STRUCTURE OF GENETIC REGULATORY NETWORKS: EVIDENCE FOR SCALE FREE NETWORKS

L. S. LIEBOVITCH

Florida Atlantic University
Center for Complex Systems & Brain Sciences
Center for Molecular Biology & Biotechnology
Department of Psychology
Department of Biomedical Sciences
777 Glades Road, Boca Raton FL, U.S.A. 33431
E-mail: liebovitch@ccs.fau.edu

V. K. JIRSA

Florida Atlantic University
Center for Complex Systems & Brain Sciences
Department of Physics
777 Glades Road, Boca Raton FL, U.S.A. 33431
E-mail: jirsa@ccs.fau.edu

L. A. SHEHADEH

Florida Atlantic University
Center for Complex Systems & Brain Sciences
Department of Biomedical Sciences
777 Glades Road, Boca Raton FL, U.S.A. 33431
E-mail: shehadeh@ccs.fau.edu

The expression of some genes increases or decreases the expression of other genes forming a complex network of interactions. Typically, correlations between the expression of different genes under different conditions have been used to identify specific regulatory links between specific genes. Instead of that "bottom up" approach, here we try to identify the global types of networks present from the global statistical properties of the mRNA expression. We do this by comparing the statistics of mRNA computed from different network models, including random and fractal, scale free networks, to that found experimentally. The novel features of our approach are that: 1) we derive an explicit form of the connection matrix between genes to represent scale free models with arbitrary scaling exponents, 2) we extend Boolean networks to quantitative regulatory connections and quantitative mRNA expression concentrations, 3) we identify possible types of global genetic regulatory networks and their parameters from the experimental data.

1. Introduction

The information encoded in deoxyribonucleic acid, DNA, in genes is transcribed into messenger ribonucleic acid, mRNA, that is then translated into proteins that perform the structural and functional characteristics of cells. Some of these proteins, called transcription factors, then bind back onto DNA increasing, or decreasing, the expression of other genes. Recent co-regulational studies have used different computational methods to evaluate the correlations in mRNA concentrations under different experimental conditions to determine the network of genetic regulation. These methods include correlation analysis, cluster analysis, neural networks, genetic algorithms, and singular value decomposition (principal component analysis) with or without additional constraints [1-18]. Instead of the "bottom up" method of those co-regulational studies, our approach here emphasizes a "top down" method to try to identify global patterns of genetic regulation. We do this in the following way: 1) We formulate different types of models of networks of genetic regulation. 2) We compute the probability density function, PDF, of the mRNA from each network. 3) We then compare those results to the experimental data from mRNA concentrations measured by cDNA microarray chips. By looking for the global features of the genetic regulatory network, as characterized by the PDF of the mRNA, we may be able to identify different types of genetic regulatory networks without the need to first identify which specific gene regulates which other specific gene.

2. Models of Genetic Regulatory Networks

We represent the genetic regulatory network of N genes by a connection matrix M, where $M(i,j)$ is the regulatory effect of the j-th gene on the i-th gene. Thus, the regulatory effects can be made quantitative which is an extension of the all or nothing (0 or 1) of Boolean models. Connection matrices with both positive (stimulatory) and negative (inhibitory) elements can cause a broad range of complex 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 only stimulatory behavior where all the elements are greater than or equal to zero so that the mRNA concentrations will always reach a steady state with a characteristic probability density function.

We studied connection matrices with random, [19] fractal scale free, and small world [20,21,22] connection topologies. Each connection matrix was 1000 x 1000 to model the interactions between 1000 genes. Interactions between genes i and j were represented by $M(i,j) = 1$, while all other $M(i,j) = 0$.

Each different model is characterized by the degree distribution of the number of genes g connected to other genes k. In the models with random connections, the degree distribution, of either the inputs into the genes or the outputs from the genes, have the Poisson distribution $g(k) = \frac{<k>^k}{k!} e^{-<k>}$ where there are on average $< k >$ interactions between genes. We formulate both Random Output and Random Input

models.

In the fractal scale free models the degree distribution has the power law form $g(k) = Ak^{-a}$. We formulate three types of such models. First, where the degree distribution of the outputs from the genes is a such a power law. In this model the statistical pattern of input regulation into each gene is the same, namely each gene has a small number of inputs from the lower order genes and ever more inputs from ever more higher order genes. We call this the Scale Free Homogeneous model since the statistical pattern of inputs into each gene is the same. In the second model, the statistical pattern of input regulation into each gene is a power law. Different genes have different average numbers of connections to other genes. We therefore call this model the Scale Free Heterogeneous model. In the third model, we adjust both the input and output degree distributions to have the same form, and therefore call this model the Scale Free Symmetric model.

To formulate the scale free models we chose the number of nonzero elements in the columns of $M(i,j)$ for the Homogeneous model, the number of nonzero elements in the rows of $M(i,j)$ for the Heterogeneous model, and the number of nonzero elements in diagonals of $M(i,j)$ for the Symmetric model as given by $Y(r) = (\frac{A}{(a-1)(N-r)})^{1/(a-1)}$ where r is the distance, respectively, from the starting column, row, or corner of the matrix. This relationship was derived by noting that the number $g(k)dk$ of genes with k connections over the interval $(k, k+dk)$ is equal to the number of nonzero elements, dr, with $Y(r) = k$ connections over the corresponding interval $(r, r+dr)$. This relationship between $g(k)$ and $Y(r)$ provides a useful explicit way to generate connection matrices corresponding to a scale free model with a given scaling exponent a without having to from the network iteratively using a growth rule to preferentially add nodes [23,24]. In all three models we varied the scaling parameter over the range $1 \leq a \leq 6$. We also computed the average path length between any pairs of genes through their connections and the clustering coefficient [29]. All of the scale free networks had both shorter path length and higher clustering coefficient than that of the random networks with the same $< k >$ indicating that these scale free networks were all "small world" networks.

In both the random and scale free models we adjusted the density of non-zero elements so that each model had the same number of average connections $< k >= 6$ between genes.

3. Statistics of the mRNA from the Different Models

We then represent the mRNA level expressed by gene i of the N genes by an element in the column vector $X(i)$. 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 $X' = MX$.

Some of the models presented here can be cast in a the form of Markovian matrices, that is, where each row (or column) has an identical sum (equal to 1 if all the elements are properly normalized). In that case, the matrix M has an eigenvalue

of unity and the iteration process $X' = MX$ converges to a steady state. However, the scale free models cannot be cast into that form. In that case, we renormalized the norm of the vector X at each time step. That is, for these models the relative values of the elements reach a steady state in the direction of the vector X, although its length continues to increase or decrease. In reality, biological constraints, such as the finite number of mRNA molecules that can be synthesized, will constrain the norm of the vector X to a finite nonzero value.

Some models, for example, the Random Output model, can be formulated as either a Markovian or non-Markovian matrix. For such models we found that the direct iteration of $X' = MX$ of the Markovian matrix and the renormalization of X at each time step of the iteration of non-Markovian matrix both produced very similar steady state statistical patterns of mRNA expression, giving credence to our renormalization procedure used to compute the models with non-Markovian matrices.

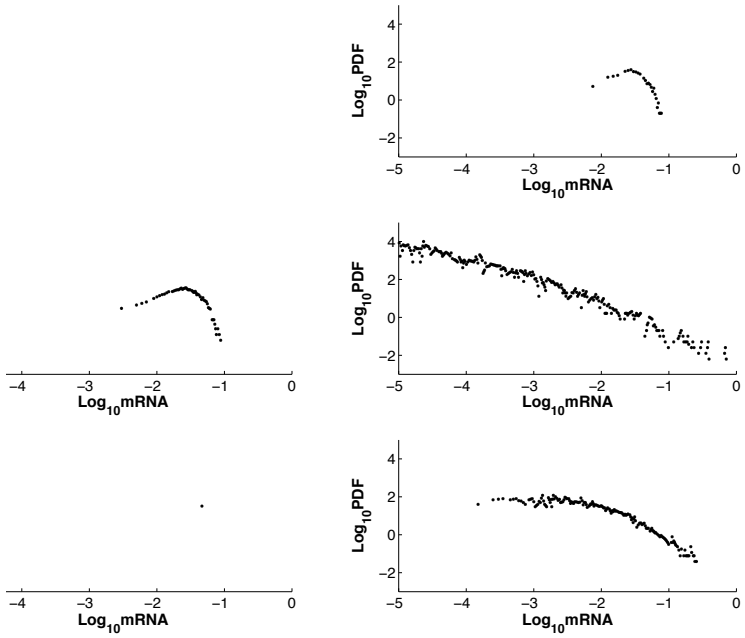

Figure 1. PDFs of mRNA concentrations computed from different models of genetic regulatory networks. Counter-clockwise starting from the top left: 1) Random Output model, 2) Random Input model, 3) Scale Free Symmetric model with scaling exponent a=2, 4) Scale Free Heterogeneous model with a=2, and 5) Scale Free Homogeneous model with a=-2.

At steady state, we then determine the PDF, the probability density function

that any of the mRNA concentrations, $X(i) = x$, is between x and $x + dx$. Our approach is to analyze the statistical form of the PDF of all the mRNA concentrations, and we do not determine which specific gene is associated with which specific mRNA concentration. That makes it possible for our approach to identify global properties of the genetic regulatory network without first needing to identify the specific regulatory connections between specific genes.

The standard method is to compute the PDF from the histogram of the number of values $n(x)$ in each interval $[x, x + dx]$, namely $PDF(x) = \frac{n(x)}{Ndx}$. Using this method, when dx is small, the PDF is well characterized at small x, but poorly characterized at large x since there are few values in each narrow bin in the tail of the distribution. When dx is large more values are captured in each bin at large x, but now the resolution of the PDF is poor at small x. We overcome these limitations by using an algorithm that combines histograms of different bins size dx so that it generates PDFs that are accurate both at small dx and at large dx [25]. The results for 2 random and 3 scale free models are shown in Fig. 1

4. Experimental Data

The mRNA concentrations expressed by thousands or tens of thousands of genes can now be measured simultaneously by cDNA microarray chips where flourescently marked mRNA binds to complementary polynucleotide sequences bound to the chip [26,27]. We calculated the PDFs of the mRNA concentrations measured by cDNA microarray chips used to study circadian rhythms in Drosophila flies [28]. Figure 2 shows the PDF determined from wild type flies kept in total darkness. This form

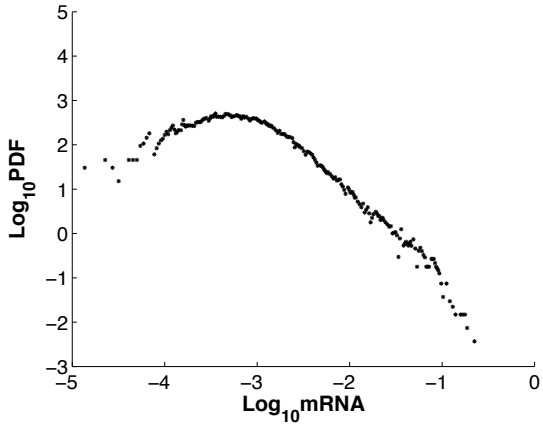

Figure 2. Typical PDF of mRNA concentrations computed from experimental data of virtually all the genes in Drospholia.

of the PDF is typical of several different light and dark conditions for both the wild

type flies and those with a mutation in the period gene that alters their circadian rhythm.

5. Conclusions

The experimental data analyzed here is most like the genetic regulatory network model of a fractal, Scale Free Symmetric network with a scaling exponent a of approximately 2 for the distributions of both the inputs into the genes and the outputs from the genes. However, since we have not shown that the PDFs of the mRNA concentrations are unique for each model, it is possible that other models that we have not studied might also provide an equally good fit to the experimental data.

The Scale Free Symmetric model means that some genes are strongly regulated by the expression of only a few other individual genes, while other genes are regulated by the average expression of a large number of genes. Moreover, the degree distribution of the input regulation into each gene is the same as the degree distribution of the output regulation from each gene.

The strength of our approach presented here is that we have been able to extract some information about the global properties of the genetic regulatory network without first determining which specific genes regulate which other specific genes. This was possible because we compared the PDF of the mRNA predictions of the models to the experimental data without considering which particular gene produced which particular mRNA concentration.

Since we have shown here that different network architectures may have different PDFs of the mRNA concentrations, this may provide a way to identify different subnetworks, for example, metabolic vs. cell cycle regulatory networks. It may also prove useful in determining if those networks change under different conditions, for example, normal vs. metastatically transformed cells. If that is the case, then changes in the PDFs may be useful in identifying such cellular transformations and the degree of those transformations. If the PDFs are not different under different conditions, it may indicate that the overall global structure of these genetic regulatory networks is so important to cell function that as some genes are turned off, other genes may be recruited to fulfill their missing function in the entire network, so that the overall structure of the network remains intact.

Some aspects of these studies may also be useful in the analysis of other network systems. For example, we derived the distribution of nonzero elements in the connection matrix corresponding to scale free distributions with different scaling exponents. This makes it possible to explicitly construct scale free networks without the need to form them iteratively by using growth rules to add nodes to the network. An important question about networks is the relationship between their structure and dynamics. For example, in many experimental cases only the activity of the nodes of the network can be measured and one seeks to use that information to determine the structure of the network. For example, there are electrical record-

ings from nerve cells and one seeks to use this information to understand how they are functionally connected, or there are traffic activity measures at routers and one seeks to understand the topology of the Internet. We found here that the tails of the PDFs of the mRNA concentrations are related to the degree distribution. The mRNA concentrations correspond to the dynamic activity at the nodes of a network and the degree distribution to the structure of a network. This relationship forms an interesting link between the dynamics and the structure of a network, that may be, or may not be, present in other types of networks as well.

6. Acknowledgments

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).

References

1. J.L DeRisi, V.R. Iver, and P.O. Brown, *Science*, **278**, 680-6 (1997). (Data online at http://cmgm.stanford.edu/pbrown/explore/index.html).
2. M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, *Proc Natl Acad Sci USA*, **95**, 14863-8 (1998).
3. G.S. Michaels, D.B. Carr, M. Askenazi, S. Fuhrman, X. Wen, R. Somogyi, *Pac Symp Biocomput* ,**3**, 42-53 (1998).
4. P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown , D. Botstein, B. Futcher, *Mol. Biol. Cell*, **9**, 3273-97 (1998). (Data online at http://cellcycle-www.stanford.edu/).
5. X. Wen, S. Fuhrman, G.S. Michaels, D.B. Carr, S. Smith, J.L. Barker, and R. Somogyi, *Proc. Natl. Acad. Sci. USA*, **95**, 334-9 (1998).
6. M.K.S Yeung, J. Tegner, and J.J. Collins, *Proc. Natl. Acad. Sci. USA*, **9**, 6163-68 (1999).
7. A.J. Hartermink, D.K. Gifford, T.S. Jaakkola,and R.A. Young, *Symp. Biocomputing*, **6**, 422-33 (2001).
8. P. D'haeseleer, X. Wen, S. Fuhrman, R. Somogyi, *Pac. Symp. Biocomputing*, **4**, 41-52 (1999).
9. S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church, *Nat. Genet*, **22**, 281-5 (1999).
10. p. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub, *Proc. Natl. Acad. Sci. USA*, **96**, 2907-12 (1999).
11. M. Wahde and J. Hertz, *Biosystems*, **55**, 129-36 (2000).
12. P. D'haeseleer, S. Liang, and R. Somogyi, *Bioinformatics*, **16**, 707-26 (2000).
13. S. Raychaudhuri, J.M. Stuart, and R.B. Altman, *Pac Symp Biocomput*, **5**, 452-63 (2000).
14. M.P. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M Jr. Ares, and D. Haussler, *Proc. Natl. Acad. Sci. USA*, **97**, 262-7 (2000).
15. O. Alter, P.O. Brown, and D. Botstein, *Proc. Natl. Acad. Sci. USA*, **97**, 10101-6 (2000).
16. N.S. Holter, A. Maritian, M. Cieplak, N.V. Fedoroff, and J.R. Banavar, *Proc. Natl. Acad. Sci. USA*, **98**, 1693-8 (2001).

17. T.G. Dewey, and D.J. Galas, *Funct. Integer Genomics*,**1**, 269-78 (2001).
18. A. Bhan, D.J. Galas, and T.J. Dewey, *Bioinformatics*, **18**, **11**, 1486-93 (2002).
19. P.P. Erdos, and A. Renyi, *Publ. Math. Inst. Hung. Acad. Sci.*, **5**:17-61 (1960).
20. A.L. Barabasi, and R.A. Albert, *Science* , **286**, 509-512 (1999).
21. H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.L. Barabasi, *Nature*, **407**, 651-4 (2000).
22. D.J. Watts and S.H Strogatz , *Nature* , **393**, 440-42 (1998).
23. B.A. Huberman and A. L. Adamic, *Nature* **401**, 131-6 (1999).
24. R. Albert and A.L. Barabasi,*Phys. Rev. Lett.* **85**, 5234-7 (2000)
25. L.S. Liebovitch, A.T. Todorov, M. Zochowski, D. Scheurle, L. Colgin, M.A. Wood, K.A. Ellenbogen, J.M. Herre, and R.C. Berstein, *Phys. Rev. E*, **59**, 3312-19 (1999).
26. M. Schena, Ed., DNA Microarrays: A Practical Approach, Oxford Univ. Press, NY (1999).
27. J.B. Rampal, Ed., DNA Arrays: Methods and Protocols, Human Press (2001).
28. 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**, 9562-7 (2002).
29. R. Albert, and A.L. Barabasi, *Revs. Mod. Phys.*, **74**, 47-97 (2002).