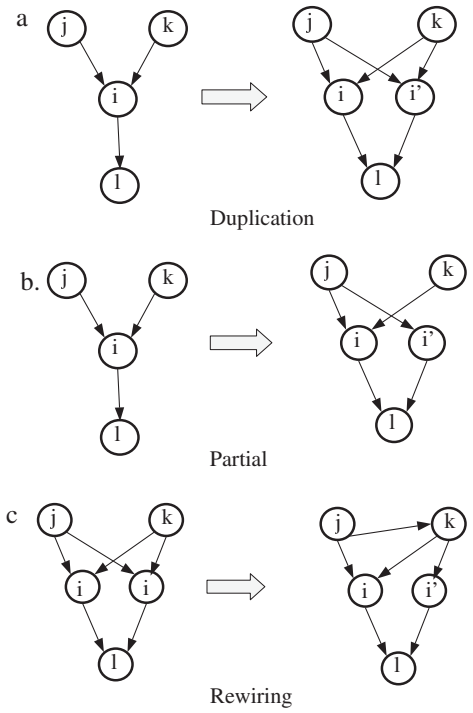

Fig. 1. Schematic representation of network growth through gene duplication. (a) Shows pure gene duplication where a new node is created by duplicating the connectivity of the parent. This results in an increase in degree of the neighboring nodes. Node i is duplicated to give i'. Nodes j, k, l are neighbors. (b) The partial duplication model where node i is duplicated to i' but not all the original connections are retained. (c) Shows a rewiring process where edge i–j is rewired to becom i–k.

RESULTS

Network properties of expression dynamics

Our analysis allowed us to determine digraphs for each set of experimental conditions. We considered three examples from the yeast data sets: the diauxic shift data, the cell cycle data (alpha factor) and the cell cycle data (cdc20). For each data set we generated a range of digraphs by varying the threshold for the linear and non-linear terms in the model.

Two global network parameters, the clustering coefficient and the characteristic path length, were determined from the networks. The results in Table 1 show that the gene expression graphs have very high clustering coefficients and relatively low average path lengths. Also, shown in Table 1 are the corresponding parameters for randomly generated graphs with an identical number of nodes and edges as the yeast networks. The clustering coefficients for the yeast networks are much higher than the equivalent random graphs while the characteristic path lengths are quite similar. The path length in random graphs depends on the ratio of edges to nodes and is low under the present conditions.

On the basis of the clustering coefficient and mean path length, it is tempting to classify the yeast expression networks as a small world network. However, when we examine the distribution of connectivity of these networks we consistently see scale-free behavior, a feature that is not seen in small world models. The complexity of the networks we derived from the expression data are, however, dependent on the threshold parameter, ε, of the analytical method. More linkages are added to the inferred network as the threshold is lowered so that the network becomes more and more complex. This raises the question of a possible analytical artifact, and suggests that we examine how the scaling of these increasingly complex networks depends on the threshold—a free parameter of the analytical method. To address this question we generated a series of networks from the same data set by varying ε and examined and compared the scaling of the resulting networks. Plotting the number of nodes as a function of the number of incident edges, the degree of the node, in Figure 2 reveals a strongly consistent scaling behavior that is independent of threshold level. In Figure 2 representative networks are shown with very different levels of complexity. This result establishes that the observation of a scale-free distribution is insensitive to changes in the analytical parameter ε. The networks in the figure are clearly very dissimilar to the eye, the bottom one containing almost ten-fold more edges and yet they exhibit the same scaling. It should also be noted that when exit edges are counted instead of incident edges ('ins' versus 'outs'), the scaling of the connectivity remains unchanged.

Figure 3 shows the overall scaling behavior for all of the data from the diverse set of experiments. As can be seen, these different data sets show the same power law over two orders of magnitude. The scaling is identical to within the error inherent in this data set—each power law yields an exponent of 3/2. The robustness of these scaling results were assessed using a randomized residual technique (Manly, 1997, see Supplementary information).

Network properties of gene duplication model

The results of the computer simulations of network growth are shown in Table 2 for a variety of growth models and for the two different starting networks (network seeds). For comparison, we also show the results for the GRN model, originally introduced by Barabasi and co-workers (Barabási and Albert, 1999). As can be seen, the GRN model produces lower cluster coefficients and longer path lengths than the experimental data. When 'evolving' the networks in the simulation, it is important to establish the initial condition or starting network. Two different seed networks were used—random and clustered. The random graph was generated by starting with a fixed

Table 1. Statistical graph parameters for gene expression networks

Data set	Λ_1 Cluster coefficient	Average path length	Λ_2^* Cluster coefficient	Average path length
Diauxic shift				
Original	0.58	3.0	0.67	2.3
Random	0.17	1.9	0.19	1.9
Cell cycle-alpha factor				
Original	0.66	2.6	0.46	3.5
Random	0.06	2.5	0.15	1.9
Cell cycle-cdc 28				
Original	0.88	2.2	0.71	2.4
Random	0.07	2.4	0.07	2.4

For each of the 3 data sets: Λ_1 and Λ_2^* were generated using four eigenvalues. Thresholds of 0.006 and 0.0012 were applied to Λ_1 and Λ_2^* respectively. The thresholds were chosen to generate adjacency matrices whose SCCs would have roughly 200 nodes (to make the other computations tractable). The SCCs of these networks were computed and the numbers in the first row are the values of the cluster coefficient and the average path length. Random graphs with roughly the same number of nodes and edges the respective SCCs were generated. The second row lists the cluster coefficient and average path lengths for these random equivalents.

Table 2. Statistical graph parameters for simulated networks

Simulation	Cluster coefficient	Average path length	Scaling exponent
Growing random network model			
Random seed	0.02 ± 0.01	7.4 ± 1.4	2.6 ± 0.3
Clustered seed	0.50 ± 0.03	2.8 ± 0.1	2.4 ± 0.1
Gene duplication model			
Random seed	0.03 ± 0.02	9.7 ± 3.8	0.13 ± 0.11 (no)
Clustered seed	0.86 ± 0.03	2.1 ± 0.1	0.31 ± 0.22 (no)
Partial gene duplication model			
Random seed	0.03 ± 0.02	19 ± 8	1.7 ± 0.2
Clustered seed	0.68 ± 0.09	2.4 ± 0.1	1.5 ± 0.3
Gene duplication with preferential rewiring			
Random seed	0.06 ± 0.03	7.6 ± 2.5	1.4 ± 0.15
Clustered seed	0.74 ± 0.06	2.2 ± 0.15	1 ± 1 (unstable)

Network growth was simulated as described in the text. Two different seed graphs were used—the 'random' seed has 70 nodes and 100 edges with a clustering coefficient of 0.017 and the 'clustered' seed has 83 nodes and 362 edges with a cluster coefficient of 0.8. The seed networks were grown for 100 iterations to generate comparable networks of tractable size. All entries represent the average and standard deviation of 100 simulations. Gene duplication model is taken to show no scaling. Gene duplication with preferential rewiring shows scaling that is extremely sensitive to initial conditions in the case of the clustered seed.

set of nodes and randomly assigning linkages between them. The clustered network was generated from our experimental data at high thresholds. For instance, the networks in Figure 2 could serve as seed networks. These networks generally have more of a hub-like structure than the random networks.

For random seeds, all of the gene duplication models showed an increase in the cluster coefficient as the network grows. When a clustered seed is used, the cluster coefficient remains fairly constant with these models. This suggests that the initial conditions in the clustered seed are closer to the stationary state of the growing network. This result is not seen with the GRN model, where the cluster coefficient actually decreases drastically with network growth. Inspection of the results in Table 2 indicates that the mixed duplication models with a clustered seed give comparable graph parameters to those observed in the experimental data (see Table 1), while the GRN is unsuccessful in reproducing these results. An examination of the scaling exponent for the mixed duplication model

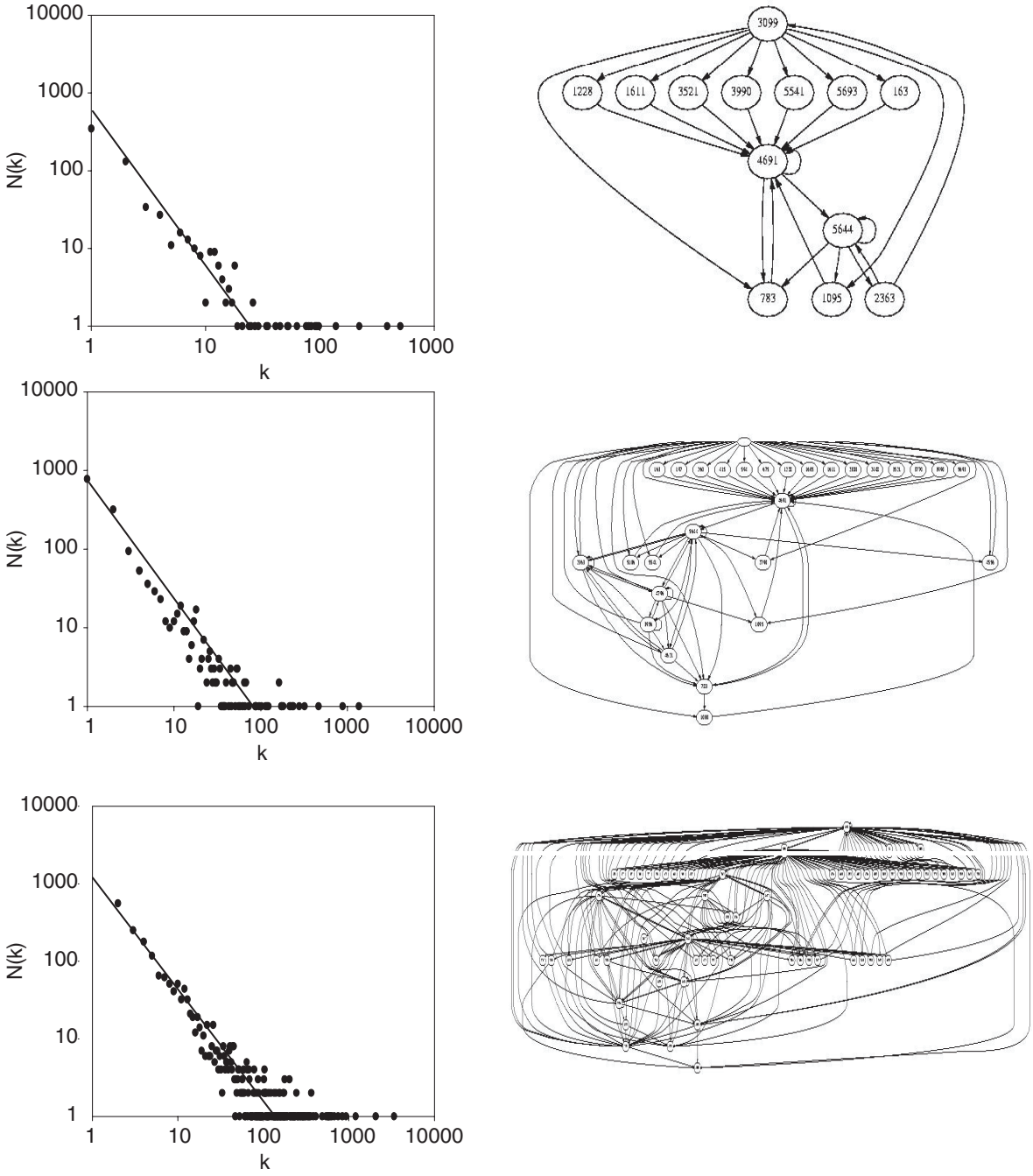

Fig. 2. Plots showing the effect of the threshold parameter on the size of a network and on the scaling of node degrees. Strongly coupled components of networks at various thresholds ϵ (right-hand side). Plot of degree distribution, $N(k)$ versus k for diauxic shift data at different thresholds (left-hand side). Plots show that scaling is independent of threshold value. Actual thresholds differ somewhat between left- and right-hand examples. Threshold was varied by a factor of 2 to generate scaling plots. The lines in right-hand plots represent functions with a slope of $-3/2$. Note that they intersect the axes at different points, however, reflecting the different numbers of nodes in the graphs. Graphs derived from the cdc data of Spellman *et al.* (1998).

also shows agreement with the experimental results (see Supplementary information). Most of the previous linear growth models (GRNs) yield values for the scaling exponent γ that lie in the range $2 < \gamma < 3$. Our analysis of the gene expression data from yeast suggests that these models are not appropriate because we observe exponents

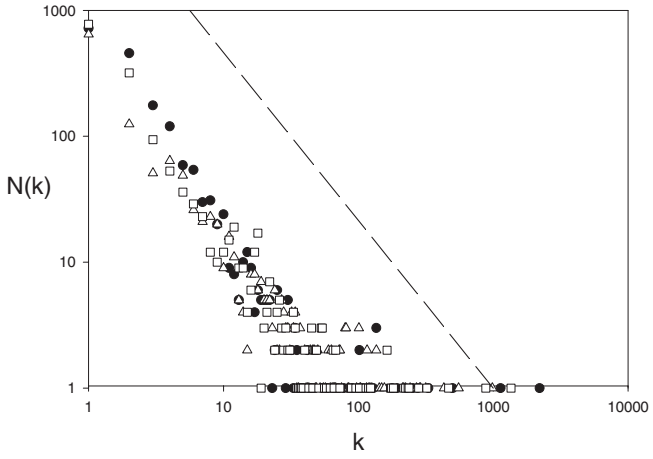

Fig. 3. Plot of the distribution of degrees with $N(k)$, number of nodes with degree k, plotted versus k. Networks for three different gene expression data sets were used—cell cycle data (cdc-28) *bullet*; cell cycle data (alpha) △; diauxic shift data (. See references DeRisi *et al.* (1997) and Spellman *et al.* (1998) for a complete description of data sets. Dashed line is drawn with a slope of $-3/2$. Threshold parameter was set so that the adjacency matrix has connections out of a possible connections.

smaller than 2. So in addition to not giving an appropriate cluster coefficient, the GRNs also do not mimic the scaling properties of expression networks.

SUMMARY AND CONCLUSIONS

In this work, we report and model inferred networks of gene expression from a dynamic analysis of time series of whole genome profiles. These networks are statistically robust and share common properties across very different data sets from yeast. This approach is an intermediate level of analysis, going beyond strictly statistical approaches such as cluster analysis or principal component analysis, to create a phenomenological, causal model of gene control based on dynamic correlations. It falls short of a full quantitative, predictive physico-chemical model. Such network models have great potential utility, however, in investigating genome-wide databases generated by high-throughput technologies of gene activity. While the power of these technologies is the ability to create a global profile of a given functionality, the quality of data and lack of statistical assessment of individual parameters is a difficulty. Given the current state of the technologies, network methods offer attractive approaches to model building and data mining.

These yeast expression networks have a number of interesting properties. They have short mean path lengths characteristic of highly connected networks and high clustering coefficients associated with very 'clique-ish' graphs. Additionally, they show a scale-free distribution of connectivities with scaling exponents that are less than 2. This combination of graph traits is unique and is not observed in other real world networks analyzed to date. These properties also present restrictive constraints for developing models of network formation. Studies of previous models for the growth of networks have elucidated the behavior of some properties of real networks like the Internet, but as we show here, they do not explain the biological networks represented by genetic regulatory networks. These models fail because they cannot yield exponents below 2 and because they often do not have either high cluster coefficients or low mean path lengths. We cannot at this time assess whether these results apply just to the yeast system or have a greater generality. Recently, the properties of a number of biological networks have been explored. Metabolic networks showing the connectivity of substrates show high cluster coefficient and a scaling exponent of 1.6 (Wagner and Fell, 2001). Other studies of metabolic networks show a higher scaling exponent of 2.2 (Jeong *et al.*, 2000). The yeast protein–protein interaction map has been reported also to have high cluster coefficients and a higher exponent of 2.5. Our analysis of the protein–protein data however, using a composite of all the existing databases (Uetz *et al.*, 2000; Ito *et al.*, 2001), gives an exponent of 1.5 (see supplementary information). The results obtained here suggest that some biological networks show lower scaling than other observed networks and may obey a $-3/2$ power law.

The graphical parameters for the experimental networks can be matched to simulated networks using a new network growth model based on gene duplication. Gene duplication provides a natural and compelling model for the growth of genetic regulatory networks. There is now abundant evidence from recent genome analysis from yeast (Seoighe and Wolfe, 1999) to human (Lander *et al.*, 2001) that Ohno's original hypothesis that new genes are almost always created by duplication is largely valid. Gene duplication is now widely accepted as the single most important mechanism for generating new functions and processes (Ohno, 1970). This evolutionary mechanism must be at work in shaping the structure and function of interactions between genes and regulatory networks. We may be seeing evidence of this in the scaling law evident in the yeast data. If this is correct and other, organism-specific evolutionary processes do not obscure the effect, this scaling should be evident in other organisms.

ACKNOWLEDGEMENTS

We gratefully acknowledge funding from the W.M. Keck Foundation, support from the Kenneth T. and Eileen L. Norris Foundation to DG, and NIH grant 1R01 GM63912-01 to TGD.

REFERENCES

Albert,R. and Barabási,A.-L. (2001) Statistical mechanics of complex networks. condmat/0106096.

Amaral,L.A.N., Scala,A., Barthélémy,M. and Stanley,H.E. (2000) Classes of small-world networks. *Proc. Natl Acad. Sci. USA*, **97**, 11149–11152.

Barabási,A.-L. and Albert,R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

Cohen,J.E. (1988) Threshold phenomena in random structures. *Discr. Appl. Math.*, **19**, 113–128.

Dewey,T.G. and Galas,D. (2001) Dynamic models of gene expression and classification. *Func. Integr. Genomics*, **1**, 269–278.

DeRisi,J., Iyer,V. and Brown,P. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Dorogovtsev,S.N. and Mendes,J.F.F. (2001) Scaling properties of scale-free evolving networks: continuous approach. *Phys. Rev. E*, **63**, 056125-1–056125-18.

Heyer,L.J., Kruglyak,S. and Yooseph,S. (1999) Exploring expression data: Identification and analysis of coexpressed genes. *Genome Res.*, **9**, 1106–1115.

Holter,N.S., Maritan,A., Cieplak,M., Fedoroff,N.V. and Banavar,J.R. (2001) Dynamic modeling of gene expression data. *Proc. Natl Acad. Sci. USA*, **98**, 1693–1698.

Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Jeong,H., Tombor,B., Albert,R., Oltvai,Z.N. and Barabási,A.-L. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Kauffman,S. (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.*, **22**, 437–467.

Krapivsky,P.L., Redner,S. and Leyvraz,F. (2000) Connectivity of growing random networks. *Phys. Rev. Lett.*, **85**, 4629–4632.

Lander,E. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Manly,B.F.J. (1997) *Randomization, Boot Strap and Monte Carlo Methods in Biology*. Chapman & Hall, London.

Ohno,S. (1970) *Evolution by Gene Duplication*. Springer, Berlin.

Rzhetsky,A. and Gomez,S.M. (2001) Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. *Bioinformatics*, **17**, 988–996.

Seoighe,C. and Wolfe,K. (1999) Updated map of duplicated regions in the yeast genome. *Gene*, **238**, 253–261.

Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Strogatz,S. (2001) Exploring complex networks. *Nature*, **410**, 268–276.

Uetz,P., Giot,L., Cagney,G., Mansfield,T., Judson,R., Knight,J. Lockshon,D. *et al.* (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

Wagner,A. (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.*, **18**, 1283–1292.

Wagner,A. and Fell,D. (2001) The small world inside large metabolic networks. *Proc. Roy. Soc. London* Series B, (in press).

Watts,D.J. (1999) *Small Worlds-the Dynamics of Networks Between Order and Randomness*. Princeton University Press, Princeton, NH.