Gene regulatory network growth by duplication

Sarah A Teichmann & M Madan Babu

We are beginning to elucidate transcriptional regulatory networks on a large scale[1] and to understand some of the structural principles of these networks[2,3], but the evolutionary mechanisms that form these networks are still mostly unknown. Here we investigate the role of gene duplication in network evolution. Gene duplication is the driving force for creating new genes in genomes: at least 50% of prokaryotic genes[4,5] and over 90% of eukaryotic genes[6] are products of gene duplication. The transcriptional interactions in regulatory networks consist of multiple components, and duplication processes that generate new interactions would need to be more complex. We define possible duplication scenarios and show that they formed the regulatory networks of the prokaryote *Escherichia coli* and the eukaryote *Saccharomyces cerevisiae*. Gene duplication has had a key role in network evolution: more than one-third of known regulatory interactions were inherited from the ancestral transcription factor or target gene after duplication, and roughly one-half of the interactions were gained during divergence after duplication. In addition, we conclude that evolution has been incremental, rather than making entire regulatory circuits or motifs by duplication with inheritance of interactions.

The basic unit of gene regulation consists of a transcription factor, its DNA binding site and the target gene or transcription unit it regulates. This basic unit can be elaborated to form a complex network in two ways: some genes may be regulated by more than one transcription factor, and some transcription factors may control more than one gene. In *E. coli* and yeast, a considerable number of regulatory interactions have been determined and are available in the RegulonDB database[7] and in the data sets in refs. 2 and 3, which we used in this analysis.

We investigated how these networks evolved to form complex systems in which >100 transcription factors regulate several hundred genes. Gene duplication and subsequent divergence is the primary mechanism for the evolution of genomes and complexity[4,5]. The rate and mechanisms of duplication in eukaryotes have been investigated in detail[8]. When new genes evolve by duplication, regulatory interactions in networks can be either conserved or lost during the divergence process. Previous theoretical analyses have addressed this at an abstract level[9–12]. Here, we investigate the role of gene duplication and

determine the extent to which duplicated genes inherit interactions from their ancestors in *E. coli* and yeast.

To find instances of gene duplication, we need to reliably detect homology among genes. We used structural domain assignments from the SUPERFAMILY database[13] to identify homology among the proteins (**Supplementary Methods** online), as this method can capture more distant relationships than sequence comparisons alone[14]. From the domain assignments by the SUPERFAMILY hidden Markov models to the transcription factors, we observed that the DNA-binding domains of *E. coli* and yeast largely come from different families, with only two families in common. Furthermore, comparison of the matches in terms of the domain architecture of the genes indicated that more than one-half of the genes with structural assignments in the *E. coli* and yeast networks are the results of gene duplication (**Table 1**; *E. coli*: (352 + 82) / (500 + 110) = 71%; yeast: (173 + 70) / (277 + 80) = 68%). In this analysis, we considered proteins with the same domain architecture to have arisen from a common ancestor (**Supplementary Methods** online).

Many transcription factors and target genes arose by gene duplication. After the duplication event, the regulatory interaction may be inherited or may be lost. In either case, a new interaction may also be gained during divergence. Taking this into account, we describe the possible mechanisms by which duplications of transcription factor genes, target genes or both might lead to the formation of new interactions in the regulatory network. Then, by inspecting the data currently available, we determined the extent to which each mechanism has contributed to the formation of the regulatory networks of *E. coli* and yeast (**Supplementary Note** and **Supplementary Methods** online).

When duplication of a transcription factor occurs (**Fig. 1a**), the new transcription factor may initially recognize the same binding site and, hence, regulate the same target gene as the original transcription factor. During subsequent divergence, the duplicated transcription factor may continue to regulate the same target genes as its ancestor but respond to a different signal (**Fig. 2a**), or it may recognize a new binding site upstream of some other target gene(s). Investigation of the known network in both organisms[2,3,7] showed that duplication of transcription factor genes followed by inheritance of interaction has contributed considerably to the growth of the regulatory network: more than two-thirds of *E. coli* (77%) and yeast (69%) transcription factors have at least one interaction in common with their duplicates (**Table 1**). This accounts for 128 interactions (10%) in *E. coli* and 188 interactions (22%) in yeast
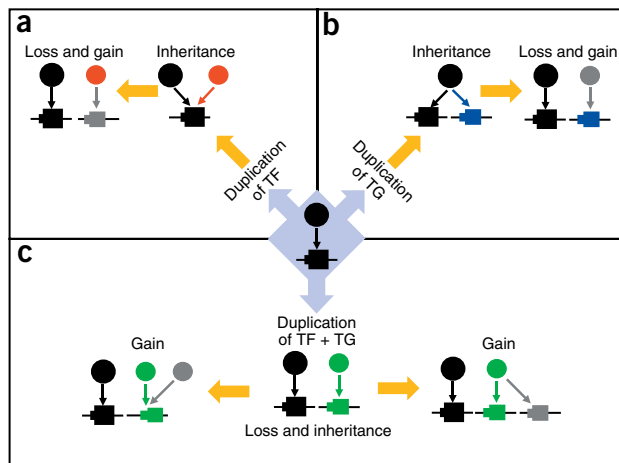
Figure 1 Duplication growth models and consequences for network evolution. The basic unit of gene regulation is shown in the center: the transcription factor (TF), the target gene (TG) and its binding site. The three panels describe the possible duplication events of this basic unit and the subsequent divergence resulting in new regulatory interactions. Duplication events are represented by light blue arrows and divergence events by orange arrows. Divergence may also result in the loss of the duplicated gene, but we consider only duplicated genes that are selected for. (**a**) Duplication of the transcription factor leads to both transcription factors regulating the same gene. Divergence can result in the duplicated transcription factor regulating the original target gene by competing for the same binding site (red arrow, duplication and inheritance of interaction) used by the ancestral transcription factor or regulating a different gene (gray arrow, duplication and gain of interaction). (**b**) Duplication of a target gene results in both genes being regulated by the same transcription factor. Divergence can lead to the duplicated gene remaining under the control of the same transcription factor (blue arrow, duplication and inheritance of interaction) or coming under the control of a different transcription factor (gray arrow, duplication and gain of interaction). (**c**) Duplication of transcription factor and its target genes gives rise to new regulatory interactions. Divergence can result in homologous transcription factors regulating homologous genes (green arrow, duplication and inheritance of interaction). Subsequent divergence of the transcription factor or the target gene can result in additional interactions (gray arrow, duplication and gain of interaction).

(**Fig. 3** and **Table 1**). This fraction is larger in yeast than in *E. coli* because many genes in yeast are regulated by two or more transcription factors, whereas many genes are regulated by only one or two transcription factors in *E. coli* (**Supplementary Note** online). As a rule, larger genomes have more transcription factors per gene[15].

In the second duplication scenario, duplication of the target gene and its upstream region can explain the evolution of new genes along with their regulatory regions (**Fig. 1b**). During divergence, the duplicated target gene may change its coding sequence to carry out a different function but conserve its upstream region, or both the coding sequence and the upstream region may diverge, resulting in recognition by a different transcription factor. The first possibility results in homologous genes being regulated by the same transcription factor[16,17] (**Fig. 2b**), and the latter results in homologous genes being regulated by different transcription factors, which is not uncommon in yeast[18]. Duplication of the target gene with inheritance of interaction contributed to 272 interactions (22%) and 166 interactions (20%) in the *E. coli* and yeast networks, respectively (**Fig. 3** and **Table 1**).

Yeast and *E. coli* show extensive duplication under both duplication scenarios discussed above, meaning that this phenomenon is not biased by prokaryotic horizontal transfer or the operon structure.

So far, we have considered duplications of transcription factors and target genes separately. But a transcription factor and its target gene could both duplicate around the same time (**Fig. 1c**), especially if they were adjacent on a chromosome. Divergence of both the transcription factor and the recognition sites in the DNA could then occur, such that the new transcription factor would regulate only the new target gene, and the old transcription factor would regulate only its original target gene. Though it might seem unlikely, this process can be traced convincingly in some cases (*e.g.*, two sugar catabolism operons in *E. coli*[17]; **Fig. 2c**). There are 74 (6%) and 31 (4%) such interactions in the *E. coli* and yeast networks, respectively (**Fig. 3** and **Table 1**).

Figure 3 and **Tables 1** and **2** provide an overview of the contribution of the different types of regulatory interactions to the entire network. The largest fraction of interactions represents cases in which either the transcription factor or target gene was duplicated, and gained new interactions after duplication during divergence, with or without loss of the original interaction (**Fig. 1**). There are 637 such interactions in *E. coli* (52%) and 365 in yeast (43%; **Fig. 3**). The second largest group of interactions comprises those inherited by transcription factors or target genes after duplication (38% and 45% in *E. coli* and yeast, respectively), and the smallest group comprises
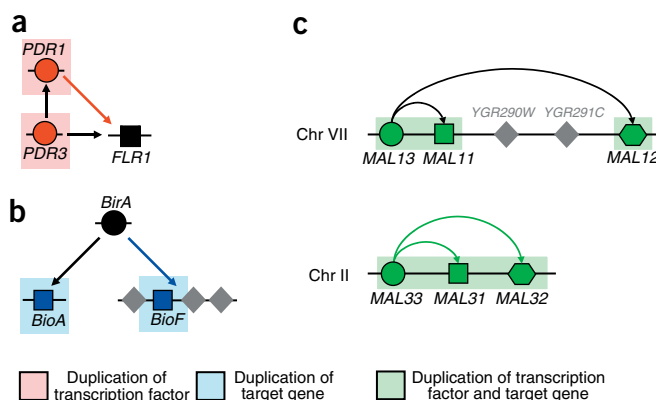

Figure 2 Duplications in the *E. coli* and yeast networks. Transcription factors and target genes that have the same domain architecture are shown as circles and squares with the same color. (**a**) Duplication of transcription factors in a feed-forward motif (FFM) in yeast. The homologous transcription factors *PDR1* and *PDR3* are involved in drug responses and regulate multidrug transporters in yeast. This FFM could have evolved by duplication according to the scheme shown in **Figure 1a**. (**b**) Duplication of target genes in a single input module (SIM) in *E. coli*. The *BioA* and *BioBFCD* operons are regulated by the *BirA* transcription factor only, a topology that is a SIM. *BioA* and *BioF* are homologous enzymes in the biotin biosynthesis pathway, and so this SIM could have evolved by duplication of target genes, as shown in **Figure 1b**. (**c**) Duplication of both a transcription factor and its target genes in yeast. This is an example in which both the transcription factor and target genes were duplicated to produce additional regulatory interactions in the network according to the scheme shown in **Figure 1c**. The simultaneous duplication of a transcription factor and two target genes is facilitated by the fact that the transcription factor and target genes are adjacent to each other on the yeast chromosome.

a Gene regulatory network in *E. coli*

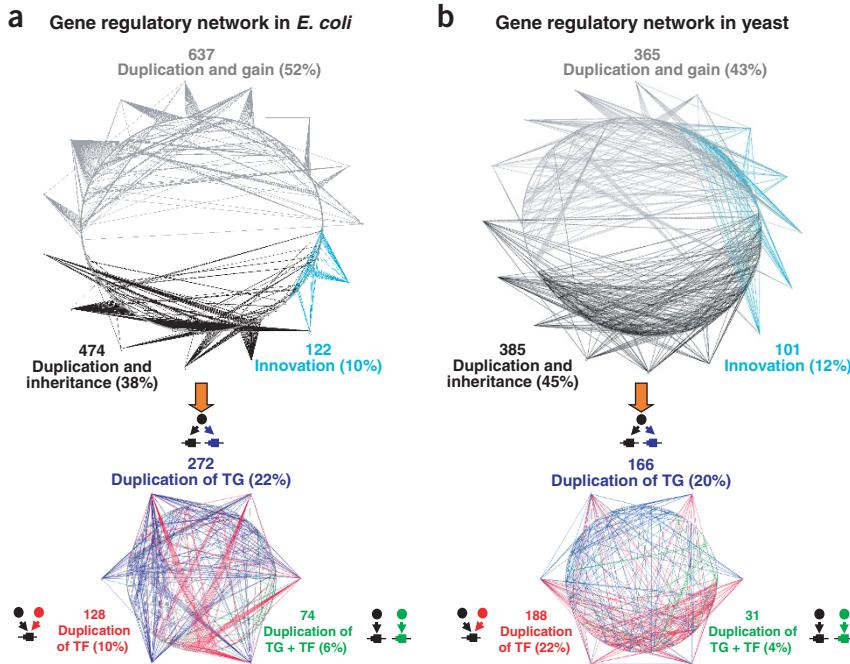**b** Gene regulatory network in yeast

Figure 3 Duplication in the gene regulatory networks in *E. coli* and yeast. In the top panels, known regulatory interactions with information about evolutionary relationships are depicted for (**a**) *E. coli* (1,233 interactions) and (**b**) yeast (851 interactions). The nodes on the outside of the circles represent transcription factors that regulate more than 10 target genes. Interactions shown in gray occur between duplicate genes without direct evidence that the interaction was inherited after duplication and thus are new interactions gained during divergence. Interactions shown in turquoise occur between genes that do not have homologs and thus are innovations. For interactions shown in black, there are homologous proteins that have either the same transcription factor or the same target gene in their interactions, as shown in the bottom panels. In the bottom panels, interactions with evidence of duplication and inheritance (shown in black in the top panels) are classified into the three types of duplications shown in **Figure 1**: duplication of the transcription factor, of the target gene and of both these elements. The different types of interaction are given on our website (see URL). This figure was generated using the Osprey network visualization system[30].

interactions that were pure innovations (10% and 12% in *E. coli* and yeast, respectively). In reality, there are probably many more duplications, as the complete network in both organisms is much larger than currently known, and there are many duplicate transcription factors and target genes that have not yet been characterized[17].

We assessed the statistical significance of the occurrence of these events in 10,000 networks with randomly assigned domain architectures (**Table 2**). These events very rarely occur by chance at the frequencies observed. We also assessed the robustness of the duplication levels and their statistical significance by artificially introducing noise

Table 1 Duplications of genes and interactions in *E. coli* and yeast regulatory networks

	E. coli	Yeast
Target genes		
All target genes	718 + 77 TgTfs	368 + 34 TgTfs
Target genes with structural assignments	500 + 70 TgTfs	277 + 27 TgTfs
Target genes with homologs to each other	352 + 55 TgTfs	173 + 21 TgTfs
Target genes with homologs and shared or homologous transcription factors	186 + 47 TgTfs[a]	115 + 10 TgTfs[a]
Percentage of homologous target genes with shared or homologous transcription factors out of all target genes with homologs	57%	64%
Transcription factors		
All transcription factors	121	109
Transcription factors with structural assignments	110	80
Transcription factors with homologs to each other	82	70
Transcription factors with homologs and shared or homologous targets	63[b]	48[b]
Percentage of homologous transcription factors with shared or homologous targets out of all transcription factors with homologs	77%	69%
Regulatory interactions evolved by duplication		
All interactions	1,409	906
Interactions with potential information about homology through structural assignments	1,233	851
Interactions of genes where at least one gene has a homolog in the network	1,111	750
All interactions of homologs with shared or homologous transcription factors or target genes	474	385
Interactions of homologous target genes with shared transcription factors[c]	272	166
Interactions of homologous transcription factors with shared target genes	128	188
Interactions of homologs where whole interaction is duplicated elsewhere	74	31
Percentage of interactions with evidence of duplication out of all interactions with at least one genes with homologs	38%	45%

[a]Of the cases enumerated, 29 target genes and 2 target genes that are also transcription factors (TgTFs) have homologous transcription factors, and the rest have shared transcription factors. In yeast, these numbers are 28 target genes and 3 TgTFs. [b]Of the cases enumerated, 41 of the *E. coli* transcription factors and 9 of the yeast transcription factors have homologous target genes, and the rest have shared target genes. [c]Some of these interactions are of homologs with both shared transcription factors and shared target genes. We count these cases as interactions of homologous target genes with shared transcription factors only.

Table 2 Statistical significance of duplication types with inheritance of interaction compared to random distribution of homologs in network

Duplication type with inheritance of interaction	Yeast			E. coli		
	Number of genes	P value	Z score	Number of genes	P value	Z score
Target genes[a]	94	<10⁻⁴	7.9	173	<10⁻⁴	4.7
Transcription factors	39	0.01	2.4	22	0.007	2.7
Transcription factors and target genes[b]	40	0.008	−2.4	79	0.04	−1.7

[a]These include some transcription factors as in **Table 1**. [b]These are only genes that do not qualify for the other two duplication types.

into the yeast regulatory network (**Supplementary Methods** online). The significance barely changed with the introduction of 5% noise but fluctuated with the introduction of 10%, 20% and 30% noise. Because we did not use results from large-scale experiments or computational predictions, the rates of false positives and negatives are probably low in our data sets.

We next asked whether duplication patterns are linked to the topology, or structure, of the networks. A number of topological features are common to the gene regulatory networks in *E. coli* and yeast[2,3]. A key common feature is that the number of target genes per transcription factor roughly obeys a power law, which is typical of 'scale-free' networks[19] (**Fig. 4a** and **Supplementary Note** online). Given the power-law distribution of target genes per transcription factor as a topological characteristic and the importance of target gene duplication as an evolutionary feature of the network, we asked whether the two are linked. If transcription factors with many target genes have a particularly high proportion of homologous genes as their targets, then the scale-free topology of the network can be ascribed, at least in part, to target gene duplications.

In both organisms, there were transcription factors with homologous target genes ranging from only two to many (**Fig. 4b,c**). There was no marked tendency for transcription factors with more target genes to have a larger fraction of homologous target genes. We found that in *E. coli* and yeast, the duplication levels were significant in 7 and 14 transcription factors, respectively (**Fig. 4b,c**). These transcription factors regulate different numbers of target genes and not just large numbers of

genes. These findings show that the power-law distribution of target genes per transcription factor is not purely a consequence of duplication and inheritance of interactions of target genes.

Different types of networks have over-represented topological elements. These are sets of interactions connected in specific patterns called 'network motifs'[1,2,20]. These motifs have been engineered artificially[21,22], but here we addressed how they were formed during evolution.

The first of the two patterns studied, the feed-forward motif (FFM), features a general transcription factor that regulates a target gene and a specific transcription factor that also regulates the target gene (**Fig. 2a**). This motif could theoretically evolve by duplication of one of the two transcription factors (**Supplementary Note** online). But none of the *E. coli* FFMs and only two pairs of transcription factors and one group of three transcription factors involved in more than one-third of the yeast FFMs can be explained this way. The second pattern, called the single input module (SIM), consists of a single transcription factor that alone regulates a group of genes (**Fig. 2b**). A SIM could evolve by duplication of target genes (**Supplementary Note** online), but target gene duplication does not occur more frequently in SIMs than in the entire network.

Our results show that none of the motifs were formed by duplication of an entire ancestral motif, similar to previous results[23] using a different data set and a different method of detecting homology. Though many of the genes and interactions in network motifs evolved by duplication, the topologies themselves are not direct products of duplication with inheritance. The reasons why these topologies are favorable are beginning to be elucidated experimentally[24,25].

In conclusion, we quantified the mechanisms of network evolution for the known gene regulatory networks of *E. coli* and yeast, two distinct networks with different protein families and topologies. In both organisms, only a small fraction (~10%) of the interactions evolved by innovation, consisting of transcription factors and target genes without homologs. Almost 90% of the interactions evolved by duplication of either a transcription factor or a target gene: roughly one-half of these interactions evolved by duplication with inheritance of interaction, and the other half by duplication with gain of new interactions. These duplications are incremental rather than modular duplications of entire motifs or regulatory circuits. Our quantification of these mechanisms has implications for artificial network evolution and design.
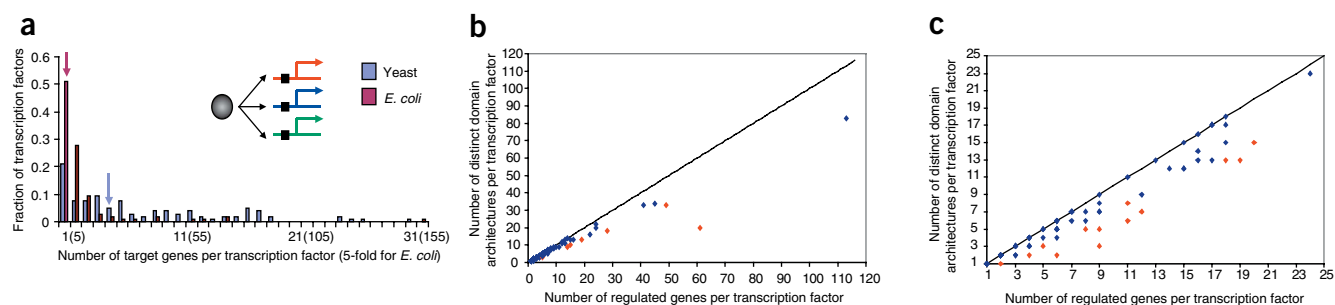

Figure 4 Target gene duplications for all *E. coli* and yeast transcription factors. (**a**) The fraction of transcription factors regulating a certain number or range of numbers of target genes is shown. Both the yeast and *E. coli* distributions are roughly power laws, and the scale is 5-fold greater for *E. coli* than yeast. The median value for each organism is marked and represents five target genes per transcription factor. (**b,c**) The number of target genes with structural assignments versus the number of distinct domain architecture families of these target genes for each transcription factor for *E. coli* (**b**) and yeast (**c**). Groups of target genes without any duplicates are on the diagonal. If there are two target genes with the same domain architecture (*i.e.*, if one of the genes evolved by a duplication event from the other), the *y* value for the transcription factor decreases by 1 and the point moves one step below the diagonal. Each further duplication event moves the transcription factor's point further below the diagonal. There are off-diagonal points for transcription factors with large and small numbers of target genes, without a marked increase for more influential transcription factors. Transcription factors with statistically significant numbers of target genes are marked in red, and these are also found across a range of target gene numbers. This shows that duplication of target genes is not the driving force for the power-law pattern of target genes per transcription factor, which is shown in **a**.

METHODS

Gene regulatory networks and motifs. We took the set of regulatory interactions for *E. coli* from the data set in ref. 2, which uses the information available in the RegulonDB database[7] and provides new interactions compiled from the literature. There were 1,409 regulatory interactions involving 121 transcription factors and 795 target genes. We found 42 FFMs and 30 SIMs in this network. We took the transcription factors and their target genes in yeast from the data set in ref. 3, which consisted of 906 interactions involving 109 transcription factors and 402 target genes. There are 131 FFMs and 29 SIMs in this network. The large number of FFMs in yeast reflects the extensive transcription factor inter-regulation in the eukaryote compared with the prokaryote. Details on this are provided in **Supplementary Note** online.

Identification of duplicated genes. Detecting homology among distant paralogous proteins in an organism is a difficult task because of sequence divergence. But it is well known that the structure of a protein is more conserved than its sequence. Thus, to reliably detect distant relationships among *E. coli* and yeast proteins, we used three-dimensional structural domain assignments of the proteins in the network as a measure of homology. If two proteins had the same domain architecture, or a series of domains from the same protein families, we assumed that they were derived from the same common ancestor, as supported by analysis of protein structures[26] and sequences[27].

We obtained domain architectures from the domain assignments in the SUPERFAMILY database[13] (version 1.61) for the protein sequences in the yeast and *E. coli* genomes. Evolutionary information about domains is inherent in the classification scheme of the SCOP database[28], and the hidden Markov models of the SUPERFAMILY database are based on these domains.

We considered domain architectures that differed only by gaps or repeats of domains to be homologous, as repeats are sometimes missed by the structural assignment method. When compared with sequence clusters found by FASTA[29] of whole sequences (E value ≤ 0.01 in a large database, match over 80% sequence), our method of comparing domain architectures never split sequence clusters. Several sequence clusters had the same domain architecture, however. To illustrate the coverage of the method, 48% of all yeast proteins in the genome had a domain assignment, whereas only ~5% can be clustered by FASTA in the manner described above.

If there was a domain assignment for only one protein in a transcription factor–regulated gene pair, we could trace duplication only if the pair was embedded in a suitable network topology. For instance, if a transcription factor lacked a domain assignment but regulated two genes that are homologous, we could still trace the evolution of such interactions (**Fig. 2b**).

Identification of duplicated edges and simulation procedure. We assessed the significance of the shared interactions among homologs by comparison with a scenario in which the domain architectures were randomly shuffled across proteins. We simulated this by retaining the topology of the real network and randomly shuffling domain architectures among those nodes with domain architecture information. We shuffled the transcription factors separately from target genes. We carried out the simulation 10,000 times, and each time we calculated the numbers of homologous transcription factors with shared targets and of homologous target genes with shared transcription factors. The fraction of homologs with shared interactions was never as high as that observed in the real network in all 10,000 iterations of the calculation (**Supplementary Methods** online).

URL. Information on the data set used and structural assignments is available at http://www.mrc-lmb.cam.ac.uk/genomes/madanm/net_evol/.

Note: Supplementary information is available on the Nature Genetics website.

1. Lee, T.I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804 (2002).
2. Shen-Orr, S.S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* **31**, 64–68 (2002).
3. Guelzim, N., Bottani, S., Bourgine, P. & Kepes, F. Topological and causal structure of the yeast transcriptional regulatory network. *Nat. Genet.* **31**, 60–63 (2002).
4. Brenner, S.E., Hubbard, T., Murzin, A. & Chothia, C. Gene duplications in *H. influenzae*. *Nature* **378**, 140 (1995).
5. Teichmann, S.A., Park, J. & Chothia, C. Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements. *Proc. Natl. Acad. Sci. USA* **95**, 14658–14663 (1998).
6. Gough, J., Karplus, K., Hughey, R. & Chothia, C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* **313**, 903–919 (2001).
7. Salgado, H. *et al.* RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.* **29**, 72–74 (2001).
8. Lynch, M. & Conery, J.S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000)
9. Wagner, A. Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization. *Proc. Natl. Acad. Sci. USA* **91**, 4387–4391 (1994).
10. Bhan, A., Galas, D.J. & Dewey, T.G. A duplication growth model of gene expression networks. *Bioinformatics* **18**, 1486–1493 (2002).
11. Vázquez, A., Flammini, A., Maritan, A. & Vespignani, A. Modeling of protein interaction networks. *Complexus* **21**, 38–44 (2003).
12. Solé, R.V., Pastor-Satorras, R., Smith, E. & Kepler, T.B. A model of large-scale proteome evolution. *Adv. Complex Systems* **5**, 43–54 (2002).
13. Gough, J. & Chothia, C. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.* **30**, 268–272 (2002).
14. Madera, M. & Gough, J. A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res.* **30**, 4321–4328 (2002).
15. van Nimwegen, E. Scaling laws in the functional content of genomes. *Trends Genet.* **19**, 479–484 (2003)
16. Rajewsky, N., Socci, N.D., Zaptocky, M. & Siggia, E.D. The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparison. *Genome Res.* **12**, 298–308 (2002).
17. Madan Babu, M. & Teichmann, S.A. Evolution of transcription factors and the gene regulatory network in *E. coli*. *Nucleic Acids Res.* **31**, 1234–1244 (2003).
18. Papp, B., Pal, C.Y. & Hurst, L.D. Evolution of cis-regulatory elements in duplicated genes of yeast. *Trends Genet.* **19**, 417–422 (2003).
19. Albert, R. & Barabasi, A.L. Statistical mechanics of complex networks. *Reviews Modern Phys.* **74**, 47–97 (2002).
20. Milo, R. *et al.* Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827 (2002).
21. Guet, C.C., Elowitz, M.B., Hsing, W. & Leibler, S. Combinatorial synthesis of genetic networks. *Science* **296**, 1466–1470 (2002).
22. Yokobayashi, Y., Weiss, R. & Arnold, F.H. Directed evolution of a genetic circuit. *Proc. Natl. Acad. Sci. USA* **99**, 16587–16591 (2002).
23. Conant, G.C. & Wagner, A. Convergent evolution of gene circuits. *Nat. Genet.* **34**, 264–266 (2003).
24. Becskei, A. & Serrano, L. Engineering stability in gene networks by autoregulation. *Nature* **405**, 590–593 (2000).
25. Gardner T.S., Cantor C.R. & Collins, J.J. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* **403**, 339–342 (2000).
26. Bashton, M. & Chothia, C. The geometry of domain combination in proteins. *J. Mol. Biol.* **315**, 927–939 (2002).
27. Apic, G., Gough, J. & Teichmann, S.A. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.* **310**, 311–325 (2001).
28. Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
29. Pearson, W.R. & Lipman, D.J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448 (1988).
30. Breitkreutz, B.J., Stark, C. & Tyers, M. Osprey: a network visualization system. *Genome Biol.* **4**, R22 (2003).