Evolution of DNA-binding Transcription Factors and Regulatory Networks in Prokaryotes

13

Ernesto Perez-Rueda, Nancy Rivera-Gomez, Mario Alberto Martinez-Nuñez and Silvia Tenorio-Salgado

Abstract

The capabilities of organisms to contend with environmental changes depend on their repertoire of genes and their ability to regulate their expression. DNA-binding transcription factors have a fundamental role in this process, because they regulate transcription positively or negatively as a consequence of environmental signals. In this chapter we briefly describe some of the most recent findings on regulatory network evolution from the perspective of DNA-binding transcription factors. We explore diverse elements associated with the evolution of regulatory networks, such as gene duplication, where new interactions can emerge together with their upstream and downstream binding sites. The chapter is divided into sections covering the evolution of transcription factors and their domains, their evolution, and a global analysis. Hypotheses concerning a comprehensive picture of how regulatory networks have evolved in prokaryotes and the role of transcription factors in this organization are discussed.

Introduction

Brian Goodwin has suggested that organisms are more than the sum of their parts. In this regard, we cannot infer the phenotype of one organism by knowing the genes associated with it, because a change in a single gene is not enough to cause a change in the complete phenotype (Goodwin, 1994). Therefore, if we could obtain such information, we could understand how the structure is made. In this direction, organismal development is intimately related to genetic regulation since, for example, organisms or metabolic responses require the concerted action of many regulatory proteins. von Hippel (1998) described in an elegant manuscript that 'regulatory mechanisms developed in all organisms appear to be almost infinite in number, but the basic principles on which they operate are relatively few.' It is plausible that the regulatory elements described in all the organisms change depending on their context; however, they will act in a similar fashion to allow or block gene expression.

In all organisms, it is well known that gene regulation occurs predominantly at the level of transcription initiation, and transcription factors (TFs) play an important role, because they determine when a gene is expressed or repressed, according to the environmental conditions (Martinez-Antonio *et al.*, 2006). Given the importance of this kind of proteins, many authors have evaluated their presence and abundance in diverse organisms. From these studies, it has been observed that the number of TFs increases from a few hundred in archaea and bacteria, such as *Pyrococcus horikoshii*, *Bacillus subtilis* and/or *Escherichia coli* K12, to over 3000 in *Homo sapiens* (Levine and Tjian, 2003; Perez-Rueda *et al.*, 2004; Perez-Rueda and Janga, 2010). This increment correlates with the hypothesis of genome maturation, (Lane and Martin, 2010) where it is proposed that it is necessary for a greater number of regulatory elements to regulate a greater number of genes. Consequently, the number of genetic circuits or regulatory networks that arises also increases (Bhardwaj *et al.*, 2010). Therefore, minor changes in single genes may propagate along such networks and may produce, in the end, quite drastic effects on gene expression in response to external stimuli and change related to development. But how do these changes occur, considering that cellular differentiation, for example, sporulation, requires the concerted interplay between sigma factors, TFs, and their binding sites?

In this chapter we summarize some of the most recent insights from studies on regulatory network evolution from the perspective of DNA-binding TFs, considering that the evolution of regulatory networks requires at least two main mechanisms, gene duplication and gene transfer. Thus, new interactions may emerge together with their upstream and downstream binding sites. We break the subject into sections, covering the evolution of TFs and their domains, the promoters, their evolution, and a global analysis. We finish with some conjectures that attempt to provide a comprehensive picture about how regulatory networks have evolved in prokaryotes and the role of TFs in this organization.

Elements involved in the regulatory process

The regulation of transcription initiation in bacteria is primarily mediated by sigma (σ) factors, which provide most of the specificity for the promoter recognition and DNA melting needed for transcription initiation (Gruber and Gross, 2003; Ishihama, 2000; Wosten, 1998). Indeed, σ factors perform these functions only when bound to the RNA polymerase (RNAP). On the other hand, DNA-binding TFs (Browning and Busby, 2004) affect gene expression, in a wider context, by blocking or allowing the access of the RNAP to the promoter, depending on the operator context and ligand-binding status (Wall *et al.*, 2004; Martinez-Antonio *et al.*, 2006; Miroslavova and Busby, 2006; Janga and Collado-Vides, 2007). Usually, most of the gene transcription in exponentially growing bacteria is initiated by the RNAP carrying a housekeeping σ factor, similar to the *E. coli* σ^{70} or *B. subtilis* σ^{A}. Alternative σ factors typically redirect the transcriptional machinery or RNAP towards a subset of genes required under specific conditions, such as the stress response or growth transitions, among others (Wosten, 1998; Ishihama, 2000; Gruber and Gross, 2003). TFs represent a class of proteins devoted to sensing and binding signals to regulate the response to specific compounds (Martinez-Antonio *et al.*, 2006; Goelzer *et al.*, 2008). In Fig. 13.1 we present a simple regulatory network composed of at least three basic components:

1 the DNA-binding TF, which can be self-regulated;
2 the regulatory region on the DNA, where the

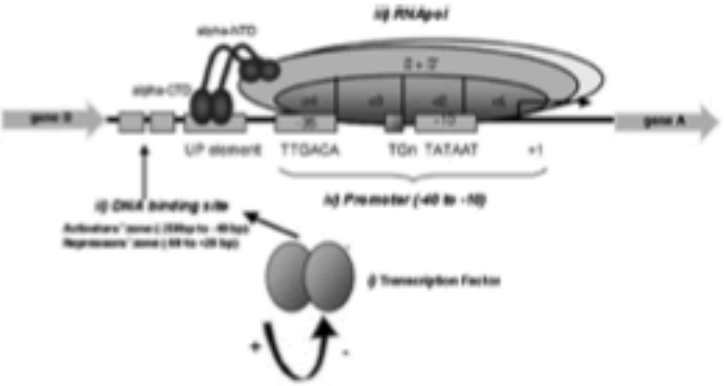

Figure 13.1 Schematic drawing of the basic construction principle of an elementary unit of a hypothetical genetic regulatory network. (i) DNA-binding TF. This protein can activate or repress gene expression. In addition, it can be positive or negatively self-regulated. (ii) A DNA-binding site, which usually is located between the −60 and +20 positions relative to the transcription start. (iii) An RNAP that consists of a protein complex necessary to start the mRNA synthesis. (iv) A sequence promoter the RNAP recognizes specifically over DNA to start the mRNA synthesis.

TF binds and by which the transcription start is modulated;

3 the promoter-binding site, where RNAP binding starts RNA synthesis.

As we will discuss in detail later in this chapter, these links are mostly 'one-to-many,' that is, each TF regulates more than one gene and most genes are controlled by some, albeit generally few, TFs. Every TF is itself regulated by another one. This combinatory perspective gives rise to regulatory networks. Each of these elements can be broken down into regulons. For example, the arabinose regulons in *E. coli* K12, composed of more than 10 different genes involved in the assimilation of arabinose is regulated by two different transcription factors, Crp and AraC (Salgado *et al.*, 2006), where Crp can be associated with a plethora of additional functions. The TF itself typically contains a DNA-binding domain and other regulatory domains, such as multimerization domains that mediate interactions with other proteins or metabolites in order to react to physiological or environmental changes.

Promoter and regulatory region

Transcription starts when the σ factor interacts with the RNAP to recognize its specific sequence promoter (Fig. 13.1). This promoter recognition stage imposes the existence of at least one σ factor per organism, which typically belongs to the σ^{70} family in bacteria (Paget and Helmann, 2003). In this context, bacterial systems could switch between different transcriptional programmes based exclusively on their repertoire of σ factors. Nonetheless, the transcriptional programmes mediated solely via σ factors would be restricted, as a result of their limited repertoire and the small collection of ligands they can recognize, such as guanosine tetraphosphate (ppGpp) (Jores and Wagner, 2003). As a consequence, σ factors exhibit a restricted ability for directly coupling responses to environmental conditions with gene transcription. In addition, σ factors have a constrained DNA-binding region in terms of the lengths and diversity of sequences they recognize, as they need to be structurally coupled to the RNAP on the promoter zone. These DNA restricted zones of action divide the universe of σ factor families into promoters recognized by σ^{70} family and those recognized by σ^{54} family, for instance, the binding zones correspond to about bp −10 to −35 for σ^{70} and bp −12 to −24 for σ^{54}, relative to the transcription start site in the bacterium *E. coli K12* (Gralla, 1996; Lloyd *et al.*, 2001).

On the other hand, TFs define a different regulatory level than do σ factors. These proteins exhibit diverse structural and functional domains, with one of them associated with binding DNA specifically, whereas the second one is devoted to sensing and binding one or more signals from endogenous and/or exogenous sources (Martinez-Antonio *et al.*, 2006). For example, *E. coli* K-12 TyrR binds to three aromatic amino acids and ATP (Pittard *et al.*, 2005); TrmB of the archaeon *Pyrococcus furiosus* binds to three different compounds (Lee *et al.*, 2003, 2005; Perez-Rueda and Janga, 2010). In addition, TFs have the ability to associate combinatorially not only with σ factors but also with a number of other TFs and DNA-binding sites (Adhya, 2003; Barnard *et al.*, 2004), thus allowing the rewiring of a transcriptional network depending on the environmental conditions. For instance, *sodA*, a gene encoding superoxide dismutase in *E. coli*, is regulated by up to eight different TFs responsible for various cellular responses, including Fur (ferric uptake regulation protein), Arc (aerobic respiratory control), and Fnr (fumarate nitrate reduction/regulator of anaerobic respiration) (Compan and Touati, 1993; Salgado *et al.*, 2006). Therefore, the diversity of sequences recognized by TFs is enormous and can occur anywhere from a few bases downstream of the promoter zone to up to hundreds of bases upstream of the transcription start site (Fig. 13.1) (Collado-Vides *et al.*, 1991; Madan Babu and Teichmann, 2003b). For instance, the *E. coli* K-12 global regulator CRP (catabolic repressor protein) can associate with four of the six possible σ factors and coregulate more than 50 different TFs (Salgado *et al.*, 2006). CcpA in *B. subtilis*, a global regulatory protein involved in catabolite repression, may act as a positive regulator of genes involved in excretion of carbon excess and can associate with three different sigma factors (σ^{A}, σ^{L} and σ^{E}) and more than 10 different TFs (Makita *et al.*, 2004;

Moreno-Campuzano *et al.*, 2006). In summary, TFs constitute a class of proteins whose space of action is more flexible than σ factors, not only in sensing diverse environmental and endogenous stimuli but also in recognizing a wide range of binding-site sequences over a larger zone on the DNA around the transcription start site.

Evolution of the repertoire of TFs

DNA-binding domains

The structures of more than 30 prokaryotic DNA-binding proteins have now been determined, and hundreds of amino acid sequences are known for many more. In general, the DNA-binding domains associated with TFs are among the most ancient domains, and they have been proposed as derived from a relatively small set of folds (Aravind and Koonin, 1999; Perez-Rueda and Collado-Vides, 2001; Madan Babu and Teichmann, 2003). These domains have been used to classify the TFs in terms of families (Perez-Rueda *et al.*, 2004). From these studies diverse principles have emerged regarding TFs; for example, the most abundant DNA-binding domain in prokaryotes is the helix–turn–helix (HTH), identified in more than 80% of TFs (Perez-Rueda and Collado-Vides, 2001; Perez-Rueda and Janga, 2011) (Fig. 13.2). Most of the archaeal and bacterial HTHs appear to have undergone a common general evolutionary pathway. The HTH might then have been a motif with general nucleic acid-binding functions that appeared early in evolution (Aravind and Koonin, 1999; Roy *et al.*, 2002) and whose subsequent radiation in the archaeal and bacterial lineages might have involved considerable loss and acquisition events. Additionally, lineage-specific duplications resulted in the accumulation of particular families in microbial species, such as the LysR family, whose members have been abundantly identified in almost all organisms. This hypothesis is consistent with the notion that a genome evolves from a set of precursor genes to a mature size by gene duplications and increasing modifications (Yanai *et al.*, 2000). Alternative DNA-binding structures, such as helix–loop–helix motifs, zinc-fingers, and σ-sheet DNA-binding structures, have been also identified, although in lower proportions, and their distributions are constrained to specific organisms. For instance, the β-sheet proteins have been identified almost exclusively in *Gammaproteobacteria*. The distribution of the RNA-binding domains (associated with cold shock proteins) suggests that they might have been acquired after the prokaryotes and eukaryotes split, probably by lateral gene transfer from eukaryotes, based on the high diversity identified in this cellular domain.

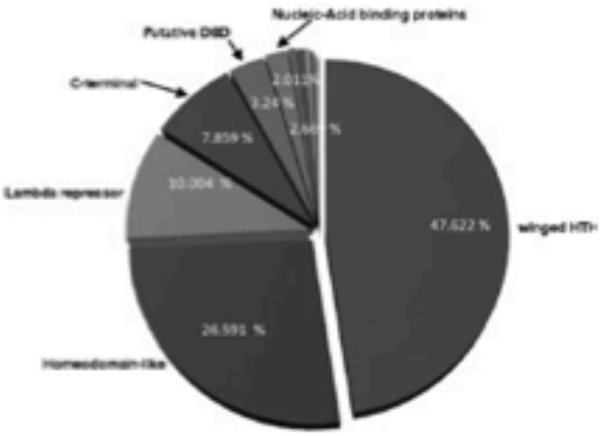

Figure 13.2 Distribution of DNA-binding domains of TFs in bacteria and archaea as defined in the Superfamily database (Madera *et al.*, 2004). The winged HTH represents 47.6% of the total repertoire of DNA-binding domains, being the most abundant structure. The lambda-repressor DNA-binding domain is present at the second highest abundance, as 26% of the repertoire. In minor proportions occur alternative DNA-binding domains, such as the homeodomain-like, with 10.0%, the C-terminal effector domain of the bipartite response regulators, at 7.8%, the putative DNA-binding domain, at 3.24%, and the nucleic acid-binding proteins, at 2.01%. In a low fraction, corresponding to 2.6% of the total DNA-binding domains, are AbrB/MazE/MraZ-like, KorB DNA-binding domain-like, TrpR-like, ACT-like, flagellar transcriptional activator FlhD, haemolysin expression-modulating protein H, a DNA-binding domain in eukaryotic TFs, DNA-binding domain, p53-like TFs, and the DNA-binding domain of the Mlu1 box-binding protein MBP1.

Abundance of TFs correlates with genome size in prokaryotes

It has been documented that organisms respond and adapt to diverse environmental conditions as a consequence of their gene repertoire and regulatory mechanisms, among other elements

(Lynch and Conery, 2003; Bengtsson, 2004; Lynch, 2006). Recent studies have shown that the evolutionary events associated with regulatory proteins, such as their expansion and contraction, contribute significantly in shaping the gene repertoire and genome size of the different lineages of prokaryotes (Perez-Rueda *et al.*, 2004; Minezaki *et al.*, 2005; Oguiza *et al.*, 2005; Rodionov, 2007). Based on comparative genomics, it has been shown that transcription factors increase in quadratic proportion with respect to genome size (van Nimwegen, 2003; Cordero and Hogeweg, 2007; Molina and van Nimwegen, 2008). In particular, this proportion is more significant when the repertoire of TFs is compared with the proportion of σ factors, being roughly ten times higher (hundreds of TFs vs. tens of σ factors) when the general profiles in all the genomes analysed are considered, suggesting a proportion on the order of 1 σ factor:10 TFs:100 annotated open reading frames per genome, although some genomes behave exceptionally. This observation suggests that a possible functional relationship between TFs and prokaryote lifestyles influences the observed trend. A plausible hypothesis is that the abundance of TFs increases with an increase in an organisms' complexity (Brown *et al.*, 2002; Changizi, 2001; Levine and Tjian, 2003; van Nimwegen, 2003; West and Brown, 2005) as a consequence of different evolutionary events, such as gene expansion, gene loss, and lateral gene transfer, among others (Levine and Tjian, 2003; Aravind *et al.*, 2005; Madan Babu *et al.*, 2006). In addition, the necessity to regulate responses to variable environments could also contribute to the abundance of TFs.

Lifestyles explain the abundance of σ factors and TFs in larger genomes

In previous sections we suggested that regulatory complexity should increase in larger genomes and might be associated with bacterial lifestyles, as the environment should influence the bacterial genome structure and function. Thus, to understand how the complexity of gene regulation depends on the number of TFs as a function of increasing genome size and how they are associated with the lifestyles, in previous work (Perez-Rueda *et al.*, 2009) we classified in four global lifestyle classes all the bacterial organisms. These included extremophiles, intracellular bacteria, pathogens, and free-living bacteria. From this analysis, it was identified that the increment of regulatory complexity in TFs contributes significantly to the regulatory complexity of prokaryotes belonging to different lifestyle groups. These results agree with previous observations that suggest that a few regulatory elements identified in small genomes would compensate for the regulation of the entire genome with an increase in the number of DNA-binding sites per element, in contrast to the large number of elements identified in large genomes, which control a smaller proportion of DNA-binding sites on average (Molina and van Nimwegen, 2008). In addition, a larger proportion of genes in small genomes are organized in operons, simplifying the transcriptional machinery necessary for gene expression, in contrast to large genomes, which have reduced number of genes in operons, which would influence the proportion of TFs in those organisms (Cherry, 2003), suggesting that complex lifestyles require a higher proportion of TFs and transcription units to better orchestrate a response to changing conditions.

Abundance of TFs does not correlate with diversity of families, and large families are not the most widely distributed

An appealing hypothesis is that the high diversity of TF families contributes significantly to the regulatory plasticity. In line with this hypothesis, the repertoire of TFs identified in bacteria has been classified into families to evaluate their distribution and abundance in all the prokaryotes. This analysis showed a reduced diversity of families in small genomes, with an increasing proportion in larger ones, especially in pathogens and free-living organisms. The diversity of families reaches a maximum in genomes with around 5000 open reading frames. The higher number of TFs in larger genomes does not necessarily imply diversity of families beyond this plateau, but instead an increase in the size of some families of TFs. Congruent with this observation, the average number of TFs per family increases linearly, with a few families of TFs expanding disproportionately (Janga and Perez-Rueda, 2009; Perez-Rueda *et* *al.*, 2009). These families comprise LysR and TetR, which represent about 25% of the total set of TFs identified. Members of these two families increase abruptly in larger genomes and coincide with the plateauing of the diversity of families in these bacterial genomes. Another feature associated with large families is that they are not widely distributed among bacteria despite their role in controlling important processes, such as cell–cell communication (LuxR), response to external conditions by two-component systems (OmpR), sensing, uptake, and metabolism of external food sources (GntR and LysR), and antibiotic resistance (TetR). Alternatively, families with few copies per genome, such as DnaA, LexA, and IHF, which have been proposed to be essential under standard growth conditions in *E. coli* and in maintaining DNA and nucleoid integrity, (Gerdes *et al.*, 2003; Yamazaki *et al.*, 2008) might be considered universal in bacteria, because they have been identified in at least 80% of the genomes, suggesting gene loss events in bacteria in which they are absent.

In summary, a TF family's abundance and distribution should be associated with the following evolutionary events in bacteria: (i) small families widely distributed among bacteria might be related to ancestral functions beyond transcriptional regulation, such as DNA organization, nucleoid integrity, or DNA salvage; (ii) large families might be associated with the regulation of dispensable or emergent processes in bacterial evolution, such as those involved in quorum sensing, belonging to the members of the LuxR family, which are widely identified in bacteria. Indeed, the evolution of this mechanism in bacteria has been proposed to be one of the early steps in the development of multicellularity (Miller and Bassler, 2001) and may be correlated with bacterial specialization.

Evolution of partner domains

DNA-binding TFs usually make contact with their DNA targets as homodimers or homotetramers, as is the case for the lactose repressor (Lewis, 2005). In this regard, there are diverse questions concerning whether multimerization precedes DNA binding. Therefore, the TFs can act as activators or repressors as a consequence of their multimerization state. Amoutzias *et al.* (2004) concluded that the ancestral TFs were probably the homodimerizing ones and that these proliferated through a series of single-gene duplications. Recent evidence supports this hypothesis, as investigators have described a high abundance of small-sized TFs, or proteins that contain a dimerization domain, but no DNA-binding domain in archaeal genomes (Perez-Rueda and Janga, 2011). The main consequence of gene duplications, mainly involving TFs, is to give rise to a complex interaction network. To the best of our knowledge, there are no studies so far that have linked genetic networks and their regulation with other networks, such as metabolism, although many effector domains are protein interaction domains. These data could help in understanding how the ligand-binding domains have been recruited to regulate gene expression. However, previous studies (Madan Babu and Teichmann, 2003; Aravind *et al.*, 2005) suggested that in the winged HTH superfamily of TFs, the partner domains contribute to the structural differentiation of duplicated genes. Therefore, the partner domains are associated with diverse functions, such as regulating allosterically the function of TFs across binding to a wide variety of functional compounds, in protein–protein interactions, or with enzymatic properties (Madan Babu and Teichmann, 2003), and they are fundamental to linking environmental conditions and the functional conformational changes in the regulators (Taraban *et al.*, 2008). Because there are few exhaustive analyses describing the partner domain repertoire in bacteria, their functional and evolutionary diversity must be evaluated (Madan Babu and Teichmann, 2003; Rivera-Gomez *et al.*, 2011). From this perspective, one might expect that the high diversity of TF families and their associated partner domains contributes significantly to the regulatory plasticity, as we previously mentioned; however, further studies are necessary. Thus, the diversity associated with dimerization domains has allowed the TFs to interact with many different partners and achieve specificity for the target gene and physiological conditions. On the other hand, multimerization domains may have been facilitated by the acquisition of a second protein interaction domain during evolution (Kaufmann *et al.*, 2005).

Evolution of promoters

Unlike coding regions, the evolution of upstream regulatory regions cannot be easily evaluated by means of sequence alignments and comparisons. Indeed, the structure, organization, and function of promoters differ widely from that of the coding sequences, resulting in totally different evolution rates and consequences of sequence variation (Rodriguez-Trelles *et al.*, 2003). In general, promoter organization underlies more variation than the coding sequences, and it is mainly influenced by DNA structure (Olivares-Zavaleta *et al.*, 2006) such as supercoiling (Martinez-Nunez *et al.*, 2010). Bacterial promoter sequences, or upstream regulatory regions, contain TF-binding sites. Several TFs often interact, depending on signals and physiological or environmental requirements. This gives rise to activating or repressing complexes (Koike *et al.*, 2004). Ishihama described that among the set of promoters under the control of the same σ factor, the level of transcription varies depending on the culture conditions or the growth phase. For that reason, it is important to evaluate the diversity of promoters and their regulation associated with all genes in an organism, for instance, in *E. coli* K-12, approximately 50% of the promoters are under the control of one specific regulator, whereas the other 50% of genes are regulated by more than two TFs. Likewise, the promoters for the genes involved in the construction of cell structures are controlled by environmental conditions and specific signals, and each is in turn monitored by a different TF. The binding sites of all these multiple factors are located in a single promoter. The most typical examples of the multifactor promoter system are the promoters for the genes encoding the master regulator for flagellum formation in *E. coli* K12, FlhCD, and the master regulator for biofilm formation, CsgD. The complexities of these promoters reflect the opposite behaviours of bacterial survival, i.e. planktonic growth as single cells (FlhCD) and biofilm formation as a bacterial community under stressful conditions in nature (Soutourina *et al.*, 1999; Ogasawara *et al.*, 2010).

Coevolution of regulatory elements

Since some proteins tend to work together in a functional context, analyses of distributions of different families in the function of the σ factors have been recently performed. Hence, the cooccurrence of the regulatory protein families (TFs and σ factors) in all the prokaryotes was evaluated. From this analysis, it was found that the distribution of the σ^{54} factor and IHF and EBP families are correlated (Fig. 13.3), supporting the functional interdependence discussed above and probable coevolution in which members and mechanisms have been preserved along the course of evolution in bacteria. A second cluster that includes σ^{70}, the ECF family of σ factors, and other highly abundant families (more than 15 members per genome) responsible for regulating diverse mechanisms of stress responses (MarR), antibiotic resistance (TetR), osmotic responses (OmpR), and the quorum-sensing response (LuxR), among other

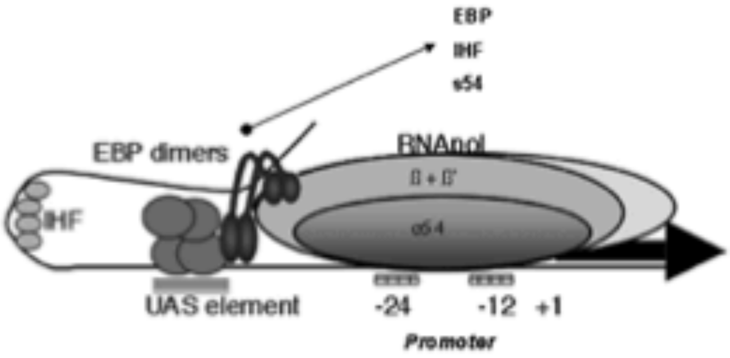

Figure 13.3 Coevolution of σ factors and TF families. A similar occurrence distribution pattern was observed for IHF, EBP, and σ^{54} families, suggesting a functional interdependence between these families. A coregulatory mode of action for these regulatory proteins is also shown. Figure modified from Perez-Rueda *et al.* (2009).

processes, was also found to be clustered, suggesting a strong functional relationship among these σ and TF families. These clusters, in addition, give insights into the functional interdependence between regulatory proteins from different families, which could help in the characterization of regulators in poorly studied organisms.

Evolution of regulatory networks

The regulation of TFs plays a key role in morphological diversity. Simple modifications within the upstream regulation region of a TF can explain both minor and major changes between species, without involving any disruption of gene structure. Therefore, evolution of regulatory regions is thought to be a major source of diversity. (Lozada-Chavez *et al.*, 2008; Perez and Groisman, 2009a,b) Duplication events of TFs are another evolutionary source that can allow diversity, permitting a more versatile adaptation of the functional divergence gained from the duplication of structural genes. Different aspects of the evolution of the regulatory networks have been examined, including the coevolution of the upstream regulatory regions and their corresponding TFs, the likely consequences of gain, loss, and replacement of TFs in the regulatory networks of duplicated genes (Teichmann and Babu, 2004; Gelfand, 2006), and also the topological and dynamic properties of the regulatory networks (Luscombe *et al.*, 2004; Madan Babu *et al.*, 2006; Balaji *et al.*, 2007).

Role of duplication events in regulatory networks

Duplication events of regulatory genes provide new interactions for the transcriptional regulatory network. These new interactions are the raw material for the generation of divergence in gene expression, which should happen in most of the copies that have remained in the genome (Teichmann and Babu, 2004). In this model, the loss and gain of regulatory interactions may occur following the duplication of either a TF or a target gene or following the duplication of both a TF and a target gene (Fig. 13.4). The evolutionary plasticity of the regulatory networks is not only the result of the duplication of TF interactions within a regulatory network, as previously proposed by Teichmann and Babu (2004), but also the result of the divergent effects of the TF interactions in activating or repressing the transcription of duplicated genes, as suggested by Martinez-Nuñez *et al.* (2010). Indeed, examples have been recently identified of regulatory systems where the TF is maintained but a different regulatory role is gained (either activation or repression) in one of the duplicated genes. This evolutionary scenario can be observed in the regulation of the *E. coli gntK* and *idnK* gluconate kinase genes, which are involved in 6-phosphogluconate synthesis in the Entner–Doudoroff and pentose phosphate pathways, respectively. Although the same TFs, CRP, GntR, and IdnR, regulate all these genes, IdnR represses the transcription of *gntK*, whereas it activates the transcription of *idnK* (Bausch *et al.*, 2004; Salgado *et al.*, 2006).

This form of diversification in regulation allows plasticity of the transcriptional regulatory network without the need to increase the number of interactions within it, if not only by varying the type of regulation (positive or negative) exerted by the TFs on their targets. It is possible that modulation is one of the first steps towards evolutionary innovation at a biochemical level, perhaps as a step towards the modification of the entire metabolic pathway (Martinez-Nunez *et al.*, 2010).

Concluding remarks

As a consequence of the abundance of data which have become recently available, the idea has emerged to conceptualize genetic networks as directed graphs with nodes corresponding to the TFs, linked by edges to their target genes. The previously mentioned regulatory elements (TFs, σ factors, upstream binding sites, downstream binding sites, and promoters) can be combined at a higher level, into so–called network motifs (Milo *et al.*, 2002; Shen-Orr *et al.*, 2002). The impacts of evolutionary forces on the topological structures of regulatory networks known as motifs have been exhaustively analysed by diverse authors, for example, biological regulatory networks (Milo *et al.*, 2002; Shen-Orr *et al.*, 2002) and duplicated gene networks (Teichmann and Babu, 2004). In

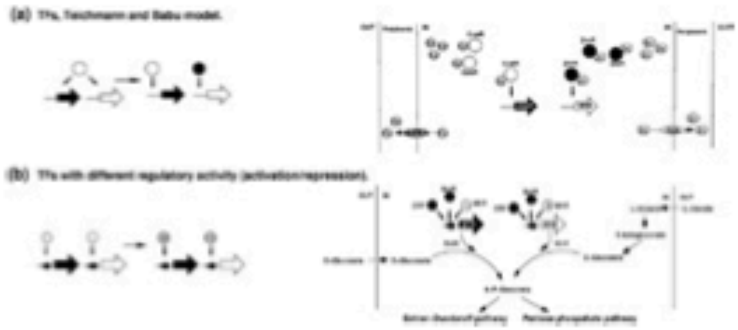
(a) TFs, Teichmann and Babu model.
(b) TFs with different regulatory activity (activation/repression).

Figure 13.4 Regulatory elements involved in the regulation of the duplicated genes. (a) Simplification of the model of Teichmann and Babu (Teichmann and Babu, 2004), where a new TF is gained to regulate one of the duplicated genes. TFs are shown as circles. (b) Extension of the Teichmann and Babu model. The differential regulation of the duplicated genes depends on the activation or repression mechanism associated with the TF on its target genes. Figure modified from Martínez-Núñez *et al.* (2010).

these studies, the authors reported that duplication of an entirely feed-forward motif (a topological structure in which a TF regulates a second TF and both TFs simultaneously regulate a target gene) has not been observed in the regulatory networks of model organisms, although single genes generated by duplication could be part of a new feed-forward or other kind of motif. For example, in *B. subtilis* the duplicated σ^F and σ^E are part of different feed-forward motifs. In the first case, σ^F forms a feed-forward motif with the anti-anti-σ factor SpoIIAA and the anti-σ factor SpoIIAB, as σ^F regulates SpoIIAA and SpoIIAB expression, whereas SpoIIAA modulates the expression of SpoIIAB. In a similar manner, the second feed-forward loop is formed by σ^E, PhoP, and PhoR, which are involved in phosphate uptake, the post-exponential growth phase, and other stress responses (Pragai *et al.*, 2004). Duplication events, or mutations in the duplicated regulatory genes or in the regulatory target sites, can generate new feed-forward motifs useful for the rewiring of the regulatory networks of the duplicated genes, which would favour the adaptation process of the organism as it responds to changes in its niche (Gelfand, 2006). Probably the most basic motif is the autoregulatory loop: a TF that regulates its own expression (Fig. 13.1).

Generally, these motifs have been considered basic architectures in the regulatory networks, because they often overlap (Browning and Busby, 2004). In this context, diverse authors have analysed the structure of both genetic and protein–protein interaction networks (Yeger-Lotem and Margalit, 2003; Yeger-Lotem *et al.*, 2004). These analyses have shown that certain topologies of small subnets are statistically very much over-represented (Shen-Orr *et al.*, 2002). Conant and Wagner introduced the notion of common ancestry for gene circuits or motifs, where two motifs share a common ancestor if every pair of genes in the two circuits is derived from a common ancestor; all pairs in the circuits must be duplicated genes (Conant and Wagner, 2003). They found that no pairs of motifs with identical topology had common ancestry, and they concluded that their emergence is the result of convergent evolution and not duplication of one or a few ancestral circuits, suggesting that convergent evolution was more likely to be important in module topology than for protein sequences.

A third level of network organization consists of transcriptional modules (Babu *et al.*, 2004). Modules represent collections of TFs that are expressed under distinct experimental or environmental conditions (Ihmels *et al.*, 2002) and are largely controlled by one (or very few) regulators, as was shown by hierarchical clustering of expression profiles (Segal *et al.*, 2003). For instance, the *E. coli* K-12 global regulator CRP can regulate the expression of more than 20 different TFs (Thieffry *et al.*, 1998; Gama-Castro *et al.*, 2008). Exhaustive analyses of global features of genetic networks and protein interaction networks have revealed a scale-free topology (Barabasi and Albert, 1999; Barabasi and Oltvai, 2004); in other words, there are few genes, or so-called hubs, that control many others, and many genes have only a few links. These hubs can be defined as global regulators, such as Crp *in E. coli* K-12. However, regulatory networks are dynamic: it was shown that large-scale topological changes have occurred in the *E. coli*, *B. subtilis*, and *Saccharomyces cerevisiae* genomes and that although a few TFs serve as permanent hubs, they act transiently only under certain conditions (Luscombe *et al.*, 2004). It is also worth noting that more complex organisms have a higher number of regulatory genes per target gene (van Nimwegen, 2003) suggesting that it is mostly the evolution by duplication and diversification of transcription factors and of their interactions that increases organismic complexity as a whole. Over the last few years, the availability of large numbers of experimental and theoretical data has significantly enhanced our understanding of evolution of complex networks and, at the same time, enabled us to transfer knowledge from better-studied model organisms (such as *E. coli* and *B. subtilis*) to those for which fewer data are available. Basically, the ancestral genetic networks we observe today were probably a small group of DNA-binding domains that, while conserving their structure, diverged into a large variety of TFs. More recently, most proteins, among them TFs, underwent many cycles of domain rearrangements (Amoutzias *et al.*, 2005). Additional dimerization and sensor domains were gained and lost at different times. Further, they evolved across a series of single-gene duplications, thus generating networks of regulatory genes that arrange into these modules. These events may be quite recent and lineage specific, as we have learned from the uneven distribution of some TF families (Perez-Rueda *et al.*, 2004). A growing number of findings suggest that structurally similar or even identical motifs can arise repeatedly and thus represent a simple level of convergent evolution. More complex modules, which may also have preferentially arisen through a series of single-gene duplications, would give rise to similar topologies. The evolution of promoter regions is less well understood, although it is of great importance. Genotypic changes at this level are probably among the main reasons why, despite minor interorganismic differences at the level of proteins, major changes in the topologies of genetic networks during development induce wide morphological differences and diversity in contemporary organisms.

In summary the mechanisms of generating diverse networks can be associated with diverse evolutionary forces, such as gene duplication, gene loss, changes in the regulatory mechanisms (regulatory role modulation), acquisition of new activities, modular rearrangements, and finally, functional divergence. Therefore, we believe that with the availability of more information, we will be able to understand in a more comprehensive fashion the evolutionary dynamics associated with regulatory networks.

Acknowledgements

We gratefully acknowledge to Alejandro Huerta-Saquero, Antonio Ibarra, and the anonymous referee for their helpful comments in the preparation of the manuscript. This work was supported by DGAPA grant IN209511 to EP-R. NR-G, MAM-N, and ST-S were financed by a PhD fellowship from CONACyT.

References

Amoutzias, G.D., Robertson, D.L., and Bornberg-Bauer, E. (2004). The evolution of protein interaction networks in regulatory proteins. Comp. Funct. Genomics *5*, 79–84.

Amoutzias, G.D., Weiner, J., and Bornberg-Bauer, E. (2005). Phylogenetic profiling of protein interaction networks in eukaryotic transcription factors reveals focal proteins being ancestral to hubs. Gene *347*, 247–253.

Aravind, L., and Koonin, E.V. (1999). DNA-binding proteins and evolution of transcription regulation in the archaea. Nucleic Acids Res. *27*, 4658–4670.

Aravind, L., Anantharaman, V., Balaji, S., Babu, M.M., and Iyer, L.M. (2005). The many faces of the helix–turn–helix domain: transcription regulation and beyond. FEMS Microbiol. Rev. *29*, 231–262.

Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M., and Teichmann, S.A. (2004). Structure and evolution of transcriptional regulatory networks. Curr. Opin. Struct. Biol. *14*, 283–291.

Balaji, S., Babu, M.M., and Aravind, L. (2007). Interplay between network structures, regulatory modes and sensing mechanisms of transcription factors in the transcriptional regulatory network of *E. coli*. J. Mol. Biol. 372, 1108–1122.

Barabasi, A.L., and Albert, R. (1999). Emergence of scaling in random networks. Science *286*, 509–512.

Barabasi, A.L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nat. Rev. Genet. *5*, 101–113.

Bausch, C., Ramsey, M., and Conway, T. (2004). Transcriptional organization and regulation of the L-idonic acid pathway (GntII system) in *Escherichia coli*. J. Bacteriol. *186*, 1388–1397.

Bengtsson, B.O. (2004). Modelling the evolution of genomes with integrated external and internal functions. J. Theor. Biol. *231*, 271–278.

Bhardwaj, N., Carson, M.B., Abyzov, A., Yan, K.K., Lu, H., and Gerstein, M.B. (2010). Analysis of combinatorial regulation: scaling of partnerships between regulators with the number of governed targets. PLoS Comput Biol 6, e1000755.

Brown, J.H., Gupta, V.K., Li, B.L., Milne, B.T., Restrepo, C., and West, G.B. (2002). The fractal nature of nature: power laws, ecological complexity and biodiversity. Philos. Trans. R. Soc. Lond. B Biol. Sci. *357*, 619–626.

Browning, D.F., and Busby, S.J. (2004). The regulation of bacterial transcription initiation. Nat. Rev. Microbiol. *2*, 57–65.

Changizi, M.A. (2001). Universal scaling laws for hierarchical complexity in languages, organisms, behaviors and other combinatorial systems. J. Theor. Biol. *211*, 277–295.

Cherry, J.L. (2003). Genome size and operon content. J. Theor. Biol. *221*, 401–410.

Conant, G.C., and Wagner, A. (2003). Asymmetric sequence divergence of duplicate genes. Genome Res. *13*, 2052–2058.

Cordero, O.X., and Hogeweg, P. (2007). Large changes in regulome size herald the main prokaryotic lineages. Trends Genet. 23, 488–493.

Gama-Castro, S., Jimenez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Penaloza-Spinola, M.I., Contreras-Moreira, B., Segura-Salazar, J., Muniz-Rascado, L., Martinez-Flores, I., Salgado, H., *et al.* (2008). RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. Nucleic Acids Res. *36*, D120–124.

Gelfand, M.S. (2006). Evolution of transcriptional regulatory networks in microbial genomes. Curr. Opin. Struct. Biol. *16*, 420–429.

Gerdes, S.Y., Scholle, M.D., Campbell, J.W., Balazsi, G., Ravasz, E., Daugherty, M.D., Somera, A.L., Kyrpides, N.C., Anderson, I., Gelfand, M.S., *et al.* (2003). Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. J. Bacteriol. *185*, 5673–5684. Goelzer, A., Bekkal Brikci, F., Martin-Verstraete, I., Noirot, P., Bessieres, P., Aymerich, S., and Fromion, V. (2008). Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of *Bacillus subtilis*. BMC Syst. Biol. 2, 20.

Goodwin, B. (1994). How the Leopard Changed its Spots: The Evolution of Complexity (Scribner). <Please provide place of publication>

Gralla, J.D. (1996). Activation and repression of *E. coli* promoters. Curr. Opin. Genet. Dev. *6*, 526–530.

Gruber, T.M., and Gross, C.A. (2003). Multiple sigma subunits and the partitioning of bacterial transcription space. Annu. Rev. Microbiol. *57*, 441–466.

von Hippel, P.H. (1998). An integrated model of the transcription complex in elongation, termination, and editing. Science *281*, 660–665.

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

Ishihama, A. (2000). Functional modulation of *Escherichia coli* RNA polymerase. Annu. Rev. Microbiol. *54*, 499–518.

Janga, S.C., and Collado-Vides, J. (2007). Structure and evolution of gene regulatory networks in microbial genomes. Res. Microbiol. *158*, 787–794.

Janga, S.C., and Perez-Rueda, E. (2009). Plasticity of transcriptional machinery in bacteria is increased by the repertoire of regulatory families. Comput. Biol. Chem. 33, 261–268.

Jores, L., and Wagner, R. (2003). Essential steps in the ppGpp-dependent regulation of bacterial ribosomal RNA promoters can be explained by substrate competition. J. Biol. Chem. *278*, 16834–16843.

Kaufmann, K., Melzer, R., and Theissen, G. (2005). MIKC-type MADS-domain proteins: structural modularity, protein interactions and network evolution in land plants. Gene *347*, 183–198.

Koike, H., Ishijima, S.A., Clowney, L., and Suzuki, M. (2004). The archaeal feast/famine regulatory protein: potential roles of its assembly forms for regulating transcription. Proc. Natl. Acad. Sci. U.S.A. *101*, 2840–2845.

Lane, N., and Martin, W. (2010). The energetics of genome complexity. Nature *467*, 929–934.

Levine, M., and Tjian, R. (2003). Transcription regulation and animal diversity. Nature 424, 147–151.

Lewis, M. (2005). The lac repressor. C. R. Biol. *328*, 521–548.

Lloyd, G., Landini, P., and Busby, S. (2001). Activation and repression of transcription initiation in bacteria. Essays Biochem. *37*, 17–31.

Lozada-Chavez, I., Angarica, V.E., Collado-Vides, J., and Contreras-Moreira, B. (2008). The role of DNA-binding specificity in the evolution of bacterial regulatory networks. J. Mol. Biol. *379*, 627–643.

Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. Nature *431*, 308–312.

Lynch, M. (2006). Streamlining and simplification of microbial genome architecture. Annu. Rev. Microbiol. *60*, 327–349.

Lynch, M., and Conery, J.S. (2003). The origins of genome complexity. Science *302*, 1401–1404.

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

Madan Babu, M., Teichmann, S.A., and Aravind, L. (2006). Evolutionary dynamics of prokaryotic transcriptional regulatory networks. J. Mol. Biol. *358*, 614–633.

Madera, M., Vogel, C., Kummerfeld, S.K., Chothia, C., and Gough, J. (2004). The SUPERFAMILY database in 2004: additions and improvements. Nucleic Acids Res. 32, D235–239.

Martinez-Antonio, A., Janga, S.C., Salgado, H., and Collado-Vides, J. (2006). Internal-sensing machinery directs the activity of the regulatory network in *Escherichia coli*. Trends Microbiol. *14*, 22–27.

Martinez-Nunez, M.A., Perez-Rueda, E., Gutierrez-Rios, R.M., and Merino, E. (2010). New insights into the regulatory networks of paralogous genes in bacteria. Microbiology 156, 14–22.

Miller, M.B., and Bassler, B.L. (2001). Quorum sensing in bacteria. Annu. Rev. Microbiol. *55*, 165–199.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. Science *298*, 824–827.

Minezaki, Y., Homma, K., and Nishikawa, K. (2005). Genome-wide survey of transcription factors in prokaryotes reveals many bacteria-specific families not found in archaea. DNA Res. *12*, 269–280.

Miroslavova, N.S., and Busby, S.J. (2006). Investigations of the modular structure of bacterial promoters. Biochem. Soc. Symp. 1–10.

Molina, N., and van Nimwegen, E. (2008). Universal patterns of purifying selection at noncoding positions in bacteria. Genome Res *18*, 148–160.

van Nimwegen, E. (2003). Scaling laws in the functional content of genomes. Trends Genet. *19*, 479–484.

Ogasawara, H., Yamada, K., Kori, A., Yamamoto, K., and Ishihama, A. (2010). Regulation of the *Escherichia coli* csgD promoter: interplay between five transcription factors. Microbiology *156*, 2470–2483.

Oguiza, J.A., Kiil, K., and Ussery, D.W. (2005). Extracytoplasmic function sigma factors in *Pseudomonas syringae*. Trends Microbiol. *13*, 565–568.

Olivares-Zavaleta, N., Jauregui, R., and Merino, E. (2006). Genome analysis of *Escherichia coli* promoter sequences evidences that DNA static curvature plays a more important role in gene transcription than has previously been anticipated. Genomics *87*, 329–337.

Paget, M.S., and Helmann, J.D. (2003). The sigma70 family of sigma factors. Genome Biol *4*, 203.

Perez, J.C., and Groisman, E.A. (2009a). Evolution of transcriptional regulatory circuits in bacteria. Cell *138*, 233–244.

Perez, J.C., and Groisman, E.A. (2009b). Transcription factor function and promoter architecture govern the evolution of bacterial regulons. Proc. Natl. Acad. Sci. U.S.A. *106*, 4319–4324.

Perez-Rueda, E., and Collado-Vides, J. (2001). Common history at the origin of the position-function correlation in transcriptional regulators in archaea and bacteria. J. Mol. Evol. *53*, 172–179.

Perez-Rueda, E., and Janga, S.C. (2010). Identification and genomic analysis of transcription factors in archaeal genomes exemplifies their functional architecture and evolutionary origin. Mol. Biol. Evol. 27, 1449–1459.

Perez-Rueda, E., and Janga, S.C. (2011). Identification and genomic analysis of transcription factors in archaeal genomes exemplifies their functional architecture and evolutionary origin. Mol. Biol. Evol. 27, 1449–1459.

Perez-Rueda, E., Collado-Vides, J., and Segovia, L. (2004). Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea. Comput. Biol. Chem. *28*, 341–350.

Perez-Rueda, E., Janga, S.C., and Martinez-Antonio, A. (2009). Scaling relationship in the gene content of transcriptional machinery in bacteria. Mol. Biosyst. *5*, 1494–1501.

Pragai, Z., Allenby, N.E., O'Connor, N., Dubrac, S., Rapoport, G., Msadek, T., and Harwood, C.R. (2004). Transcriptional regulation of the phoPR operon in *Bacillus subtilis*. J. Bacteriol. *186*, 1182–1190.

Rivera-Gomez, N., Segovia, L., and Perez-Rueda, E. (2011). The diversity and distribution of TFs and their partner domains play an important role in the regulatory plasticity in bacteria. Microbiology. <Please provide volume number and page range>

Rodionov, D.A. (2007). Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. Chem. Rev. *107*, 3467–3497.

Rodriguez-Trelles, F., Tarrio, R., and Ayala, F.J. (2003). Evolution of cis-regulatory regions versus codifying regions. Int. J. Dev. Biol. *47*, 665–673.

Roy, S.W., Fedorov, A., and Gilbert, W. (2002). The signal of ancient introns is obscured by intron density and homolog number. Proc. Natl. Acad. Sci. U.S.A. *99*, 15513–15517.

Salgado, H., Gama-Castro, S., Peralta-Gil, M., Diaz-Peredo, E., Sanchez-Solano, F., Santos-Zavaleta, A., Martinez-Flores, I., Jimenez-Jacinto, V., Bonavides-Martinez, C., Segura-Salazar, J., *et al.* (2006). RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acids Res. *34*, D394–397.

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

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

Soutourina, O., Kolb, A., Krin, E., Laurent-Winter, C., Rimsky, S., Danchin, A., and Bertin, P. (1999). Multiple control of flagellum biosynthesis in *Escherichia coli*: role of H-NS protein and the cyclic AMP-catabolite activator protein complex in transcription of the flhDC master operon. J. Bacteriol. *181*, 7500–7508.

Taraban, M., Zhan, H., Whitten, A.E., Langley, D.B., Matthews, K.S., Swint-Kruse, L., and Trewhella, J. (2008). Ligand-induced conformational changes and conformational dynamics in the solution structure of the lactose repressor protein. J. Mol. Biol. *376*, 466–481.

Teichmann, S.A., and Babu, M.M. (2004). Gene regulatory network growth by duplication. Nat. Genet. *36*, 492–496.

Thieffry, D., Huerta, A.M., Perez-Rueda, E., and Collado-Vides, J. (1998). From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. Bioessays *20*, 433–440.

Wall, M.E., Hlavacek, W.S., and Savageau, M.A. (2004). Design of gene circuits: lessons from bacteria. Nat. Rev. Genet. 5, 34–42.

West, G.B., and Brown, J.H. (2005). The origin of allometric scaling laws in biology from genomes to ecosystems: towards a quantitative unifying theory of biological structure and organization. J. Exp. Biol. *208*, 1575–1592.

Wosten, M.M. (1998). Eubacterial sigma-factors. FEMS Microbiol. Rev. 22, 127–150.

Yamazaki, Y., Niki, H., and Kato, J. (2008). Profiling of *Escherichia coli* Chromosome database. Methods Mol. Biol. *416*, 385–389.

Yanai, I., Camacho, C.J., and DeLisi, C. (2000). Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. Phys. Rev. Lett. *85*, 2641–2644.

Yeger-Lotem, E., and Margalit, H. (2003). Detection of regulatory circuits by integrating the cellular networks of protein–protein interactions and transcription regulation. Nucleic Acids Res. *31*, 6053–6061.

Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R.Y., Alon, U., and Margalit, H. (2004). Network motifs in integrated cellular networks of transcription-regulation and protein–protein interaction. Proc. Natl. Acad. Sci. U.S.A. *101*, 5934–5939.