Structure and evolution of transcriptional regulatory networks

M Madan Babu[1,2*], Nicholas M Luscombe[3*], L Aravind[4], Mark Gerstein[3] and Sarah A Teichmann[1,5]

The regulatory interactions between transcription factors and their target genes can be conceptualised as a directed graph. At a global level, these regulatory networks display a scale-free topology, indicating the presence of regulatory hubs. At a local level, substructures such as motifs and modules can be discerned in these networks. Despite the general organisational similarity of networks across the phylogenetic spectrum, there are interesting qualitative differences among the network components, such as the transcription factors. Although the DNA-binding domains of the transcription factors encoded by a given organism are drawn from a small set of ancient conserved superfamilies, their relative abundance often shows dramatic variation among different phylogenetic groups. Large portions of these networks appear to have evolved through extensive duplication of transcription factors and targets, often with inheritance of regulatory interactions from the ancestral gene. Interactions are conserved to varying degrees among genomes. Insights from the structure and evolution of these networks can be translated into predictions and used for engineering of the regulatory networks of different organisms.

Addresses
[1]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK
[2]e-mail: madanm@mrc-lmb.cam.ac.uk
[3]Department of Molecular Biophysics and Biochemistry, Yale University, PO Box 208114, New Haven, Connecticut 06520-8114, USA
[4]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA
[5]e-mail: sat@mrc-lmb.cam.ac.uk
*These authors contributed equally to this work

Current Opinion in Structural Biology 2004, **14**:283–291

This review comes from a themed issue on
Sequences and topology
Edited by Peer Bork and Christine A Orengo

Available online 19th May 2004

0959-440X/$ – see front matter

DOI 10.1016/j.sbi.2004.05.004

Introduction

The biological characteristics of an organism emerge largely as a result of the dynamic interplay between its gene repertoire and the regulatory apparatus. In this review, we discuss our current understanding of the structural organisation of transcriptional regulatory systems from a network perspective and their possible evolutionary histories.

Structure of the transcriptional regulatory network

The assembly of regulatory interactions linking transcription factors to their target genes in an organism can be viewed as a directed graph, in which the regulators and targets represent the nodes, and the regulatory interactions are the edges (Figure 1). This resulting network is a complex, multilayered system that can be examined at four levels of detail. At the most basic level, the network comprises a collection of transcription factors, downstream target genes and the binding sites in the DNA (Figure 1a). At the second level, these basic units are organised into recurrent patterns of interconnections called network motifs, which appear frequently throughout the network (Figure 1b) [1••,2•]. At the third level, the motifs cluster into semi-independent transcriptional units called modules (Figure 1c). Finally, at the top level, the regulatory network consists of interconnecting interactions among the modules, to build up the entire network (Figure 1d).

It should be noted that much of the work on regulatory networks has focused on *Escherichia coli* and the yeast *Saccharomyces cerevisiae*, for which data are most abundant. The individual regulatory interactions in *E. coli* have been collected manually from the literature in the RegulonDB database [3••]. In yeast, on the other hand, manually curated data [4] have been greatly augmented by the output of large-scale DNA-binding data from chromatin immunoprecipitation-chip (ChIp-chip) experiments [2•,5].

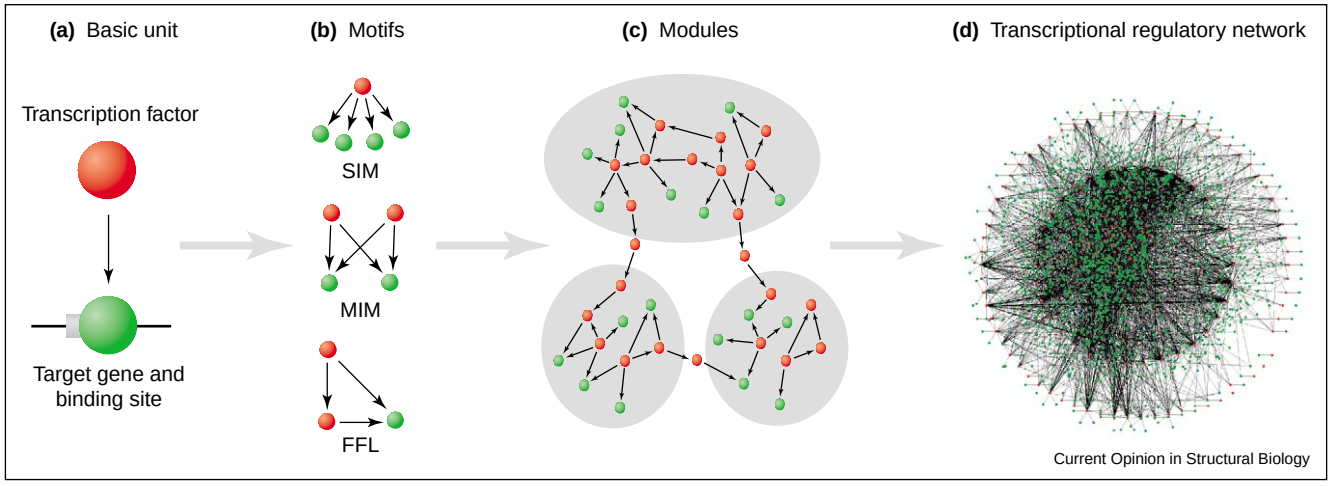
In the following sections, we will survey the different computational studies that have characterised the structural organisation of these regulatory networks.

Motifs

At a local level, the transcriptional network can be broken down into a series of regulatory motifs. These represent the simplest units of network architecture, in which there are specific patterns of inter-regulation between transcription factors and target genes. Motifs do not often represent independent units that are functionally separable from the rest of the network. However, they have been shown theoretically and experimentally to possess particular kinetic properties that determine the temporal program of expression of the target genes [6•].

We show schematics of three prevalent motifs in Figure 1b: single input, multiple input and feed-forward loop motifs. The first two comprise direct-acting motifs,

Figure 1



Structural organisation of transcriptional regulatory networks. **(a)** The 'basic unit' comprises the transcription factor, its target gene with DNA recognition site and the regulatory interaction between them. **(b)** Units are often organised into network 'motifs', which comprise specific patterns of inter-regulation that are over-represented in networks. Examples of motifs include single input (SIM), multiple input (MIM) and feed-forward loop (FFL) motifs. **(c)** Network motifs can be interconnected to form semi-independent 'modules', many of which have been identified by integrating regulatory interaction data with gene expression data, and imposing evolutionary conservation. **(d)** The entire assembly of regulatory interactions constitutes the 'transcriptional regulatory network', which provides the blueprint for regulation of gene expression in an organism.

whereby a single or multiple transcription factors regulate their targets. Yu *et al.* [7•] showed that target genes belonging to the same single and multiple input motifs tend to be co-expressed, and that the level of co-expression is higher when multiple transcription factors are involved. The feed-forward loop is composed of two transcription factors, whereby the first regulates the second and both regulate a final target gene. Further motifs identified by Lee *et al.* [2•] in yeast represent patterns of interconnections of variable complexity, such as the autoregulatory and regulatory chain motifs.

Modules

The organisation of the regulatory network can also be captured at an intermediate level by examining its modularity. Intuitively, one might expect distinct cellular processes to be conveniently regulated by discrete and separable modules. Indeed, Guelzim *et al.* [8•] reported the global fragmentation of the regulatory network in yeast. The clustering coefficient — a measure of the propensity for nodes to form 'cliques' — was fivefold higher than would be expected for a random network.

There have been several different approaches to identifying modules and these studies have provided distinct outcomes with respect to the resulting modules. The main conclusion, however, is that regulatory networks tend to be highly interconnected and very few modules are entirely separable from the rest of the network. In fact, many identified modules are nested within each other in a hierarchical organisation at differing levels of connectivities.

Dobrin *et al.* [9] showed that many of the multiple input and feed-forward loop motifs in *E. coli* overlap, so that they share transcription factors or target genes. Thus, many small, highly connected motifs group into a few larger modules, which in turn integrate into even larger ones. These nested modules are interconnected through local regulatory hubs. Such an organisation combines the capacity for rapid regulatory changes through regulatory hubs with integration of the regulatory processes across several modules.

Other approaches to identifying modules have incorporated further data sources, such as gene expression data sets. Typical analyses have applied clustering algorithms to gene expression data to find sets of co-expressed genes. In one of the original studies by Tavazoie *et al.* [10], it was reported that some of the major co-expression clusters coincided with functional groupings of genes. In an ambitious extension of Teichmann and Babu's [11] work, Stuart *et al.* [12] recently clustered over 3000 microarray experiments on four eukaryotic genomes and identified 22 163 gene pairs whose co-expression is conserved across all organisms. They grouped sets of orthologues into modules, suggesting that co-expression of gene pairs over large evolutionary distances implies a selective advantage for co-regulation, perhaps because the genes are functionally related.

In another interesting study, Ihmels *et al.* [13] added a different perspective by taking the experimental conditions into account when defining the gene clusters. Their 'signature' algorithm identifies clusters according to the

experimental conditions in which the expression patterns of genes are most significantly correlated. The authors identified 86 transcriptional modules and the experimental conditions in which they operate. Segal *et al.* [14] used a probabilistic algorithm to partition gene modules first based on their expression profiles, and then identified specific regulatory genes that are predicted to control the modules by comparing the expression profiles of candidate regulators and the gene modules. They were able to identify 50 different modules with distinct regulatory programs. Particularly illuminating was the formation of higher order groupings by the individual modules, which are regulated by partly overlapping but distinct regulators.

Bar-Joseph *et al.* [15] improved previous algorithms by explicitly linking gene expression data with the regulatory interaction data produced by Lee *et al.* [2•] through ChIp-chip experiments. In this way, the authors were able to partition 655 distinct genes and 68 transcription factors into 106 regulatory modules. Many of the identified modules could be linked to particular cellular processes.

Global network organisation

At a global level, the overall structure or topology of the gene regulatory network can be described by parameters derived from graph theory. The incoming connectivity is the number of transcription factors regulating a target gene, which quantifies the combinatorial effect of gene regulation. A recent study by Guelzim *et al.* [8•] reported that the fraction of target genes with a given incoming connectivity decreases exponentially. The exponential behaviour indicates that most target genes are regulated by similar numbers of factors (93% of genes are regulated by 1–4 factors in yeast) and presumably reflects the molecular limits on the number of transcription factors that can affect a target gene simultaneously, which are imposed by protein and DNA structural constraints at promoters.

The outgoing connectivity, which is the number of target genes regulated by each transcription factor, is distributed according to a power law, contrary to the incoming connectivity parameter. This is indicative of a hub-containing network structure, in which a select few transcription factors participate in the regulation of a disproportionately large number of target genes. These hubs can be viewed as 'global regulators', as opposed to the remaining transcription factors that can be considered 'fine tuners'. Global regulators can be defined based on the number of genes they regulate [1••,16•]. In the transcriptional network in yeast, regulatory hubs have a propensity to be lethal if removed [17]. Martinez-Antonio and Collado-Vides [18•] defined global regulators by taking into account additional factors, such as the number of co-regulators and the number of conditions.

One must bear in mind that the gene regulatory network of an organism is a dynamic entity and different sections of the network will be active under different conditions. Gutierrez-Rios *et al.* [19] have shown that the expression levels of transcription factors and target genes in *E. coli* under different experimental conditions correlate with the known *E. coli* regulatory network. More recently, Luscombe *et al.* (unpublished results) have used yeast expression data sets to extract the active subnetworks in yeast under different conditions. They found that, under conditions in which the cell is responding quickly to a change in external conditions (such as DNA damage, diauxic shift or stress response), the topology of the network is simple, with few cascades of transcription factors. In multistage processes, such as cell cycle or sporulation, the opposite applies, presumably because complex serial regulatory processes are required to drive the cell through sequential phases.

Evolution of the gene regulatory network

So far, we have described the structure of the transcriptional regulatory network at differing levels of detail and complexity. We will now address the evolution of these networks, from the lowest level in terms of the repertoire of transcription factors to the global level of the entire network.

Transcription factor families

Evolutionary relationships among transcription factors can be detected through local alignment sequence searches for close homologues, through sequence profile searches with PSI-BLAST score matrices [20] or hidden Markov models [21] for conserved DNA-binding domains. The latter methodology is particularly useful in identifying very distant relationships, which may elude conventional sequence searches. The demography of transcription factors in the completely sequenced genomes of various organisms can be estimated using assignments of hidden Markov models from the Pfam [22] and SUPERFAMILY [23] databases.

We display estimates of the predicted transcription factors in five genomes in Table 1, ranging from about 300 in *E. coli* to over 3000 in humans. These constitute between 6% (in *E. coli* and yeast) and 8% (in human) of the encoded proteins in these organisms. van Nimwegen's [24•] earlier observation that larger genomes tend to have more transcription factors per gene is in accordance with the trend seen in eukaryotes (Table 1).

The assignment of the DNA-binding domains also allows us to assess the evolutionary relationships among transcription factors. Independent studies in *E. coli* [16•,25], archaea [26], plants and animals [27–29] have consistently demonstrated that the repertoires of transcription factors draw their DNA-binding domains from a relatively small, ancient conserved repertoire.

Table 1

Numbers of DNA-binding transcription factors in five organisms[a].

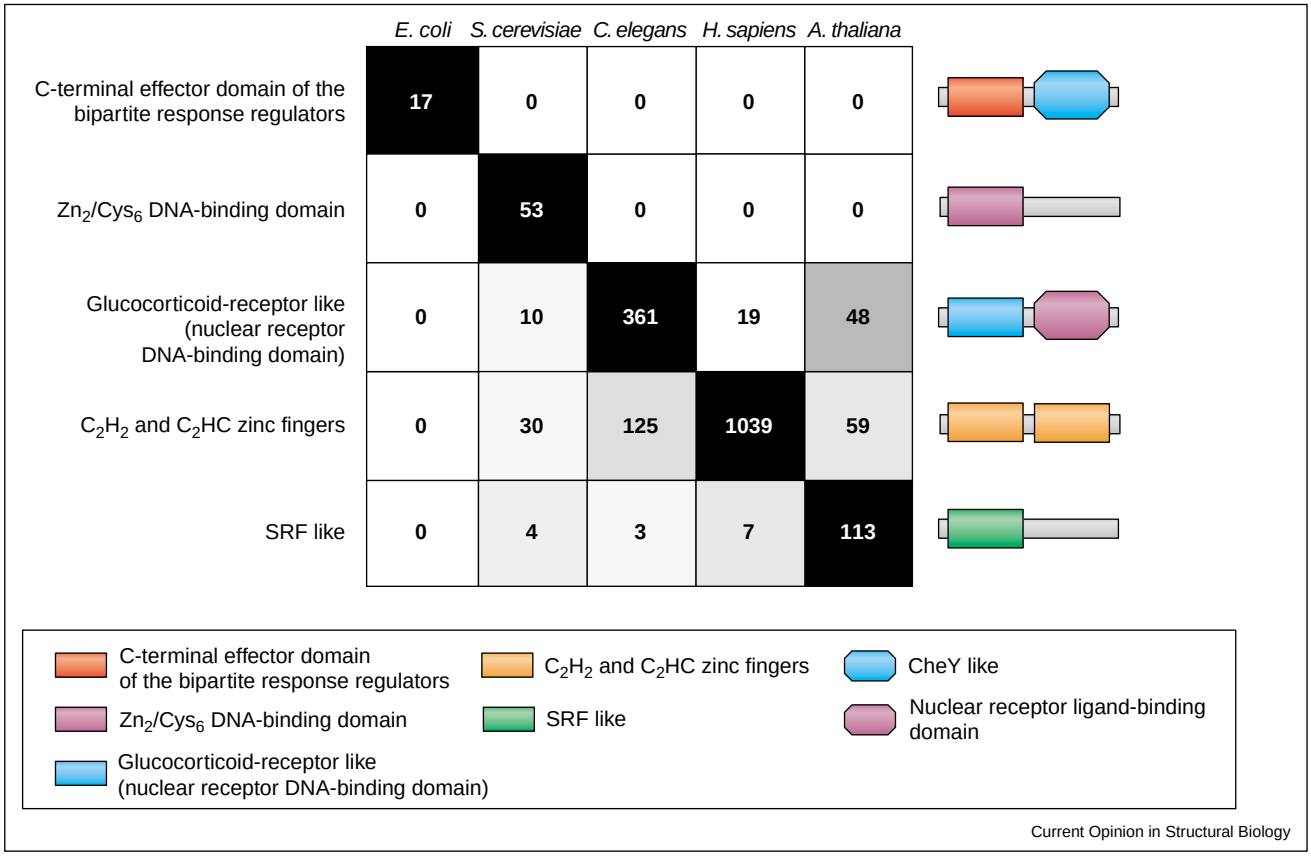
Organism	Number of transcripts	Number of proteins with DNA-binding domains	Percentage of transcripts containing DNA-binding domains
E. coli	4280	267	6.2
S. cerevisiae	6357	245	3.9
C. elegans	31 677	1463	4.6
H. sapiens	32 036[b]	2604	8.1
A. thaliana	28 787	1667	5.7

[a]DNA-binding domain assignments from Pfam and SUPERFAMILY are used to establish the repertoire of DNA-binding transcription factors in five model organisms. An expectation value threshold of 0.002 was used in making the assignments. Co-regulators that do not bind DNA directly are excluded. [b]Predicted by Ensembl v19.34a [42].

Of these, the Winged-helix domain and the Zinc ribbon are encountered in all three principal superkingdoms of life [26]. The Ribbon-Helix-Helix (MetJ/Arc) domain is found only in the prokaryotes [26], whereas the crown group eukaryotes display a proliferation of several novel DNA-binding domains, such as the C_2H_2 zinc fingers, the AT hooks, the HMG1 domain and the MADS box [30].

In Figure 2, we provide examples of some of the most common binding domains in the five genomes listed in Table 1. The DNA-binding domain families were chosen to emphasize that many families are specific to individual phylogenetic groups or greatly expanded in some genomes. For example, the nuclear hormone receptor family transcription factors are very abundant in *Caenorhabditis elegans* compared with other organisms, whereas the Zn_2/

Figure 2



Lineage-specific expansion of DNA-binding domain families. Examples of DNA-binding domain families of transcription factors that are prevalent in one of the five genomes, but are rare in the others. The genomic occurrence of each family is provided in the table and we depict their most common domain architectures alongside. SRF, serum response factor.

Cys_6 fungal-type zinc finger is expanded in the fungi, but absent elsewhere. In contrast to the high level of conservation of other regulatory and signalling systems across the crown group eukaryotes, some of the transcription factor families are dramatically different in the various lineages. This suggests a major role for recurrent, massive and lineage-specific expansions in the evolution of transcription factors in the crown group eukaryotes [31,32]. In prokaryotes, several orthologous groups of transcription factors show a much wider spread across phylogenetically diverse organisms, suggesting a role for horizontal transfers, in addition to diversification through a lower level of lineage-specific duplications.

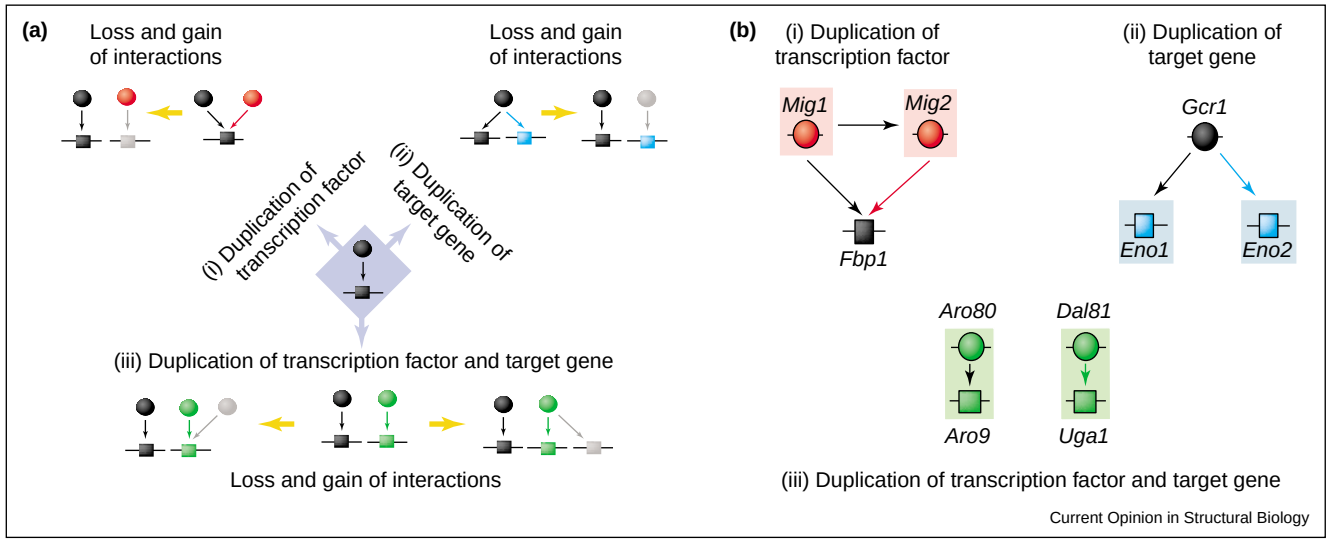
The functional role of transcription factor families in prokaryotes has also been examined. Focusing on the regulatory hubs, global regulators in *E. coli* encompass many distinct protein families and the importance of the transcription factor in the regulatory network is not associated with a particular family [33]. Furthermore, members of *E. coli* transcription factor families can comprise a mixture of activators and repressors, and so the nature of the DNA-binding domain does not necessarily establish the regulatory effect of the factor. Rather, it appears that this is determined by the location of the binding site relative to the transcription start site; activators tend to bind upstream, and repressors very close to or downstream of the start site [33]. In eukaryotes, on the other hand, there is a greater role for epigenetic regulation via chromatin. Specific transcriptional regulation seems to be overlaid on top of such epigenetic regulation. Further support for the above statement is provided by several parasitic eukaryotes, such as the apicomplexans, that lack the diverse panoply of transcription factors seen in the crown group eukaryotes, but have a well-developed apparatus for chromatin modification and basal transcription [34,35].

Regulatory interactions within an organism

The effect of gene duplication on the structure of the transcriptional regulatory network has also been examined. This can be assessed from three possible scenarios (Figure 3a): duplication of the transcription factor, duplication of the target gene with its regulatory region, and duplication of both the factor and target. Following duplication of a transcription factor, both copies of the factor will regulate the same target genes, until regulatory interactions are gradually gained or lost. Similarly, after duplication of a target gene, both copies will be regulated by the same set of transcription factors. Figure 3b illustrates examples of each type of duplication in yeast.

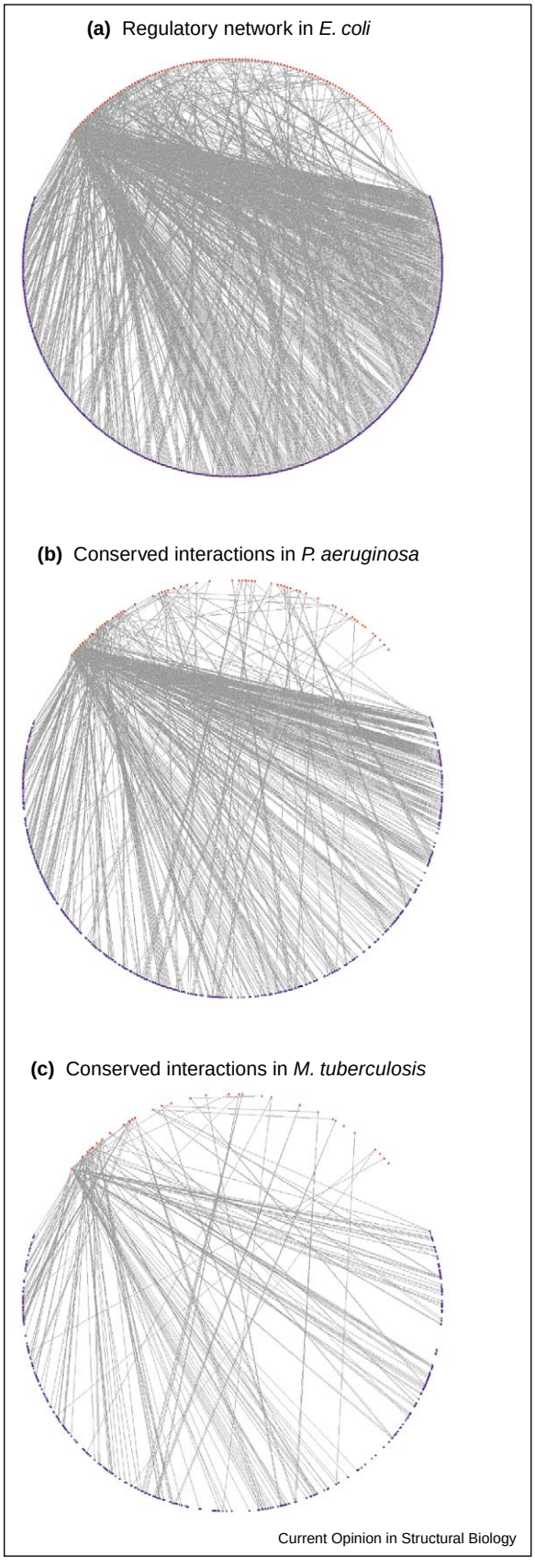
Several studies have compared the regulation of duplicated target genes and have concurred that there is significant similarity in their transcriptional regulation

Figure 3



Regulatory network growth by gene duplication. **(a)** Possible scenarios for the evolution of the basic unit are duplication of (i) the transcription factor, (ii) the target gene and (iii) both. Transcription factor duplication results in both copies regulating the same target. During divergence, new or existing regulatory interactions may be gained or lost. Similarly, target gene duplication results in both copies being regulated by the same transcription factor. Divergence may result in gain or loss of regulators. **(b)** Examples of the three scenarios in the yeast regulatory network. (i) Duplication of transcription factor. *Mig1* and *Mig2* are repressors of glucose metabolism. Both proteins recognise the same DNA binding sites, which suggests that there is redundancy in their regulatory roles. *Mig1* regulates *Mig2* as well as the target gene, forming a feed-forward loop. (ii) Duplication of target gene. *Gcr1* is a global regulator that controls the expression of *Eno1* and *Eno2*. Both targets are enolases that have probably inherited the regulator during duplication. (iii) Duplication of both. *Aro80* and *Dal81* are transcription factors that are homologous to each other, as are their respective targets, *Aro9* and *Uga1*. Homologous pairs of interactions such as these are rare, and may have evolved by duplication of a single chromosomal region encoding both the transcription factor and target.

Figure 4



Conservation of the *E. coli* regulatory network in other bacterial genomes. Transcription factors are represented as red circles, target genes as blue circles and regulatory interactions as grey lines. **(a)** The network of regulatory interactions in *E. coli*. **(b)** Predicted regulatory interactions in *P. aeruginosa*. Interactions are inferred through the presence of orthologous transcription factors and target genes. As a close relative of *E. coli*, many of the interactions are probably conserved. **(c)** Predicted regulatory interactions in *M. tuberculosis*. As a more distant relative, fewer interactions are conserved.

compared to unrelated genes. Papp *et al.* [36] considered closely related target genes in yeast and found that the number of shared regulatory motifs decreases with evolutionary time. Maslov *et al.* [37] also analysed pairs of related target genes in yeast and found that the fraction of shared transcription factors diminishes with decreasing sequence identity.

Teichmann and Babu [38••] defined homologous transcription factors and target genes in *E. coli* and yeast according to their structural domain assignments so that distant relationships could be identified. They showed that gene duplication played a major role in the evolution of the regulatory network and that about 45% of regulatory interactions in both organisms arose by duplication with inheritance of interaction. From this insight, one might think that certain network motifs emerged as a by-product of the duplicative process described above and depicted in Figure 3b. However, this does not appear to be the case and there are only very few examples in which motifs were created by duplication of its constituent elements [38••,39].

Thus, the detailed topology of regulatory networks is not the result of duplication of transcription factors and target genes with inheritance of interaction. There is plenty of loss and gain of interactions after gene duplication, which can result in network motifs and the particular connectivity distributions at the global network level. However, the simple mechanism of duplication and inheritance is responsible for a large fraction of the interactions in regulatory networks.

Regulatory interactions across organisms

Next we turn to network evolution across genomes and ask whether regulatory interactions are conserved among organisms. In other words, if there is a known regulatory interaction between a transcription factor and target gene in one organism, does the orthologous transcription factor regulate the orthologous target gene in another organism? There is evidence that this is the case. For instance, Cripps and Olsen [40•] have shown that the regulatory network for cardiac development in fly is evolutionarily conserved and has been elaborated upon in higher organisms.

More recently, Yu *et al.* [41••] assessed the extent to which regulatory interactions are conserved in yeast,

worm and fly. They showed that orthologous transcription factors and target genes tend to share the same regulatory interaction if the sequences of the regulators are sufficiently similar (generally sequence identities >30–60% depending on the protein family), and called these interactions 'regulogs'. This regulog concept is a useful prediction tool, as regulatory interactions can be transferred from one organism to another as long as orthologous transcription factors and target genes exist.

As expected, the conservation of genes and interactions is related to the phylogenetic distance between the organisms. By applying the regulog principle, we illustrate conservation based on 1293 regulatory interactions in *E. coli* (Figure 4a). For *Pseudomonas aeruginosa* (Figure 4b), a pathogenic γ-proteobacterium that is related to *E. coli* and is less characterised in experimental terms, we can predict many regulatory interactions because a large proportion of orthologous regulators and targets are present, corresponding to 632 interactions. By contrast, only 224 interactions can be mapped onto *Mycobacterium tuberculosis* (Figure 4c), a bacterial pathogen of the actinomycete lineage, because it has fewer orthologues. Preliminary results from these genomes suggest that target genes tend to be more conserved than transcription factors. This is in agreement with results from Maslov *et al.* [37], who found that the rate of evolutionary differentiation of transcriptional regulatory interactions proceeds faster than that of target genes and their protein interactions [37].

Interestingly, there is no bias towards conservation of network motifs, and regulatory interactions in motifs are lost or retained at the same rate as the other interactions in the network (MM Babu *et al.*, unpublished results). Thus, the transcriptional regulatory network appears to evolve in a step-wise manner, with loss and gain of individual interactions probably playing a greater role than loss and gain of whole motifs or modules.

Conclusions

We have surveyed recent findings on the structure and evolution of transcriptional regulatory networks at four levels of detail. The most basic elements of these networks are the component transcription factors and target genes, and the regulatory interactions between them. Transcription factors belong to a limited repertoire of DNA-binding domain families and the relative abundance of these families varies among phylogenetic groups.

The second structural level in regulatory networks is the motif. These over-represented patterns of regulatory interactions are associated with particular kinetic properties that govern the gene expression program. Though all the regulatory interactions must be present in a motif to confer the kinetic properties, there do not appear to be special evolutionary constraints on the constituent elements. Motifs have not evolved by duplication of complete sets of interactions [39] and they are not preferentially conserved across organisms.

The third structural level is the module, which has been defined using several competing approaches by integrating gene expression data. Though there is no consensus on the precise groups of genes and interactions that form modules, it is clear that transcriptional regulatory networks possess a modular structure. Future work may further elucidate the functions and biological significance of these modules.

Finally, at the global level, transcriptional regulatory networks display a scale-free distribution in the outgoing connectivity parameter. This is indicative of the presence of hubs and we discussed the occurrence of global regulators that are central to the structure of regulatory networks. Several recent publications have shown that both transcription factors and target genes can evolve by duplication, with inheritance of the regulatory interactions [36,37,38••]. About 45% of the interactions in the known *E. coli* and yeast networks can be attributed to this mechanism. Other interactions are created by duplication and divergence, and yet others by mechanisms such as lineage-specific innovation of transcription factors, domain shuffling, and recombining DNA-binding domains with different sensor and signalling domains.

It is evident that genes and regulatory interactions are conserved to varying degrees in closely related organisms, and this can be exploited to reconstruct the regulatory networks of poorly characterised organisms [41••]. It is notable that transcription factors are less conserved than target genes, which suggests that regulation of genes evolves faster than the genes themselves.

There has been great progress in our understanding of the structure and evolution of transcriptional regulatory networks over the past couple of years. These mostly computational results have been spurred on by the compilation and publication of the results of functional genomics experiments, such as that of Lee *et al.* [2•]. These insights into the regulatory network structure and its evolution can be translated into prediction of transcription factors and engineering of their regulatory interactions.

Acknowledgements

We are grateful to Sarah K Kummerfeld and Martin Madera for providing the Pfam domain assignments. MMB acknowledges support from Trinity College, Cambridge and the Cambridge Commonwealth Trust.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

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

The authors determine the network motifs that are over-represented in the regulatory network of *E. coli* relative to random networks. They find feed-forward loops, single input modules and dense overlapping regulons.

2. • Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I *et al.*: **Transcriptional regulatory networks in *Saccharomyces cerevisiae***. *Science* 2002, **298**:799-804.
The authors describe the use of chromatin immunoprecipitation-chip (ChIp-chip) experiments to unravel the gene regulatory network for 106 transcription factors in yeast. Analysis of the network identifies six network motifs.

3. •• Salgado H, Gama-Castro S, Martinez-Antonio A, Diaz-Peredo E, Sanchez-Solano F, Peralta-Gil M, Garcia-Alonso D, Jimenez-Jacinto V, Santos-Zavaleta A, Bonavides-Martinez C *et al.*: **RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12**. *Nucleic Acids Res* 2004, **32**:D303-D306.
RegulonDB is a comprehensive literature-curated database containing information on various aspects of transcription in *E. coli*. In addition to the data from the literature, it also contains computationally predicted promoters, binding sites and transcriptional units.

4. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV *et al.*: **TRANSFAC: transcriptional regulation, from patterns to profiles**. *Nucleic Acids Res* 2003, **31**:374-378.

5. Horak CE, Luscombe NM, Qian J, Bertone P, Piccirillo S, Gerstein M, Snyder M: **Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae***. *Genes Dev* 2002, **16**:3017-3033.

6. • Mangan S, Zaslaver A, Alon U: **The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks**. *J Mol Biol* 2003, **334**:197-204.
The authors show experimentally that the feed-forward loop serves as a sign-sensitive delay element. The motif's main feature is to make decisions based on noisy input by filtering out fluctuations in the input stimuli, but still allowing rapid response.

7. • Yu H, Luscombe NM, Qian J, Gerstein M: **Genomic analysis of gene expression relationships in transcriptional regulatory networks**. *Trends Genet* 2003, **19**:422-427.
This paper describes the relationships of the expression patterns between transcription factors and target genes in the yeast regulatory network, taking into consideration inverted and time-shifted relationships. It includes a description of the expression patterns in network motifs.

8. • Guelzim N, Bottani S, Bourgine P, Kepes F: **Topological and causal structure of the yeast transcriptional regulatory network**. *Nat Genet* 2002, **31**:60-63.
The authors reveal that the gene regulatory network in yeast shows an exponential distribution for the number of incoming connections, but a power-law distribution for the outgoing connections.

9. Dobrin R, Beg QK, Barabasi AL, Oltvai ZN: **Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network**. *BMC Bioinformatics* 2004, **5**:10.

10. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture**. *Nat Genet* 1999, **22**:281-285.

11. Teichmann SA, Babu MM: **Conservation of gene co-regulation in prokaryotes and eukaryotes**. *Trends Biotechnol* 2002, **20**:407-410.

12. Stuart JM, Segal E, Koller D, Kim SK: **A gene-coexpression network for global discovery of conserved genetic modules**. *Science* 2003, **302**:249-255.

13. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N: **Revealing modular organization in the yeast transcriptional network**. *Nat Genet* 2002, **31**:370-377.

14. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data**. *Nat Genet* 2003, **34**:166-176.

15. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA *et al.*: **Computational discovery of gene modules and regulatory networks**. *Nat Biotechnol* 2003, **21**:1337-1342.

16. • Madan Babu M, Teichmann SA: **Evolution of transcription factors and the gene regulatory network in *Escherichia coli***. *Nucleic Acids Res* 2003, **31**:1234-1244.
The authors point out that three-quarters of all transcription factors in *E. coli* have arisen by gene duplication and describe 15 global regulators in the known regulatory network of *E. coli*.

17. Yu H, Greenbaum D, Lu H, Zhu X, Gerstein M: **Genomic analysis of essentiality within protein networks**. *Trends Genet* 2004, in press.

18. • Martinez-Antonio A, Collado-Vides J: **Identifying global regulators in transcriptional regulatory networks in bacteria**. *Curr Opin Microbiol* 2003, **6**:482-489.
The authors define global regulators by explicitly including various criteria, such as number of target genes, co-regulators, sigma factors and the conditions in which they exert control.

19. Gutierrez-Rios RM, Rosenblueth DA, Loza JA, Huerta AM, Glasner JD, Blattner FR, Collado-Vides J: **Regulatory network of *Escherichia coli*: consistency between literature knowledge and microarray profiles**. *Genome Res* 2003, **13**:2435-2443.

20. Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF: **IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices**. *Bioinformatics* 1999, **15**:1000-1011.

21. Eddy SR: **Hidden Markov models**. *Curr Opin Struct Biol* 1996, **6**:361-365.

22. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL *et al.*: **The Pfam protein families database**. *Nucleic Acids Res* 2004, **32**:D138-D141.

23. 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* 2001, **313**:903-919.

24. • Van Nimwegen E: **Scaling laws in the functional content of genomes**. *Trends Genet* 2003, **19**:479-484.
This work shows that the number of transcription factors increases with genome size and that larger genomes have more transcription factors per target gene.

25. Perez-Rueda E, Collado-Vides J: **The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12**. *Nucleic Acids Res* 2000, **28**:1838-1847.

26. Aravind L, Koonin EV: **DNA-binding proteins and evolution of transcription regulation in the archaea**. *Nucleic Acids Res* 1999, **27**:4658-4670.

27. Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR *et al.*: ***Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes**. *Science* 2000, **290**:2105-2110.

28. Ledent V, Vervoort M: **The basic helix-loop-helix protein family: comparative genomics and phylogenetic analysis**. *Genome Res* 2001, **11**:754-770.

29. Morgenstern B, Atchley WR: **Evolution of bHLH transcription factors: modular evolution by domain shuffling?** *Mol Biol Evol* 1999, **16**:1654-1663.

30. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al.*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860-921.

31. Lespinet O, Wolf YI, Koonin EV, Aravind L: **The role of lineage-specific gene family expansion in the evolution of eukaryotes**. *Genome Res* 2002, **12**:1048-1059.

32. Coulson RM, Enright AJ, Ouzounis CA: **Transcription-associated protein families are primarily taxon-specific**. *Bioinformatics* 2001, **17**:95-97.

33. Madan Babu M, Teichmann SA: **Functional determinants of transcription factors in *Escherichia coli*: protein families and binding sites**. *Trends Genet* 2003, **19**:75-79.

34. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL: **The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum***. *PLoS Biol* 2003, **1**:E5.

35. Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C, Widmer G, Tzipori S: **Complete genome sequence of the apicomplexan, *Cryptosporidium parvum***. *Science* 2004, **304**:441-445.

36. Papp B, Pal C, Hurst LD: **Evolution of cis-regulatory elements in duplicated genes of yeast**. *Trends Genet* 2003, **19**:417-422.

37. Maslov S, Sneppen K, Eriksen KA, Yan KK: **Upstream plasticity and downstream robustness in evolution of molecular networks**. *BMC Evol Biol* 2004, **4**:9.

38. •• Teichmann SA, Madan Babu M: **Gene regulatory network growth by duplication**. *Nat Genet* 2004, **36**:492-496.
This work addresses the role of duplication in the evolution of the known networks in *E. coli* and yeast. The authors establish to what extent different duplication scenarios have contributed to network growth. They conclude that duplication of transcription factors and target genes with inheritance of interaction contributes a major fraction of the network in both organisms. However, these mechanisms alone cannot explain the evolution of network motifs and the scale-free topology.

39. Conant GC: **Wagner A: Convergent evolution of gene circuits**. *Nat Genet* 2003, **34**:264-266.

40. • Cripps RM, Olson EN: **Control of cardiac development by an evolutionarily conserved transcriptional network**. *Dev Biol* 2002, **246**:14-28.
The authors show that the transcriptional network controlling the development of heart formation in fruit fly has been evolutionarily conserved and elaborated upon in higher eukaryotes.

41. •• Yu H, Luscombe N, Lu H, Zhu X, Xia Y, Han J, Bertin N, Chung S, Goh C, Vidal M, Gerstein M: **Annotation transfer for genomics: assessing the transferability of protein-protein and protein-DNA interactions between organisms**. Genome Res 2004, in press.
This work is a comprehensive assessment of the extent to which protein–protein interactions and transcriptional regulatory interactions are conserved among orthologues. This is interesting from an evolutionary point of view and vital for informed prediction of interactions based on conservation.

42. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T *et al.*: **The Ensembl genome database project**. *Nucleic Acids Res* 2002, **30**:38-41.