Evolutionary Dynamics of Prokaryotic Transcriptional Regulatory Networks

M. Madan Babu1,2*, Sarah A. Teichmann1 and L. Aravind2*

1National Center for Biotechnology Information, National Institutes of Health, MD 20894 USA
2MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB22QH, UK

The structure of complex transcriptional regulatory networks has been studied extensively in certain model organisms. However, the evolutionary dynamics of these networks across organisms, which would reveal important principles of adaptive regulatory changes, are poorly understood. We use the known transcriptional regulatory network of Escherichia coli to analyse the conservation patterns of this network across 175 prokaryotic genomes, and predict components of the regulatory networks for these organisms. We observe that transcription factors are typically less conserved than their target genes and evolve independently of them, with different organisms evolving distinct repertoires of transcription factors responding to specific signals. We show that prokaryotic transcriptional regulatory networks have evolved principally through widespread tinkering of transcriptional interactions at the local level by embedding orthologous genes in different types of regulatory motifs. Different transcription factors have emerged independently as dominant regulatory hubs in various organisms, suggesting that they have convergently acquired similar network structures approximating a scale-free topology. We note that organisms with similar lifestyles across a wide phylogenetic range tend to conserve equivalent interactions and network motifs. Thus, organism-specific optimal network designs appear to have evolved due to selection for specific transcription factors and transcriptional interactions, allowing responses to prevalent environmental stimuli. The methods for biological network analysis introduced here can be applied generally to study other networks, and these predictions can be used to guide specific experiments.

© 2006 Published by Elsevier Ltd.

Keywords: transcriptional regulatory network; evolution; network; network motif; regulation

Introduction

Of the several steps at which the flow of information from a gene to its protein product is controlled, regulation at the transcriptional level is a fundamental mechanism observed in all organisms. This form of regulation is typically mediated by a DNA-binding protein (transcription factor) that binds to target sites in the genome and, either singly or in combination with other factors, regulates the expression of one or more target genes. The sum total of such transcriptional interactions in an organism can be conceptualised as a network, and is termed the transcriptional regulatory network.1–11 In such a network, nodes represent genes and edges represent regulatory interactions. Studies on the transcriptional regulatory network at an abstract level have shown that they have architectures resembling scale-free networks, with striking structural and topological similarity to other networks from biological and non-biological systems. They are characterized by the recurrence of small patterns of interconnections, called network motifs,12–16 which were first defined in Escherichia coli,3 and were subsequently found in yeast and other organisms.2,4

Even though the general structural properties of transcriptional networks are well understood, there are several fundamental questions regarding the provenance and evolution of transcriptional regulatory networks that remain unanswered: What are the trends of conservation of transcription factors,
target genes and regulatory interactions in the network? How do interactions that specify topologically equivalent motifs within the network evolve? How does the global structure of the network evolve? We addressed these questions by using the experimentally determined transcriptional regulatory network of E. coli as a reference network.3,17 and performing a comparative genomic analysis to predict components of the regulatory network for 175 prokaryotes with completely sequenced genomes, from diverse lineages of the bacterial and archaeal kingdoms (a list of the genomes is provided as Supplementary Data).

Reconstruction of transcriptional regulatory networks

While there has been considerable progress in unravelling the regulatory networks of various model organisms such as E. coli, the extrapolation of this information to poorly studied organisms, whose complete genome sequences are now available, remains a major challenge. Target genes for a transcription factor can be identified by using sequence profiles of known binding sites across different organisms and by arriving at a set of genes with conserved regulatory sequences. However, this method requires prior knowledge about binding sites, and is applicable only to closely related genomes, since orthologous transcription factors may regulate orthologous target genes through very divergent binding sites in distantly related organisms.18–21

Alternatively, using an experimentally characterized transcriptional network as template, one can infer transcriptional targets of a regulator in a genome of interest by identifying orthologues of transcription factors and their target genes. It is now generally accepted that in the majority of cases, orthologous transcription factors regulate orthologous target genes. This procedure of transferring information about transcriptional regulation from a genome with known regulatory interactions to another genome by identifying orthologous proteins was assessed recently by Yu et al.,22 and was found to be a fairly robust method for predicting such interactions in eukaryotes. In fact similar approaches, based on orthologue detection using a bi-directional best-hit procedure,23–28 have been developed successfully to transfer information on interactions to other organisms and have proved to be useful in predicting new interactions.

Detecting orthologues is a non-trivial process. After testing various orthologue detection procedures (e.g. bi-directional best-hit and best hits with defined e-value cut-offs), we arrived at a hybrid procedure, which was used to identify orthologous proteins in a genome. Using this approach, we predicted components of the transcriptional networks. With the best-characterized transcriptional regulatory network currently available, that of E. coli with 755 genes (129 transcription factors) and 1295 transcriptional interactions, as a template, we used our orthologue detection procedure and predicted transcriptional interaction networks, for the first time, for 175 prokaryotic genomes (Figure 1; Supplementary Data M1, M2 and S3). Our method is based on identifying the orthologues of E. coli transcription factors and the orthologues of E. coli target genes in each of the prokaryote genomes using protein sequence comparisons, without considering conservation of the transcription factor-binding sites in the DNA. We chose the orthology-based method rather than using binding site information because using binding sites (i) would place an additional constraint and would drastically reduce the size of the E. coli network that we considered (only a few transcription factors in the network have enough experimentally characterized binding site information to build a reliable profile) and (ii) would limit the number of genomes that can be compared, as reliable detection of these short binding sites in distantly related organisms is not possible, and may hence bias our analysis.20 It should be noted that the method we develop can, in principle, be applied to any reference network, and we chose the E. coli network because it is the most comprehensive when compared to networks available for other prokaryotes.

The transcriptional regulatory network for E. coli is shown in Figure 1, along with the networks for a Gram-positive mammalian pathogen, Bacillus anthracis and a free-living organism, Streptomyces coelicolor, which were reconstructed using the E. coli network as the reference network. We wish to emphasize that we do not predict new regulatory interactions that have been gained in the other organisms, and no current computational methods allow us to do it in the absence of other external information such as gene expression data, etc. However, we can still predict potentially conserved interactions, apart from conserved regulatory components and genes, which can shed light on network evolution (see Supplementary Data S17).

Orthologue detection, which is the foundation for our network reconstruction procedure can be confounded by rapid duplication, divergence and loss of genes. Hence, in this study, we assessed our procedure using the expression data available for Vibrio cholerae and the known regulatory network of Bacillus subtilis. We studied the extent to which target genes with the same set of known and predicted transcription factors have similar expression profile.29 We found that co-regulated genes in E. coli, for which the transcriptional network is known, and in V. cholerae, for which predictions were based on the reconstructed network, tend to be strongly co-expressed (Supplementary Data S2). Ideally, one would want to carry out such an assessment for as many genomes as possible; however, the availability of meaningful gene expression data for other organisms limits us to restrict this analysis to V. cholerae. This result supports the validity of reconstructing
Figure 1. Known transcriptional regulatory network for *E. coli* and reconstructed transcriptional regulatory networks for a pathogen, *Bacillus anthracis* and a free-living organism, *Streptomyces coelicolor*. Transcription factors and target genes are represented as red and blue circles, and transcriptional interactions are represented as black lines. The transcription factors are ordered according to their connectivity to emphasize the scale-free-like structure of the network at the global level, where there are few dominant regulatory hubs that control many target genes and many transcription factors that regulate few target genes. Information about the number of transcription factors, target genes, regulatory interactions and regulatory network motifs is provided at the right. This Figure illustrates how we can reconstruct transcriptional networks for genomes, including organisms, that are poorly characterised but are important. The table below shows the number of transcription factors that have a DNA-binding domain belonging to a particular family. It is clear by comparing the numbers that each genome has evolved its own set of transcriptional regulators, and hence regulatory interactions, by using the same DNA-binding domains to different extents.
transcriptional networks by inferring regulatory interactions between orthologous transcription factors and orthologous target genes in prokaryotes. Additionally, the experimentally determined transcriptional regulatory network of B. subtilis (a bacterium very distantly related to our reference organism E. coli) shows a good degree of congruence with the interactions predicted by our analysis, thereby lending support to the validity of our reconstruction procedure (Supplementary Data S11; note that the experimentally determined B. subtilis network is far from complete and is much smaller than the E. coli network and, hence, this comparison should not be treated as a gold standard to get false positive and false negative estimates).

**Results and Discussion**

**Evolution of genes and their regulatory interactions**

Transcription factors evolve rapidly and independently of their target genes

To assess the evolutionary trends in the network conservation, we asked if transcriptional regulators and their targets are conserved differentially in evolution. When we quantified the extent of conservation of the 755 genes in the 175 genomes, we found that transcription factors are less conserved across genomes during evolution than their target genes (Figure 2(a)). We assessed the significance of this observed bias in the conservation patterns by simulating network evolution (Supplementary Data M3), and found it to be statistically significant \((p < 10^{-4})\). This suggests that the evolution of major phenotypic differences between organisms is a consequence of replacements of transcription factors, which provide regulatory inputs, rather than changes in the gene repertoire itself. The above observation is consistent with the fact that metabolic enzymes, which constitute a major set of the target genes and thus the metabolic network, are well conserved across the different organisms.\(^{30}\)

Because a regulatory interaction involves a transcription factor and its target gene, one might expect them to be present or absent as pairs. Indeed, studies on the conservation of physically interacting proteins across the different proteomes suggest that proteins forming a complex, especially interacting pairs, tend to be conserved or lost in a concerted manner.\(^{30,31}\) To test this, we first created two sets of gene pairs using the transcriptional regulatory network: (i) the interacting set, which consisted of all pairs of genes that are known to interact in the transcriptional network and (ii) the non-interacting set, which consisted of all possible pairs of genes in the network that are known not to interact in the original network. We then analysed the conservation pattern across the 175 organisms for all the pairs of genes in both sets. If interacting proteins are preferentially conserved or lost, one would expect the trends obtained for the two sets to be vastly different. On the other hand, if interacting proteins are conserved to the same extent as any pair of non-interacting proteins, one would expect the trends obtained for the two sets to be similar. Interestingly, in the transcriptional regulatory network studied here, the relative conservation of pairs from the two sets is close to 1 across all genomes, suggesting that there is no strong preference for pairs of interacting transcription factors and target genes to be conserved in the network (Supplementary Data M6d; and Figure 2(b)). This observation suggests that the forces of natural selection act relatively independently to retain or discard transcription factors and their targets, as opposed to protein pairs that are involved in physical interactions.

**Organisms have evolved their own set of transcription factors**

Several bacteria, such as B. anthracis and S. coelicolor (Figure 1), have few transcription factors with orthologues in the E. coli reference network. However, these genomes are relatively large and must contain a complex transcriptional regulatory network. Hence, we searched these proteomes with sensitive sequence profile methods to identify proteins with DNA-binding domains typically observed in known transcription factors.\(^{32}\) In most genomes, we could identify other transcription factors belonging to the same repertoire of DNA-binding domain families as in E. coli (Figure 1), but not orthologous to the E. coli proteins. This observation indicates that new lineage-specific transcription factors, and thereby new interactions, are gained constantly during evolution. However, our current level of understanding prevents a systematic prediction of the newly gained interactions (Supplementary Data S14). For small genomes, typically those of obligate parasites with low fractions of transcription factors, no additional transcription factors were detectable (Supplementary Data S4). Consistent with recent observations by van Nimwegen, Ranea et al. and by us,\(^{34–36}\) the increase in the number of regulatory proteins with genome size is non-linear (Figure 2(c)). This implies that as genome size increases, a greater than proportional increase in the numbers of transcription factors is required for controlling the newly added genes. This tendency might correlate with the need to regulate specialized groups of genes individually, or with the need for integration of distinct inputs and introducing more layers in the regulatory hierarchy of the metabolically or organizationally complex bacteria with large genomes.\(^{35}\) These observations, taken together with the higher degree of conservation of target genes, suggest certain general principles that operate in the evolution of transcriptional networks. (1) In small parasitic genomes, transcription factors have been lost due to the absence of selective pressures for...
Figure 2. Conservation of the transcription regulatory network and regulatory interactions. (a) For each of the 175 genomes, the fraction of target genes conserved (x-axis) is plotted against the fraction of transcription factors conserved (y-axis). The diagonal, shown in green, represents equal conservation of transcription factors and target genes, and every point represents a genome. The graph shows that more target genes are conserved than transcription factors in the genomes considered ($y = 1.1173x - 16.609$, $R^2 = 0.9223$, $p < 10^{-4}$). The significance of this trend was assessed by simulating network evolution, which involved neutral removal of different sets of genes 10,000 times (Supplementary YJMBI 58077—25/2/2006—SATHYA—202231—XML – pp. 1–20 / GH)
regulation. (2) In larger genomes, target genes are often controlled by additional regulators or regulators that are non-orthologous, so that genes are responsive to a variety of different inputs that are typically dependent on the environmental niche of the organism. For example, in a number of phylogenetically distant free-living bacteria, including proteobacteria, B. subtilis and Streptomyces, there is an expansion of so-called one-component transcription factors of the LysR family, which sense a wide range of small-molecule ligands. However, there are no homologues of such transcription factors in any of their close relatives, which are obligate pathogens. This is consistent with the need in all of these evolutionarily weakly related, free-living, physiologically complex bacteria to be able to sense the same set of metabolites in their environment.

Organisms with similar lifestyles preserve homologous regulatory interactions

Given that the conservation of transcription factors and target genes is poorly correlated, we sought to address the extent to which transcriptional interactions are conserved across different organisms. To do this, we developed an algorithm to deduce the relationship between different genomes based on conservation of specific sets of transcriptional interactions (Supplementary Data M5). We first represented the presence or the absence of genes and interactions of the reference E. coli transcriptional network in each of the 175 prokaryotic genomes as a binary interaction conservation profile. The organisms were then hierarchically clustered on the basis of their interaction conservation profiles (Figure 3(a)).

The clustering of organisms based on the interactions present and depicted in Figure 3(b) reveals that transcriptional interactions appear to be shaped by a set of disparate forces. At low to moderate evolutionary distance from the reference organism E. coli, namely proteobacteria, the clustering generally mirrors organism-based relationships constructed using highly conserved gene sequences, suggesting that the regulatory network retains a noticeable phylogenetic signal. Beyond the proteobacteria, a number of other effects appear to dominate upon the background of the weak phylogenetic signal. These include the similarities in genome size (Supplementary Data S15) and ecological adaptations (e.g., for example, bacteria with comparable genome size, such as several species of Bacillus, Corynebacterium and Mycobacterium, whose principal habitat is the soil, form a cluster (Figure 3(b)). Likewise, the obligate or intracellular parasites from diverse bacterial clades, namely Mycoplasma, Rickettsiae and Chlamydiae, group together in this analysis.

This observation motivated us to investigate whether our finding that organisms with similar lifestyles conserve similar interactions was a general principle, or was limited to anecdotal examples. First, we systematically defined the lifestyle of each organism as a combination of four properties: oxygen requirement, optimal growth temperature, environmental condition, and whether it is a pathogen. For example, E. coli K12 would belong to the lifestyle class called facultative/mesophilic/host-associated. Then, we compared the similarity between organisms belonging to different lifestyle classes as a function of the interactions they have in common (Supplementary Data S12 and M10). Each element in the matrix shown in Figure 4(a) corresponds to the normalized average similarity in the interaction content across all pairs of organisms belonging to the lifestyle classes considered. The values along the diagonal, which reflect the average similarity among organisms belonging to the same lifestyle class, tend to be much greater than the off-diagonal elements. This lends support to our finding that organisms with similar lifestyles do indeed tend to conserve similar interactions.

We defined an index, called the lifestyle similarity index (LSI), to measure the strength of this trend. The LSI is calculated as the ratio of the average similarity among the diagonal elements to the average similarity of the off-diagonal elements. LSI values greater than 1 mean that organisms within the same lifestyle class have more similarity of interactions compared to organisms belonging to a different lifestyle class. Amongst the organisms included in our study, we obtained an LSI value that is far above 1 and is statistically significant (LSI = 1.42; \( p < 10^{-3} \); Z-score = 2.96). This shows that organisms belonging to the same lifestyle have a significantly higher number of interactions in common in comparison to organisms from other lifestyle classes. The enrichment in similarity in some off-diagonal elements, corresponding to organisms belonging to different lifestyle classes, is associated primarily with certain mesophilic organisms. This arises due to the fact that these lifestyle classes contain organisms with large genomes that are phylogenetically related to other specialized organisms spanning the entire spectrum.
of lifestyle classes. For example, *B. subtilis*, a mesophile, may be related to a thermophile like *Bacillus stearothermophilus*, a pathogen like *B. anthracis*, anaerobe like *Thermoanaerobacter tengcongensis* and microaerophiles such as Lactobacillus. We emphasize here that each phylogenetic group does have organisms with different lifestyles and, for each lifestyle class that we define, there are organisms belonging to different phylogenetic groups. Thus, the LSI values are not overly biased by evolutionary relatedness for a lifestyle class or by the enrichment of a particular lifestyle class within a phylogenetic group (Supplementary Data S13). This proposition applies also to a subsequent analysis of lifestyle classes and motif conservation, described later.

Thus, on the basis of the information about conservation of specific transcription factors and their interactions, we were able to potentially reconstruct the presence of transcriptional response pathways similar to those in the reference network in various other genomes, and understand their evolution ( Supplementary Data S5, S10, M4, M5). In combination with other context-based methods,42,43 these reconstructions could aid in probing poorly characterized or experimentally intractable organisms, including key human pathogens. For example, important pathogens like *Yersinia pestis*, *Pseudomonas syringae* and a nitrogen-fixing symbiont of the soybean plant, *Bradyrhizobium japonicum*, have conserved the GntR, KdgR, ExuR and UxuR transcriptional regulators. These regulators can sense different hexuronates and hexuronides, suggesting the conservation of the pathway that can catabolize these compounds. Examination of target gene conservation shows that they are indeed conserved, thus allowing us to predict the presence of these response pathways and their transcriptional regulators. Extending our approach with other reference networks in future could identify potential targets for therapeutic intervention in unrelated pathogens that share a similar ecological niche.

**Evolution of local network structure**

At the local level, the transcription regulatory network shows recurrent topological patterns of interconnections called network motifs, which can be viewed as building blocks of the network.2–4 Such network motifs have been suggested to carry out specific information processing tasks and hence dictate specific patterns of gene expression.6,44 For example, the function of feed-forward motifs is to respond only to persistent signals, and that of single-input motifs is to bring about a quick and coordinated change in gene expression.44 We investigated whether natural selection acts at the level of...
motifs and whether the interactions, which are conserved across organisms, correspond to network motifs in the reference network.

In the E. coli network, there are 277 feed-forward motifs, 43 single-input motifs and 70 multiple-input motifs. To evaluate motif conservation in the network, we devised a new method (Supplementary Data M8; and Figure 5) that clusters topologically equivalent motifs according to their conservation profile. First, we generated a motif conservation profile for each genome, and then carried out a two-way clustering procedure: a k-means clustering of all motifs with similar distribution profile across genomes, followed by a hierarchical clustering of genomes on the basis of their motif conservation profile (Supplementary Data S6 and S7).

Contrary to our expectation, the analysis showed a surprising result, that there is no significant conservation of whole motifs when compared to random networks of similar size. Interestingly, within coherent monophyletic lineages of prokaryotes with different lifestyles, particular network motifs may not be conserved, whereas genomes belonging to unrelated phylogenetic groups with similar lifestyle conserve these motifs. For example, Fnr (a global regulator, activated during low levels of oxygen), NarL (transcriptional regulator of a two-component signal transduction system) and NuoN (subunit of the NADH dehydrogenase complex I) form a feed-forward motif,
Figure 4. Organisms with similar lifestyles show similarity in interaction and motif content. Each element in the matrix represents (a) the average similarity in the interaction content or (b) the average similarity in the motif content among all pairs of organisms belonging to the two lifestyle classes considered. The lifestyle similarity index (LSI) is the ratio of average similarity between organisms in the same lifestyle class to the average similarity between organisms belonging to a different lifestyle class, i.e. the ratio of the average value of the diagonal elements to the average value of the off-diagonal elements. LSI > 1 suggests that organisms with similar lifestyles tend to show similar interactions or network motif contents. The LSI values are 1.42 and 1.34 for the similarity based on (a) interactions and (b) network motifs. The matrix shown is not symmetric because each element has been row-normalized to highlight the trend. This is done for illustrative purposes only, and will not affect the LSI values. (For details, see Supplementary Data S12 and M10.)
Figure 5. Network motif conservation in genomes. (a) A procedure to build and cluster motif conservation profiles to identify topologically equivalent motifs that are conserved in the different genomes. (b) Cluster diagram of the 277 feed-
which is not conserved in other γ proteobacteria, whereas it is found in several distantly related genomes (Figure 6(a) and (b)). These results suggested that organisms belonging to the same lifestyle have conserved similar network motifs. To test the generality of this observation, we computed the similarity in motif conservation among organisms belonging to different lifestyles (Supplementary Data S12). Figure 4(b), which represents the similarity in network motif conservation among organisms belonging to different lifestyles, again shows that the diagonal elements (average similarity in motif content among organisms within the same lifestyle class) are much greater than the off-diagonal elements (average lifestyle similarity among organisms belonging to different lifestyle class). In this case, the LSI value is 1.34 (p < 10^{-3}, Z-score = 2.83), showing that organisms with the same lifestyle share common motifs when compared to organisms of different lifestyles. This suggests that, apart from the phylogenetic component in retaining interactions, organisms with similar lifestyles tend to conserve network motifs as a general principle.

Though motifs tend to be conserved within lifestyle classes, we sought to know whether interactions in motifs are conserved preferentially compared to other interactions in the network regardless of lifestyle class. To determine this, we computed a motif conservation index (C.I) for each organism (Supplementary Data M9). C.I is defined as the logarithm of the ratio of the fraction of conserved interactions that forms a motif in E. coli to the fraction of all interactions conserved. We then carried out network simulation experiments where we: (a) selected for interactions in motifs; (b) neutrally selected interactions in the network; and (c) selected against interactions in motifs and obtained the trends for C.I. As shown in Figure 6(c), the observed trend of conservation index for the 175 genomes is closest to a model of unbiased removal of interactions, rather than preferential conservation or loss of interactions in motifs. Thus, we find that there is no preferential conservation of whole motifs or parts thereof (Supplementary Data S8). These results indicate that interactions in motifs and whole motifs are not conserved preferentially when compared to other interactions in the network, which was unexpected. All these findings suggest that motif formation is dynamic during evolution, and can easily be reshaped during adaptation to changes in lifestyle.

Orthologous genes can come under the control of different motifs in different organisms

Careful examination of partially conserved motifs provided a possible answer as to why interactions in motifs are not specifically conserved. Our analysis revealed that orthologous genes can become part of different types of motifs in the regulatory network of different organisms by losing or gaining specific transcription factors (Figure 6(d); and Supplementary Data S9). This observation implies that an orthologous gene in a different organism may acquire a different pattern of gene expression in order to adapt better to changing environments. In this context, it is interesting to note that a more recent study by Dekel and Alon has demonstrated that E. coli strains grown in different lactose environments optimize levels of LacZ expression to increase its growth-rate by keeping the cost of production low. Thus, our results lend strong support (at a genomic level across many organisms) that different organisms arrive at different solutions by tinkering with specific regulatory interactions to optimize expression levels rather than porting whole blocks of pre-existing transcriptional interactions. For example, a gene that may need to be regulated tightly (as a feed-forward motif) in one organism, might need to be regulated directly and quickly (as in a single-input motif) in a different genome due to changes in lifestyle, development or environment. Specific cases are shown in Figure 6(d) (a complete list for every genome can be obtained from the supplementary website; see Supplementary Data). An example of this is E. coli, which is adapted to a life with fixed aerobic and anaerobic phases, and will not express the fumarate reductase genes (FrdB and FrdC, which converts fumarate to succinate under anaerobic conditions to derive energy) unless there is persistent signal for lack of oxygen (through a feed-forward motif involving both Fnr and NarL). In contrast, Haemophilus influenzae, which encounters rapid redox fluctuations during host infection and needs to regulate the fumarate reductase genes more quickly than E. coli, appears to depend solely on Fnr for the response (by employing a single-input motif).
Figure 6. Evolution of local network structure. (a) A feed-forward motif formed by Fnr, NarL and NuoN genes in *E. coli* is completely conserved in a closely related genome, *Salmonella typhi*, but not in other γ-proteobacterial genomes. Fnr is a regulatory hub and is not always conserved in genomes within the same phylogenetic group. This suggests that condition-specific regulatory hubs can either be lost or displaced by a non-orthologous protein in closely related genomes. (b) Distantly related organisms that have conserved all interactions in the regulatory motif and that have conserved the regulatory hub, Fnr. (c) This is a plot of fraction of genes conserved (x-axis) against conservation index (y-axis), which is a measure of the extent to which interactions in a motif are conserved. The conservation index (C.I.) is calculated as the logarithm of the ratio of the fraction of conserved interactions that forms a motif in *E. coli* to the fraction of all interactions conserved. C.I. > 0 means that interactions in motifs are selected for; C.I. close to 0 suggests total neutral conservation of motifs.
Evolution of global network structure

We next sought to address how the fine-scale, independent evolution of interactions in network motifs affects the global structure of the network. Transcriptional regulatory networks have a hierarchical structure best approximated by the scale-free network model, in which the outgoing connectivity of transcription factors follows a power law.\( y = ax^{-\gamma} \), where \( y \) is the number of transcription factors, \( x \) is the number of target genes and \( \gamma \) is the scale-free exponent. In other words, there are few transcription factors that regulate many target genes. These influential transcription factors play the role of regulatory hubs in the network.

Transcriptional regulatory hubs evolve like other transcription factors in the network

It was postulated that scale-free behaviour should hold good for randomly selected parts of networks of this form.\(^4\) In fact, the reconstructed transcriptional networks for a majority of the genomes follow a power law distribution in their outgoing connectivity (Figure 7(a); and Supplementary Data S16). We then asked if this is because transcription factors occupying hubs in the network are preferentially conserved.

Earlier studies on protein–protein interaction networks have suggested that the proteins with the most interactions tend to be more conserved across organisms.\(^50,5^1\) We explored this possible link between connectedness of a hub and its conservation by comparing the observed retention of regulatory interactions to simulated models of network evolution where transcription factors are lost in an unbiased manner, and models where hubs are preferentially conserved or lost. The observed pattern is closest to the unbiased removal of transcription factors and is independent of the number of their target genes in the \( E. \) coli network. This implies that there is no correlation between the degree of connectedness of a particular transcription factor and its conservation across genomes, as shown in Figure 7(b). Table 1 provides specific examples of transcription factors with many target genes that are poorly conserved across genomes and transcription factors with low connectivity that are conserved in many genomes.

A possible explanation for the absence of any bias for conservation of influential transcription factors is suggested by our recent investigation of regulatory hubs in yeast. Most of these regulatory hubs are condition-specific, i.e. they regulate many genes only in particular conditions, but remain silent in other conditions.\(^5^2,5^3\) Most transcription factors in prokaryotes are condition-specific, in that they respond to a specific environmental signal.\(^5^4\) If the lifestyle of the organism does not involve a particular condition, then that regulatory protein can be lost. For example, in the opportunistic pathogen \( P. \) aeruginosa, which actively utilizes phenolic compounds, the transcriptional regulators MhpR, HcaR and FeaR can sense the compounds and activate target genes that encode enzymes involved in their catabolism. However, more obligate pathogens like \( S. \) aureus and \( C. \) jejuni, which do not typically face phenolic compounds in their natural niches, lack both the regulators and their target genes for the neutrality towards maintenance of motifs and C.I. < 0 suggests selection against such interactions. The C.I. trends obtained by carrying out network simulation where interactions in motifs were selected for, neutrally selected and selected against are shown in blue, green and red respectively. The plot of C.I. for the 175 genomes (in black) shows that interactions that form the network motif are not selected for or against in evolution and are conserved to the same extent as any other interaction in the network. This implies that evolution acts by tinkering to form different motifs using the same interaction in different genomes. (d) Analysis of partially conserved motifs revealed that by losing (or gaining) specific transcription factors, orthologous genes in different genomes could be expressed in different ways according to specific requirements. Genes in a feed-forward motif (FFM) in \( E. \) coli can be regulated as a part of a single-input module (SIM) by losing a transcription factor (shown in grey). Feed-forward motif regulation ensures that target gene expression is not sensitive to fluctuations in input signals. Whereas a SIM regulation ensures expression of target genes as long as there is some input signal, like small molecules. Genes in a multiple-input motif (MIM) in \( E. \) coli can be regulated as a single-input motif (SIM) by losing a transcription factor. Genes regulated as a part of a MIM ensures that target genes are transcribed only when two input signals cross particular thresholds independently. Thus, by losing a transcription factor, genes that are tightly regulated can be regulated in a simple manner. Genes regulated as a part of an FFM can be regulated as a part of an MIM by the loss of a transcription factor. Thus, evolution tinkers with specific regulatory interactions to create different motifs in genomes when orthologous genes need to be expressed differently. Generically, the Figure illustrates that by bringing about temporal expression of specific transcription factors, the same target gene can follow different patterns of gene expression, be it in a different cell type or in different organisms.
utilization of these aromatic compounds. These observations suggest an important principle of adaptation at the regulatory level: the adaptive value of an orthologous transcription factor might be different across different organisms, depending on the environment, thereby affecting the number of genes regulated by it and the overall network structure. These findings support an interesting

Figure 7. Evolution of global network structure. (a) The fraction of genes conserved (x-axis) is plotted against the scale-free exponent value (γ) of the conserved networks (y-axis) for the different genomes, in black. γ is the scale-free exponent in the power law equation $y = ax^{-γ}$, where $y$ is the number of transcription factors and $x$ is the number of target genes. The average value of $γ$ obtained after simulating a neutral model of network evolution 10,000 times is shown in green. This plot shows that for genomes that conserve about 30–60% of the genes, the exponent value ($γ$) is higher than in random networks, suggesting a more pronounced hub-like behaviour. (b) The fraction of transcriptional regulatory interactions conserved in the network (x-axis) is plotted against the fraction of the genes conserved (y-axis) for each of the 175 genomes ($y = 0.007x^2 - 1.7843x + 100, R^2 = 0.95$). To understand what the observed trend means, we simulated the process of network evolution by incorporating different rules: (i) remove genes with low connectivity first, implying selection to conserve highly connected transcription factors, shown in blue; (ii) remove genes neutrally with no special emphasis on connectivity, implying neutral evolution, shown in green; and (iii) remove nodes with high connectivity first, implying selection, against highly connected regulatory proteins, shown in red. We find that the observed trend is more similar to the trend that we obtain while simulating the neutral condition. (c) Reconstructed transcriptional regulatory networks for H. influenzae and B. pertussis is shown to illustrate that the scale-free-like structure is still conserved, even though regulatory hubs can be lost or replaced. For example, the distribution for the outgoing connection for the conserved network in H. influenzae is $y = 214.8x^{-0.49}$ even though a regulatory hub NarL is lost.
Table 1. Conservation and connectivity profile for transcription factors

<table>
<thead>
<tr>
<th>Transcription factor</th>
<th>GI number</th>
<th>Low connectivity (k ≤ 15)</th>
<th>High conservation (≥ 50%)</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Proteins with few target genes (connectivity, k ≤ 15) but conserved in more than 50% of the genomes studied</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ada</td>
<td>16130150</td>
<td>5</td>
<td>74.86</td>
<td>Transcriptional regulator of DNA repair</td>
</tr>
<tr>
<td>BirA</td>
<td>16131807</td>
<td>5</td>
<td>89.71</td>
<td>Biotin operon repressor</td>
</tr>
<tr>
<td>MarR</td>
<td>33347561</td>
<td>7</td>
<td>54.29</td>
<td>Repressor of multiple antibiotic resistance (Mar) operon</td>
</tr>
<tr>
<td>DnaA</td>
<td>16131570</td>
<td>10</td>
<td>87.43</td>
<td>Transcriptional regulator of house-keeping genes</td>
</tr>
<tr>
<td>CpxR</td>
<td>16131752</td>
<td>14</td>
<td>55.43</td>
<td>Regulator of a 2CST^a</td>
</tr>
<tr>
<td>OmpR</td>
<td>16131282</td>
<td>14</td>
<td>57.14</td>
<td>Regulator of adaptive response (2CST^a)</td>
</tr>
<tr>
<td>NarP</td>
<td>16130130</td>
<td>15</td>
<td>60.57</td>
<td>Regulator of aerobic respiration (2CST^a)</td>
</tr>
<tr>
<td>B. Proteins with many target genes (connectivity, k &gt; 15) but conserved in less than 50% of the genomes studied</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FhlA</td>
<td>16130638</td>
<td>16</td>
<td>34.29</td>
<td>Formate hydrogen lyase activator</td>
</tr>
<tr>
<td>Rob</td>
<td>16132213</td>
<td>17</td>
<td>24.00</td>
<td>Transcriptional activator for antibiotic resistance</td>
</tr>
<tr>
<td>CysB</td>
<td>16129236</td>
<td>18</td>
<td>25.14</td>
<td>Regulator of cysteine biosynthesis</td>
</tr>
<tr>
<td>FruR</td>
<td>16128073</td>
<td>25</td>
<td>21.14</td>
<td>Fructose operon repressor</td>
</tr>
<tr>
<td>Hns</td>
<td>16129198</td>
<td>26</td>
<td>14.29</td>
<td>General regulator</td>
</tr>
<tr>
<td>PurR</td>
<td>16129616</td>
<td>29</td>
<td>33.14</td>
<td>Purine biosynthesis repressor</td>
</tr>
<tr>
<td>Fis</td>
<td>16131149</td>
<td>34</td>
<td>28.00</td>
<td>Regulator of rRNA and tRNA operons</td>
</tr>
<tr>
<td>Lrp</td>
<td>16128856</td>
<td>53</td>
<td>44.57</td>
<td>Leucine responsive regulatory protein</td>
</tr>
<tr>
<td>NarL</td>
<td>16129184</td>
<td>66</td>
<td>36.00</td>
<td>Regulator of anaerobic respiration</td>
</tr>
<tr>
<td>ArcA</td>
<td>16132218</td>
<td>70</td>
<td>40.00</td>
<td>Regulator of aerobic respiration</td>
</tr>
<tr>
<td>Ihf</td>
<td>16129668</td>
<td>96</td>
<td>37.71</td>
<td>General regulator</td>
</tr>
<tr>
<td>Fnr</td>
<td>16129295</td>
<td>110</td>
<td>49.14</td>
<td>Transcriptional regulator of aerobic, anaerobic respiration and osmotic balance</td>
</tr>
</tbody>
</table>

* 2CST, two-component signal transduction system.

The binding affinity and specificity of a transcription factor and its target site can be affected by relatively small changes in the DNA-binding interface of the transcription factor, or in the binding site.57,58 As a result, DNA-binding domains could evolve new target sites relatively easily, resulting in rapid de novo emergence of new transcriptional interactions. This could be another reason for the high level of variability of transcriptional regulatory interactions in evolution. Protein–protein interactions, on the other hand, involve larger interfaces and hence more mutations are needed to alter them. Indeed, Maslov et al.59 have shown that paralogous proteins in yeast that participate in transcriptional regulatory interactions evolve new interactions rapidly when compared to paralogous proteins that participate in protein–protein interactions.

Scale-free structure emerges independently during evolution

Our analysis of the reconstructed networks reveals that even though particular regulatory hubs may be lost or possibly replaced, as shown in Figure 7(c) for Haemophilus influenzae and Bordetella pertussis, the distribution of outgoing connectivity is still best approximated by a power-law function. It is evident from Figure 7(a) that the exponent (y in the fit $y = ax^{-g}$) increases in genomes where the fraction of conserved nodes is less than 60%, and this is statistically significant (Supplementary Data M7). This implies that a hierarchical structure, with a connectivity distribution approximated by a power law, is still retained in poorly conserved networks (Figure 6(c)). The fact that the reconstructed networks are still power-law-like in their distribution, but with a different exponent value when compared to random networks of a similar size, suggests selection for different proteins to be regulatory hubs. This may result in the independent emergence of new network structures that resemble scale-free networks. However, for phylogenetically close genomes that have conserved more than 60% of the genes, the values of the exponent are similar to that of the ancestral network (Figure 7(a)).

To further investigate the prevalence of network topology resembling a scale-free structure in other bacteria, we analysed the extent to which the structure of the regulatory network changes in B. subtilis. For this, we compiled information about the known transcriptional regulatory network of B. subtilis from DBTBS.60 By comparing it with the E. coli network, we found that even though they share a similar set of target genes, the individual transcriptional factors have been replaced to
Figure 8. Comparison of the known *B. subtilis* and the *E. coli* transcriptional regulatory network. Regulatory hubs for each network are shown separately. The number in parentheses represents the number of target genes for the transcription factor. Relevant information and the outgoing connectivity distribution are given below each network. (a) The *B. subtilis* network has a scale-free structure with CcpA as its major regulatory hub. The closest hit in *E. coli* is the CytR protein (cytidine repressor belonging to the LacI family of repressors), which has only ten target genes. CcpA is the major carbon catabolite repressor protein that interacts with its co-repressor Hpr-P protein and controls genes involved in carbon metabolism. (b) The *E. coli* network has Crp as its major regulatory hub and it has a scale-free structure. Crp is a functional analogue of the CcpA protein in *B. subtilis*. It regulates genes involved in carbon metabolism but is not evolutionarily related to CcpA. There is no Crp orthologue in *B. subtilis* and this organism does not sense cAMP like *E. coli*. The closest match to Crp from *E. coli* in *B. subtilis* is Fnr, which is known to regulate ten target genes in *B. subtilis*. Thus, we see how two organisms with the ability to carry out specific carbon metabolism have evolved their own set of regulatory hubs and maintained the scale-free structure. This supports the argument that the scale-free-like structure evolved independently in the two genomes.
Figure 9. Evolution of network structure at three levels of organization. Three levels of organization of network structure. (a) At the level of genes and interactions, transcription factors tend to evolve more rapidly than their target genes. This, coupled with the observation that different genomes evolve their own transcription factors, may mean that they sense and respond to different signals in their changing environments. At the level of regulatory interactions, organisms with similar lifestyles conserve similar regulatory interactions, indicating the strong influence of environment on gene regulation. (b) At the local level, interactions in motifs evolve like any other interaction in the network. By losing or gaining TFs, orthologous target genes can come under different motifs. Organisms with similar lifestyle conserve similar motifs. (c) At the global level, regulatory hubs that are lifestyle-specific are lost as rapidly as other transcription factors in the network. New regulatory interactions could also evolve due to the relative ease of diversification of DNA-protein contacts. Scale-free structure emerged independently in evolution with different TFs emerging as hubs as dictated by lifestyle.

**Network evolution at three distinct levels of organization**

TFs evolve rapidly and independently of the TGs

Motifs do not evolve as rigid units

Condition-specific regulatory hubs may be lost or replaced due to changing environments

Organisms evolve new TFs possibly to sense new signals in changing environments

By losing or gaining TFs, orthologous target genes can come under different motifs

New regulatory interactions could also evolve due to the relative ease of diversification of DNA-protein contacts

Organisms with similar lifestyle conserve similar interactions

Organisms with similar lifestyle conserve similar motifs

Scale-free structure emerged independently in evolution with different TFs emerging as hubs as dictated by lifestyle.

All observations suggest that transcriptional regulatory networks in prokaryotes are very flexible and adapt rapidly to changes in environment by tinkering individual interactions to arrive at an optimal design.
a considerable extent. However, the overall power law-like distribution in the outgoing connectivity of the network has been maintained in the *B. subtilis* network. This comparison revealed that different proteins act as regulatory hubs in the two genomes. For example, CcpA and Crp are the two regulatory hubs in *B. subtilis* and *E. coli*, respectively controlling many genes involved in carbon metabolism. Both have very different modes of regulation and are not evolutionarily related, suggesting independent innovation of regulatory hubs to regulate orthologous target genes. We find also that proteins that are regulatory hubs in the *E. coli* reference network are either absent or regulate very few target genes in other organisms. These observations provide additional support for the suggestion that the hierarchical structure of these networks has converged to an architecture similar to scale-free networks, albeit with independently recruited regulatory hubs (Figure 8).

**Conclusions**

We present the first comprehensive analysis of the evolution of transcriptional regulatory networks at three distinct levels of organization by comparing the conservation of an experimentally established reference network of 1295 interactions across 175 microbial genomes (computationally analyzing ~500,000 protein sequences). At the level of individual genes, we show that target genes are more conserved across genomes than transcription factors, and the conservation of a target gene and its transcription factor are uncoupled, unlike the tight correlation of conservation patterns in physically interacting protein pairs. Each organism has evolved its own set of transcription factors, suggesting that a major factor in adaptation to new environment is the emergence of distinct repertoires of transcription factors, which probably integrate new inputs (Figure 9(a)). At the local level, there is no preferential conservation of interactions within network motifs, or of whole network motifs. However, it appears that by losing or conserving specific transcriptional regulators, orthologous genes in different genomes can be incorporated within different regulatory contexts and can thereby easily exhibit different patterns of gene expression. We find also that organisms with similar lifestyle have similar motif content and interactions in their conserved networks (Figure 9(b)). Thus, natural selection appears to tinker with individual interactions to arrive at an optimal design for a given organism. At the level of global network topology, we see that conservation of transcription factors is independent of the number of target genes they regulate, and depends on the lifestyle of the organism rather than the phylogenetic distance from *E. coli* (Figure 9(c)). Additionally, a comparison of the known regulatory networks of *E. coli* and *B. subtilis* confirms the above observations and reveals that different proteins act as regulatory hubs in the two genomes. This, coupled with our observation that the reconstructed networks have different power law exponents, suggests that the overall tendency towards a scale-free like behaviour is an emergent property in evolution.

With advancement in large-scale experimental studies to identify transcriptional regulatory networks, we believe that the general methods we present here will be useful in studying any set of networks and genomes. These results and predictions can serve as a scaffold for experimental studies on transcriptional control in poorly characterized genomes, and could be relevant for designing ChIP-chip experiments for pathogens and in engineering of regulatory interactions for organisms with biotechnological value.

**Materials and Methods**

Detailed descriptions of the methods are given in the Supplementary Data.

**Dataset**

The *E. coli* transcriptional regulatory network consisting of 112 transcription factors, 711 target genes and 1295 regulatory interactions was assembled from the RegulonDB,‡ and from data in the literature.† Protein sequences for the 176 completely sequenced organisms were downloaded from the NCBI website. The expression dataset for benchmarking predicted regulatory interactions was downloaded from the Stanford Microarray Database. The known transcriptional regulatory network for *B. subtilis* was obtained from DRTBSS. Information about lifestyle for the organisms was obtained from the NCBI genome information website and from the Bergeys manual of Bacteriology.

**Identification and analysis of conserved transcriptional regulatory networks**

The *E. coli* regulatory network was used as the reference network to study network evolution in the 175 other organisms. The detailed algorithm to reconstruct transcriptional networks and the procedure to identify orthologous genes are available as Supplementary Data (methods M1 and M2). Validations for the interaction transfer procedure are available as Supplementary Data (S2 and S11). Network motifs were identified using standard algorithms. The algorithm to compare networks and the procedures of the relevant statistical tests to measure significance are available as Supplementary Data (M5 to M10). The procedure to compare lifestyle similarity and motif or interaction conservation for the 175 organisms is described in Supplementary Data (S12).

**Uncited Reference**

32.

‡ http://genome-www5.stanford.edu/
Acknowledgements

M.M.B. and L.A. gratefully acknowledge the Intramural research program of National Institutes of Health, USA for funding their research. M.M.B. acknowledges the MRC Laboratory of Molecular Biology, Trinity College, Cambridge, Cambridge Commonwealth Trust and the National Institute of Health Visitor Program for financial support. We thank Dr Nakai and Yuko Makita for sending us information on the B. subtilis network. We thank Drs N. Luscombe, C. Chothia, P. Ten Wolde, L. LoConte, L. M. Iyer, V. Pisupati, S. Balaji, S. Maslau and members of our groups for reading the manuscript.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2006.02.019

Supplementary methods, information and the predictions are available at: http://www.mrc-lmb.cam.ac.uk/genomes/madanm,evdy/

References


Edited by R. Ebrigt

(Received 1 December 2005; received in revised form 6 February 2006; accepted 7 February 2006)