Network motifs in natural and artificial transcriptional regulatory networks

Wolfgang Banzhaf * **and P. Dwight Kuo**

Memorial University of Newfoundland, St. John's, NL, Canada A1B 3X5

We show that network motifs found in natural regulatory networks may also be found in an artificial regulatory network model created through a duplication/divergence process. It is shown that these network motifs exist more frequently in a genome created through the aforementioned process than in randomly generated genomes. These results are then compared with a network motif analysis of the gene expression networks of *Escherichia coli* and *Saccharomyces cerevisiae*. In addition, it is shown that certain individual network motifs may arise directly from the duplication/divergence mechanism.

1. INTRODUCTION

Analysis of network motifs has recently become of interest with respect to transcriptional regulation networks. Methods are mainly based on searching for connexion patterns among small numbers of nodes. Here we shall introduce a class of artificial regulatory networks which can be used to compare results obtained through the same methods as have been applied to natural regulatory networks of *Escherichia coli* and *Saccharomyces cerevisiae*. We shall see that the high frequency of certain network motifs detected in natural systems can be found in artificial systems as well, provided they are generated by a gene duplication and divergence process. This leads us to believe that the actual frequency distribution of motifs ("motif fingerprint") in natural regulatory networks is as much if not more a consequence of the process of network generation than of subsequent evolutionary selection.

The artificial regulatory network model presented here has previously been shown to generate networks which exhibit scale-free and small world network topologies [12]. Specifically, if the network generation process is one of duplication and divergence (similar to that presented in [17], though working on an actual genome level) we can show such global connectivity statistics. In this paper, we extend those observations to network motifs and demonstrate that certain motifs frequent in natural regulatory systems also occur repeatedly in this model.

It has also been shown in the past that the regulatory network model is able to reproduce dynamic phenomena found in natural genetic regulatory networks, for instance shifts in onset and offset of gene expression (heterochrony) based on single bit-flip mutations [3]. As such, this model can relate changes in time and intensity 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.

2. BACKGROUND

2.1 Regulatory networks

Regulatory networks are an important new research area in biology [6, 8]. With the realization that in higher organisms only a tiny fraction of DNA is translated into proteins, the question of determining the function of the remaining DNA becomes all the more pertinent. A reasonable answer for the function of this remaining unexpressed DNA appears to be regulation. According to Neidhardt et al. [16], 88% of the genome of the bacterium *E. coli* is expressed with 11% suspected to contain regulatory information (also see Thomas [20]). Given the selective pressures on bacterial genomes, this would point to a very prominent rôle for regulation in general.

In addition, it has been recognized that understanding the differences between species and thus the key to evolution lies in the DNA information controlling gene expression [11]. Since many evolutionary effects can be traced back to their regulatory causes, regulatory networks mediate between development and evolution and thus serve to help unfold the patterns and shapes of organism morphology and behaviour [10, 2].

Studying models of regulatory networks can help us to understand some of these mechanisms by providing lessons for both natural and artificial systems under evolution.

2.2 Network motifs

There has recently been significant interest in studying static network motifs as a tool for under-

*Corresponding author. Tel: (709) 737-8652, fax: (709) 737-2009, e-mail: banzhaf@cs.mun.ca

Journal of Biological Physics and Chemistry **4** (2004) 85–92
Received 27 April 2004

standing regulatory networks [15, 14, 18, 24]. Complex networks have previously been classified by global characteristics such as scale-free [1, 4, 5, 9, 23] and small world network connexion topologies [21, 22]. In order to investigate networks further beyond their global features requires an understanding of the potential basic structural elements that make up complex networks.

It has been proposed that studying so-called "network motifs" can lead towards such an understanding [13, 15, 18]. Network motifs may be defined as the structural elements (subgraphs) which form the basic elements of more complex networks. Whereas an edge usually connects two nodes, network motif analysis starts with three nodes and their corresponding connexions.

Interestingly, certain network motifs occur with significantly higher probability in natural regulatory networks than in random networks [15]. It has also been shown that network motifs may be conserved over evolutionary time, for instance in the yeast protein interaction network [24].

In order to detect all n-node network motifs, we have implemented an algorithm similar to one devised by Milo et al. [15]. The algorithm scans all rows of the adjacency matrix M of connexions between nodes searching for non-zero elements (i, j) which represent a connexion from node i to node j. The algorithm then recursively traverses the neighbouring vertices connecting vertex i and j until a specific n-node motif is detected. The constituent vertices and edges of a motif are then compared to previously found motifs in order to ensure that none have been overcounted. It must also be noted that the total number of motifs of a given type is counted and possible isomorphisms are considered to be the same motif type. Table 1 in the Appendix lists all 3-node connexion patterns in directed graphs, including auto-connexions, up to isomorphism. We shall later refer to particular motifs with their motif ID (given in the table) only.

3. ARTIFICIAL REGULATORY NETWORK MODEL

The artificial regulatory network (ARN) model presented here is based on work by one of the authors [3, 2]. The ARN consists of a bit string representing a genome with direction (i.e. 5′→ 3′ in DNA) and mobile "proteins" which interact with the genome through their constituent bit patterns. In this model, proteins are able to interact with the genome most notably at "regulatory" sites located upstream from genes, see Figure 1. Attachment to these sites produces either inhibition or activation of the corresponding protein. It can thus be interpreted as a regulatory network with proteins acting like transcription factors.

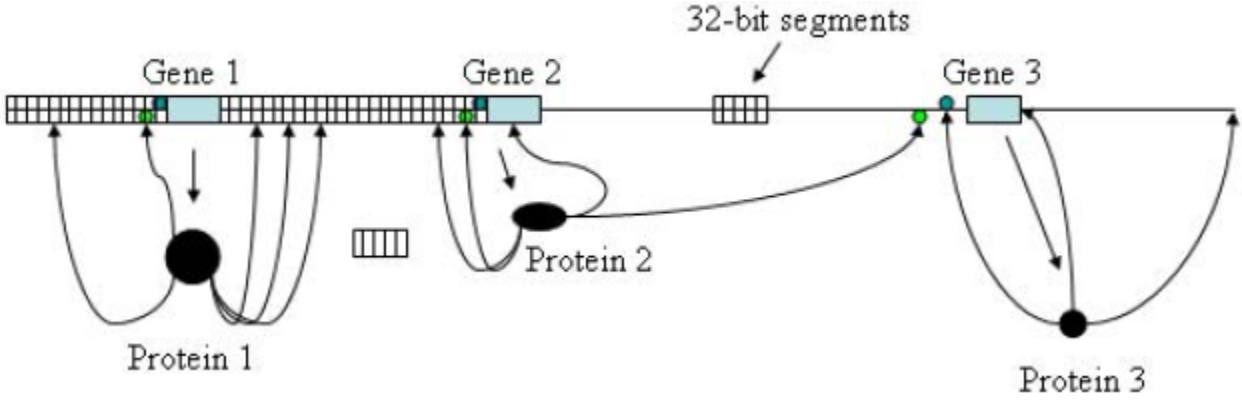

Figure 1. The interaction between proteins and the genome.

The genome itself can be created through a series of duplication/divergence events. First, a random 32 bit string is generated. This string is then used in a series of length duplications followed by mutations in order to generate a genome of length L_G. A "promoter" bit sequence of 8 bits was then arbitrarily selected to signal the start of a gene on the genetic string analogous to an open reading frame (ORF) on DNA. The actual gene length is set to a fixed length of l_g= 5 32-bit integers which results in an expressed bit pattern of 160 bits per gene. Therefore, genes can 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 promoter sites 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). In this model, we assume only one regulatory site for the increase of expression and one site for the decrease of expression

of proteins. This is a radical simplification since natural genomes may have 5–10 regulatory sites that may even be occupied by complexes of proteins [2].

Processes such as transcription, and elements such as introns, RNA-like mobile elements and translation procedures resulting in a different alphabet for proteins are neglected in this model. 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 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_g) for a given bit position are resolved by chance.

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 bits on both patterns 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 [2].

By making the simplifying assumption that the occupation of both of a gene's regulatory sites modulates the expression of its corresponding protein, we may deduce a gene-protein interaction network comprising the different genes and proteins which can be parametrized by strength of match.

By examining the interaction networks at different matching strengths (we call them thresholds) we may obtain different network topologies for the same connected network components. 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 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 threshold levels of 21 and 22. Here and later we do not discern between enhancer and inhibitor sites, although such an analysis would be necessary to understand the actual function of motifs.

It must be stressed that although the actual genome has not changed, by simply changing the threshold parameter we can obtain different network topologies. It may be noted by the more astute reader that the diagrams in Figs. 2 and 3 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 parametrized 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 here.

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 connexions 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.

The end result of this process would be firstly isolated pairs of nodes, then nodes without connexions, which would disappear from the network completely.

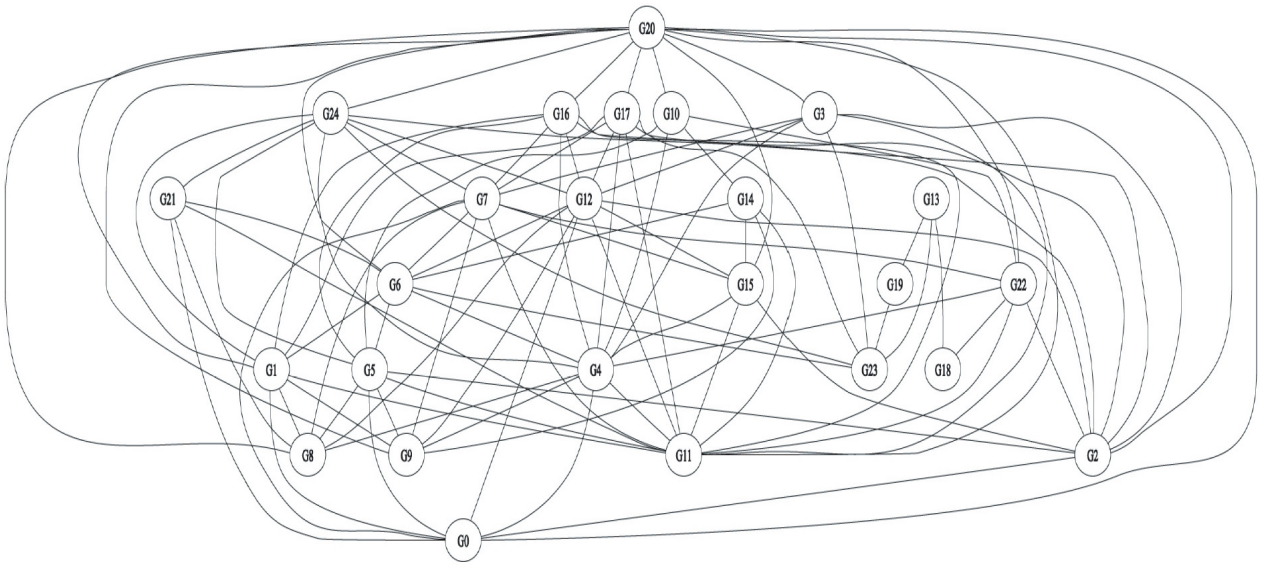

Figure 2. Sample of a gene-protein interaction network for a duplication/divergence genome at a threshold of 21 bits.

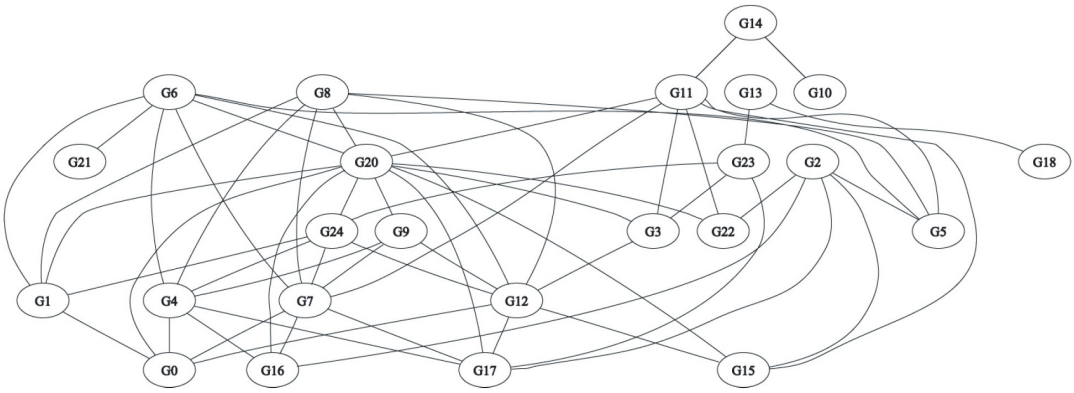

Figure 3. The same gene-protein interaction network at a threshold of 22 bits.

4. RESULTS

The network motif finding algorithm was applied to 800 instances of the artificial regulatory model generated by the duplication/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 motif counting are shown in Figs 4 and 5. For both methods of network generation, the genome length was set at 131072 (12 duplication events in the case of duplication/divergence). For networks generated by duplication/divergence the mutation rate was set at 1%. In both cases the threshold had to be determined. We observed that 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 our artificial regulatory network framework, the threshold was chosen by iteratively raising the threshold until the network generated had a ratio that was equal to or less than 2 to 1.

This was then compared to the results of applying the algorithm to the transcriptional networks of *Escherichia coli* [18, 19] and *Saccharomyces cerevisiae* [7], see Figs 6 and 7. Milo et al. defined network motifs as *n*-node subgraphs which occur significantly more than at random [15]. We prefer a more general definition here, and speak instead of a characteristic motif fingerprint if talking about the count with regard to a particular network. It can be seen in Figs 4–7 that the most frequent natural motifs (ID 22 and ID 12) are both well represented in duplication/divergence type artifical networks whereas only one of them can be detected in fully random networks.

Table 2 in the Appendix lists all regulatory networks looked at for this paper, and lists the distribution of all motifs. Note that for artificial networks we have chosen average numbers of counts, whereas there is only one example each for the natural regulatory systems.

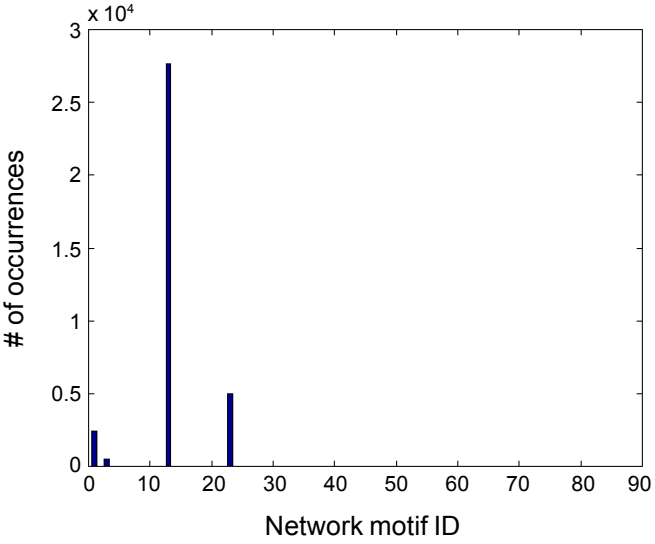

Figure 4. Average of frequency of occurrence of network motifs in 800 instances of the artificial network model generated by a duplication/divergence procedure.

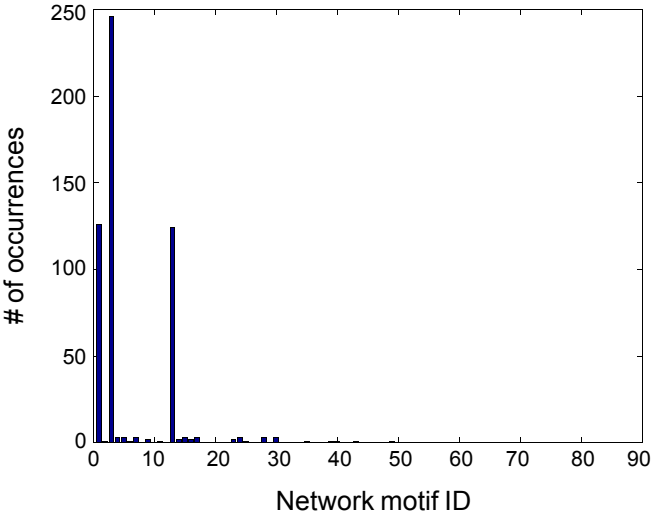

Figure 5. Average of frequency of occurrence of network motifs in 800 randomly generated instances of the artificial network model.

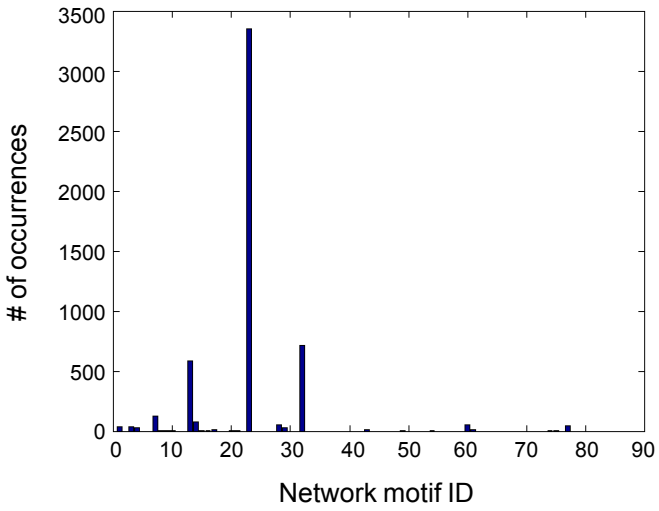

Figure 6. Frequency of occurrence of network motifs in the transcriptional network of *Escherichia coli.*

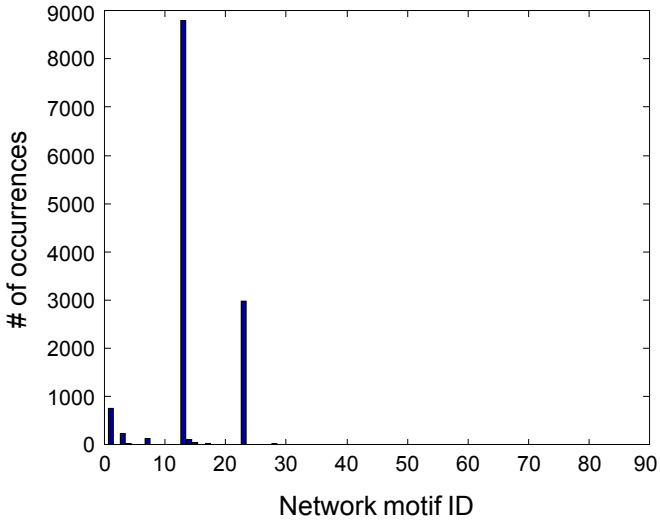

Figure 7. Frequency of occurrence of network motifs in the transcriptional network of *Saccharomyces cerevisiae*.

5. CONCLUSIONS

A look at the table might convince us that there is a clear relation between the natural distribution of motifs and the distribution in artificial networks generated by a duplication and divergence process. No evolutionary selection pressure has been applied in artificial systems, though. Thus it can be stated that the distribution outcome is more a reflexion of the mechanism of its generation than a result of evolutionary pressures (although evolutionary pressures are certainly responsible for fine tuning of distributions). The duplication/divergence events might interweave in a complex and history-dependent way with selective pressure for function. Unless this pressure favours the absolute minimal network to achieve a certain function, the relationship between motif structure and motif function is not so straightforward as has been suggested in the literature [14].

In pondering the generation process for networks, it might be pointed out that from the simplest elements of a network, two nodes connected by a link (or one node connected by a link to itself), an arbitrary number of patterns can be generated by duplication and divergence events. The Bi-Fan motif (4 node) found in abundance in natural regulatory networks can be easily generated from the above element by simple duplication. In the same vein, the feed forward loop consisting of 3 nodes (ID 14) can be generated from a 2 node motif with an autoconnexion by a partial duplication and two mutations effecting a loss of the autoconnexions.

So far we have not examined the case of enhancing and inhibiting connexions, a fact that further complicates motif analysis. It remains to be seen whether dividing networks into their smallest components will teach us something useful about the overall structure of regulatory networks.

ACKNOWLEDGMENTS

The authors would like to kindly thank François Képès of the Atelier de Génomique Cognitive, CNRS and Nadav Kashtan of the Weizmann Institute of Science for helpful discussions and suggestions.

REFERENCES

1. R. Albert, H. Jeong & A. Barabási. The internet's achilles' heel: error and attack tolerance of complex networks. *Nature (Lond.)* **406** (2000) 378–382.
2. W. Banzhaf. On the dynamics of an artificial regulatory network. In: *Advances in Artificial Life— Proceedings of the 7th European Conference on Artificial Life (ECAL)* (eds W. Banzhaf, T. Christaller, P. Dittrich, J.T. Kim & J. Ziegler), vol. 2801 of *Lecture Notes in Artificial Intelligence*, pp. 217–227. Berlin: Springer (2003).
3. W. Banzhaf. Artificial regulatory networks and genetic programming. In: *Genetic Programming Theory and Practice* (eds R. L. Riolo & B. Worzel), Ch. 4, pp. 43–62. Dordrecht: Kluwer (2003).
4. A. L. Barabási & R. Albert. Emergence of scaling in random networks. *Science* **286** (1999) 509–512.
5. A. L. Barabási, H. Jeong, R. Ravasz, Z. Neda, T. Vicsek & A. Schubert. Evolution of the social network of scientific collaborations. *Physica A* **311** (2002) 590–614.
6. *Computational Modeling of Genetic and Biochemical Networks* (eds J.M. Bower & H. Boulouri). Cambridge, Mass.: MIT Press (2001).
7. M. C. Costanzo, M. E. Crawford, J. E. Hirschman, J. E. Kranz, P. Olsen, L. S. Robertson, M. S. Skrzypek, B. R. Braun, K. L. Hopkins, P. Kondu, C. Lengieza, J. E. Lew-Smith, M. Tillberg & J. I. Garrels. YPD™, PombePD™ and WormPD™: model organism volumes of the BioKnowledge™ library, an integrated resource for protein information. *Nucleic Acids Research* **29** (2001) 75–79.
8. E. H. Davidson. *Genomic Regulatory Systems*. San Diego: Academic Press (2001).

9. M. Faloutsos, P. Faloutsos & C. Faloutsos. On power-law relationships of the internet topology. In: *SIGCOMM*, pp. 251–262 (1999).
10. F. Harold. *The Way of the Cell*. Oxford: University Press (2001).
11. L. Hood & D. Galas. The digital code of DNA. *Nature (Lond.)* **421** (2003) 444–448.
12. P. D. Kuo & W. Banzhaf. Scale-free and small world network topologies in an artificial regulatory network model. In: *Artificial Life IX*, (in press).
13. S. Mangan & U. Alon. Structure and function of the feed forward loop network motif. *Proc. Natl Acad. Sci. USA* **100** (2003) 11980–11985.
14. R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. S. Shen-Or, I. Ayzenshtat, M. Sheffer & U. Alon. Superfamilies of evolved and designed networks. *Science* **303** (2004) 1538–1542.
15. R. Milo, S. Shen-Or, S. Itzkovitz, N. Kashtan, D. Chklovskii & U. Alon. Network motifs: simple building blocks of complex networks. *Science* **298** (2002) 824–827.
16. F. C. Neidhardt. *Escherichia coli and Salmonella typhimurium*. Washington, DC: ASM Press (1996).
17. P. S. Romualdo, E.D. Smith & R.V. Solé. Evolving protein interaction networks through gene duplication. *J. theor. Biol.* **222** (2003) 199–210.
18. S. S. Shen-Or, R. Milo, S. Mangan & U. Alon. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics* **31** (2002) 64–68.
19. D. Thieffry, A. M. Huerta, E. Pérez-Rueda & J. Collado-Vides. From specific gene regulation to global regulatory networks: a characterisation of the *Escherichia coli* transcriptional network. *BioEssays* **20** (1998) 433–440.
20. G. H. Thomas. Completing the *E. coli* proteome: a database of gene products characterised since completion of the genome sequence. *Bioinformatics* **7** (1999) 860–861.
21. D.Watts. *Small Worlds: the Dynamics of Networks between Order and Randomness*. Princeton: University Press (2003).
22. D. Watts & S. Strogatz. Collective dynamics of small-world networks. *Nature (Lond.)* **363** (1998) 202–204.
23. S.Wuchty. Scale-free behavior in protein domain networks. *Mol. Biol. Evolution* **18** (2001) 1694–1702.
24. S.Wuchty, Z.N. Oltvai & A.-L. Barabási. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genetics* **35** (2003) 176–179.

APPENDIX

Table 1: Network motifs and their identification (ID).

Motif										
Motif ID	0	1	2	3	4	5	6	7	8	9
Motif										
Motif ID	10	11	12	13	14	15	16	17	18	19
Motif										
Motif ID	20	21	22	23	24	25	26	27	28	29
Motif										
Motif ID	30	31	32	33	34	35	36	37	38	39
Motif										
Motif ID	40	41	42	43	44	45	46	47	48	49
Motif										
Motif ID	50	51	52	53	54	55	56	57	58	59
Motif										
Motif ID	60	61	62	63	64	65	66	67	68	69
Motif										
Motif ID	70	71	72	73	74	75	76	77	78	79
Motif										
Motif ID	80	81	82	83	84	85				

Table 2. Network motifs and their distribution. D/D: Duplication/Divergence genomes; Rand: Random genomes. AlonIDs shown as A are autoconnected nodes without a code in Alon's system.

Network IDs		Count in				Network IDs		Count in			
ID	AlonID	D/D	Rand	*E. coli*	*S. cerv*	ID	AlonID	D/D	Rand	*E. coli*	*S. cerv*
0	6	2424	126	35	751	45	A	1	0	0	0
1	A	4	1	0	1	46	110	0	0	0	0
2	12	490	246	40	246	47	A	0	0	0	0
3	A	11	3	26	24	48	A	0	1	3	0
4	14	6	3	0	0	49	A	0	0	0	0
5	A	0	1	0	0	50	A	0	0	0	0
6	A	12	3	124	138	51	A	0	0	0	1
7	A	0	0	8	0	52	A	0	0	0	0
8	A	0	2	1	0	53	A	0	0	1	0
9	A	0	0	2	0	54	A	0	0	0	0
10	A	0	1	0	0	55	A	0	0	0	0
11	A	0	0	0	0	56	A	0	0	0	0
12	36	27659	124	587	8800	57	A	0	0	0	0
13	A	8	2	76	104	58	A	0	0	0	0
14	38	15	3	2	44	59	A	0	0	54	4
15	A	0	2	1	1	60	A	0	0	12	0
16	A	20	3	11	22	61	A	0	0	0	0
17	46	0	0	0	1	62	A	0	0	0	0
18	A	0	0	0	0	63	A	0	0	0	0
19	A	0	0	2	1	64	A	10	0	0	0
20	A	0	0	1	0	65	A	0	0	0	0
21	A	0	0	0	0	66	A	0	0	0	0
22	A	5016	2	3353	2987	67	A	0	0	0	0
23	74	36	3	0	18	68	238	0	0	0	0
24	A	5	1	0	0	69	A	0	0	0	0
25	78	3	0	0	0	70	A	0	0	0	0
26	A	0	0	0	0	71	A	0	0	0	0
27	A	6	3	53	25	72	A	0	0	0	0
28	A	0	0	32	0	73	A	0	0	6	0
29	A	0	3	0	0	74	A	0	0	3	0
30	A	0	0	0	0	75	A	0	0	0	0
31	A	14	0	713	0	76	A	0	0	46	0
32	A	0	0	0	0	77	A	0	0	0	0
33	A	3	0	0	0	78	A	0	0	0	0
34	A	0	1	0	0	79	A	0	0	0	0
35	A	0	0	0	0	80	A	0	0	0	0
36	A	0	0	0	0	81	A	0	0	0	0
37	A	0	0	0	0	82	A	0	0	0	0
38	98	0	1	0	0	83	A	0	0	0	0
39	A	0	1	0	0	84	A	0	0	0	0
40	102	0	0	0	0	85	A	0	0	0	0
41	A	0	0	0	0						
42	A	6	1	14	3						
43	A	0	0	0	0						
44	108	0	0	0	0						