Small World and Scale–Free Network Topologies in an Artificial Regulatory Network Model

P. Dwight Kuo and Wolfgang Banzhaf
Memorial University of Newfoundland, St. John's, NL, Canada A1B 3X5
{kuo,banzhaf}@cs.mun.ca
http://www.cs.mun.ca/∼{kuo,banzhaf}
Tel: (709) 737-8652 Fax: (709) 737-2009

Abstract

Small world and scale–free network topologies commonly exist in natural and artificial systems. Many mechanisms for producing these topologies have been presented in the literature. We present an artificial regulatory network model generated by a duplication / divergence process on a randomly generated genetic string and show that networks with small world and scale–free topologies can be produced with some regularity.

Introduction

Recently, there has been significant interest in small world and scale–free network topologies and potential methods or processes which may generate them (Romualdo et al., 2003; Valverde et al., 2002; Barabasi et al., 2001; Barabasi and Albert, 1999). In the majority of these contributions, the mechanisms for generating such topologies are based on preferential attachment (Romualdo et al., 2003; Valverde et al., 2002; Barabasi and Albert, 1999). In this contribution we work within the framework of a model of an artificial regulatory network first presented by Banzhaf (Banzhaf, 2003a; Banzhaf, 2003b) generated by a duplication / divergence process similar to that presented in (Romualdo et al., 2003). However, their model operates directly on the nodes and edges of the model, bypassing any genetic–type representation of the network.

Here, we show that scale–free and small world network topologies appear with some regularity in the gene–protein network interaction diagram generated by parameterizing the networks by the degree of matching between genes and proteins and discuss possible implications. Duplication and divergence are performed directly on the genetic–string, not on the actual nodes and edges of the interaction network.

It has also been shown that this model can reproduce phenomena found in natural genetic regulatory networks such as heterochrony (Banzhaf, 2003a). As such, this model can relate changes in the timing and intensity of gene expression to tiny pattern changes on bit strings which could possibly provide the algorithmic " missing link " between genotypes subject to constant evolutionary changes and the remarkably stable phenotypes found in the real world.

Background

Regulatory Networks

Regulatory networks are an important new research area in biology (Bower and Boulouri, 2001; Davidson, 2001; Kitano, 2001). With the realization that in higher organisms only a tiny fraction of DNA is translated into proteins, the question of what the rest of the DNA is actually doing becomes all the more pressing. Regulation appears to be a very reasonable answer for a functional role for unexpressed DNA. According to Neidthardt et al. (Neidhardt, 1996), 88% of the genome of the bacterium *E. Coli* is expressed with 11% suspected to contain regulatory information (also see Thomas (Thomas, 1999)).

In addition, it has been recognized that the DNA information controlling gene expression is the key to understanding differences between species and thus to evolution (Hood and Galas, 2003).

The three major genetic mechanisms, all tied to regulation (Davidson, 2001) which allow such a variety of reactions of living organisms to the pressure for survival are:

1. Interactions between the products of genes
2. Shifts in the timing of gene expression (heterochrony)
3. Shifts in the location of gene expression (spatial patterning)

These mechanisms allow nature to set up and control the mechanisms of evolution, development and physiology. Since many evolutionary effects can be traced back to their regulatory causes, regulatory networks mediate between development and evolution thus unfolding the patterns and shapes of organism morphology and behaviour (Davidson, 2001; Banzhaf, 2003b).

Studying models of regulatory networks can help us understand some of these mechanisms providing lessons for biology and in the area of artificial evolution.

Scale–Free Network Topologies

It has been found that a high degree of self–organization may characterize the large–scale properties of complex networks

(Barabasi and Albert, 1999). Many researchers have shown that the probability $P(k)$ that the number of nodes connected to k (vertex degree) other nodes in a network decays as a power law, following: $P(k) \sim k^{-\gamma}$ in systems as diverse as the internet (Faloutsos et al., 1999), protein interaction networks (Wuchty, 2001), the electrical power grid of the western United States of America (Watts, 2003), the neuronal network of the worm *Caenorhabditis Elegans* (Watts, 2003), and the network of citations of scientific papers (Barabasi et al., 2002).

It has thus been suggested that scale–free networks emerge in the context of a dynamic network with the addition of new vertices connecting preferentially to vertices which are highly connected in the network (Barabasi and Albert, 1999), as well as through explicit optimization (Valverde et al., 2002).

Small World Network Topologies

Small world graphs can be defined as any graph with n vertices and average vertex degree k that exhibits $L \approx L_{random}(n,k) \sim \frac{\ln(n)}{\ln(k)}$, and $C \gg C_{random} \sim \frac{k}{n}$ for $n \gg k \gg \ln(n) \gg 1$ (Watts, 2003). C is referred to as the clustering coefficient (if vertex v has k_v neighbors, $C = \frac{2}{n}\sum_{v=1}^{n}\left(\frac{k_v(k_v-1)}{2}\right)$) of the network while L is the characteristic path–length of the network (average number of links connecting two nodes). L_{random} and C_{random} refer respectively to the characteristic path–length and clustering coefficient for a completely random graph with the same k and n.

Like scale–free network topologies, the small world topology has also been noted in many networks (including those with scale–free topology) such as the electrical power grid of the western United States of America (Watts, 2003), the neuronal network of the worm *Caenorhabditis Elegans* (Watts, 2003), and the network of film actors who have acted in the same films (Watts, 2003).

Artificial Regulatory Network Model

Our artificial regulatory network (ARN) model is based on work by Banzhaf (Banzhaf, 2003a; Banzhaf, 2003b). In this model, the ARN consists of a genome represented by a bit string with direction (i.e. 5' $\rightarrow$ 3' in DNA) and mobile "proteins" which are equipped with bit patterns for interactions with the genome. The proteins are able to wander about in order to interact with the genome, notably at "regulatory" sites located upstream (3' $\rightarrow$ 5' direction) from genes. Attachment to these sites inhibits or activates the production of the corresponding protein thereby demonstrating the mechanisms of activation and inhibition.

Creation of the genome commences with the generation of a random 32–bit string. This string is then used in a series of whole length duplications similar to those found in nature (Wolfe and Shields, 1997) followed by mutations in order to generate a genome of length L_G. A "promotor" bit sequence of 8–bits was then arbitrarily selected to be "01010101". In a genome generated by randomly choosing "0" s and "1" s, this one–byte pattern can be expected to appear with probability $2^{-8} = 0.39\%$. Since the promotor pattern itself is repetitive, overlapping promotors or periodic extensions of the pattern are not allowed, i.e. a bit sequence of "0101010101" (10–bits) is detected as a single promotor site starting at the first bit. The promotor signals the beginning of a gene on the bit string which is analogous to an open reading frame (ORF) on DNA - a long sequence of DNA that contains no "stop" codon and therefore encodes all or part of a protein. This gene is set to a fixed length of $l_{gene} = 5$ 32–bit integers which results in an expressed bit pattern of 160 bits for each gene. Therefore, genes can thus be created by complete duplications of previously created genes, mutation, and / or combinations of the end and starting sequences of the genome during duplication.

Immediately upstream from the promotor exist two 32–bit segments which represent the enhancer and inhibitor sites. As previously mentioned, attachment of proteins to these sites results in changes to protein production for the corresponding genes. In this model, we assume only one regulatory site for the expression and one site for the suppression of protein production. This is a radical simplification since natural genomes may have 5–10 regulatory sites that may even be occupied by complexes of proteins (Davidson, 2001; Banzhaf, 2003b).

The model presented here completely disregards processes such as transcription, and neglects elements such as introns, RNA–like mobile elements and translation procedures resulting in a different alphabet for proteins. This last mechanism is replaced as follows: Each protein is a 32–bit sequence which results from a many–to–one mapping of its corresponding gene which contains five 32–bit integers. The protein sequence is created by performing the majority rule on each bit position of these five integers so as to arrive at a 32–bit protein. Ties (not possible with an odd number for l_{gene}) for a given bit position are resolved by chance.

These proteins may then be examined to see how they may "match" with the genome. This comparison is implemented by using the XOR operation which returns a "1" if both inputted bits are complementary. In this scheme, the degree of match between the genome and the protein bit patterns is specified by the number of bits set to "1" during an XOR operation. In general it can be expected that a Gaussian distribution results from measuring the match between proteins and bit sequences in the random genome (Banzhaf, 2003b).

If we make the simplifying assumption that the occupation of two regulatory sites per gene modulates the expression of the corresponding protein, we may deduce an interaction network comprising the different genes and proteins which can be parameterized by strength of match.

By examining the interaction networks at different matching strengths (thresholds) we may obtain different network topologies for the same connected network components. An example is shown in Figs. 1 and 2. Each node in the diagram represents a gene found in the genome along with its corresponding protein forming a gene–protein pair. Edges in the diagram represent some form of influence of one gene's protein on another gene. For the diagrams presented, a random genome was created by the previously mentioned duplication and mutation procedure with the network interaction diagrams being created at thresholds of 21 and 22.

It must be stressed that although the actual genome has not changed, by simply changing the threshold parameter, we have obtained a different network topology. It may be noted by the more astute reader that the diagrams in Figs. 1 and 2 possess different numbers of genes and proteins. This is due to the fact that only connected gene–protein pairs are displayed in the diagrams. Should a change in the parameterized threshold lead to the creation of an isolated node, it is deleted from the diagram. Also note that only the largest network of interactions is displayed.

It is possible to have multiple clusters of gene–protein interactions that are not interconnected. This is likely to occur as the threshold level is increased. As connections between gene–protein pairs are lost due to the threshold, each cluster of gene–protein pairs begins to become isolated from the others. This often occurs abruptly indicating a phase transition between sparse and full network connectivity.

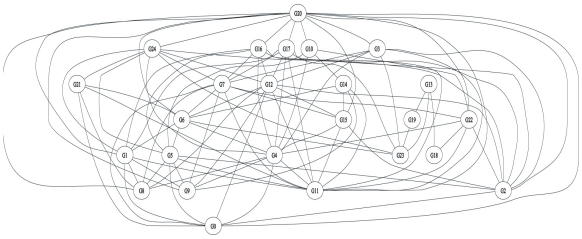

Figure 1: Gene-protein interaction network for a random genome at a threshold of 21 bits.

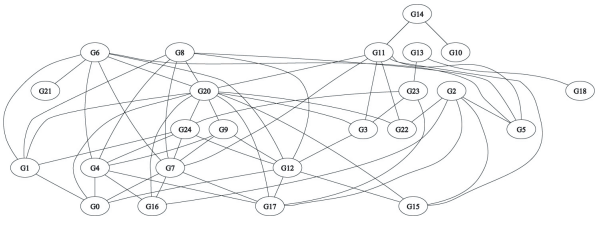

Figure 2: Gene-protein interaction network for a random genome at a threshold of 22 bits.

Results

At mutation rates of 1% and 5%, 200 genomes were generated by 12 duplication events per genome leading to individual genomes of length $L_G = 131072$. From these genomes, the number of genes were then determined based on the number of promotor patterns present. The distribution of the number of genes present in the genome of size L_G is shown in Figs. 3 & 4.

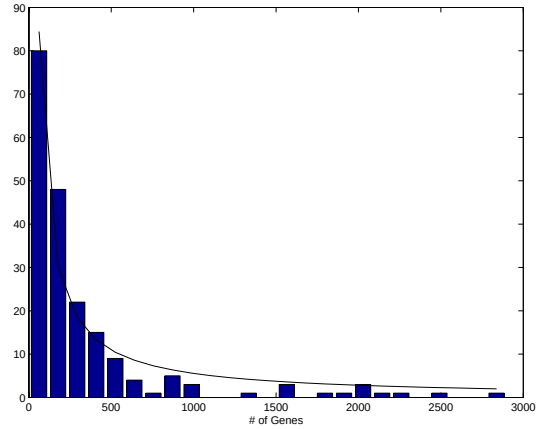

Figure 3: Histogram of the number of genes in each genome (200 genomes) fitted to a power law: $P(g) \sim g^{-\gamma}$ for a mutation rate of 1.0%. γ was calculated to be 0.9779.

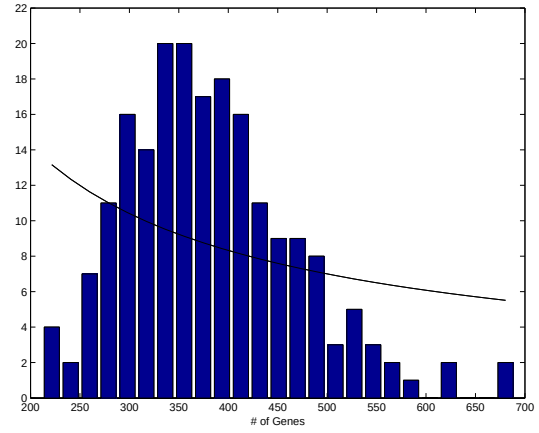

Figure 4: Histogram of the number of genes in each genome (200 genomes) fitted to a power law: $P(g) \sim g^{-\gamma}$ for a mutation rate of 5.0%.

It can be observed that the distribution of the number of genes in Fig. 3 follows a power–law distribution. As well, if we turn the mutation rate lower to 0.1% (results not shown) the distribution of the number of genes again shows a scale–free like distribution.

However, in Fig. 4 the apparent distribution is disrupted. This is attributed to the higher rate of mutation. At such a

mutation rate, the rewiring of the network becomes so prevalent that it begins to disrupt the duplication of nodes leading to a randomly connected network. For an 8 bit promotor, the probability that it remains intact after one duplication event is only 66% at a mutation rate of 5%. Therefore, it can be expected that many of the genes copied during the duplication process will be subsequently destroyed in later duplication steps. However, there will also be other genes which arise from this higher mutation rate. But, these new genes will also be easily destroyed via mutation. Genomes which start with very large numbers of genes are disrupted early on in the duplication process by mutation, while those with few genes obtain additional genes through mutation.

To test this explanation, we created genomes of length L_G completely at random without the use of duplication / divergence. The distribution of these completely randomly generated networks are shown in Fig. 5. As can be seen, this distribution is quite similar to that generated in Fig. 4 lending additional support to the hypothesis that at 5% mutation the network topology becomes randomized. Therefore, we may use the distribution of the number of genes in networks generated by duplication / divergence as an estimate of the effect of mutation rate on the network as compared to randomly generated genomes.

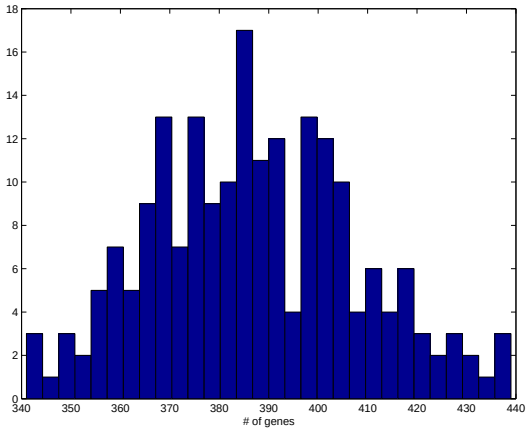

Figure 5: Histogram of the number of genes in 200 genomes whose bits have been chosen at random.

In general, the duplication process, despite being performed directly on the genetic string can be considered to be similar to the mechanism of preferential attachment.

Consider the duplication process on a string which contains multiple genes while neglecting the effects of mutation. For the case of this argument, we also assume that no additional genes are created from a duplication event by joining the end and beginning of one genome string. We start with a network of 5 gene–protein pairs connected as shown on the left side of Fig. 6 and proceeding through a single duplication event generating the network shown on the right side.

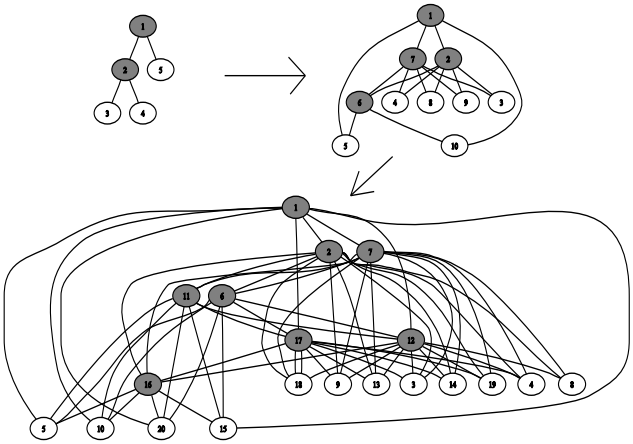

Figure 6: An example of the effect of two duplication events. Highly connected (shaded) nodes become even more higly connected (preferential attachment). Each node represents a gene / protein pair; each edge represents an interaction between gene / protein pairs.

It can be seen that the more highly connected nodes on the left, nodes 1 and 2 and their copies 6 and 7 (shown in gray), become even more highly connected after a single duplication event. This can again be seen in the third diagram which shows the result of another duplication event. As the number of duplication events increases, the difference in the number of connections between highly connected nodes and less connected nodes increases. This can be thought of as a form of preferential attachment since nodes that are already highly connected will become even more so after subsequent duplication events. Preferential attachment has been shown to be a mechanism which can generate scale–free networks (Barabasi and Albert, 1999; Romualdo et al., 2003).

However, this neglects the mechanism of mutation. Mutation may be thought of as an operator which reorganizes the network. If mutations should occur on a gene, this may either change the gene–protein pair's binding site, or the generated protein thus reorganizing a portion of the network. The other possibilities are that mutations may either disrupt the promotor pattern in effect deleting a gene–protein pair from the network, or create a new gene-protein pair by creating a new promotor site.

With these considerations in mind, we may then examine the networks generated by these genomes to see whether their topologies may be considered scale–free and / or small world.

The network of gene–protein interactions was parameterized by the threshold value leading to a maximum of 32 possible networks for each genome. The histograms of the probability of being connected with k components were fitted to the equation $\alpha k^{-\gamma}$ for each threshold value using the sum of

least squares method. The threshold value which produced a γ value closest to 2.5 was kept. It has been found that a large number of networks which have displayed scale–free behaviour exhibit values of $2 < \gamma \leq 3$ (Goh et al., 2002).

Values for the parameter γ characterizing scale–free networks were also calculated for each of the genomes and are shown in Figs. 7 & 8.

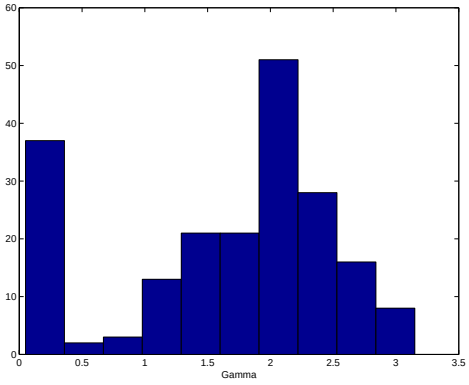

Figure 7: Distribution of values of γ for the best fit of $P(k) \sim k^{-\gamma}$ with a mutation rate of 1.0%.

It can be seen that there exist many genomes created at random which may be considered to satisfy the definition of a scale–free network. In Fig. 7 there is a large number of networks whose coefficient γ is close to zero. This can be attributed to the fact that since the mutation rate is low, the probability of discovering new promotor patterns through subsequent duplication / divergence steps is not high. Therefore, if there were few promotors in the initial starting string, there will often be few genes in the overall genome. With a small number of genes, the scale–free coefficient γ will often be of small magnitude.

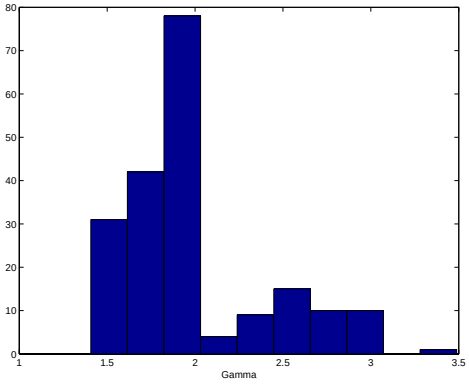

Figure 8: Distribution of values of γ for the best fit of $P(k) \sim k^{-\gamma}$ with a mutation rate of 5.0%.

For each network the clustering coefficient, C, and the characteristic path length, L, were calculated and compared to the corresponding metric for a randomly connected network of the same size and vertex degree distribution. The threshold value that produced a network with the smallest absolute difference $\mid L - L_{random} \mid$ that also satisfied $C \gg C_{random}$ were taken to be those most characteristic of the small world network topology. The additional constraint that $L > 1.3$ was also enforced so as to try to exclude graphs that were close to being fully connected. The distributions for

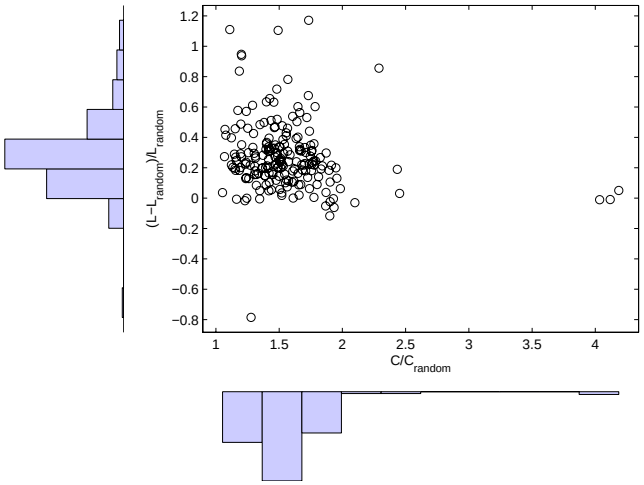

Figure 9: Scatter plot and histograms of values of $\frac{C}{C_{random}}$ and $\frac{L_{random}-L}{L_{random}}$ for each of the randomly generated genomes (200 genomes) with a mutation rate of 1.0%.

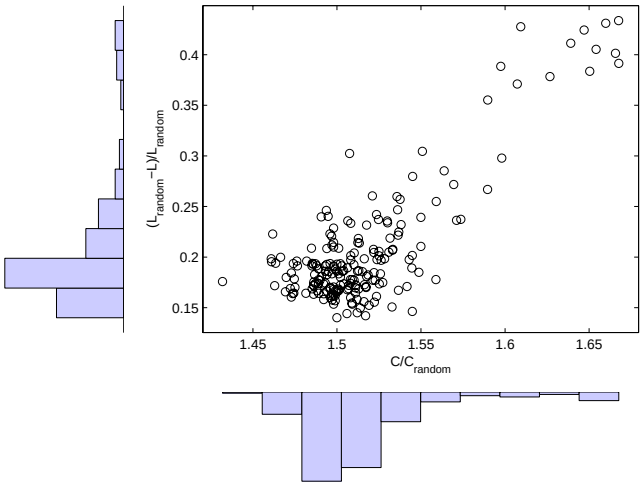

Figure 10: Scatter plot and histograms of values of $\frac{C}{C_{random}}$ and $\frac{L_{random}-L}{L_{random}}$ for each of the randomly generated genomes (200 genomes) with a mutation rate of 5.0%.

the clustering coefficient and the characteristic path length obtained from the 200 genomes for both rates of mutation are shown in Figs. 9 & 10.

From these figures, it can be seen that in the majority of genomes, there exists a threshold at which the interaction network approaches or satisfies the definition of a small world network topology.

Conclusions

A model of an artificial regulatory network model has been presented. The construction of such a network using a simple whole genome duplication process directly on a genetic–string representation of the genome produces a network construction scheme similar to preferential attachment. The addition of a mutation operator introduces a kind of rewiring of the network topology by changing activation / inhibition sites, creating / destroying gene–protein pairs and changing the configuration of proteins which the genes code for. Examining networks generated in this way by varying the threshold at which genes and proteins may interact shows that many of these regulatory networks display the charateristics of small world and scale–free network topologies with some regularity.

Note that we have assumed that duplication proceeds by duplicating the whole genome which occurs relatively rarely in nature (Wolfe and Shields, 1997; Nadeau and Sankoff, 1997). Future work may include investigating the effects of shorter length duplication events on regulatory network topologies.

Acknowledgements

The authors would like to kindly thank François Képès of Atelier de Génomique Cognitive, CNRS for helpful discussions and suggestions.

References

Banzhaf, W. (2003a). Artificial regulatory networks and genetic programming. In Riolo, R. L. and Worzel, B., editors, *Genetic Programming Theory and Practice*, chapter 4, pages 43–62. Kluwer.

Banzhaf, W. (2003b). On the dynamics of an artificial regulatory network. In Banzhaf, W., Christaller, T., Dittrich, P., Kim, J. T., and Ziegler, J., editors, *Advances in Artificial Life – Proceedings of the 7th European Conference on Artificial Life (ECAL)*, volume 2801 of *Lecture Notes in Artificial Intelligence*, pages 217–227. Springer Verlag Berlin, Heidelberg.

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

Barabasi, A. L., Erzsebet, R., and Vicsek, T. (2001). Deterministic scale-free networks. *Physica A*, 299:559–564.

Barabasi, A. L., Jeong, H., Ravasz, R., Neda, Z., Vicsek, T., and Schubert, A. (2002). Evolution of the social network of scientific collaborations. *Physica A*, 311:590–614.

Bower, J. and Boulouri, H., editors (2001). *Computational Modeling of Genetic and Biochemical Networks*. MIT Press, Cambridge, MA.

Davidson, E. H. (2001). *Genomic Regulatory Systems*. Academic Press, San Diego, CA.

Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999). On power-law relationships of the internet topology. In *SIGCOMM*, pages 251–262.

Goh, K. I., Oh, E., Jeong, H., Kahng, B., and Kim, D. (2002). Classification of scale–free networks. *Proceedings of the National Academy of Sciences, USA*, 99(20):12583–8.

Hood, L. and Galas, D. (2003). The digital code of DNA. *Nature*, 421:444–448.

Kitano, H., editor (2001). *Foundations of Systems Biology*. MIT Press, Cambridge, MA.

Nadeau, J. and Sankoff, D. (1997). Comparable rates of gene loss and functional divergence after genome duplications early in verebrate evolution. *Genetics*, 147:1259–1266.

Neidhardt, F. C. (1996). *Escherichia Coli and Salmonella typhimurium*. ASM Press, Washington, DC.

Romualdo, P., Smith, E., and Solé, R. (2003). Evolving protein interaction networks through gene duplication. *Journal of Theoretical Biology*, 222:199–210.

Thomas, G. H. (1999). Completing the e.coli proteome: a database of gene products characterised since completion of the genome sequence. *Bioinformatics*, 7:860–861.

Valverde, S., Ferrer Cancho, R., and Solé, R. (2002). Scale–free networks from optimal design. *Europhysics Letters*, 60:512–517.

Watts, D. (2003). *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton, NJ: Princeton University Press.

Wolfe, K. and Shields, D. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387:708–713.

Wuchty, S. (2001). Scale-free behavior in protein domain networks. *Molecular Biology & Evolution*, 18(9):1694–1702.