Available online at www.sciencedirect.com
SCIENCE DIRECT®
BioSystems xxx (2006) xxx–xxx
www.elsevier.com/locate/biosystems

# Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence

P. Dwight Kuo[a,*], Wolfgang Banzhaf[a], André Leier[b]

[a] Department of Computer Science, Memorial University of Newfoundland, St. John's, NL, Canada A1B 3X5
[b] Advanced Computational Modelling Centre, University of Queensland, Brisbane, Qld 4072, Australia

Received 18 October 2005; received in revised form 21 January 2006; accepted 23 January 2006

**Abstract**

Topological measures of large-scale complex networks are applied to a specific artificial regulatory network model created through a whole genome duplication and divergence mechanism. This class of networks share topological features with natural transcriptional regulatory networks. Specifically, these networks display scale-free and small-world topology and possess subgraph distributions similar to those of natural networks. Thus, the topologies inherent in natural networks may be in part due to their method of creation rather than being exclusively shaped by subsequent evolution under selection.

The evolvability of the dynamics of these networks is also examined by evolving networks in simulation to obtain three simple types of output dynamics. The networks obtained from this process show a wide variety of topologies and numbers of genes indicating that it is relatively easy to evolve these classes of dynamics in this model.

*Keywords:* Regulatory networks; GRNs; Network motifs; Scale-free; Small-world; Duplication and divergence

## 1. Introduction

Regulatory networks have become an important new area of research in the biological and biomedical sciences (Bower and Bolouri, 2001; Davidson, 2001; Kitano, 2001). Specifically, the DNA information controlling gene expression (i.e. regulation) is the key to understanding differences between species and to evolution (Hood and Galas, 2003). Taking these regulatory interactions as a whole, a network of interactions (a so-called regulatory network) can be visualized where genes interact by regulating other genes and their products to produce and regulate a myriad of cellular processes and functions. This allows nature to set up and control the mechanisms of evolution, development and physiology. Studying models of regulatory networks can help us to understand some of these mechanisms providing valuable lessons for biology.

This contribution uses an artificial genetic regulatory network model to pose questions regarding the topological organization of regulatory networks. Specifically, ensembles of this network model are investigated to determine whether they may be classified as scale-free, small-world and possess network motifs. In addition, the networks are then evolved toward simple output dynamics.

* Corresponding author. Tel.: +1 8582435763; fax: +1 7097397026.
*E-mail addresses:* kuo@cs.mun.ca, pdkuo@ucsd.edu (P. Dwight Kuo), banzhaf@cs.mun.ca (W. Banzhaf), leier@maths.uq.edu.au (A. Leier).

0303-2647/$ – see front matter © 2006 Published by Elsevier Ireland Ltd.
doi:10.1016/j.biosystems.2006.01.004

## 2. Background

### 2.1. Topological measures

Since one of the most basic features of any complex network is its structure, it is natural to investigate network connectivity. The structure of networks is often constrained and shaped by the growth processes that create them (including evolution in the case of natural networks). Studying the topology of natural networks allows an understanding of the structures and dynamics which have been exploited by nature. By comparing the topologies of artificial networks with natural networks, questions regarding the benefits of one topology over another can be answered. In addition, some insights into the growth processes which create particular topologies may be gained.

Typically, nodes in such an abstraction represent individual genes and their associated proteins while the directed edges which connect the nodes represent one gene's effect (excitatory or inhibitory) on another.

#### 2.1.1. Scale-free network topologies

A topological feature often found in large complex networks is the so-called "scale-free" topology. In networks of such a topology, the vertex degree distribution, $P(k)$, decays as a power-law. This has been shown for a variety of biological systems (Wuchty, 2001; Watts, 2003; Jeong et al., 2000; Guelzim et al., 2002; van Noort et al., 2004; Babu et al., 2004). A scale-free network topology can emerge in the context of a growing network with the addition of new vertices connecting preferentially to vertices which are highly connected in the network (Barabási and Albert, 1999), as well as through explicit optimization (Valverde et al., 2002) and duplication and divergence (Romualdo et al., 2003; Kuo and Banzhaf, 2004).

#### 2.1.2. Small-world network topologies

Another topological feature found in large complex networks is the so-called "Small-world" topology. Watts (2003) defines a Small-world graph as any graph with $n$ vertices and average vertex degree $k$ that exhibits $L \approx L_{\text{random}}(n, k) \sim \frac{\ln(n)}{\ln(k)}$ and $C \gg C_{\text{random}} \sim \frac{k}{n}$ for $n \gg k \gg \ln(n) \gg 1$. $C$ is the clustering coefficient which is defined as follows: if vertex $v$ has $k_v$ neighbours, $C = \frac{2}{n}\sum_{v=1}^{n}\left(\frac{k_v(k_v-1)}{2}\right)$, where $L$ is the characteristic path-length of the network (average number of links connecting two nodes). $L_{\text{random}}$ and $C_{\text{random}}$ refer to the characteristic path-length and clustering coefficient for a random graph with the same $k$ and $n$, respectively. Small-world topology has also been noted in biological networks (Watts, 2003; van Noort et al., 2004).

#### 2.1.3. Network motifs

The previous two topological measures characterize networks at the global level. Local graph properties of networks have also been investigated such as static network motifs (Milo et al., 2002, 2004; Shen-Or et al., 2002; Wuchty et al., 2003; Yeger-Lotem et al., 2004; Dobrin et al., 2004; Mangan and Alon, 2003; Vazquez et al., 2004; Banzhaf and Kuo, 2004).

Network motifs are defined as the structural elements (subgraphs) which occur in statistically significant quantities in the networks under consideration as compared to random networks (Milo et al., 2002). The implication of having certain subgraphs being found in greater abundance than would be expected in similar random networks is that these local network motifs may convey a functional advantage to the system. It is believed that studying network motifs can lead to a better understanding of the potential basic structural elements which make up complex networks. Several motifs such as the bi-fan (Kashtan et al., 2004), the feed-forward loop (Mangan and Alon, 2003) and the feedback loop (Kashtan et al., 2004) have been the subject of study.

Tables A.1 (three-nodes), A.2 and A.3 (four-nodes) show connection patterns in directed graphs including auto-regulatory connections. A presentation of all four-node connection patterns is impractical due to space limitations.

### 2.2. Artificial regulatory network model

The artificial regulatory network (ARN) model considered here (Banzhaf, 2003a,b; Banzhaf and Kuo, 2004; Kuo and Banzhaf, 2004; Kuo et al., 2004) consists of a bit string representing a genome with direction (i.e. $5' \rightarrow 3'$ in DNA) and mobile "proteins" which interact with the genome through their constituent bit patterns. Proteins are able to interact with the genome, most notably at "regulatory" sites located upstream from genes. Attachment to these sites produces either inhibition or activation of the corresponding protein. These interactions may be interpreted as a regulatory network with proteins acting as transcription factors.

A "promoter" signals the beginning of a gene on the bit string 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. Each gene is set to a fixed length of $l_{\text{gene}} = 5$ 32-bit integers which results in an expressed bit pattern of 160-bits. A promoter bit sequence of 8-bits was arbitrarily selected to be “01010101”. By randomly choosing “0”s and “1”s to generate a genome, any one-byte pattern can be expected to appear with probability $2^{-8} = 0.39\%$. Since the promoter pattern itself is repetitive, overlapping promoters or periodic extensions of the pattern are not allowed, i.e. a bit sequence of “0101010101” (10-bits) is detected as a single promoter site starting at the first bit. However, regions associated with one gene may overlap with another should a promoter pattern also exist within a portion of the coding region of a gene. In such cases, each gene is treated independently.

Immediately upstream from the promoter exist two additional 32-bit segments which represent the enhancer and inhibitor sites. As previously mentioned, attachment of proteins (transcription factors) to these sites results in changes to protein production for the corresponding genes (regulation). It is assumed that only one regulatory site exists for the increase of expression and one site for the decrease of expression of a given protein. This is a radical simplification since natural genomes may have 5–10 regulatory sites per gene that may even be occupied by complexes of proteins (Banzhaf, 2003a).

Processes such as transcription, diffusion, spatial variations and elements such as introns, RNA-like mobile elements and translation procedures resulting in a different alphabet for proteins are neglected. This last mechanism is replaced as follows. Each protein is a 32-bit sequence constructed by a many-to-one mapping of its corresponding gene which contains five 32-bit sequences. The protein sequence is created by performing the majority rule on each bit position of these five sequences so as to arrive at a 32-bit protein. Ties (not possible with an odd number for $l_g$) for a given bit position are resolved by chance.

Proteins may then be examined to see how they “match” with the genome at the regulatory sites. This comparison is implemented using the XOR operator which returns a “1” if bits on both patterns are complementary. 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, a Gaussian distribution results from measuring the match between proteins and bit sequences in a randomly generated genome (Banzhaf, 2003a). By making the simplifying assumption that the occupation of both of a gene’s regulatory sites modulates the expression of its corresponding protein, a gene–protein interaction network may be deduced comprising the different genes and proteins parameterized by strength of match. The bit-string for one gene is shown in Fig. 1.

The rate at which protein $i$ is produced is given by:

$$\frac{\mathrm{d}c_i}{\mathrm{d}t} = \frac{\delta(e_i - h_i)c_i}{\sum_j c_j} \tag{1}$$

$$e_i, h_i = \frac{1}{N}\sum_j^N c_j \exp(\beta(u_j - u_{\max})) \tag{2}$$

where $e_i$ and $h_i$ represent the excitation and inhibition of the production of protein $i$, $u_j$ represents the number of matching bits between protein $j$ and activation or inhibition site $i$, $u_{\max}$ represents the maximum match (in this case, 32), $\beta$ and $\delta$ are positive scaling factors, and $c_i$ is the concentration of protein $i$ at time $t$. The concentrations of the various proteins are required to sum to 1. This ensures competition between binding sites for proteins.

The effect of one gene’s products on another can be investigated in the ARN model by looking at the degree of match between one gene’s protein and another’s regulatory sites (one excitatory and one inhibitory site). At different matching strengths (thresholds), different network topologies are obtained. An example is shown in Figs. 2 and 3. Each node in the diagram represents a gene found in the genome along with its corresponding pro-

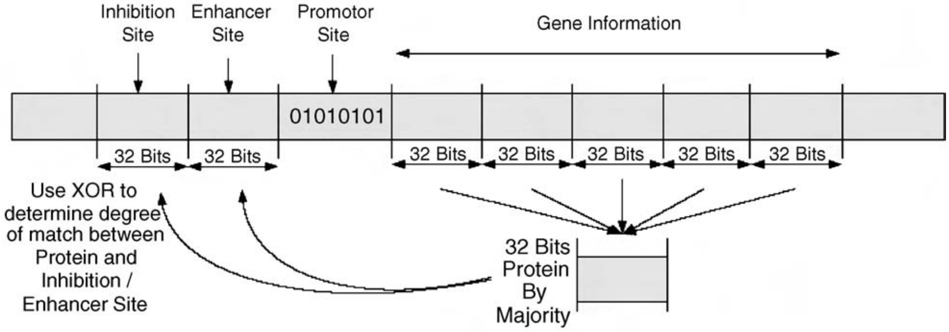


Fig. 1. Bit string for one gene in the ARN model.

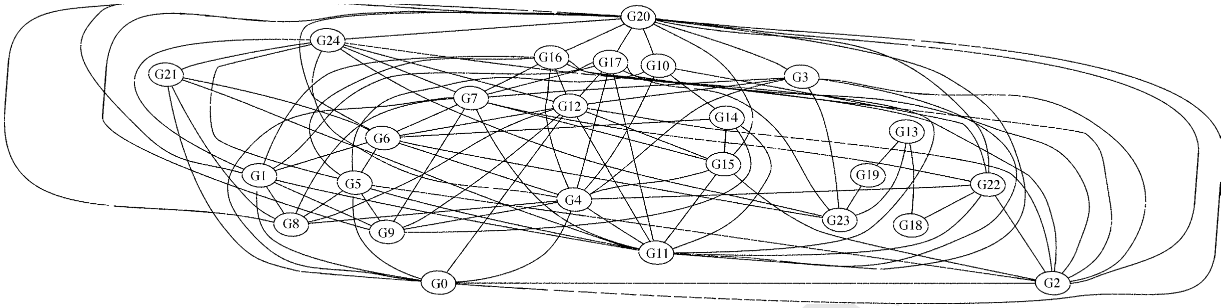


Fig. 2. Gene–protein interaction network for a random genome at a threshold of 21 bits.

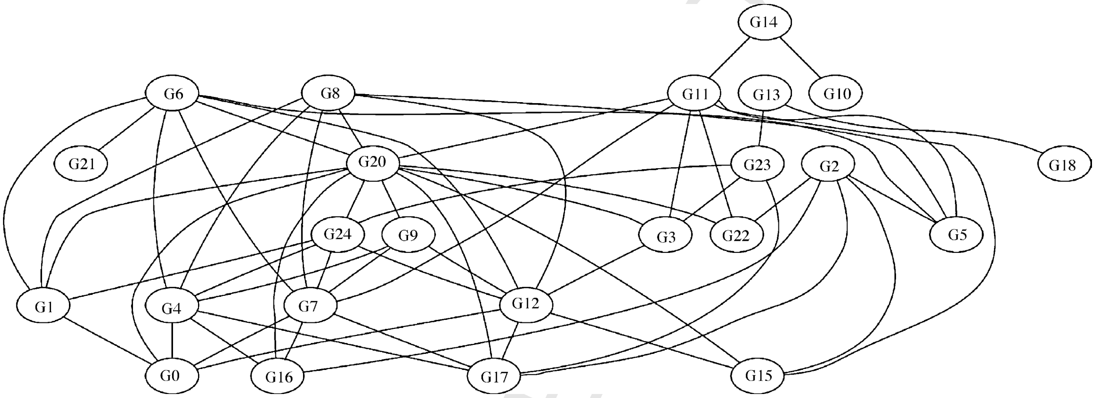


Fig. 3. Gene–protein interaction network for a random genome at a threshold of 22 bits.

tein forming a gene–protein pair. Edges in the diagram represent a regulatory influence of one gene's protein on another gene. For the diagrams presented, the network interaction diagrams at thresholds of 21 and 22 are shown. Fig. 3 is in fact a subgraph of Fig. 2.

Although the actual genome has not changed, by simply changing the threshold parameter, different network topologies are obtained. Figs. 2 and 3 also possess different numbers of genes since only connected gene–protein pairs are displayed. Should a change in the parameterized threshold lead to the creation of an isolated node, it is deleted from the diagram. 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 becomes isolated from the others. This often occurs abruptly indicating a phase transition between sparse and full network connectivity. The relationship between the number of edges in the graph and the threshold is shown in Fig. 4 for a sample of 200 networks. As the threshold increases from 0 to 32 (the *x*–axis), the fraction of edges in the graph over the number of edges in a fully connected network of the same number of nodes (also the number of edges in any ARN graph at threshold 0) goes from 1.0 to 0.0. There is a sharp transition from full connectivity to no connectivity.

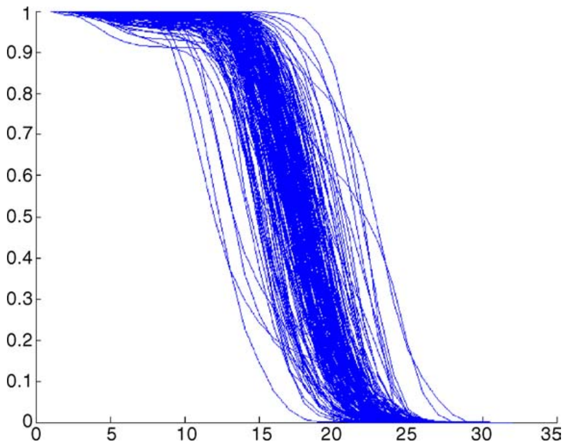


Fig. 4. Diagram showing the fraction of edges in a graph at a given threshold (*x*–axis) compared to a fully connected graph for 200 networks.

### 2.3. Whole genome duplication and divergence

Whole genome duplication might be an important evolutionary mechanism for generating novelty in the genome and additionally might give a reasonable explanation for speciation (Ohno, 1970). When whole genome duplication occurs, pairs of functionally redundant paralogous genes are created. Since only one gene of a pair of paralogous genes is required to retain its original function, the second is free to diverge. This might lead to the second gene being lost or acquiring a novel function through subsequent mutations. A review of the role of gene duplication in the creation of novel proteins can be found in Hughes (2005).

Evidence for either whole genome duplications or substantial gene duplication events exist in the literature. Specifically, there has been evidence for gene duplications in *Saccharomyces cerevisiae* (Wolfe and Shields, 1997; Friedman and Hughes, 2001; Teichmann and Babu, 2004; Dujon et al., 2004; Kellis et al., 2004) (and in simulation by van Noort et al. (2004)), *Escherichia coli* (Babu and Teichmann, 2003; Friedman and Hughes, 2001; Teichmann and Babu, 2004; Babu et al., 2004), vertebrates (Nadeau and Sankoff, 1997) and other organisms. More generally, three quarters of the transcription factors in *E. coli* have arisen from gene duplication (Babu and Teichmann, 2003) and at least 50% of prokaryotic genes and over 90% of eukaryotic genes are created by gene duplication (Teichmann and Babu, 2004). A review of the mechanisms facilitating gene duplications can be found in Zhang (2004).

## 3. Network topologies in the ARN model

With the ARN, duplication and divergence can be more directly investigated due to its implementation on the genetic string as opposed to an examination at the network level (i.e. where gene duplication happens on the genome level in nature) as is the case in other abstract regulatory network models (i.e. differential equation models, Boolean models). In addition, topological relationships can be easily investigated by parameterization of the threshold. Specifically, the presence of scale-free, Small-world and network motif topologies can be observed in the ARN model. In Sections 3.1–3.3, we summarize our findings previously published in parts in Banzhaf and Kuo (2004) and Kuo and Banzhaf (2004).

### 3.1. Gene duplication and the ARN model

The ARN genome is created through a series of whole length duplication and divergence events. First, a random 32-bit string is generated. This string is then used in a series of whole length duplications followed by mutations to generate a genome of length $L_G$.

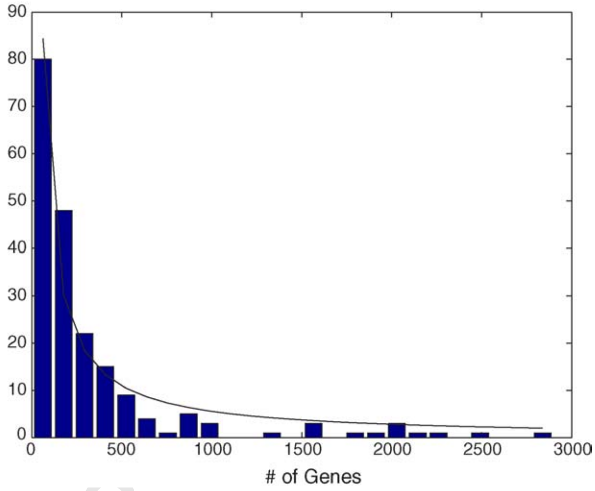


Fig. 5. Histogram of the number of genes in each genome (200 genomes) fitted to a power-law: $P(g) \sim g^{-\gamma}$ for a mutation rate of 1.0%. $\gamma$ was calculated to be 0.9779.

To generate such networks, a divergence (or mutation) rate for the duplication and divergence mechanism must be chosen. First, mutation rates of 1% and 5% were examined. Two-hundred genomes were generated by 12 duplication events per genome leading to individual genomes of length $L_G = 2^{12} \times 32 = 131{,}072$. From these genomes, the number of genes were then determined based on the number of promoter patterns present.

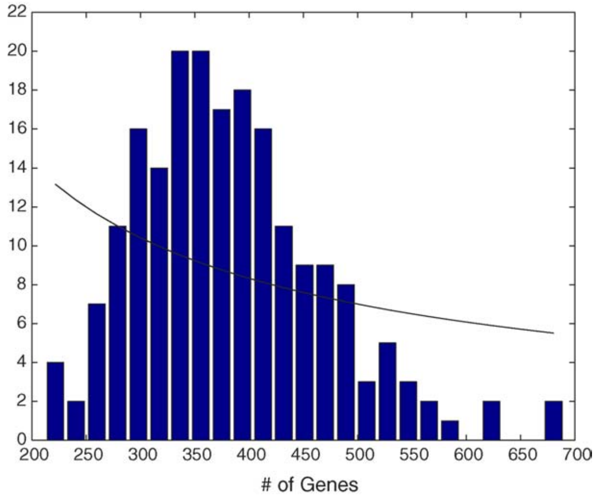


Fig. 6. Histogram of the number of genes in each genome (200 genomes) fitted to a power-law: $P(g) \sim g^{-\gamma}$ for a mutation rate of 5.0%.

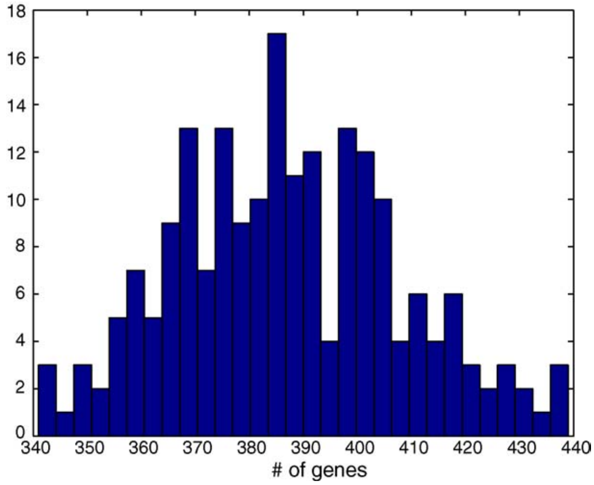


Fig. 7. Histogram of the number of genes in 200 genomes whose bits have been chosen at random.

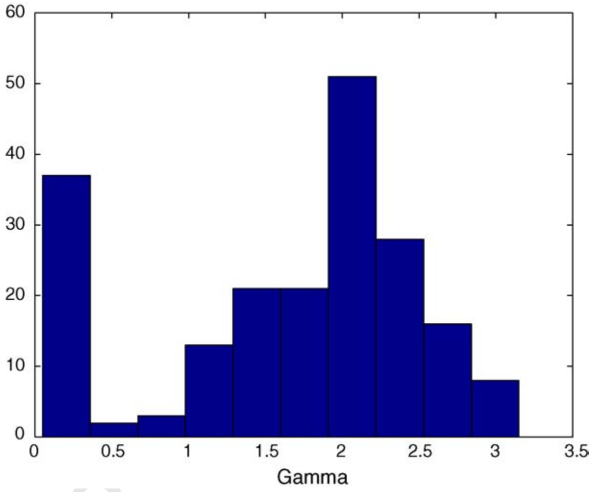


Fig. 8. Distribution of values of $\gamma$ for the best fit of $P(k) \sim k^{-\gamma}$ with a mutation rate of 1.0%.

The distribution of the number of genes present in the genome of size $L_G$ is shown in Figs. 5 and 6.

The distribution of the number of genes in Fig. 5 follows a power-law-like distribution. However, in Fig. 6 the distribution is disrupted. This is attributed to the higher rate of mutation. At such a mutation rate, the disruption of the network becomes so prevalent that it begins to disrupt the duplication of nodes leading to a network with a random number of genes.

For an 8-bit promoter, the probability that it remains intact after one duplication event is only 66% at a mutation rate of 5%. Therefore, many of the genes copied during the duplication process will be subsequently destroyed (by disruption of the promoter) 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, genomes of length $L_G$ were created completely at random without the use of duplication and divergence. The distribution of these completely randomly generated networks are shown in Fig. 7. This distribution is quite similar to that generated in Fig. 6 lending additional support to the hypothesis that at 5% mutation the network topology becomes effectively randomized.

In the case of no mutations (0% probability of mutation) during the duplication process, a large number of networks either have zero genes (where there are no 01010101 patterns in the original 32-bit starting string), or have $2^{(\text{\# of duplications})}$ genes (due to the presence of a 01010101 pattern in the original 32-bit starting string). We wish to obtain a network which shows a topology primarily due to the effects of duplication. Therefore, the distribution of the number of genes in networks generated by duplication and divergence may be used as an estimate of the effect of mutation rate on the network as compared to randomly generated genomes. Obtaining a power-law-like distribution of the number of genes accomplishes this goal. That distribution is sufficiently randomized so as not to resemble the case of 0% mutation while not being dominated by mutational effects (as shown by its lack of similarity to the Gaussian-like distributions shown in Figs. 6 and 7). With these considerations in mind, the networks generated by 1% divergence may be examined with respect to their topologies.

### *3.2. Scale-free and small-world topologies in the ARN model*

The network of gene–protein interactions is parameterized by the threshold value leading to 32 possible networks for each genome (although the case of zero connectivity and full connectivity are neglected). The histograms of the vertex degree distribution were fitted to the equation $P(k) = \alpha k^{-\gamma}$ for each threshold value, using the sum of least squares method. The threshold value which produced a $\gamma$ value closest to 2.5 was kept (a large number of networks which have displayed scale-free behavior exhibit values of $2 < \gamma \leq 3$ (Goh et al., 2002)). Values for the parameter $\gamma$ characterizing scale-free networks were calculated for 200 genomes and are shown in Figs. 8 and 9.

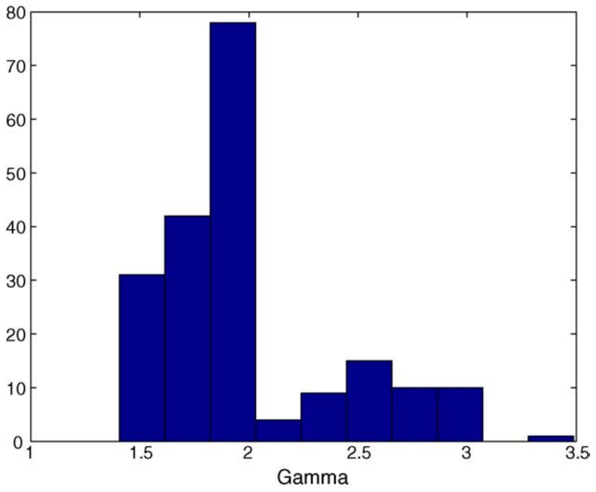


Fig. 9. Distribution of values of $\gamma$ for the best fit of $P(k) \sim k^{-\gamma}$ with a mutation rate of 5.0%.

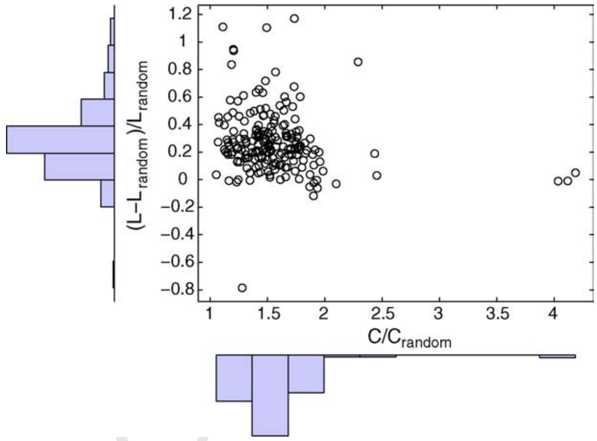


Fig. 11. Plot of $C/C_{\text{random}}$ and $(L_{\text{random}} - L)/L_{\text{random}}$ for each of the randomly generated genomes (200 genomes) with a mutation rate of 1.0%.

There exist many genomes created by duplication and divergence which may be considered to satisfy the definition of a scale-free network. Fig. 10 shows an example of one network's vertex degree distribution fit to a power-law distribution. It does obey a distribution similar to a power-law (scale-free) distribution.

In Fig. 8, there is a large number of networks whose coefficient $\gamma$ is close to 0, which would seem to be at odds with the previous statement. However, it can be attributed to the fact that with a low mutation rate the probability of discovering new promoter patterns through subsequent duplication and divergence steps is also low. Therefore, if there were few promoters in the initial string, there will often be few genes in the overall genome. With a small number of genes, the scale-free coefficient $\gamma$ will often be of small magnitude. In addition, from the distribution of $\gamma$ in Fig. 9, the majority of the networks created by 5% mutation cannot be classified as scale-free. This again, reinforces the previous finding that a mutation rate of 5% or higher during the duplication and divergence process generates networks that are close to having random connectivity.

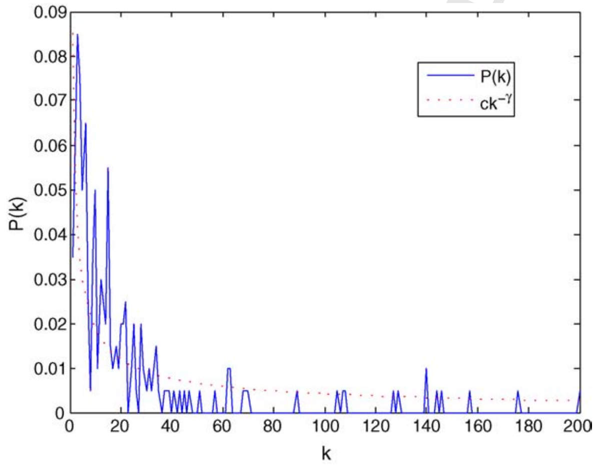
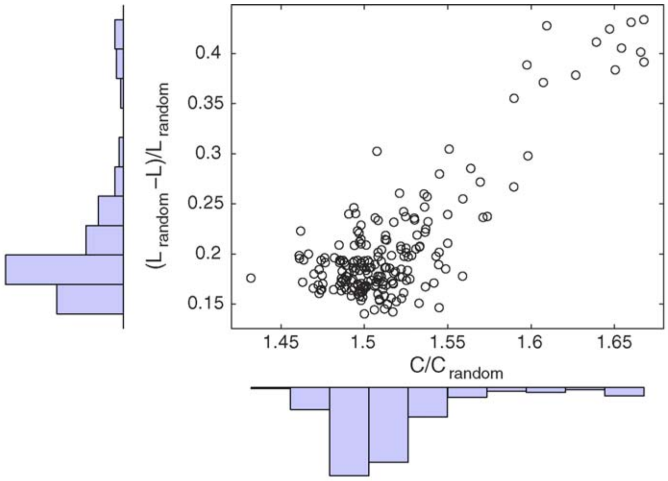


Fig. 10. Degree distribution of a network generated by duplication and divergence with 1% mutation.

To test whether these networks could also be classified as having small-world topology, the clustering coefficient, $C$, and the characteristic path-length, $L$, were calculated and compared to a randomly connected network of the same size and vertex degree distribution. The threshold value that produced a network with the smallest absolute difference, $| L - L_{\text{random}} |$, that also satisfied $C \gg C_{\text{random}}$ were taken to be those most characteristic of the Small-world network topology. The additional constraint, $L > 1.3$, was also enforced to exclude graphs that were close to being fully connected.

The distributions for the clustering coefficient and the characteristic path-length obtained from the 200 genomes for 1% mutation are shown in Fig. 11. It can be derived from the figure, that a majority of genomes has a threshold at which the interaction network approaches or satisfies the definition of a small-world network topology. All graphs considered as having scale-free and small-world topology were found in the transition areas of Fig. 4.

Why does whole genome duplication create scale-free and small-world topologies? Part of the answer is that the duplication process, despite being performed directly on the genetic string can be considered to be similar to the mechanism of preferential attachment at the network level.



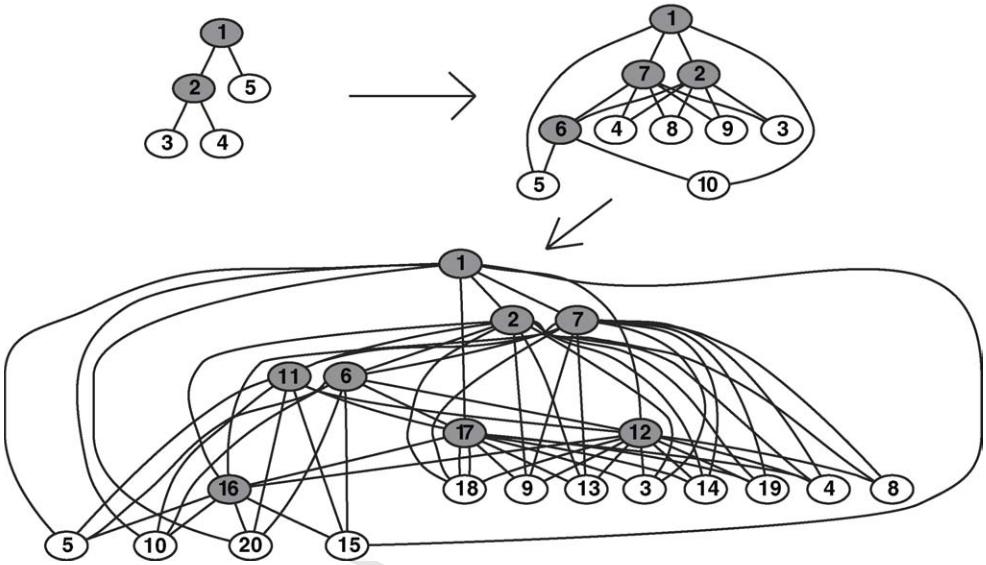


Fig. 12. An example of the effect of two duplication events. Highly connected (shaded) nodes become even more highly connected (preferential attachment). Each node represents a gene protein pair; each edge represents an interaction between gene–protein pairs.

Consider the duplication process on a string which contains multiple genes while neglecting the effects of mutation. For simplicity, it is assumed that no additional genes are created from a duplication event by joining the end of one genome and the beginning of its copy. On the left of Fig. 12, a network of five gene–protein pairs is shown that proceeds through a single duplication event generating the network shown on the right side.

The more highly connected nodes on the left (the original nodes and their copies—all shown in grey) become even more highly connected after a single duplication event. This can again be seen in the third part of the diagram which shows the result of a further 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 (Barabási and Albert, 1999; Romualdo et al., 2003).

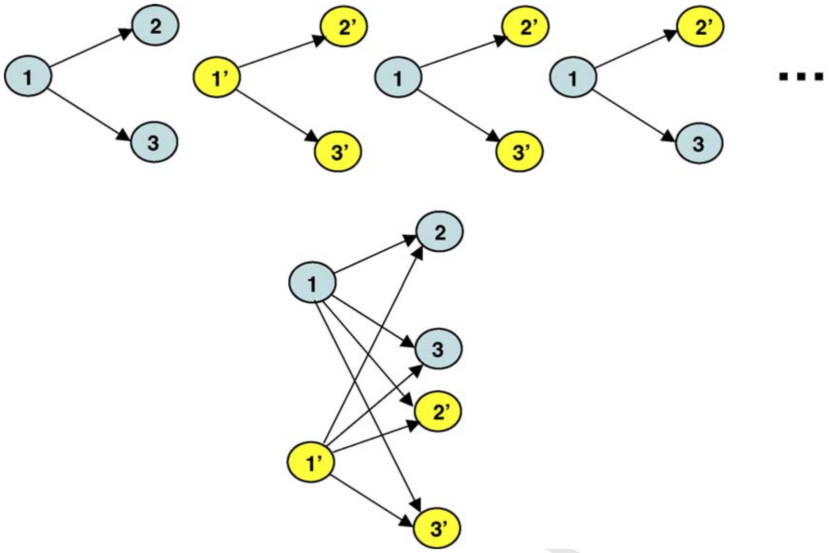


Fig. 13. Decomposition of a six-node graph created by duplication. Demonstrates that any of the nodes in the original topology can be replaced with its copy without changing the topology and vice versa. If we replace any node in the original graph (nodes 1, 2, and 3) with its copy (nodes 1′, 2′, and 3′) and its associated edges to the original graph, the overall topology remains identical.

However, this part of the answer neglects the mechanism of mutation. Mutation may be thought of as an operator which reorganizes the network. If mutations 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 promoter pattern in effect deleting a gene–protein pair from the network, create a new gene–protein pair by creating a new promoter site, or are neutral. The topology of the network as measured by the number of genes in the system is dominated by the effects of duplication, not divergence. Thus, the scale-free distribution observed is due to the duplication mechanism, acting similar to preferential attachment.

How can the small-world topologies found in the ARN model be explained? If we examine the definition of a small-world network more closely, it colloquially states that a network is highly clustered but that there are many links between these clusters which effectively reduce the overall diameter of the network. Frequently, hubs also appear in small-world networks (Watts, 2003). Hubs also appear in the ARN model through the duplication process (analogous to preferential attachment to more highly connected nodes). However, because of the way the duplication process works (assuming no mutation), the maximum distance[1] between any two nodes before and after a duplication remains constant. This happens because the duplication step effectively makes a copy of all nodes and all edges simultaneously. It is self-evident that the maximum distance between any two nodes in only the original graph and the copied portion of the network are the same (if we discount the edges which connect the original nodes with the copied nodes). Thus, the path-length between any two nodes in the original graph is the same as in the copy.

This shows that the maximum path-length is invariant to duplication and thus generally remains small (see Fig. 13). Therefore, the average path-length will always be bounded by the maximum path-length and will never increase. As the network grows via the duplication process, its characteristic path-length might only grow very slowly – if at all – due to mutations.

The clustering coefficient of the network is quite high again as a result of the duplication process. Because of the regularity of the connection patterns, nodes in the network remain highly connected and increase in connectivity with each duplication event. Mutation only serves to perturb the topology partially randomizing some of the edges in the graph. Thus, the formation of small-world topologies is consistent with the network creation method of whole genome duplication and divergence.

### *3.3. Network motifs in the ARN model*

Tables A.1 (three-nodes), A.2 and A.3 (four-nodes) show connection patterns in directed graphs including

[1] The number of edges traversed to get from node “*a*” to “*b*”.

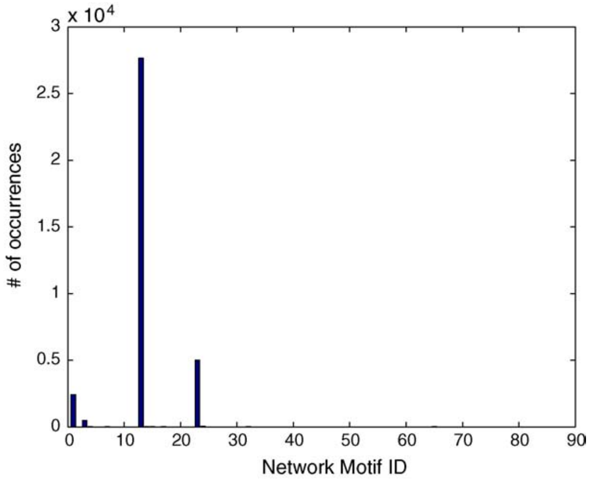


Fig. 14. Average frequency of occurrence for subgraphs of size three in 800 instances of the artificial regulatory network model generated by a duplication and divergence procedure.

auto-regulatory connections up to isomorphism. This list includes networks with auto-regulatory connections (those which have edges which begin and end at the same node) which have been previously ignored by others (Milo et al., 2002, 2004; Wuchty et al., 2003; Yeger-Lotem et al., 2004; Dobrin et al., 2004; Mangan and Alon, 2003). We believe that such connectivity may be important.

To detect all $n$-node subgraphs, a subgraph finding algorithm similar to one devised by Milo et al. (2002) was implemented. The algorithm was applied to 800 instances of the artificial regulatory model generated by the duplication and divergence process. As a control, it was additionally applied to 800 networks whose genomes were generated randomly (by choosing the full number of bits at random). Results of applying the subgraph counting algorithm to the two cases are shown in Figs. 14 and 15. For both methods of network generation, the genome length was set at $2^{17} = 131{,}072$ (12 duplication events in the case of duplication and divergence). For networks generated by duplication and divergence, the mutation rate was set at 1% since this creates networks dominated by duplication effects.

In both cases, the threshold had to be determined. The ratio of the number of edges to the number of vertices for the two natural regulatory networks was approximately 2 to 1. Therefore, in the ARN framework, the threshold was chosen by iteratively raising the value until the network generated had a ratio that was equal to or less than 2 to 1.

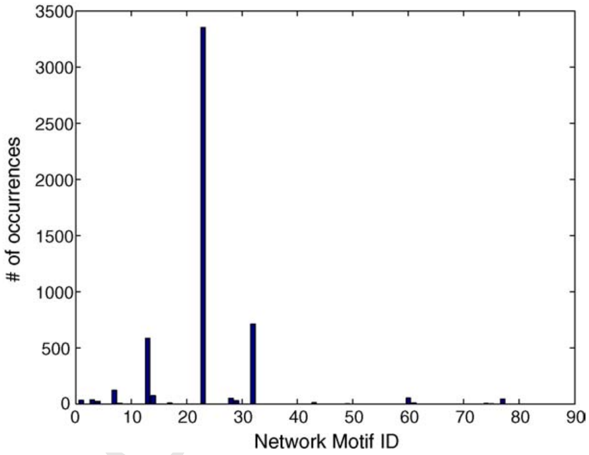


Fig. 16. Frequency of occurrence for subgraphs of size three in the transcriptional network of *Escherichia coli*.

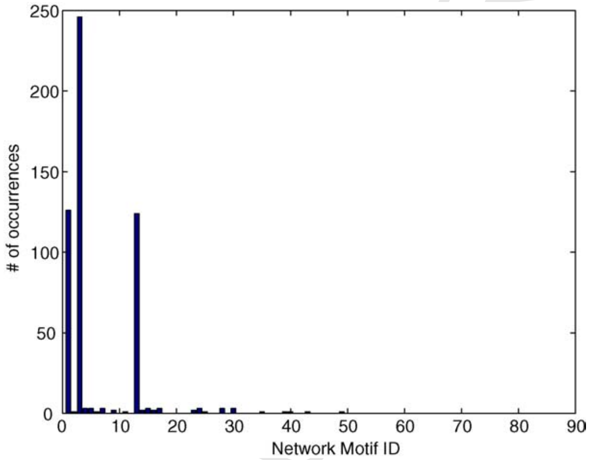


Fig. 15. Average frequency of occurrence for subgraphs of size three in 800 randomly generated instances of the artificial regulatory network model.

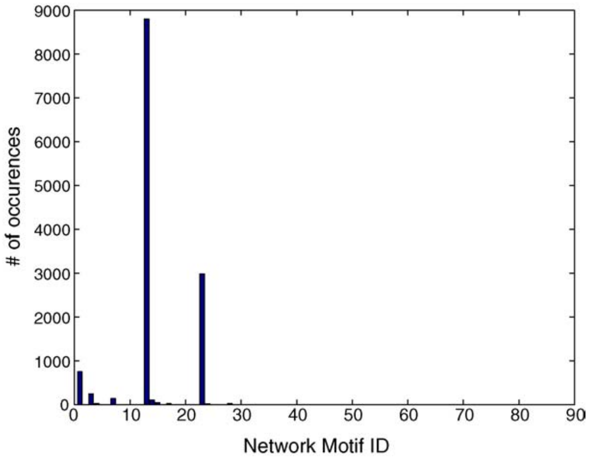


Fig. 17. Frequency of occurrence for subgraphs of size three in the transcriptional network of *Saccharomyces cerevisiae*.

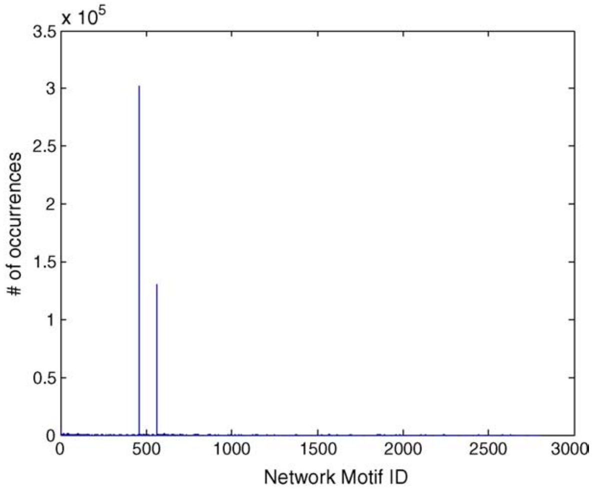


Fig. 18. Average frequency of occurrence for subgraphs of size four in 200 instances of the artificial regulatory network model generated by a duplication and divergence procedure.

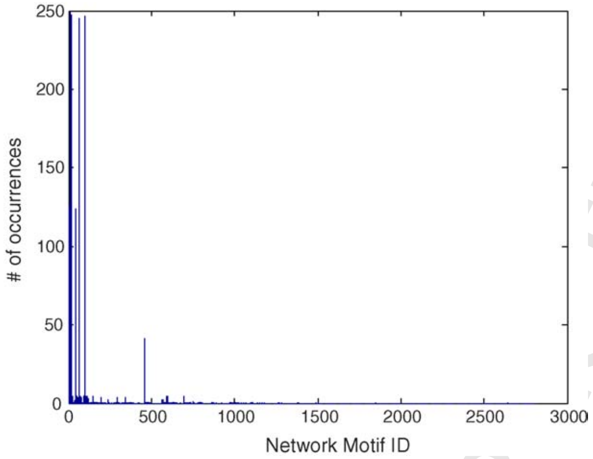


Fig. 19. Average frequency of occurrence for subgraphs of size four in 200 randomly generated instances of the artificial regulatory network model.

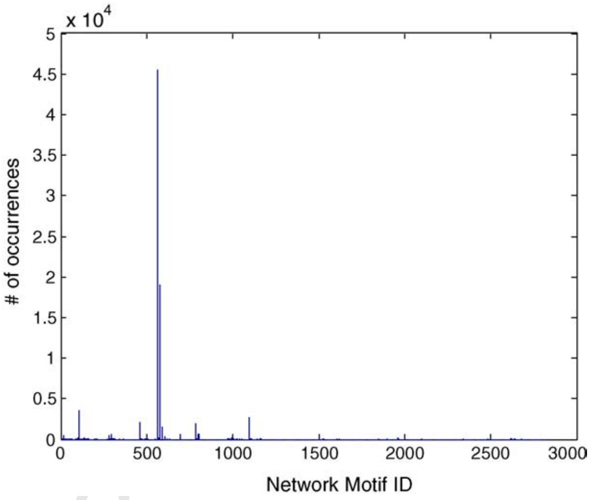


Fig. 20. Frequency of occurrence for subgraphs of size four in the transcriptional network of *Escherichia coli*.

This was then compared to the results of applying the algorithm to two natural transcriptional networks[2], *E. coli* (Shen-Or et al., 2002) and *S. cerevisiae* (Milo et al., 2002). The results can be seen in Figs. 16 and 17. In Figs. 14–17, the most frequent natural subgraphs (ID-22 and ID-12) are both well represented in duplication and divergence-generated artificial networks whereas only one can be detected in fully random networks.

The subgraph counts for subgraphs of size three and four for all types of regulatory networks investigated are presented in Tables A.1 and A.3. For artificial networks, average numbers of counts are shown, whereas for natural regulatory systems only one network each is investigated.

Using the sum of square error (SSE) criterion, the similarity between the distributions of subgraphs for the four types of networks was calculated. The similarity is shown for both three and four node subgraphs in Table 1.

The network distributions obtained from duplication and divergence (D&D) are quite similar to that of *S. cerevisiae* for subgraph sizes of both three and four according to the SSE criterion. In contrast, the distributions of the

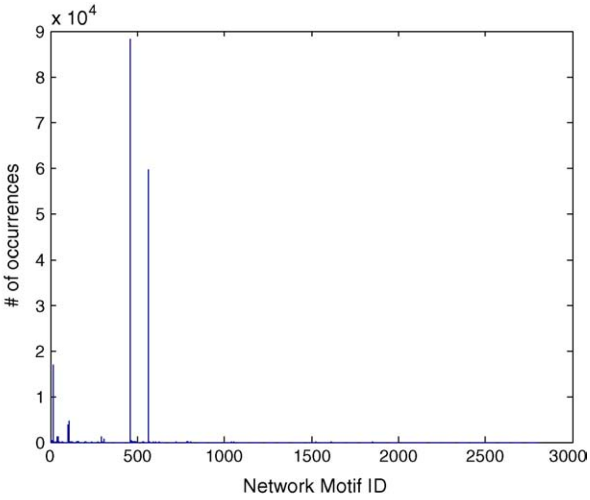


Fig. 21. Frequency of occurrence for subgraphs of size four in the transcriptional network of *Saccharomyces cerevisiae*.

[2] Obtained from http://www.weizmann.ac.il/mcb/UriAlon/.

Table 1
Sum of square error (SSE) between the distributions of subgraph counts (for subgraph size three/four) for the four types of networks examined

| | D&D | Rand | *E. coli* | Yeast |
|---|---|---|---|---|
| D&D | 0 | – | – | – |
| Rand | 1.5348/5.3093 | 0 | – | – |
| *E. coli* | 1.0844/1.4227 | 2.2392/5.6148 | 0 | – |
| Yeast | 0.0072/0.0984 | 1.4886/5.1497 | 1.1693/1.2356 | 0 |

Each distribution has been normalized such that the maximum count of any individual subgraph is 1.0.

randomly generated networks were not similar to any of the three other network types investigated. Networks created by duplication and divergence and the regulatory networks of *E. coli* and *S. cerevisiae* are all more similar to each other than to the randomly generated networks.

Because gene duplication is considered a more important mechanism of evolution in eukaryotes than in prokaryotes, it is interesting that the duplication and divergence networks are more similar to the eukaryotic *S. cerevisiae* rather than the prokaryotic *E. coli*. This might suggest that the topology has been shaped by duplication events in *S. cerevisiae*'s evolutionary history. Teichmann and Babu (2004) suggest that over 90% of eukaryotic genes are created by gene duplication. Our observations support this argument: It is striking how similar the distributions of subgraphs are for these three networks as compared to the randomly created topologies.

We can further investigate the individual subgraphs well represented in these networks. From Figs. 14, 16 and 17, motifs with IDs 12 and 22 are present in substantial numbers. These motifs correspond to the so-called single input module (Milo et al., 2002). This is also the case when examining subgraphs of size four in Figs. 18–21 where network motif IDs 459 and 563 are well represented. However, in counts of both three and four node subgraphs, the single input modules were not well represented in randomly created graphs.

How is the single-input module created by duplication and divergence? We can examine the effect of duplication on the simplest of gene interactions, where one gene has a regulatory influence on another. If these genes and their connections are duplicated we can obtain the so-called single input module network motif.

Fig. 22 shows the effects of two duplications on the simplest of regulatory influences. As can be seen two types of subgraphs should be created with equal probability, the single-input module and the so-called single-output module. However, from examining the motif counts for both natural and artificial networks the counts yield asymmetrical number. In Leier et al. (2005) we will show why this is a natural consequence of the duplication and divergence process.

## 4. Evolving dynamics in the ARN model

In the previous section, the topology of the ARN model was investigated. Topology, however, is only one of the aspects of a genetic regulatory network. It is the dynamics of the network that gives rise to the myriad of functions observed in natural systems. Here we examine the dynamics of our ARN model by attempting to evolve simple time series.

If we try to evolve time series in the ARN model, the evolvability of the ARN model can be looked at with some possible relevance to the evolvability of natural systems. The types of analysis and search mechanisms relevant to such processes could also be important to the field of synthetic biology where synthetic genetic regulatory networks have been evolved in vivo toward dynamics such as oscillations (Yokobayashi et al., 2002) in silico (Mason et al., 2004) and in numero (François and Hakim., 2004). Such an investigation also provides a framework in which we can begin to study the interplay between network dynamics, evolution and topology (see also Kuo et al. (2004)).

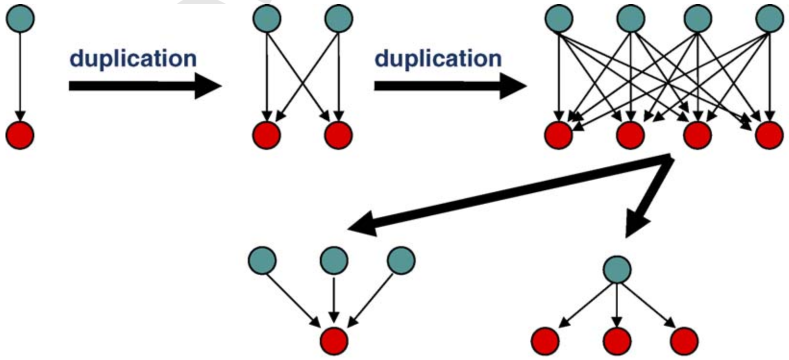


Fig. 22. The effect of whole genome duplication on the simplest possible interaction between two genes.

### 4.1. Extracting a signal from the ARN model

Simulation of the ARN model produces the dynamics of the protein concentrations in the system. However, the system has no assigned semantics—protein concentrations have no meaning outside the system (they perform no cellular function other than regulation). Additionally, since the protein concentrations must sum to 1 (i.e. $\sum c_i = 1$), certain functions are excluded (e.g. two sinusoids with the same phase and frequency).

In order to use the ARN framework to obtain more arbitrary dynamics, a mapping is required. We have chosen to do this by adding an additional transcription factor binding site to the genome. Remember that proteins acting as transcription factors can bind to transcription factor binding sites influencing the transcription of adjacent genes. The rate of transcription of this new site is taken to be similar to a protein concentration which has no other effects on the system. It is the dynamics of this particular site that will be evolved toward specific dynamics.

This is done by randomly choosing an additional 64-bit sequence along the genome. The first 32-bits specify a transcription factor binding site representing an inhibition site while the second 32-bits specify a transcription factor binding site for activation. The proteins in the system are free to bind to these two additional regulatory sites (which can be thought of as a gene with no protein of its own or promoter). The levels of activation and inhibition produced at these two sites are calculated in the same way as in Eq. (2) and are modulated by the proteins in the system. However, instead of calculating a “concentration” of a protein generated from this site (which generates no actual protein of its own) as is the case for a gene, the activity at this site is simply summed and used directly as an output function, $s(t) = \sum_i (e_i - h_i)$. Normalization of $s(t)$ between $-1$ and 1 generates the dynamics of this site which are taken to be the dynamics extracted from this network. Without this normalization step, it is difficult to match the scaling of the desired dynamics. However, since the scaling is effectively arbitrary, this is not a problem.

The additional binding sites added to the genome are a method to extract dynamics from the changes in protein concentrations of the ARN model. This can be visualized as a network like the ones presented in Figs. 2 and 3 except where each protein is linked to an additional node representing the new inhibition/activation site (that does not generate a protein of its own). Additional inhibition/activation sites may also be added to the genome for the extraction of additional signals.

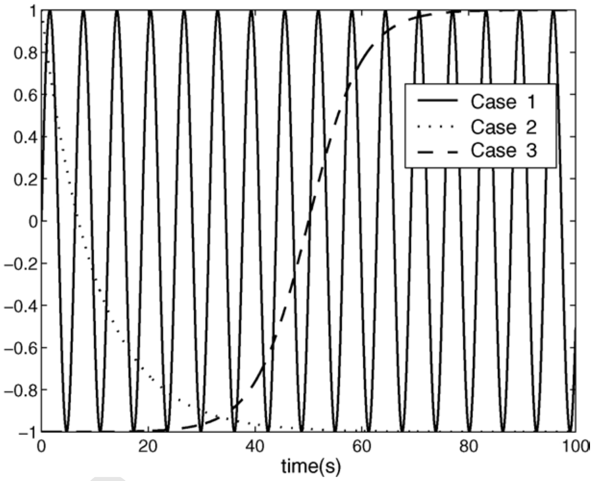


Fig. 23. Plot of the three time series.

### 4.2. Optimization and simulation details

A simple (50 + 100)-Evolutionary Strategy (ES) is used to evolve the solution, $s(t)$ (Beyer and Schwefel, 2002). Genomes were generated by 10 duplication events per genome subject to 1% mutation leading to individual genomes of length $L_G = 32{,}768$. Each generation, 100 new individuals are created from the current population using 1% single-point (bit-flip) mutation (i.e. on average, 328 mutations per genome). The fitness of these solutions was calculated and the best 50 of 150 (parents + children) proceed to the next generation. The ES was terminated when the best solution found was not improved upon for 250 generations.

The objective is to minimize the fitness function calculated as the mean square error (MSE) between the desired function and the evolved function. The following cases were examined and are shown in Fig. 23: $f(t) = \sin(t)$ (Case #1), $f(t) = 2\exp(-0.1t) - 1$ (Case #2) and $f(t) = \frac{2}{1+\exp(-0.2t+10)} - 1$ (Case #3). These cases represent oscillatory, decaying exponential and sigmoidal dynamics which are all relatively simple yet biologically important.

All solutions were generated with a time step of $\mathrm{d}t = 0.1\,\mathrm{s}$. The constant step size facilitates the quick comparison of dynamics between solutions. In addition, since the dynamics of the system do not change quickly with respect to this particular step size (i.e. the second derivative of the function is small), it is an appropriate choice for the three cases. The initial protein concentrations (the initial conditions for the differential equation) are set to $\frac{1}{\#\,\text{of genes}}$. In addition, the first 100 time steps (10 s) are ignored in order to exclude the startup dynamics of the model. Thus, for calculation of the fitness function, the normalized output generated by the ARN model from time $t = 10, \ldots, 110$ s is compared with the time series $f(t)$ from time $t = 0, \ldots, 100$ s.

### 4.3. Results

Table 2 summarizes the results of 10 evolutionary runs for each of the 3 fitness cases. Fig. 24 shows the progress of the best evolutionary run for each case.

The ARN model accurately generates dynamics approximating the sinusoid, the exponential and the sigmoid functions with good accuracy for all runs. In all fitness cases and evolutionary runs, the MSE calculated was less than 0.00588654. Additional support for the success of these simulations can be seen in the final population fitness averages shown in Table 2. The average population fitness values (MSE) are relatively small with low standard deviation indicating that the population is such that all individuals generate solutions that closely approximate the respective objective functions.

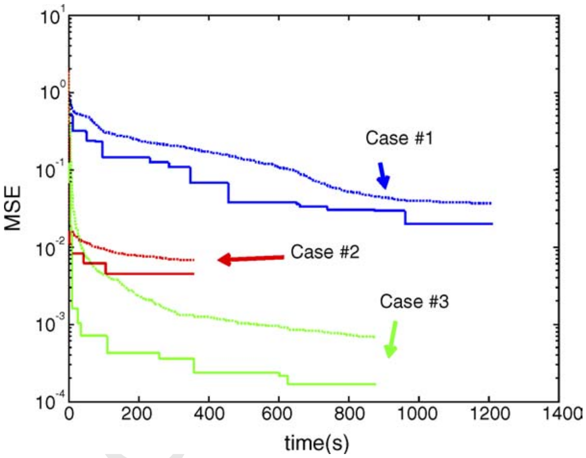


Fig. 24. Fitness plot of the best solutions and the average fitnesses using (50 + 100)-ES for each case.

Table 2
Results of 10 runs of (50 + 100)-ES on each case

| Case-run | Best MSE | #Gens. | #Genes | Avg. MSE (Pop.) | Avg. #Genes (Pop.) |
|---|---|---|---|---|---|
| 1-1 | 0.001445217 | 731 | 47 | 0.00287 (7.7e−4) | 45.31(5.72) |
| 1-2 | 0.001165628 | 381 | 74 | 0.00316 (7.8e−4) | 76.92(3.42) |
| 1-3 | 0.000614281 | 1214 | 105 | 0.00114 (1.5e−4) | 117.59(4.57) |
| 1-4 | 0.000747053 | 835 | 234 | 0.00291 (8.2e−4) | 244.00(13.2) |
| 1-5 | 0.001861556 | 428 | 63 | 0.00326 (6.8e−4) | 75.08(9.34) |
| 1-6 | 0.000640149 | 1077 | 101 | 0.00186 (3.5e−4) | 102.49(4.08) |
| 1-7 | 0.001561523 | 315 | 26 | 0.00440 (8.5e−4) | 32.78(5.55) |
| 1-8 | 0.000151746 | 1040 | 124 | 0.00058 (1.3e−4) | 135.63(6.32) |
| 1-9 | 0.000519559 | 933 | 71 | 0.00134 (3.4e−4) | 92.88(53.2) |
| 1-10 | 0.000846462 | 858 | 55 | 0.00270 (4.5e−4) | 48.57(3.22) |
| 2-1 | 0.00411971 | 708 | 133 | 0.00447 (1.3e−4) | 142.83(5.88) |
| 2-2 | 0.00478168 | 642 | 166 | 0.00554 (2.5e−4) | 185.95(13.5) |
| 2-3 | 0.00363873 | 354 | 27 | 0.00641 (5.5e−4) | 52.22(7.00) |
| 2-4 | 0.00441011 | 359 | 20 | 0.00660 (6.1e−4) | 31.95(7.38) |
| 2-5 | 0.00381064 | 747 | 97 | 0.00505 (3.0e−4) | 106.81(5.71) |
| 2-6 | 0.00402240 | 877 | 63 | 0.00464 (1.8e−4) | 58.83(4.17) |
| 2-7 | 0.00426413 | 501 | 128 | 0.00574 (3.5e−4) | 116.14(8.75) |
| 2-8 | 0.00537858 | 287 | 176 | 0.00661 (4.6e−4) | 164.40(11.1) |
| 2-9 | 0.00511630 | 466 | 58 | 0.00688 (5.6e−4) | 54.26(3.73) |
| 2-10 | 0.00588654 | 519 | 45 | 0.00643 (1.7e−4) | 45.65(3.10) |
| 3-1 | 0.00101533 | 1235 | 154 | 0.00150 (1.3e−4) | 147.59(20.6) |
| 3-2 | 0.00035992 | 557 | 36 | 0.00068 (1.2e−4) | 39.22(2.40) |
| 3-3 | 0.00001843 | 758 | 100 | 0.00004 (1.0e−5) | 102.45(2.93) |
| 3-4 | 0.00001732 | 721 | 96 | 0.00004 (1.0e−5) | 96.55(2.80) |
| 3-5 | 0.00011328 | 617 | 97 | 0.00025 (6.0e−5) | 102.78(4.02) |
| 3-6 | 0.00002073 | 825 | 104 | 0.00013 (5.0e−5) | 109.78(5.03) |
| 3-7 | 0.00005429 | 465 | 108 | 0.00044 (1.8e−4) | 112.37(11.4) |
| 3-8 | 0.00016598 | 879 | 177 | 0.00047 (2.2e−4) | 186.02(9.87) |
| 3-9 | 0.00005034 | 575 | 195 | 0.00031 (1.2e−4) | 212.16(9.57) |
| 3-10 | 0.00002219 | 987 | 39 | 0.00006 (1.0e−5) | 39.49(2.42) |

The standard deviation is given in parenthesis.

A wide variety of networks with differing numbers of genes were found to generate equivalent dynamics for the three time series. The numbers of genes used to obtain solutions was usually large, due to a lack of a penalty on the number of genes during evolution. The algorithm was then reapplied with the addition of a penalty on the number of genes. Because penalty functions are typically arbitrary and problem dependent (since they directly affect the search space), a simple approach was taken. Instead of penalizing the number of genes in the system, networks with more than 10 genes were set to have a fitness of 4.0. In this way, the fitness landscape of each time series is not as directly impacted. Regions of the search space which have 10 or less genes are completely unaffected while regions with more than 10 genes are equally penalized. In this way, we can be sure that we have not drastically altered the entire search space when performing search. In other words, the solutions found using this new fitness function could also be found with the original fitness function and would have the same fitness—which allows direct comparison of solutions.

Results of 10 runs on each time series are shown in Table A.2. The algorithm was terminated when the best fitness obtained was less than $5.0 \times 10^{-3}$ rather than after 250 generations of fitness stagnation. Use of the previous termination criterion can lead to algorithm termination before a good solution has been obtained. In all runs, networks were obtained which have 10 or less genes and can generate the desired dynamics with MSE $< 5.0 \times 10^{-3}$.

What would be the minimum number of genes required to generate equivalent dynamics for each time series? For the sinusoid, a simple oscillator can be written in the matrix form:

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & \omega \\ -\omega & 0 \end{bmatrix} \mathbf{x}(t)$$

which leads to $x_1 = -\sin(\omega t)$ and $x_2 = -\cos(\omega t)$. We can take the vector $x$ to be the concentrations of gene–protein pairs.

If this equation was to be implemented in the ARN model how would it look? There would be two gene–protein pairs represented by nodes, “1” and “2”. The first equation ($\dot{x}_1 = \omega x_2$) can be implemented by node “2” having an inhibitory relationship with node “1”. The second equation, likewise, can be implemented with an excitatory relationship between node “1” and node “2”. In this way, the simple oscillator can be implemented. For the ARN dynamic model to extract this oscillatory dynamic, it would simply have to have higher connectivity with one of the protein products of either node “1” or “2”. Therefore, the minimum possible number of genes required to generate an oscillator in the ARN model is 2.

The requirements to generate a decaying exponential in the ARN model are decidedly simpler. In the dynamical equations the effects of excitation and inhibition on one gene are exponential in nature. Therefore, we simply would need one gene in the system whose protein product binds with greater strength to the inhibitory rather than the excitatory site from which the dynamics are extracted. So, one gene is required to create the dynamics of a decaying exponential.

The situation is somewhat more complicated in the case of the sigmoid-type function. A means of deriving the minimum requirements for this function to a canonical form as was done for the previous two types of dynamics was not found. However, it can be reasoned that the minimum number of genes required must be greater than one since a network with only one gene leads to exponential-type dynamics. To show that the sigmoid dynamics can be generated with two genes, the algorithm was rerun such that networks with more than two genes had a fitness of 4.0. Fig. 25 shows examples of

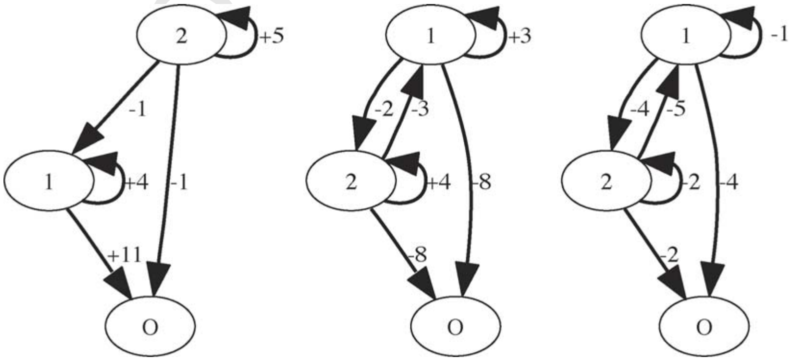


Fig. 25. Three two-gene networks that generate sigmoid dynamics. The “O” node denotes the additional site used to extract the network dynamics.

three different network topologies which can generate the sigmoid dynamics.

Therefore, the minimum number of genes required to generate a sigmoid is two.

In all of these cases, the number of genes actually used by the ARN is far higher than the minimum requirement. This has a bearing on evolvability. Provided a large number of degrees of freedom is cheaply available to the system, AND provided that the overall interaction of these degrees of freedom allows reaching a goal incrementally, a large number might have an advantage over a small number in terms of search efficiency and evolvability. We conjecture that in such a case that once a good solution has been found, a gradual decline in the number of degrees of freedom with a simultaneous readjustment of the remaining degrees is a far better strategy than employing parsimony from the beginning.

## 5. Conclusion

The ARN model first proposed by Banzhaf (2003a) was studied from the perspective of network topology and the evolution of dynamics. We addressed questions raised in both artificial evolutionary processes and network biology. Specifically, the model was examined from the perspective of the scale-free, small-world and network motif topological properties when created using a whole genome duplication and divergence process. This process was chosen since it has been previously implicated as an important factor in the evolution of genomes and due to its simplicity.

Networks generated from this processes can indeed be classified as being scale-free and small-world. Although many researchers have claimed that the presence of scale-free and Small-world network topologies are hallmarks of evolution, we believe that these properties follow naturally from the processes of generation of the networks. In addition, these networks were also found to have subgraph distributions similar to those found in the transcriptional regulatory networks of *E. coli* and *S. cerevisiae* unlike those of random networks.

For the examination of static network topology, evolution was not included among the processes. Therefore, the topologies obtained are directly related to the method of construction. This might indicate that such topologies in natural networks may be a result of the way they are created rather than being explicitly molded by evolution. In other words, the node and vertex distribution outcomes are a reflection of the generation mechanism rather than the result of evolutionary pressures. It may be the case that the motif distributions in these natural networks are to a large part also the result of other organizing forces such as duplication and divergence (although evolutionary pressures are certainly responsible for fine-tuning of distributions). Therefore, it may be more interesting to investigate transcriptional regulatory network topology with regard to the methods of network creation. Efforts in this direction are just beginning.

Further, the evolution of the dynamics of this model has been investigated. It was demonstrated that the dynamics of this model can be evolved toward simple time series behaviors such as the sinusoid, sigmoid and decaying exponential time series. Examining the networks generated in different genomes shows that many different networks give good approximations to each of the prescribed behaviors. This indicates that within the ARN framework there exist an extensive number of functionally equivalent topologies which may be progressively evolved.

Due to the way in which genes are specified in the model, there are plenty of opportunities for individuals in the population to acquire neutral mutations beneficial to their further evolution (Ohta, 2002). Since extensive noncoding regions exist in these genomes, neutral mutations are free to accumulate new genes that might suddenly appear when a new promoter pattern has been created through mutation.

An open question within this framework is how the number of genes affects the ability to generate functions of a given type. From the results presented, we deduce that it is quite easy to evolve the ARN model toward simple time series. Evolvability is helped in our case by more degrees of freedom. In addition, it was observed that each solution evolved for any of the time series differed substantially from run to run. A huge number of different topologies can generate equivalent dynamics. Is this the trick nature used to provide good, yet individual solutions to organisms?

## Acknowledgements

The authors would like to thank François Képès, Leon Glass, Theodore Perkins, Peter Swain, Mike Hallett and William Langdon for helpful discussions.

## Appendix A. Additional data

See Figs. A.1–A.3 and Tables A.1–A.3.

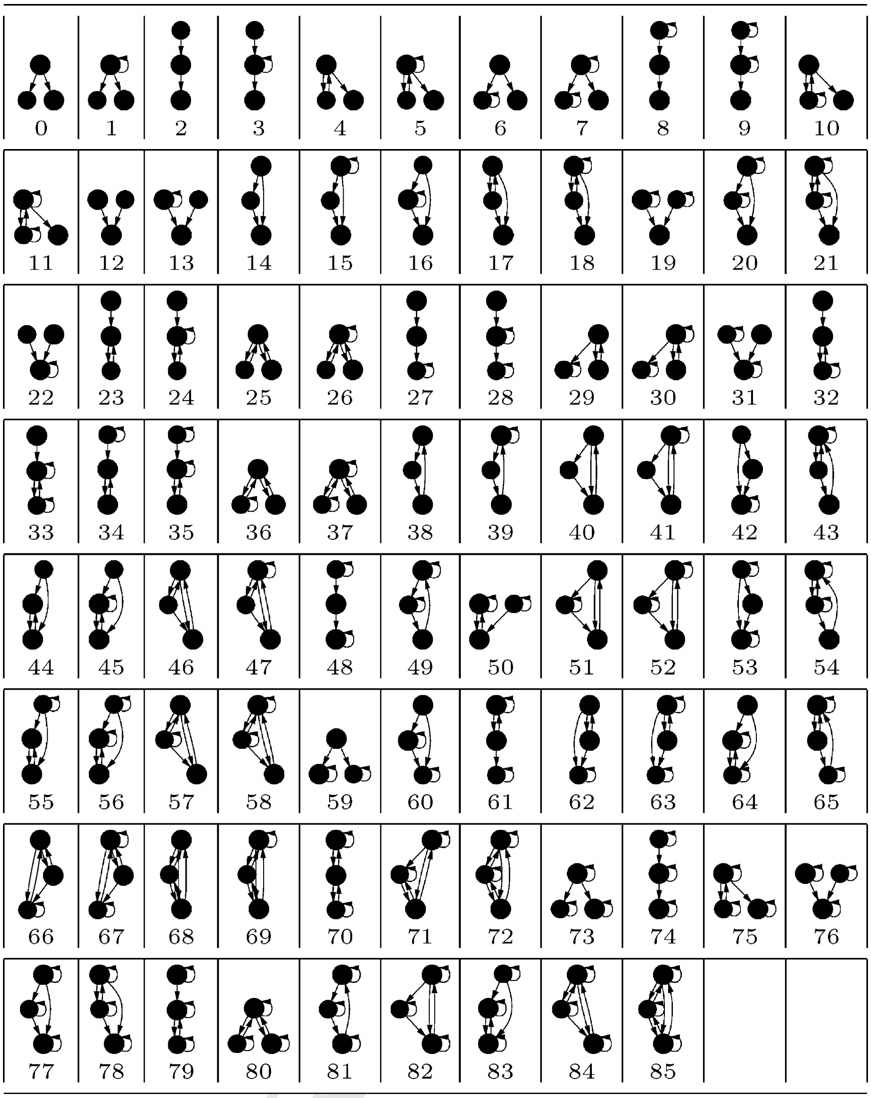


Fig. A.1. Network motifs of size three and their ID.

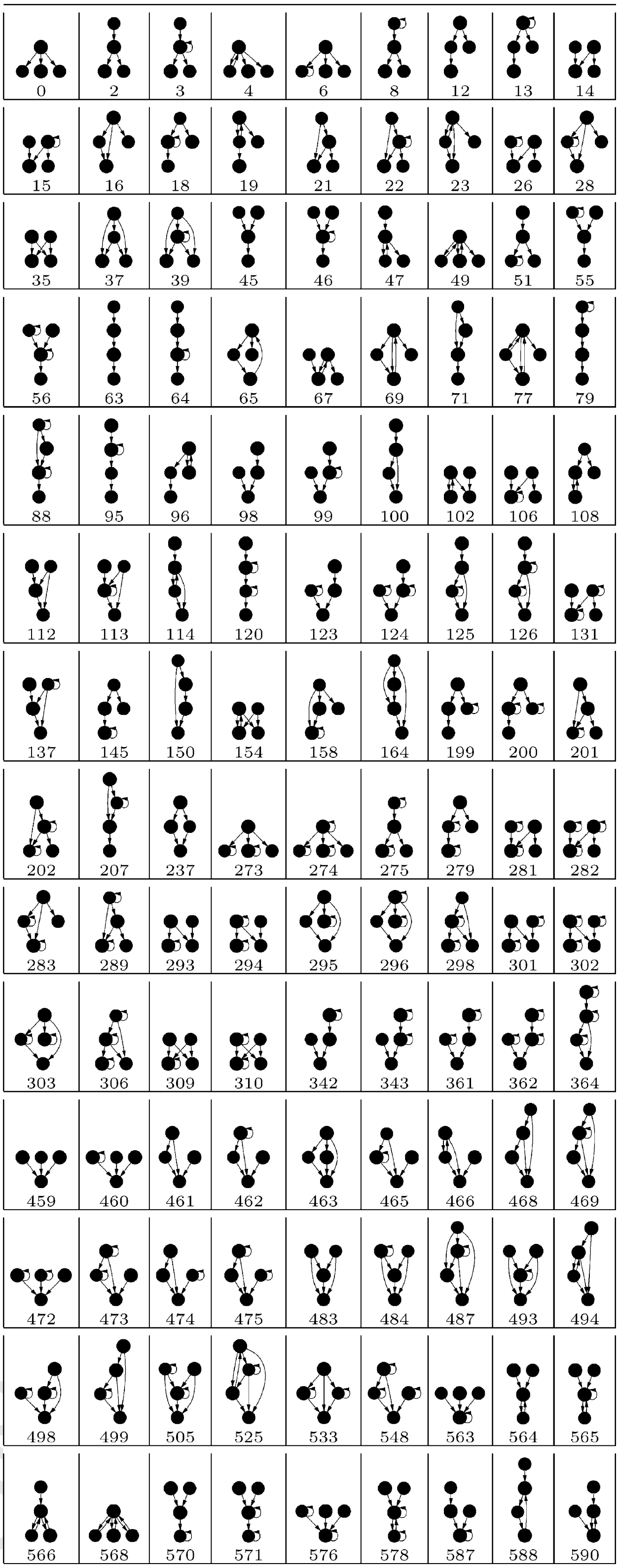


Fig. A.2. Subgraphs of size four and their ID. Only motifs which were present in at least one of the four cases are shown. All other motifs have been omitted.

Table A.1
Subgraphs of size three and their distribution

| Net. | | Count in | | | | Net. | | Count in | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | ID* | D&D | Rand | *E. coli* | *S. cerv* | ID | ID* | D&D | Rand | *E. coli* | *S. cerv* |
| 0 | 6 | 2424 | 76 | 35 | 751 | 43 | A | 0 | 0 | 0 | 0 |
| 1 | A | 4 | 0 | 0 | 1 | 44 | 108 | 0 | 0 | 0 | 0 |
| 2 | 12 | 490 | 271 | 40 | 246 | 45 | A | 1 | 0 | 0 | 0 |
| 3 | A | 11 | 0 | 26 | 24 | 46 | 110 | 0 | 0 | 0 | 0 |
| 4 | 14 | 6 | 0 | 0 | 0 | 47 | A | 0 | 0 | 0 | 0 |
| 5 | A | 0 | 0 | 0 | 0 | 48 | A | 0 | 0 | 3 | 0 |
| 6 | A | 12 | 0 | 124 | 138 | 49 | A | 0 | 0 | 0 | 0 |
| 7 | A | 0 | 0 | 8 | 0 | 50 | A | 0 | 0 | 0 | 0 |
| 8 | A | 0 | 0 | 1 | 0 | 51 | A | 0 | 0 | 0 | 1 |
| 9 | A | 0 | 0 | 2 | 0 | 52 | A | 0 | 0 | 0 | 0 |
| 10 | A | 0 | 0 | 0 | 0 | 53 | A | 0 | 0 | 1 | 0 |
| 11 | A | 0 | 0 | 0 | 0 | 54 | A | 0 | 0 | 0 | 0 |
| 12 | 36 | 27659 | 0 | 587 | 8800 | 55 | A | 0 | 0 | 0 | 0 |
| 13 | A | 8 | 0 | 76 | 104 | 56 | A | 0 | 0 | 0 | 0 |
| 14 | 38 | 15 | 0 | 2 | 44 | 57 | A | 0 | 0 | 0 | 0 |
| 15 | A | 0 | 0 | 1 | 1 | 58 | A | 0 | 0 | 0 | 0 |
| 16 | A | 20 | 0 | 11 | 22 | 59 | A | 0 | 0 | 54 | 4 |
| 17 | 46 | 0 | 0 | 0 | 1 | 60 | A | 0 | 0 | 12 | 0 |
| 18 | A | 0 | 0 | 0 | 0 | 61 | A | 0 | 0 | 0 | 0 |
| 19 | A | 0 | 0 | 2 | 1 | 62 | A | 0 | 0 | 0 | 0 |
| 20 | A | 0 | 0 | 1 | 0 | 63 | A | 0 | 0 | 0 | 0 |
| 21 | A | 0 | 0 | 0 | 0 | 64 | A | 10 | 0 | 0 | 0 |
| 22 | A | 5016 | 0 | 3353 | 2987 | 65 | A | 0 | 0 | 0 | 0 |
| 23 | 74 | 36 | 0 | 0 | 18 | 66 | A | 0 | 0 | 0 | 0 |
| 24 | A | 5 | 0 | 0 | 0 | 67 | A | 0 | 0 | 0 | 0 |
| 25 | 78 | 3 | 0 | 0 | 0 | 68 | 238 | 0 | 0 | 0 | 0 |
| 26 | A | 0 | 0 | 0 | 0 | 69 | A | 0 | 0 | 0 | 0 |
| 27 | A | 6 | 0 | 53 | 25 | 70 | A | 0 | 0 | 0 | 0 |
| 28 | A | 0 | 0 | 32 | 0 | 71 | A | 0 | 0 | 0 | 0 |
| 29 | A | 0 | 0 | 0 | 0 | 72 | A | 0 | 0 | 0 | 0 |
| 30 | A | 0 | 0 | 0 | 0 | 73 | A | 0 | 0 | 6 | 0 |
| 31 | A | 14 | 0 | 713 | 0 | 74 | A | 0 | 0 | 3 | 0 |
| 32 | A | 0 | 0 | 0 | 0 | 75 | A | 0 | 0 | 0 | 0 |
| 33 | A | 3 | 0 | 0 | 0 | 76 | A | 0 | 0 | 46 | 0 |
| 34 | A | 0 | 0 | 0 | 0 | 77 | A | 0 | 0 | 0 | 0 |
| 35 | A | 0 | 0 | 0 | 0 | 78 | A | 0 | 0 | 0 | 0 |
| 36 | A | 0 | 0 | 0 | 0 | 79 | A | 0 | 0 | 0 | 0 |
| 37 | A | 0 | 0 | 0 | 0 | 80 | A | 0 | 0 | 0 | 0 |
| 38 | 98 | 0 | 0 | 0 | 0 | 81 | A | 0 | 0 | 0 | 0 |
| 39 | A | 0 | 0 | 0 | 0 | 82 | A | 0 | 0 | 0 | 0 |
| 40 | 102 | 0 | 0 | 0 | 0 | 83 | A | 0 | 0 | 0 | 0 |
| 41 | A | 0 | 0 | 0 | 0 | 84 | A | 0 | 0 | 0 | 0 |
| 42 | A | 6 | 0 | 14 | 3 | 85 | A | 0 | 0 | 0 | 0 |

D&D: Duplication and divergence genomes; Rand: Random genomes. ID* are the subgraph designations given by Milo et al. (2002). IDs shown as A are subgraphs with self-regulatory connections which do not have a designation in Milo et al. (2002).

Table A.2
Results of 10 runs of (50 + 100)-ES on each case with a penalty function

| Case-run | Best MSE | #Gens. | #Genes | Avg. MSE (Pop.) | Avg. #Genes (Pop.) |
|---|---|---|---|---|---|
| 1-1 | 0.00287157 | 89122 | 10 | 0.00734 (1.1e−3) | 9.73(0.54) |
| 1-2 | 0.00444153 | 13643 | 8 | 0.00912 (8.1e−4) | 7.29(0.43) |
| 1-3 | 0.00486211 | 401417 | 9 | 0.01027 (2.3e−4) | 9.18(0.18) |
| 1-4 | 0.00470516 | 133229 | 10 | 0.00707 (6.1e−4) | 10.20(0.20) |
| 1-5 | 0.00356387 | 21205 | 10 | 0.01493 (4.7e−3) | 10.20(0.20) |
| 1-6 | 0.00493755 | 99553 | 10 | 0.00870 (1.5e−3) | 9.92(0.49) |
| 1-7 | 0.00398828 | 11342 | 10 | 0.02751 (1.3e−2) | 10.00(0.49) |
| 1-8 | 0.00472991 | 23091 | 10 | 0.00989 (2.4e−3) | 10.20(0.20) |
| 1-9 | 0.00480238 | 395 | 9 | 0.30263 (7.5e−2) | 9.47(0.56) |
| 1-10 | 0.00281274 | 1664 | 8 | 0.20032 (7.5e−2) | 9.59(0.89) |
| 2-1 | 0.00484099 | 639 | 8 | 0.00811 (5.4e−4) | 7.02(2.08) |
| 2-2 | 0.00492588 | 2799 | 9 | 0.00714 (6.2e−4) | 9.02(0.98) |
| 2-3 | 0.00418354 | 820 | 5 | 0.00659 (5.0e−4) | 6.32(1.69) |
| 2-4 | 0.00478972 | 5336 | 9 | 0.00636 (4.9e−4) | 9.33(1.02) |
| 2-5 | 0.00497284 | 1676 | 9 | 0.00759 (4.2e−4) | 9.31(0.71) |
| 2-6 | 0.00490717 | 468 | 9 | 0.00810 (6.9e−4) | 8.82(1.01) |
| 2-7 | 0.00430360 | 642 | 10 | 0.00785 (6.5e−4) | 8.51(1.49) |
| 2-8 | 0.00472030 | 3529 | 10 | 0.00577 (2.6e−4) | 9.67(0.73) |
| 2-9 | 0.00467765 | 10112 | 10 | 0.00601 (2.6e−4) | 10.18(0.25) |
| 2-10 | 0.00413019 | 241 | 5 | 0.00798 (9.1e−4) | 7.00(1.66) |
| 3-1 | 0.00345716 | 35 | 6 | 0.05491 (1.8e−2) | 8.84(1.35) |
| 3-2 | 0.00375144 | 61 | 9 | 0.04274 (1.5e−2) | 8.80(1.05) |
| 3-3 | 0.00425317 | 8 | 6 | 0.13660 (7.1e−2) | 7.71(1.66) |
| 3-4 | 0.00149893 | 15 | 8 | 0.10153 (4.1e−2) | 8.41(1.62) |
| 3-5 | 0.00373932 | 21 | 10 | 0.07446 (3.5e−2) | 8.44(1.42) |
| 3-6 | 0.00299901 | 208 | 8 | 0.01359 (4.0e−3) | 8.92(0.99) |
| 3-7 | 0.00341115 | 32 | 7 | 0.03841 (1.1e−2) | 8.55(1.16) |
| 3-8 | 0.00492678 | 109 | 10 | 0.01886 (6.7e−3) | 8.49(1.25) |
| 3-9 | 0.00101274 | 4 | 6 | 0.39698 (1.8e−1) | 7.73(1.84) |
| 3-10 | 0.00423338 | 19 | 9 | 0.07139 (3.1e−2) | 8.59(1.40) |

The standard deviation is given in parenthesis.

Table A.3
Subgraphs of size four and their distribution

| Net. | Count in | | | | Net. | Count in | | | |
|---|---|---|---|---|---|---|---|---|---|
| IDs | D&D | Rand | *E. coli* | *S. cerv* | IDs | D&D | Rand | *E. coli* | *S. cerv* |
| 0 | 4137 | 43 | 4 | 843 | 462 | 2 | 0 | 8 | 23 |
| 2 | 56 | 125 | 10 | 116 | 463 | 1 | 0 | 0 | 1 |
| 3 | 0 | 1 | 0 | 5 | 465 | 1 | 0 | 46 | 346 |
| 4 | 1716 | 2 | 0 | 0 | 466 | 0 | 0 | 0 | 9 |
| 6 | 3 | 2 | 38 | 150 | 468 | 0 | 0 | 0 | 1 |
| 8 | 0 | 2 | 0 | 0 | 469 | 0 | 0 | 0 | 1 |
| 12 | 61 | 249 | 3 | 329 | 472 | 0 | 0 | 17 | 6 |
| 13 | 0 | 3 | 0 | 0 | 473 | 0 | 0 | 9 | 0 |
| 14 | 1531 | 247 | 510 | 16925 | 474 | 0 | 0 | 3 | 2 |
| 15 | 0 | 3 | 0 | 31 | 475 | 0 | 0 | 2 | 0 |
| 16 | 9 | 5 | 0 | 75 | 483 | 4 | 0 | 0 | 120 |
| 18 | 0 | 3 | 5 | 19 | 484 | 0 | 0 | 1 | 1 |
| 19 | 0 | 2 | 0 | 0 | 487 | 0 | 0 | 0 | 1 |
| 21 | 0 | 4 | 1 | 11 | 493 | 5 | 0 | 16 | 33 |
| 22 | 0 | 0 | 0 | 3 | 494 | 0 | 0 | 0 | 17 |
| 23 | 1 | 0 | 0 | 0 | 498 | 0 | 0 | 1 | 4 |
| 26 | 0 | 3 | 36 | 157 | 499 | 0 | 0 | 0 | 15 |
| 28 | 0 | 0 | 2 | 10 | 505 | 0 | 0 | 1 | 0 |
| 35 | 1337 | 1 | 8 | 1105 | 525 | 0 | 0 | 0 | 1 |
| 37 | 0 | 0 | 0 | 5 | 533 | 0 | 0 | 0 | 2 |

Table A.3 (*Continued*)

| Net. | Count in | | | | Net. | Count in | | | |
|---|---|---|---|---|---|---|---|---|---|
| IDs | D&D | Rand | *E. coli* | *S. cerv* | IDs | D&D | Rand | *E. coli* | *S. cerv* |
| 39 | 0 | 0 | 0 | 1 | 548 | 0 | 0 | 1 | 0 |
| 45 | 1451 | 123 | 118 | 1246 | 563 | 130570 | 0 | 45585 | 59569 |
| 46 | 0 | 1 | 72 | 81 | 564 | 521 | 2 | 0 | 121 |
| 47 | 10 | 4 | 0 | 0 | 565 | 34 | 0 | 0 | 0 |
| 49 | 530 | 0 | 0 | 0 | 566 | 11 | 0 | 0 | 0 |
| 51 | 0 | 4 | 58 | 4 | 568 | 54 | 0 | 0 | 0 |
| 55 | 0 | 3 | 1 | 0 | 570 | 16 | 2 | 191 | 129 |
| 56 | 0 | 0 | 6 | 0 | 571 | 0 | 0 | 103 | 0 |
| 63 | 10 | 245 | 0 | 92 | 576 | 161 | 0 | 19077 | 0 |
| 64 | 0 | 3 | 8 | 0 | 578 | 20 | 0 | 0 | 0 |
| 65 | 0 | 4 | 0 | 0 | 587 | 410 | 3 | 1606 | 150 |
| 67 | 0 | 4 | 0 | 0 | 588 | 8 | 4 | 0 | 0 |
| 69 | 1 | 0 | 0 | 0 | 590 | 24 | 2 | 0 | 32 |
| 71 | 0 | 5 | 0 | 11 | 594 | 3 | 4 | 0 | 0 |
| 77 | 1 | 0 | 0 | 0 | 602 | 1028 | 0 | 415 | 24 |
| 79 | 0 | 4 | 0 | 0 | 606 | 27 | 0 | 0 | 0 |
| 88 | 0 | 0 | 1 | 0 | 617 | 0 | 0 | 90 | 0 |
| 95 | 0 | 4 | 7 | 0 | 622 | 0 | 0 | 0 | 16 |
| 96 | 1 | 4 | 0 | 0 | 632 | 0 | 0 | 5 | 0 |
| 98 | 1293 | 246 | 188 | 3859 | 647 | 3 | 0 | 0 | 0 |
| 99 | 0 | 3 | 167 | 528 | 654 | 2 | 0 | 0 | 0 |
| 100 | 0 | 5 | 0 | 51 | 658 | 20 | 0 | 0 | 0 |
| 102 | 1 | 4 | 0 | 0 | 691 | 0 | 0 | 624 | 0 |
| 106 | 291 | 3 | 3569 | 4618 | 692 | 0 | 4 | 6 | 0 |
| 108 | 2 | 4 | 0 | 16 | 693 | 0 | 0 | 8 | 0 |
| 112 | 1 | 4 | 1 | 195 | 695 | 0 | 0 | 7 | 0 |
| 113 | 0 | 0 | 39 | 83 | 722 | 0 | 0 | 0 | 1 |
| 114 | 0 | 0 | 0 | 1 | 750 | 0 | 1 | 0 | 0 |
| 120 | 0 | 0 | 12 | 0 | 786 | 0 | 0 | 1950 | 118 |
| 123 | 0 | 3 | 18 | 43 | 787 | 2 | 0 | 96 | 3 |
| 124 | 0 | 0 | 1 | 0 | 788 | 0 | 0 | 11 | 0 |
| 125 | 0 | 0 | 0 | 5 | 801 | 167 | 0 | 659 | 0 |
| 126 | 0 | 0 | 1 | 0 | 803 | 75 | 0 | 0 | 0 |
| 131 | 0 | 0 | 259 | 0 | 804 | 0 | 0 | 0 | 1 |
| 137 | 0 | 0 | 1 | 0 | 974 | 0 | 0 | 18 | 0 |
| 145 | 1 | 4 | 10 | 27 | 978 | 0 | 0 | 15 | 0 |
| 150 | 2 | 4 | 0 | 10 | 979 | 0 | 0 | 9 | 0 |
| 154 | 1 | 0 | 0 | 0 | 987 | 0 | 0 | 2 | 0 |
| 158 | 10 | 0 | 7 | 14 | 988 | 0 | 0 | 202 | 0 |
| 164 | 0 | 0 | 0 | 1 | 989 | 0 | 0 | 81 | 0 |
| 199 | 0 | 3 | 6 | 28 | 998 | 0 | 0 | 281 | 0 |
| 200 | 0 | 0 | 14 | 0 | 1001 | 0 | 0 | 1 | 0 |
| 201 | 0 | 0 | 5 | 3 | 1017 | 0 | 0 | 1 | 0 |
| 202 | 0 | 0 | 1 | 0 | 1025 | 0 | 0 | 1 | 0 |
| 207 | 0 | 0 | 5 | 0 | 1041 | 0 | 0 | 15 | 1 |
| 237 | 39 | 2 | 0 | 6 | 1053 | 0 | 0 | 9 | 1 |
| 273 | 0 | 0 | 40 | 2 | 1094 | 0 | 0 | 2710 | 0 |
| 274 | 0 | 0 | 6 | 0 | 1105 | 0 | 0 | 124 | 0 |
| 275 | 0 | 0 | 1 | 0 | 1145 | 0 | 0 | 61 | 0 |
| 279 | 0 | 0 | 9 | 0 | 1160 | 0 | 0 | 13 | 0 |
| 281 | 0 | 0 | 508 | 0 | 1521 | 44 | 0 | 26 | 3 |
| 282 | 0 | 0 | 30 | 0 | 1526 | 5 | 0 | 0 | 0 |
| 283 | 0 | 0 | 1 | 0 | 1531 | 0 | 0 | 9 | 0 |
| 289 | 0 | 0 | 1 | 0 | 1606 | 0 | 0 | 6 | 0 |
| 293 | 1 | 4 | 704 | 1261 | 1612 | 0 | 0 | 0 | 1 |
| 294 | 0 | 0 | 16 | 0 | 1618 | 0 | 0 | 5 | 0 |
| 295 | 0 | 0 | 0 | 2 | 1846 | 0 | 0 | 57 | 1 |
| 296 | 0 | 0 | 1 | 0 | 1847 | 43 | 0 | 7 | 0 |
| 298 | 0 | 0 | 1 | 0 | 1855 | 354 | 0 | 0 | 0 |
| 301 | 0 | 0 | 43 | 14 | 1897 | 0 | 0 | 14 | 0 |
| 302 | 0 | 0 | 3 | 0 | 1898 | 0 | 0 | 4 | 0 |
| 303 | 0 | 0 | 7 | 0 | 1957 | 0 | 0 | 208 | 0 |
| 306 | 0 | 0 | 1 | 0 | 1958 | 0 | 0 | 1 | 0 |
| 309 | 6 | 0 | 125 | 737 | 1968 | 0 | 0 | 99 | 0 |
| 310 | 0 | 0 | 5 | 0 | 2094 | 0 | 0 | 14 | 0 |

Table A.3 (*Continued*)

| Net. | Count in | | | | Net. | Count in | | | |
|---|---|---|---|---|---|---|---|---|---|
| IDs | D&D | Rand | *E. coli* | *S. cerv* | IDs | D&D | Rand | *E. coli* | *S. cerv* |
| 342 | 0 | 4 | 4 | 0 | 2339 | 0 | 0 | 1 | 0 |
| 343 | 0 | 0 | 11 | 0 | 2486 | 0 | 0 | 8 | 0 |
| 361 | 0 | 0 | 1 | 0 | 2579 | 1 | 0 | 0 | 0 |
| 362 | 0 | 0 | 1 | 0 | 2619 | 0 | 0 | 4 | 0 |
| 364 | 0 | 0 | 1 | 0 | 2623 | 0 | 0 | 30 | 0 |
| 459 | 301970 | 41 | 2052 | 88321 | 2634 | 0 | 0 | 1 | 0 |
| 460 | 8 | 1 | 391 | 1085 | 2643 | 0 | 0 | 18 | 0 |
| 461 | 157 | 4 | 25 | 729 | 2677 | 0 | 0 | 120 | 0 |

D&D: Duplication and divergence genomes; Rand: Random genomes. Only motifs which were present in at least one of the four cases are shown.

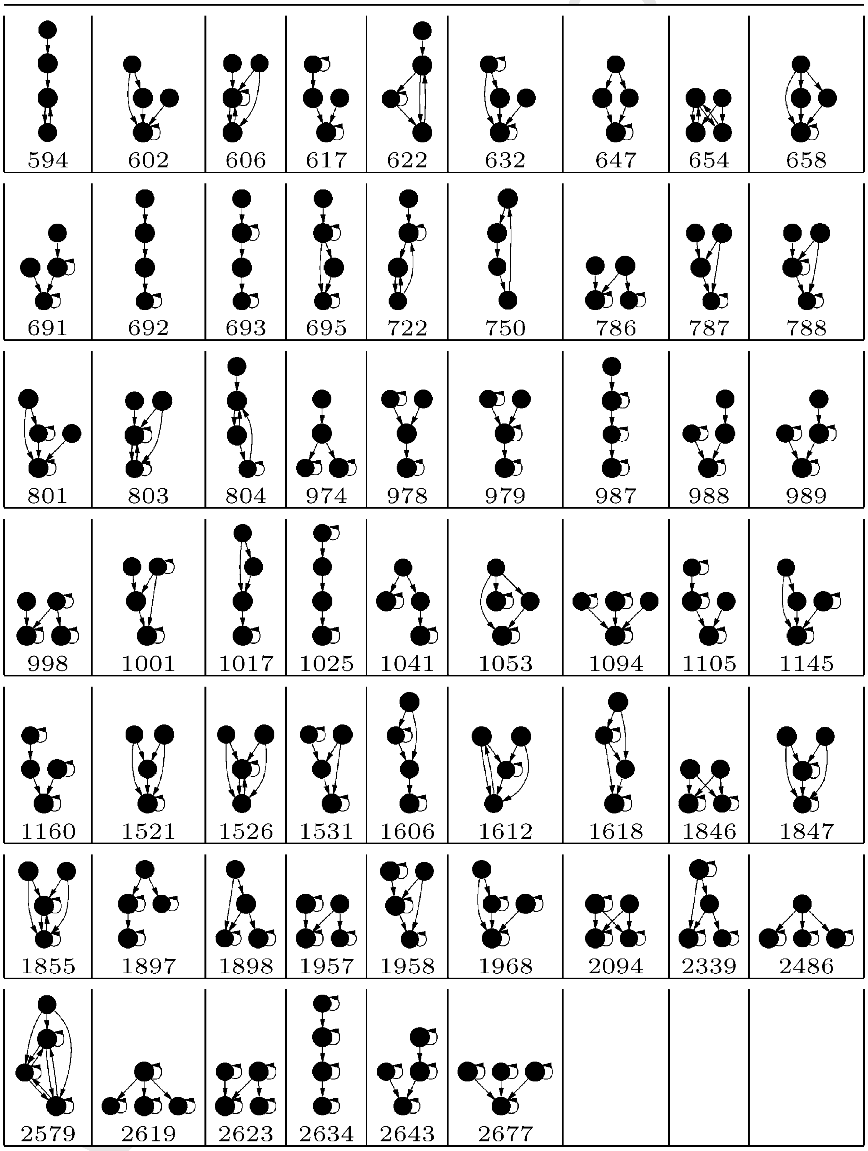


Fig. A.3. Subgraphs of size four and their ID. Only motifs which were present in at least one of the four cases are shown. All other motifs have been omitted.

## References

Babu, M., Teichmann, S.A., 2003. Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. Nucl. Acids Res. 31 (4), 1234–1244.

Babu, M., Luscombe, N., Aravind, L., Gerstein, M., Teichmann, S.A., 2004. Structure and evolution of transcriptional regulatory networks. Curr. Opin. Struct. Biol. 14, 283–292.

Banzhaf, W., 2003. On the dynamics of an artificial regulatory network. In: Banzhaf, W., Christaller, T., Dittrich, P., Kim, J.T., Ziegler, J. (Eds.), Advances in Artificial Life—Proceedings of the Seventh European Conference on Artificial Life (ECAL), vol. 2801 of Lecture Notes in Artificial Intelligence. Springer-Verlag, pp. 217–227.

Banzhaf, W., 2003. Artificial regulatory networks and genetic programming. In: Riolo, R.L., Worzel, B. (Eds.), Genetic Programming Theory and Practice. Kluwer, pp. 43–62 (Chapter 4).

Banzhaf, W., Kuo, P., 2004. Network motifs in artificial and natural transcriptional regulatory networks. J. Biol. Phys. Chem. 4 (2), 85–92.

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

Beyer, H.-G., Schwefel, H.-P., 2002. Evolution strategies: a comprehensive introduction. Nat. Comput. 1 (1), 3–52.

Bower, J., Bolouri, H. (Eds.), 2001. Computational Modelling of Genetic and Biochemical Networks. MIT Press, Cambridge, MA.

Davidson, E., 2001. Genomic Regulatory Systems. Academic Press, San Diego, CA.

Dobrin, R., Beg, Q., Barabási, A.-L., Olvai, Z., 2004. Aggregation of topological motifs in the *E. coli* transcriptional regulatory network. BMC Bioinformat. 5 (10).

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuveglise, C., Talla, E., Goffard, N., Frangeul, L., Aigle, M., Anthouard, V., Babour, A., Barbe, V., Barnay, S., Blanchin, S., Beckerich, J., Beyne, E., Bleykasten, C., Boisrame, A., Boyer, J., Cattolico, L., Confanioleri, F., De Daruvar, A., Despons, L., Fabre, E., Fairhead, C., Ferry-Dumazet, H., Groppi, A., Hantraye, F., Hennequin, C., Jauniaux, N., Joyet, P., Kachouri, R., Kerrest, A., Koszul, R., Lemaire, M., Lesur, I., Ma, L., Muller, H., Nicaud, J., Nikolski, M., Oztas, S., Ozier-Kalogeropoulos, O., Pellenz, S., Potier, S., Richard, G., Straub, M., Suleau, A., Swennen, D., Tekaia, F., Wesolowski-Louvel, M., Westhof, E., Wirth, B., Zeniou-Meyer, M., Zivanovic, I., Bolotin-Fukuhara, M., Thierry, A., Bouchier, C., Caudron, B., Scarpelli, C., Gaillardin, C., Weissenbach, J., Wincker, P., Souciet, J., 2004. Genome evolution in yeasts. Nature 430 (6995), 35–44.

François, P., Hakim, V., 2004. Design of genetic networks with specified functions by evolution in silico. Proc. Natl. Acad. Sci. 101 (2), 580–585.

Friedman, R., Hughes, A.L., 2001. Gene duplication and the structure of eukaryotic genomes. Genome Res. 11 (3), 373–381.

Goh, K., Oh, E., Jeong, H., Kahng, B., Kim, D., 2002. Classification of scale-free networks. Proc. Natl. Acad. Sci. 99 (20), 12583–12588.

Guelzim, N., Bottani, S., Bourgine, P., Képès, F., 2002. Topological and causal structure of the yeast transcriptional regulatory network. Nat. Genet. 31, 60–63.

Hood, L., Galas, D., 2003. The digital code of DNA. Nature 421 (6921), 444–448.

Hughes, A.L., 2005. Gene duplication and the origin of novel proteins. Proc. Natl. Acad. Sci. 102 (25), 8791–8792.

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

Kashtan, N., Itzkovitz, S., Milo, R., Alon, U., 2004. Topological generalizations of network motifs. Phys. Rev. E 70, 031909.

Kellis, M., Birren, B., Lander, E., 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature 428, 617–624.

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

Kuo, P., Banzhaf, W., 2004. Scale-free and small world network topologies in an artificial regulatory network model. Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems (ALIFE), pp. 404–409.

Kuo, P., Leier, A., Banzhaf, W., 2004. Evolving dynamics in an artificial regulatory network model. In: Yao, X., Burke, E., Lozano, J., Smith, J., Merelo-Guervós, J., Bullinaria, J., Rowe, J., Tino, P., Kabán, A., Schwefel, H.-P. (Eds.), Proceedings of the Eighth Conference on Parallel Problem Solving from Nature (PPSN), vol. 3242 of Lecture Notes in Computer Science, Springer-Verlag,s pp. 571–580.

Leier, A., Kuo, P., Banzhaf, W. Analysis of preferential network motif generation in an artificial regulatory network model created by duplication and divergence. Adv. Complex Syst., in preparation.

Mangan, S., Alon, U., 2003. Structure and function of the feed-forward loop network motif. Proc. Natl. Acad. Sci. 100 (21), 11980–11985.

Mason, J., Linsay, P., Collins, J., Glass, L., 2004. Evolving complex dynamics in electronic models of genetic networks. Chaos 14 (3), 707–715.

Milo, R., Shen-Or, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U., 2002. Network motifs: simple building blocks of complex networks. Science 298, 824–827.

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., Alon, U., 2004. Superfamilies of evolved and designed networks. Nature 303, 1538–1542.

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

Ohno, S., 1970. Evolution by Gene Duplication. Springer, Berlin.

Ohta, T., 2002. Near-neutrality in evolution of genes and gene regulation Proc. Natl. Acad. Sci. 99 (25) 16134–16137

Romualdo, P., Smith, E., Solé, R., 2003. Evolving protein interaction networks through gene duplication. J. Theor. Biol. 222, 199–210.

Shen-Or, S., Milo, R., Mangan, S., Alon, U., 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat. Genet. 31, 64–68.

Teichmann, S.A., Babu, M., 2004. Gene regulatory network growth by duplication. Nat. Genet. 36 (5), 492–496.

Valverde, S., Ferrer Cancho, R., Solé, R., 2002. scale-free networks from optimal design. Europhys. Lett. 60, 512–517.

van Noort, V., Snel, B., Huynen, M.A., 2004. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. EMBO Rep. 5 (3), 280–284.

Vazquez, A., Dobrin, R., Sergi, D., Eckmann, J.-P., Oltvai, Z.N., Barabasi, A.-L., 2004. The topological relationship between the large-scale attributes and local interaction patterns of complex networks. Proc. Natl. Acad. Sci. 101 (52), 17940–17945.

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

Wolfe, K., Shields, D., 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387 (6634), 708–713.

Wuchty, S., 2001. Scale-free behavior in protein domain networks. Mol. Biol. Evol. 18 (9), 1694–1702.

Wuchty, S., Oltvai, Z., Barabási, A.-L., 2003. Evolutionary conservation of motif constituents in the yeast protein interaction network. Nat. Genet. 35 (2), 176–179.

Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R.Y., Alon, U., Margalit, H., 2004. Network motifs in integrated cellular networks of transcription-regulation and protein–protein interaction. Proc. Natl. Acad. Sci. 101 (16), 5934–5939.

Yokobayashi, Y., Weiss, R., Arnold, F., 2002. Directed evolution of a genetic circuit. Proc. Natl. Acad. Sci. 99 (26), 16587–16591.

Zhang, J., 2004. Evolution by gene duplication: an update. Trends Ecol. Evol. 18, 292–298.