Advances in Complex Systems, Vol. 10, No. 2 (2007) 155–172

World Scientific
www.worldscientific.com

ANALYSIS OF PREFERENTIAL NETWORK MOTIF GENERATION IN AN ARTIFICIAL REGULATORY NETWORK MODEL CREATED BY DUPLICATION AND DIVERGENCE

ANDRÉ LEIER

Advanced Computational Modelling Centre, University of Queensland, Brisbane, QLD 4072, Australia
leier@maths.uq.edu.au

P. DWIGHT KUO

Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093-0412, USA
pdkuo@ucsd.edu

WOLFGANG BANZHAF

Department of Computer Science, Memorial University of Newfoundland, St. John's, NL A1B 3X5, Canada
banzhaf@cs.mun.ca

Received 29 March 2006
Revised 24 July 2006

Previous studies on network topology of artificial gene regulatory networks created by whole genome duplication and divergence processes show subgraph distributions similar to gene regulatory networks found in nature. In particular, certain network motifs are prominent in both types of networks. In this contribution, we analyze how duplication and divergence processes influence network topology and preferential generation of network motifs. We show that in the artificial model such preference originates from a stronger preservation of protein than regulatory sites by duplication and divergence. If these results can be transferred to regulatory networks in nature, we can infer that after duplication the paralogous transcription factor binding site is less likely to be preserved than the corresponding paralogous protein.

Keywords: Gene duplication; network motif; gene regulatory networks; artificial regulatory networks.

1. Introduction

Understanding genetic regulatory networks (GRNs) is essential to gain understanding of cell development, differentiation and organization; cell functions and malfunctions; and, ultimately, cell control aiming at the discovery of drugs against specific diseases. For the structural analysis of GRNs, the study of network motifs has become a useful approach to identify the network's basic building blocks [22,28,34].

Network motifs are subgraphs which occur in a given network significantly more often than in random networks, i.e. they are over-represented compared to what is to be expected on average [23]. The over-representation of certain network motifs is usually associated with a functional advantage for the system. In this sense, the study of network motifs might help to bridge the gap between the structure and function of regulatory networks. Mostly, basic network motifs have been studied so far, consisting of three or four nodes [28].

This contribution studies the origin of particular network motifs in an artificial regulatory network (ARN) model first introduced by Banzhaf [3, 4]. This model system has been examined earlier in terms of network topology (scale-free/small world topologies and network motifs) [5,18,20] and dynamics [19,20] when created through a whole genome duplication and divergence process. Of particular note, the ARN model was explored in terms of its subgraph distribution and compared to distributions of the genetic transcriptional regulatory networks of *Escherichia coli* and *Saccharomyces cerevisiae* [5, 20]. Results show that there is a similarity between natural subgraph distribution and the subgraph distribution for ARNs, provided these are created by duplication and divergence processes as opposed to completely randomly generated networks.

In particular, certain network motifs detected in GRNs in high numbers can be found in the ARN model in high numbers as well. It is interesting to note that duplication and divergence networks are more similar to the eukaryotic *Saccharomyces cerevisiae* than to the prokaryotic *Escherichia coli* [20]. This suggests that the topology has been shaped by duplication events in *Saccharomyces cerevisiae*'s evolutionary history. In fact, there are hints that over 90% of eukaryotic genes are created by gene duplication [32].

We have previously argued that the distribution of motifs is essentially shaped by the underlying process of network generation, i.e. by duplication and divergence, rather than by evolutionary selection pressure for function, which might be a secondary force shaping networks [5]. It is believed that gene duplication with subsequent functional divergence provided novel genes for organisms, thus facilitating adaptation, and entailing diversification. There is currently evidence for gene duplications in yeast [8, 9, 15, 32, 33], *Escherichia coli* [1, 2, 9] and vertebrates [10,11,24,27,29]. In particular, whole genome duplication has been proposed as a mechanism for generating novelty in the genome and may give a reasonable explanation for speciation [25,30]. Specifically, in yeast there is evidence that whole genome duplication has effectively rewired the yeast network allowing for rapid anaerobic growth via the loss of a specific cis-regulatory element from dozens of genes [13]. However, it is still an open question as to how the variety of functions and the inherent modularity seen in present day organisms could emerge from duplication processes [12]. Therefore, we here embark on a study to understand how such processes can influence network topology.

In Refs. 5 and 20, genomes created by whole genome duplication and divergence reveal that particular network motifs, identified as ID-12 and ID-22 (see Fig. 1)

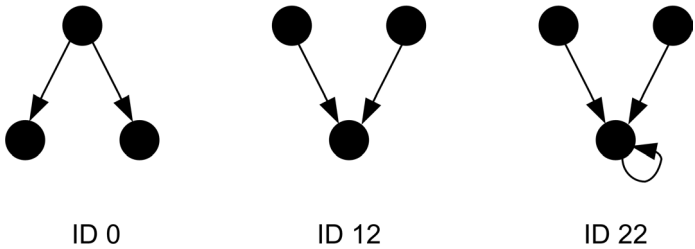

Fig. 1. From left to right: ID-0 (single-output motif), ID-12 (single-input motif) and ID-22 (equivalent to ID-12 with an additional auto-regulatory connection at the single input node).

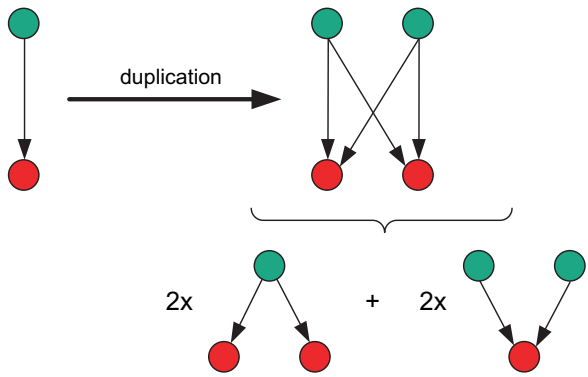

Fig. 2. The effect of one whole genome duplication on the simplest regulatory interaction. The resulting 4-node network with two input and two output nodes is the so-called *bi-fan motif*.

occur much more often than motif ID-0. This is, however, not the case in ARNs generated by randomly selecting the genome [5]. One would expect both motifs to occur with nearly the same frequency if they were solely created by an unbiased duplication of the simplest two-node motif (one protein interacting with another gene). Figure 2 depicts one DD-step on such a network showing that both single-input and single-output modules should be generated in (approximately) the same numbers.

Since this was also found to be the case in the natural GRNs of *Escherichia coli* and *Saccharomyces cerevisiae*, it is informative to study why whole genome duplication and divergence demonstrates a preference for the generation of ID-12 and 22. It should also be noted that ID-12 can be considered part of the bi-fan motif [14], which can be created by repeated duplications of the aforementioned two-node motif.

A recent study [25] has shown that network motifs that provide for reliable dynamics in the presence of noise are more abundant in biological networks. Although in this study ID-0 and ID-12 have the same "graph reliability," indicating that a selective advantage through robustness cannot (solely) explain the motif preference discussed here, it suggests that specific network motifs may be favored

in evolution due to an associated functional advantage. However, we believe that not only functionality shapes the topology of biological networks and contributes to preferential motif generation but also the underlying mechanisms that create the network topology.

Research in Refs. 1 and 32 proposed that network motifs are not created by duplication events but are built by incremental evolution of gene interactions. It was also suggested in Ref. 7 that network motifs are found through convergent evolution — not through any duplication processes. Even if this would in fact be the case, understanding why a whole genome duplication and divergence procedure preferentially generates motifs of ID-12 and ID-22 rather than ID-0 is valuable. It could shed light on possible explanations for why such preference occurs in nature and why there is a similarity between subgraph distributions in GRNs and ARNs [5]. An examination of this question is the subject of the present contribution.

2. The Artificial Regulatory Network Model

Our artificial regulatory network model tries to captures essentials of natural regulatory networks. It consists of a string of binary numbers (bits) representing a genome with a directional readout comparable to the $5' \rightarrow 3'$ direction in DNA. This artificial genome carries genes, whose beginning is marked by a unique promoter pattern (consisting of a particular sequence of bits). In our model, genes are all of identical length, which is not a severe restriction to the model since a termination sequence could be easily introduced as well. In order to keep things simple, however, such a signal which would cause alternative transcription products, has not been introduced yet. Genes are "transcribed" into another bit pattern, of a mobile protein variety. To this end, the information on the gene string is subjected to a many-to-one-mapping producing a smaller protein bit pattern. Further on, proteins can attach to the genome by virtue of complementary pattern matching between the bits carried by the protein and the bits residing on the genome string. In particular, there are two sites upstream of the promoter signal of a gene which are considered regulatory sites for enhancing or inhibiting expression of the corresponding gene. There are a couple of parameters of the model, like length of promoter (in bits), length of gene (in bits, bytes or words/integers), length of proteins and exact mapping rule, and length of regulatory sites and their distance from the promoter which need to be fixed in order to make the model work.

In our case, the 8-bit sequence 01010101 is defined as a "promoter" signalling the start of a gene analogous to an open reading frame (ORF) on DNA. Each gene has a fixed length of 160 bits divided into five 32-bit integers. A gene might overlap with another (if another promoter sequence can be found in the coding region of a gene). However, for overlapping promoters, only the first promoter (starting at the leftmost bit of a periodically extended promoter) marks a gene. Immediately upstream from the promoter are the two 32-bit segments identified as regulatory

sites for increase (enhancer site) and decrease (inhibitor site) of expression. In total, a gene is characterized by 232 bits.

Proteins, in turn, are 32-bit sequences constructed from the corresponding gene by a many-to-one mapping. This mapping is implemented by performing the majority rule on each of the the 32 bit positions in the five 32-bit integers. The matching strength between a protein and a regulatory site is determined by applying the XOR operator on each bit pair of the two 32-bit segments and summing the number of 1s resulting from these operations. A match is established if the matching strength is greater or equal to a given threshold. Thus, the gene–protein interaction network as it is determined by the genome string is parametrized by the threshold value. Figure 3 shows the schematic of a gene and its corresponding protein.

For the purpose of this contribution, we discern matchings between proteins and regulatory sites as gene–protein interactions in general but do not employ the different functionality of enhancer and inhibitor sites. Properly discerning these sites will lead to differential transcription of the corresponding genes, and to varying protein concentrations, a topic that was examined elsewhere [4, 19]. In many respects, our model is a radical simplification of real gene regulatory networks. Yet the hope is to be able to capture key features of natural systems, even under these simplified conditions.

The genome itself (a long bitstring) is created through a series of duplication and divergence (DD) steps. Starting with a randomly generated 32-bit sequence, during every DD-step, the whole genome is first duplicated and subsequently mutated with a certain mutation rate, p. That is, the newly created genome is twice as large as its "ancestral" genome and every bit in the new genome mutates with probability p. Figure 4 illustrates the DD-process for the first five DD-steps.

Naturally, each ARN can be visualized by a directed graph where nodes represent gene/protein pairs and edges represent interactions between them. An edge points from one node to another if the protein corresponding to the first node (source) matches a regulatory site of the gene corresponding to the second node (destination). It has been shown that artificial regulatory networks created by whole

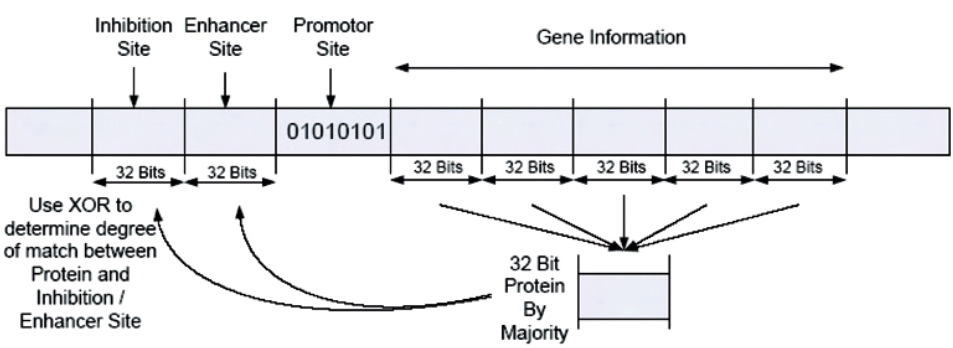

Fig. 3. Schematic view of a gene and the protein generation mechanism in the ARN model.

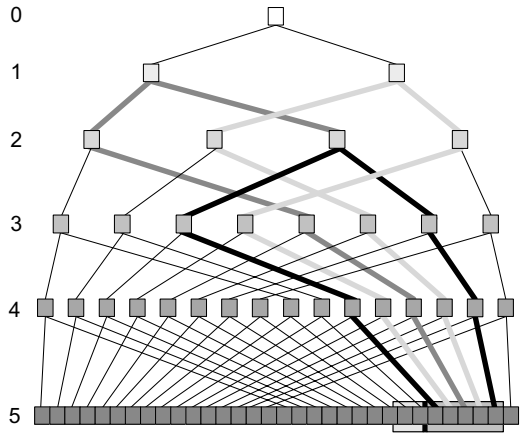

Fig. 4. Schematic view of five DD-steps. Each square represents 32 bits. The original bit string is shown as a white square. Edges in the tree refer to the "phylogeny" by means of consecutive DD-steps. The squares' grey values correspond to the number of DD-steps. A gene is marked in the genome after five DD-steps. Bold edges illustrate the paths to the most recent common ancestor between two 32-bit blocks in the gene (unless this ancestor is the original 32 bit sequence).

genome duplication and divergence at certain thresholds of matching strength display the characteristics of small-world [20] and scale-free network topologies [18] and carry network motifs [5].

Note that by changing the threshold, the genome itself does not change. It is in fact the connectivity or interactivity between genes and proteins (the topology of the network) that changes with varying threshold levels. In this work, ARN genomes were generated by 12 whole genome DD-steps with a mutation rate of 1% on a random 32-bit string. The threshold was determined by iteratively increasing its value until the ratio of the number of edges to the number of vertices in the network became equal to or less than two to one which was the approximate observed ratio for two regulatory networks found in nature [5].

3. Analysis

To understand how preferential motif generation arises as a result of the duplication and divergence process, we need to have a closer look at the creation, derivation and ancestry of genes.

Let G_M denote the genome after M DD-steps and let $l_M = 32 * 2^M$ be its length. Moreover, let x_i^M denote the bit at position i in G_M. In a single DD-step, $G_M \rightarrow G_{M+1}$, the current genome is duplicated and mutated. Every bit x_i^M, $i = 1, \ldots, l_M$, in G_M is the template or *direct ancestor* for exactly two bits in the newly generated genome G_{M+1}, namely, for x_i^{M+1} and $x_{i+l_M}^{M+1}$. In other words, x_i^M is the direct ancestor of x_j^{M+1} if and only if $j \equiv i \pmod{l_M}$. The entire DD-process is a series of DD-steps $G_0 \rightarrow G_1 \rightarrow \cdots \rightarrow G_N$, starting with the original 32-bit genome sequence G_0 ($l_0 = 32$). A specific bit x_i^N at position i in G_N is successively

derived from $N-1$ bits $x^0_{j_0}, x^1_{j_1}, \ldots, x^{N-1}_{j_{N-1}}$, where $x^k_{j_k}$ is the direct ancestor of $x^{k+1}_{j_{k+1}}$. These bits will be designated as *ancestors of* x^N_i.

Two bits x^N_i and x^N_j, $i \neq j$, have a common ancestor in G_M, $0 \leq M < N$, if and only if $i \equiv j \pmod{l_M}$. The common ancestor is bit x^M_k with $1 \leq k \leq l_M$ and $k \equiv i \equiv j \pmod{l_M}$. If x^M_k is an ancestor of x^N_i and x^N_j with $i, j < l_N$ and $k < l_M$ then x^M_{k+1} is a common ancestor of x^N_{i+1} and x^N_{j+1}. In case of $k = l_M$, x^M_1 is the common ancestor of x^N_{i+1} and x^N_{j+1}. The *latest* (*most recent*) *common ancestor* of two arbitrary bits x^N_i and x^N_j is the common ancestor x^M_k in G_M such that neither x^{M+1}_k nor $x^{M+1}_{k+l_M}$ is a common ancestor in G_{M+1}. In this case, the difference $N-M$ is a measure for the "phylogenetic distance" in the number of DD-steps between the bits at position i and j in G_N.

Using the definition of ancestry for single bits the definition of ancestry for bit sequences and genes is straightforward. It goes without saying that genes have ancestors which are not genes because they do not have full gene size (232 bits) or an ORF. In fact, at the earliest after three DD-steps (genome length of 256) it is possible that a valid gene is generated. Ignoring their real nature, we will still call these ancestors genes. Let $g^N_i = x^N_i x^N_{i+1} \cdots x^N_{i+L-1}$ and $g^N_j = x^N_j x^N_{j+1} \cdots x^N_{j+L-1}$ be two genes of length L starting at positions i and j, $i \neq j$, in genome G_N. They both share a common ancestor in G_M with $0 \leq M < N$, if and only if any (and hence every) two bits x^N_{i+k}, x^N_{j+k}, with $0 \leq k < L$, have a common ancestor in G_M. The latest common ancestral gene is defined analogous to the latest common ancestral bit. With G_M being the latest common ancestor of g^N_i and g^N_j, $N-M$ is the phylogenetic distance, that is, the number of DD-steps since the single common phylogenetic lineage of the two genes divided into two separate lineages creating two paralogous genes.

According to the size of the most recent common ancestor we distinguish between *total ancestry* (full gene size, cf. Fig. 5) and *partial ancestry* (less than full gene size, cf. Fig. 6). For total ancestry, the phylogenetic distance between

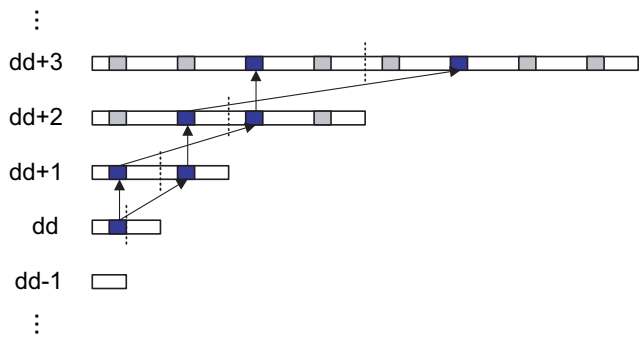

Fig. 5. Example of total ancestry of two genes from a single gene. Arrows mark the "phylogeny." Light-grey boxes represent other sequences/genes derived from the earliest ancestor which is generated after *dd* duplications. Divergence is not considered in this scheme.

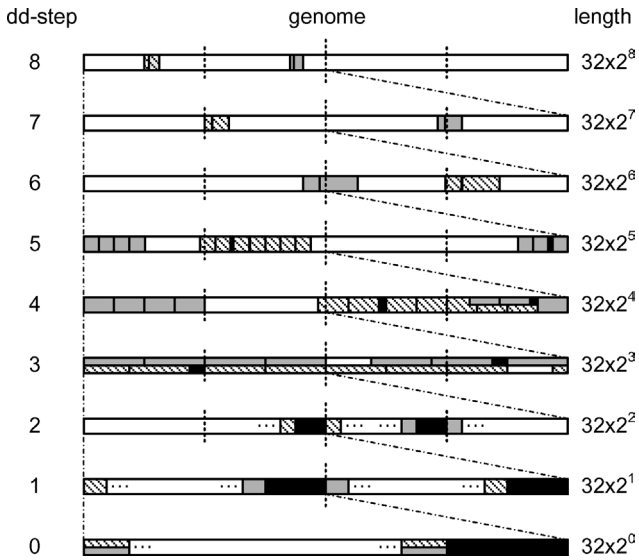

Fig. 6. Example of partial ancestry. The latest common ancestor of the two genes, marked grey and striped, is the original 32-bit sequence. Those bits representing the ORFs are marked black. Before the third duplication, ancestors overlap themselves several times. After six duplications, the two separate, non-overlapping ancestors of full size are created. In this representation, genomes are depicted with constant length. Their real length is shown on the right. Divergence is not considered in this scheme.

genes is limited by $N - 3$ due to the limited genome size. In the case of partial ancestry there is a lower limit for the phylogenetic distance of two. It might also happen that the latest common ancestor is folded multiple times (its bits are the ancestors for several bits in the two genes) and that after division the separated ancestors overlap each other. Ancestry of genes, either total or partial, will be determined by the position of the ancestor's promoter region. The differentiation between non-ancestry, partial and total ancestry will simplify our analysis.

3.1. *Robustness and similarity*

Given two bit sequences of the same length, their similarity can be measured by the number of bit positions with equal bits. For this analysis, however, it is more suitable to define similarity between two bits x_i and y_i as the probability $P(x_i = y_i)$ that they are equal. Thus, similarity $S_{x,y}$ between two bit sequences $x = x_1, \ldots, x_m$ and $y = y_1, \ldots, y_m$ of length m can be defined as the normalized sum of bit similarities for every bit position:

$$S_{x,y} = \frac{1}{m} \sum_{i=1}^{m} P(x_i = y_i).$$

In the following section, we will show that strong similarities between pairs of genes or proteins, respectively, with partial or total ancestry play a crucial role in

the generation of preferential motifs. In particular, due to the majority rule (which is an example of a many-to-one mapping) proteins are better preserved and thus more similar than regulatory sites. We argue that this form of robustness mechanism will inevitably lead to a preferential generation of single-input motifs.

The similarity between two genes g_i^N, g_j^N and, thus, between the two corresponding proteins, in G_N depends on the level of ancestry (non, partial, total), the number of DD-steps N_b until the latest common ancestor is generated in G_{N_b} and the phylogenetic distance $N_a = N - N_b$. For two uncorrelated genes without common ancestor the similarity of regulatory sites and of the corresponding proteins is 50%. This is so because every two bits (at the same relative position in the genes) are generated independent of each other and the generation of 1s or 0s in the protein is unbiased.

For genes with total ancestry we have to calculate the similarity of protein pairs and regulatory site pairs differently. We refer to the appendix for a more detailed discussion of the main steps of this calculation. Table 1 lists the similarities between two related bits in the genes/regulatory sites for a given phylogenetic distance N_a (cf. Appendix A.3). Table 2 shows the similarities between two related protein bits for certain values of N_a and N_b (cf. Appendix A.4). Here, the similarity depends also on N_b because the distribution of 5-bit pattern in the latest common ancestor in genome G_{N_b} depends on it (cf. Appendix A.2).

Tables 1 and 2 reveal that the probabilities for equal bits in proteins are much larger compared to the probabilities for equal bits in regulatory sites. For example, a gene g might be generated after the fifth DD-step ($N_b = 5$). After another $N_a = 7$ DD-steps, two proteins p_1 and p_2 (built by two genes g_1 and g_2 with latest common ancestor g) have a similarity of 97.88%. The similarity of the two corresponding regulatory sites, however, is just 87.68%.

In the event of partial ancestry of two genes, we have to consider the possibility of overlappings, that is, there is no common ancestor in the genome but

Table 1. Probability (in %) that two bits which are derived by N_a DD-steps from the same original (latest common ancestor) bit, X, are both equal to X, both equal to $Y = \text{NOT } X$ or different. The last column shows the probabilities that the two bits are identical.

N_a	$P(XX)$	$P(YY)$	$P(XY) = P(YX)$	$P(XX) + P(YY)$
1	98.01	0.01	0.99	98.02
2	96.08	0.04	1.94	96.12
3	94.21	0.09	2.85	94.29
4	92.39	0.15	3.73	92.54
5	90.62	0.23	4.57	90.85
6	88.91	0.33	5.38	89.24
7	87.25	0.43	6.16	87.68
8	85.63	0.56	6.91	86.19
9	84.07	0.69	7.62	84.76
10	82.54	0.84	8.31	83.38

Table 2. Probability that two protein bits, both derived by N_a DD-steps from an ancestral 5-bit gene pattern created after N_b DD-steps from the original 32-bit sequence are equal. Probabilities are given in % and are rounded to two decimal places.

		N_b							
		3	4	5	6	7	8	9	10
	1	99.82	99.79	99.76	99.73	99.68	99.64	99.58	99.53
	2	99.63	99.58	99.51	99.44	99.35	99.26	99.15	99.04
	3	99.43	99.34	99.24	99.13	99.00	98.85	98.70	98.54
	4	99.21	99.09	98.95	98.79	98.62	98.43	98.23	98.02
N_a	5	98.96	98.80	98.62	98.42	98.21	97.98	97.73	97.48
	6	98.69	98.49	98.27	98.03	97.77	97.49	97.21	96.91
	7	98.38	98.14	97.88	97.59	97.29	96.98	96.65	96.32
	8	98.03	97.75	97.45	97.13	96.79	96.44	96.07	95.70
	9	97.65	97.33	96.98	96.62	96.25	95.86	95.46	95.06
	10	97.22	96.87	96.49	96.09	95.68	95.26	94.82	94.39

the two separated ancestors still share genetic material. An ancestor's sequence can be shifted by between one and seven 32-bit blocks with respect to the other. There is only one possible overlapping (a shift by 32-bits) such that the regulatory site (the inhibition site) of one gene matches the regulatory site (the enhancer site) of the other gene. In this case, four out of five protein coding 32-bit blocks are equal and, due to the majority rule, proteins will still be better preserved in the ongoing DD-process than regulatory sites. Thus, the similarity between regulatory sites of the two genes is less affected by overlappings than the similarity between the corresponding proteins. Generally speaking, for partial ancestry the difference in similarities between protein and regulatory site pairs will be slightly increased compared to total ancestry.

It is already intuitively clear that the difference in similarity between proteins and regulatory sites causes a bias toward the creation of ID-12 motifs: if protein p_1 interacts with gene g_1, it is on average more likely (because of the higher similarity between proteins) that another protein p_2 also interacts with g_1 than that p_1 interacts with another gene g_2.

3.2. *Simulation results*

The following simulation results confirm the theoretical analysis previously presented: we analyzed 10,000 ARN genomes each generated by 12 DD-steps under a 1% mutation rate with regard to the similarity between proteins and regulatory sites. Figure 7 shows the average similarities between proteins and between regulatory sites separately for genes with total and partial ancestry. The difference between the average similarities of pairs of proteins and pairs of regulatory sites of total ancestry increases with the phylogenetic distance (first nine columns). The similarity between proteins is over 97% for all nine categories. For the eight phylogenetic distances for which genes with total and partial ancestry exist, the similarities for pairs of proteins and regulatory sites resulting from partial ancestry (columns

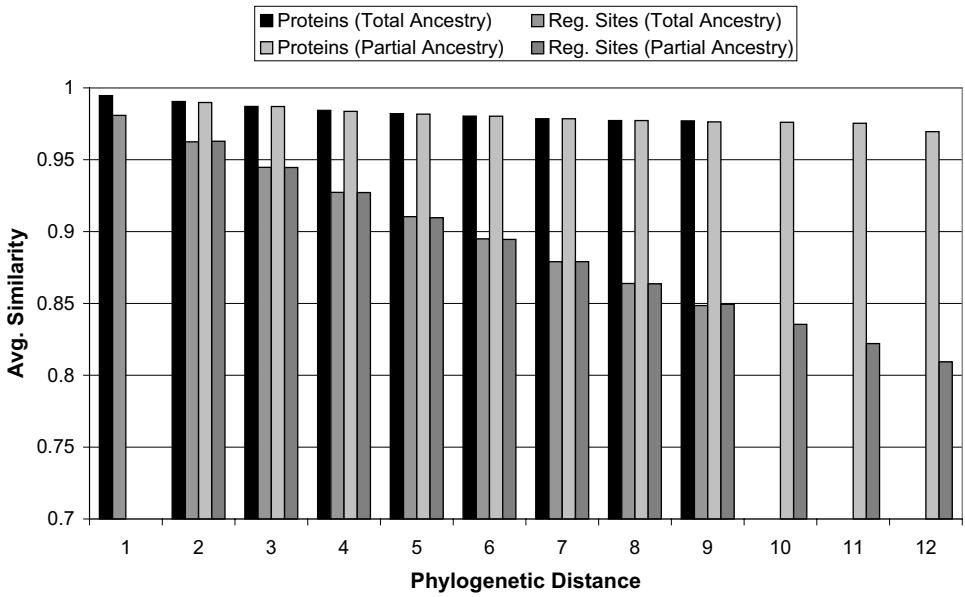

Fig. 7. Average similarity between pairs of proteins/regulatory sites of total (left two bars) and partial ancestry (right two bars) of same phylogenetic distance in genomes created by 12 DD-steps with 1% mutation.

2–12) are almost identical to those resulting from total ancestry. For higher phylogenetic distances ($9 \leq 12$) for gene pairs with partial ancestry, the trend towards larger differences in similarity continues. In other words, more divergence steps lead to a larger difference in the similarities between the robust proteins and the "unprotected" regulatory sites. This is consistent with our analysis.

The average similarity between unrelated proteins and between regulatory sites whose corresponding genes do not have common ancestors is 50%. This is so because the bits being compared originate from two different randomly generated bits in the 32-bit starting sequence. The proportion of the number of gene pairs being in either total or partial ancestry for the different phylogenetic distances is shown in Fig. 8. The sum of the proportions for both ancestries increases rapidly with the phylogenetic distance and hence the number of DD-steps. The largest contribution comes from the category representing gene pairs with partial ancestry and a phylogenetic distance of 12. In total, about 55% of all gene pairs (7.6×10^4 pairs per ARN on average) can be classified by the sort of ancestry and the phylogenetic distance. The remaining 45% are unrelated pairs of genes. It was expected that related pairs occur in significant numbers since they contribute to preferential motif generation.

3.3. *Combinatorics at work*

The duplication and divergence process induces a combinatorial effect on the motif counts, which increases the number of preferential motifs. Combinatorics is at play

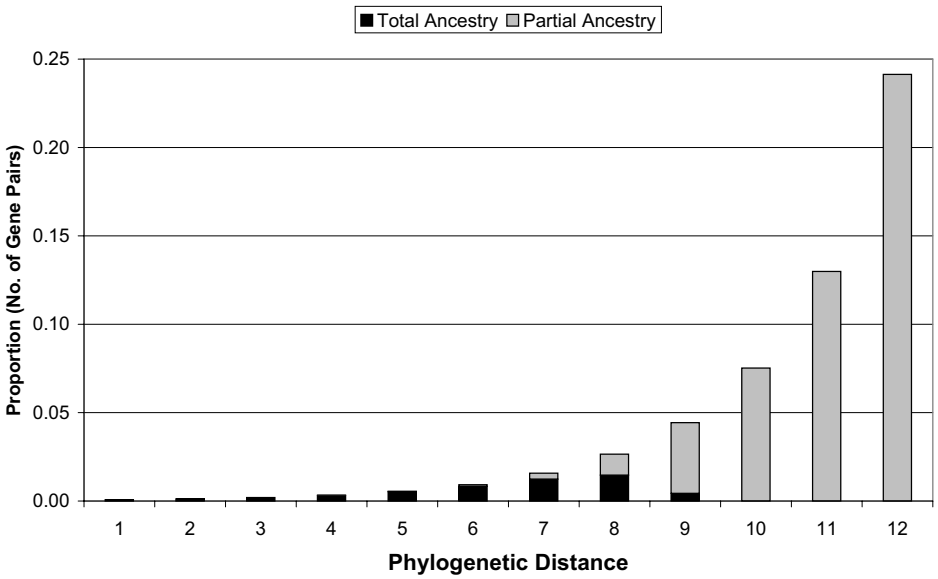

Fig. 8. Proportion of the number of gene pairs with total and partial ancestry according to the phylogenetic distance.

because duplication acts like preferential attachment that tends to create hubs [18, 20]: by connecting a node with a hub, this node, the hub, and *any* third node connected to the hub form a three-node motif which contributes to the motif count. Thus, with every node added to the existing hub-subnetwork, a multitude of new motifs is created.

The combinatorial impact on the preferential motif generation is easily understood when looking at the two different motifs and adding a "sufficiently similar" protein P (cf. Fig. 9). Here, sufficient means that P is similar enough to one or more proteins in the motifs to establish identical regulatory interactions. For an ID-0 motif — protein P_1 interacting with two genes G_1 and G_2 — the additional protein interacts with G_1 and G_2 as well and *one* new ID-0 motif is created. However, when added to an ID-12 motif — two proteins P_1 and P_2 interacting with gene G_1 — protein P would generate *two* new ID-12 motifs, one in combination with each of the two proteins P_1 and P_2. With every additional sufficiently similar protein, the number of ID-0 motifs increases by one while the number of ID-12 motifs increases by the number of proteins already interacting with the corresponding gene (assuming that there are no other gene–protein interactions preventing the creation of the motif). Hence, when a single input node becomes a hub, the number of ID-12 motifs multiplies rapidly by sharing edges with other motifs.

Of course, when adding sufficiently similar genes instead of proteins, we would observe the opposite effect, namely, that instead of ID-12 motifs, ID-0 motifs are produced in larger numbers. But we already know from our analysis above that it is much more likely that another protein will match the same gene than that another

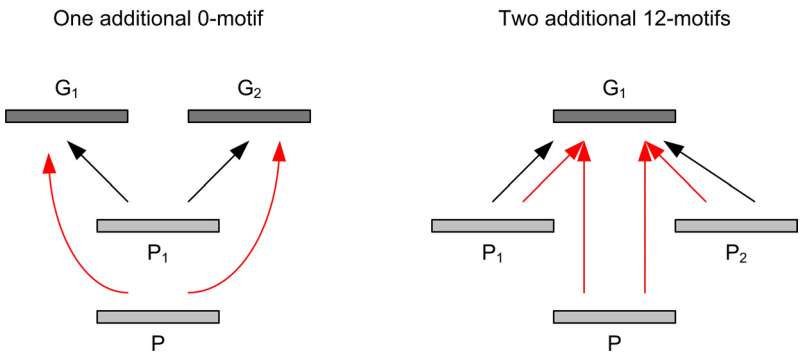

Fig. 9. Left: adding a protein P which is "sufficiently similar" to protein P_1 participating in the given ID-0 motif creates one additional ID-0 motif. Right: adding a protein P which is "sufficiently similar" to protein P_1 (and therefore also to P_2) participating in the given ID-12 motif creates two additional ID-12 motifs.

gene will be matched by the same protein. Therefore, due to combinatorics, the number of ID-12 motifs will increase. In summary, it can be stated that combinatorics amplifies the effect of preferential network motif generation. By counting motifs with the same hub (e.g. several proteins matching to a gene's regulatory site) as one motif, we might deduce the specific effects of combinatorics on preferential network motif generation.

Simulations showed that mainly hubs and not multiple occurrences of single-motifs were found in the subgraph counts. This also corresponds to assumptions formulated in Ref. 21 according to which network motifs in a regulatory network of *Saccharomyces cerevisiae* are not isolated but concentrated, thereby sharing edges and nodes.

4. Conclusions

In this contribution, explanations for the observation of preferential network motif generation in the ARN model created by a whole genome duplication and divergence process have been investigated. It was found that a protein generation mechanism based on the majority rule introduces robustness into the DD-process: mutations in the genome do not necessarily lead to changes in the resulting protein and, in fact, the probability for such a change is rather small. The difference in preservation of proteins on the one hand and regulatory sites on the other by the duplication and divergence mechanism results in preferential generation of motifs ID-12 and ID-22 rather than ID-0. The larger the number of DD-steps the stronger the robustness of the majority rule mechanism and subsequent preferential motif generation. Due to the method in which those motifs are counted combinatorics becomes effective and amplifies the number of preferential motifs.

The detected similarity between subgraph distributions in both GRNs and ARNs inevitably poses the question: Is there an analog to this artificial mechanism

in natural systems? This would require a protein generation mechanism more tolerant of mutations in the genetic exon sequences than in regulatory sites (transcription factor binding sites). Mechanisms do exist in nature that would lead to the same effects found here. For instance, the degeneracy of the codon-to-amino acid mapping performs a similar function. Changes in the constituents of the triplet codon do not always lead to the generation of a different amino acid. Such degeneracy provides a form of robustness to the system analogous to the majority rule of the ARN model. While proteins often contain critical regions essential to function and hence are conserved, an estimated 20–30% of human proteins have coding sequences which have variants in sequence within the population. However, such changes often have little or no effect on function. In fact proteins with vastly different sequences can share the same function and a common evolutionary history [17]. In particular, proteins which carry out the same function across distantly related species can vary considerably in terms of size and amino acid sequence. This could be argued to be a second source of robustness for proteins.

An argument can also be made that specific network motifs are favored due to an associated functional advantage. For example, Ref. 25 found that network motifs producing a higher reliability of information processing by suppressing effects of fluctuations occur significantly more often in natural than in random networks. Therefore, it is suggested that the topology of biological networks is shaped by the selective advantage of a robust dynamics ensuring a reproducible behavior in the presence of noise. However, the mechanisms creating network topology such as the one presented here surely have some influence on network organization. Such mechanisms can make it more likely that evolution can select such functional network motifs. In fact, it could be argued that such mechanisms litter the evolutionary history of organisms for the very reason that their occurrence often leads to beneficial functional changes. Put another way, mechanisms such as whole genome duplication shape networks not solely through topological means but because such changes also lead to functional advantages by biasing the way in which evolution samples the possible space of network configurations.

If the results of this analysis hold true for natural systems, we can infer that after a duplication event the paralogous transcription factor binding site is less likely to be preserved than the corresponding paralogous protein. In other words, subsequent divergence after duplication would first be effected through loss of a binding site (this creates ID-22). This might be reasonable since it can be postulated that the paralogous protein might be more conserved because it has multiple functions. Indeed, evidence is mounting in the molecular biology community that supports this point of view [6, 26, 31].

Acknowledgments

W. Banzhaf acknowledges financial support by NSERC under Discovery Grant RGPIN 283304-04. P. D. Kuo acknowledges support through NSERC PGS D3-317560-2005.

Appendix A. Analysis Details

A.1. *Probability for protein bit flips in 5-bit patterns*

Let P_n be the probability for a flip in a protein bit after divergence (as part of a DD-step) where n is the maximal number of equal bits (either 0 or 1) in the 5-bit gene pattern from which the protein bit is calculated before divergence happens. Irrespective of n, for each 5-bit pattern exist 16 possible mutations leading to patterns with the opposite majority rule outcome, thus causing a bit flip in the resulting protein bit (e.g. 00011 → 011010). P_n can be readily calculated by adding up the probabilities that these mutations occur:

$$P_3 = 6 * p^5 - 15 * p^4 + 16 * p^3 - 9 * p^2 + 3 * p,$$

$$P_4 = -6 * p^5 + 15 * p^4 - 14 * p^3 + 6 * p^2,$$

$$P_5 = 6 * p^5 - 15 * p^4 + 10 * p^3,$$

with p being the probability for a single bit mutation. For $p = 0.01$, we obtain $P_5 \approx 0.001\%$, $P_4 \approx 0.059\%$ and $P_3 \approx 2.912\%$.

A.2. *Pattern distribution and bit flip probabilities*

Since the probability for a bit-flip obviously depends on the 5-bit pattern for which it is supposed to occur, we must also consider the distribution of patterns, which is strongly shaped by duplication and divergence. The probabilities for a certain maximum number of equal bits in a 5-bit pattern are calculated simply by evaluating all possibilities for 5-bit patterns to occur after a certain number of DD-steps N_b. Three DD-steps are required until 5-bit patterns (in a 32 bit interval) exist. Each further DD-step merely results in mutations of the five bits. These probabilities and the overall probabilities for a bit flip in the protein after $N_b + 1$ DD-steps are shown in Table 3.

A.3. *Probability for two related genome bits to be equal*

To calculate the probability that two bits are equal (XX or YY, $Y = \text{NOT } X$) when derived by N_a DD-steps (the phylogenetic distance) from their latest common ancestor bit X, we form the transition matrix

$$\mathbf{T} = \left(\begin{array}{c|cccc} & XX & XY & YX & YY \\ \hline XX & (1-p)^2 & (1-p)p & (1-p)p & p^2 \\ XY & (1-p)p & (1-p)^2 & p^2 & (1-p)p \\ YX & (1-p)p & p^2 & (1-p)^2 & (1-p)p \\ YY & p^2 & (1-p)p & (1-p)p & (1-p)^2 \end{array}\right)$$

Table 3. Columns 2–4 (labeled by the number of equal bits b) list the probabilities (in %) rounded to two decimal places that a 5-bit pattern belongs to one of three classes assuming N_b DD-steps have occured. Given a pattern with five, four or three equal bits, the last row contains the probability of a bit flip in the majority rule outcome after another DD-step $G_{N_b} \rightarrow G_{N_b+1}$ (cf. Appendix A.1). The fifth column shows the overall probability P_{bf} (in %) for a bit flip in the protein assuming that $N_b + 1$ DD-steps have already occurred. The last column specifies the similarities S between the two protein bits (before and after the final DD-step).

N_b	$b = 5$	$b = 4$	$b = 3$	P_{bf}	S
3	89.57	7.35	3.08	0.09	99.91
4	85.25	11.35	3.39	0.11	99.89
5	81.18	14.96	3.85	0.12	99.88
6	77.35	18.21	4.44	0.14	99.86
7	73.73	21.13	5.14	0.16	99.84
8	70.32	23.74	5.94	0.19	99.81
9	67.11	26.08	6.81	0.21	99.79
10	64.07	28.17	7.76	0.24	99.76
11	61.20	30.03	8.77	0.27	99.73
12	58.49	31.68	9.83	0.31	99.69
	0.001	0.059	2.912		

with p being the mutation probability. The probabilities for the four different states of the two bits after N_a DD-steps can be read off in the first column of $\mathbf{T}^{N_a}$ (in the first DD step the two copies of X are generated and mutated; the probabilities for the four outcomes XX, XY, YX and YY correspond to the first column of $\mathbf{T}$). Table 1 in Sec. 3 lists those probabilities for the phylogenetic distances $N_a = 1 \cdots 10$.

A.4. *Probability for two related protein bits to be equal*

Here, we calculate the probabilities that related bits in two proteins (the corresponding genes have common ancestors) with a phylogenetic distance N_a are identical. Let $\mathbf{T}_{p2p}$ be the matrix containing the probabilities for each 5-bit pattern to mutate into another 5-bit pattern by a single DD-step. Table 4 shows the matrix coefficients. The tensor square of this matrix, $\mathbf{T}_{pp2pp} = \mathbf{T}_{p2p} \otimes \mathbf{T}_{p2p}$, gives us the probabilities for every combination of two types (according to the maximal number of equal bits and the majority rule outcome) of 5-bit patterns. This matrix is raised

Table 4. Probabilities (in % and rounded to three decimal places) for divergence of 5-bit patterns. The patterns are classified by the number of equal bits, b, and the majority rule outcome p. The matrix is non-symmetrical and only columns but not rows add up to 1. The column specifies the pattern before divergence, the row the pattern after divergence.

$b;\, p$	5; 0	4; 0	3; 0	5; 1	4; 1	3; 1
5; 0	95.099	0.961	0.010	0.000	0.000	0.000
4; 0	4.803	95.138	1.921	0.000	0.000	0.029
3; 0	0.097	3.843	95.157	0.001	0.058	2.882
5; 1	0.000	0.000	0.000	95.099	0.961	0.010
4; 1	0.000	0.000	0.029	4.803	95.138	1.921
3; 1	0.001	0.058	2.882	0.097	3.843	95.157

to the N_a-th power to calculate the probabilities for the case that N_a DD-steps are carried out after the ancestral gene is build in the process of duplication and divergence. From this matrix, we extract the probabilities for the two protein bits to be equal.

We can now calculate the probability that two protein bits, both derived by N_a DD-steps from an ancestral 5-bit gene pattern which was created after N_b DD-steps from the original 32-bit sequence, are equal. Therefore, we just multiply the probabilities extracted from $\mathbf{T}_{pp2pp}^{N_a}$ with those from Table 3 (cf. Appendix A.2) specifying the probabilities that the ancestral pattern has a certain pattern type (with five, four or three equal bits) after $N_b = 3 \cdots 10$ DD-steps.

References

[1] Babu, M., Luscombe, N., Aravind, L., Gerstein, M. and Teichmann, S. A., Structure and evolution of transcriptional regulatory networks, *Curr. Opin. Struc. Biol.* **14** (2004) 283–292.

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

[3] Banzhaf, W., Artificial regulatory networks and genetic programming, in *Genetic Programming Theory and Practice*, eds. Riolo, R. L. and Worzel, B. (Kluwer, 2003), pp. 43–62.

[4] Banzhaf, W., On the dynamics of an artificial regulatory network, in *Advances in Artificial Life — Proc. 7th European Conf. Artificial Life* (*ECAL*), eds. Banzhaf, W., Christaller, T., Dittrich, P., Kim, J. T. and Ziegler, J., Lecture Notes in Artificial Intelligence, Vol. 2801 (Springer, 2003), pp. 217–227.

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

[6] Carroll, S. B., Grenier, J. K. and Weatherbee, S. D., *From DNA to Diversity*, 2nd edn. (Blackwell Publishing, 2005).

[7] Conant, G. and Wagner, A., Convergent evolution of gene circuits, *Nature Genet.* **34** (2003) 264–266.

[8] Dujon, B. *et al.*, Genome evolution in yeasts, *Nature* **430**(6995) (2004) 35–44.

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

[10] Gibson, T. J. and Spring, J., Evidence in favour of ancient octaploidy in the vertebrate genome, *Biochem. Soc. Trans.* **128** (2000) 259–264.

[11] Holland, P. H. W., Gene duplication: Past present and future, *Semin. Cell. Dev. Biol.* **10** (1999) 541–547.

[12] Hughes, A. L., Gene duplication and the origin of novel proteins, *Proc. Nat. Acad. Sci. (USA)* **102**(25) (2005) 8791–8792.

[13] Ihmels, J. *et al.*, Rewiring of the yeast transcriptional network through the evolution of motif usage, *Science* **309**(5736) (2005) 938–940.

[14] Kashtan, N., Itzkovitz, S., Milo, R. and Alon, U., Topological generalizations of network motifs, *Phys. Rev. E* **70** (2004) 031909.

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

[16] Klemm, K. and Bornholdt, S., Topology of biological networks and reliability of information processing, *Proc. Nat. Acad. Sci. (USA)* **102** (2005) 18414–18419.

[17] Koehl, P., Protein structure similarities, *Curr. Opin. Struct. Biol.* **11**(3) (2001) 348–353.
[18] Kuo, P. D. and Banzhaf, W., Scale-free and small world network topologies in an artificial regulatory network model, in *9th Int. Conf. Simulation and Synthesis of Living Systems (ALIFE)*, ed. Pollack, J. (MIT Press, 2004), pp. 404–409.
[19] Kuo, P. D., Leier, A. and Banzhaf, W., Evolving dynamics in an artificial regulatory network model, in *Proc. 8th Conf. Parallel Problem Solving from Nature (PPSN)*, eds. Yao, X. *et al.*, Lecture Notes in Computer Science, Vol. 3242 (Springer, 2004), pp. 571–580.
[20] Kuo, P. D., Banzhaf, W. and Leier, A., Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence, *Biosystems* (2006), in press.
[21] Mazurie, A., Bottani, S. and Vergassola, M., An evolutionary and functional assessment of regulatory network motifs, *Genome Biol.* **6**(4) (2005) R35.
[22] Milo, R. *et al.*, Superfamilies of evolved and designed networks, *Nature* **303** (2004) 1538–1542.
[23] Milo, R. *et al.*, Network motifs: Simple building blocks of complex networks, *Science* **298** (2002) 824–827.
[24] Nadeau, J. and Sankoff, D., Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution, *Genetics* **147** (1997) 1259–1266.
[25] Ohno, S., *Evolution by Gene Duplication* (Springer, 1970).
[26] Rodriguez-Trelles, F., Tarrio, R. and Ayala, F. J., Evolution of cis-regulatory regions versus codifying regions, *Int. J. Dev. Biol.* **47** (2003) 665–673.
[27] Sankoff, D., Gene and genome duplication, *Curr. Opin. Genet. Dev.* **11** (2001) 681–684.
[28] Shen-Or, S., Milo, R., Mangan, S. and Alon, U., Network motifs in the transcriptional regulation network of *Escherichia coli*, *Nature Genet.* **31** (2002) 64–68.
[29] Sidow, A., Gen(om)e duplication in the evolution of early vertrebrates, *Curr. Opin. Genet. Dev.* **6** (1996) 715–722.
[30] Stellwag, E. J., Are genome evolution, organism complexity and species diversity linked?, *Integr. Comp. Biol.* **44** (2004) 358–365.
[31] Stone, J. R. and Wray, G. A., Rapid evolution of cis-regulatory sequences via local point mutations, *Mol. Biol. Evol.* **18** (2001) 1764–1770.
[32] Teichmann, S. A. and Babu, M., Gene regulatory network growth by duplication, *Nature Genet.* **36**(5) (2004) 492–496.
[33] Wolfe, K. and Shields, D., Molecular evidence for an ancient duplication of the entire yeast genome, *Nature* **387**(6634) (1997) 708–713.
[34] Wuchty, S., Oltvai, Z. and Barabási, A.-L., Evolutionary conservation of motif constituents in the yeast protein interaction network, *Nature Genet.* **35**(2) (2003) 176–179.