



INSTITUT PASTEUR

Research in Microbiology xx (2007) 1–8



www.elsevier.com/locate/resmic

# Structure and evolution of gene regulatory networks in microbial genomes

Sarath Chandra Janga\*, J. Collado-Vides\*

Program of Computational Genomics, CCG-UNAM, Apdo Postal 565-A, Cuernavaca, Morelos, 62100 Mexico

Received 7 June 2007; accepted 17 September 2007

## Abstract

With the availability of genome sequences for hundreds of microbial genomes, it has become possible to address several questions from a comparative perspective to understand the structure and function of regulatory systems, at least in model organisms. Recent studies have focused on topological properties and the evolution of regulatory networks and their components. Our understanding of natural networks is paving the way to embedding synthetic regulatory systems into organisms, allowing us to expand the natural diversity of living systems to an extent we had never before anticipated.

© 2007 Elsevier Masson SAS. All rights reserved.

**Keywords:** Gene regulatory networks; Transcription; Protein–DNA interactions; Prokaryotes; Evolution

## 1. Introduction

One of the greatest challenges in the post-genomic era is to elucidate the complete set of gene expression programs in an organism for all possible stimuli to which it can respond. Although the number of completely sequenced genomes is mounting rapidly, our knowledge of transcription regulation is limited to a few model organisms. Organisms devote a considerable fraction of their DNA to encoding *cis*-regulatory elements, and a significant fraction of protein coding genes encode transcription factors (TFs), both of which play an important role in controlling and coordinating gene expression at the level of transcription. Unraveling the principles and organization of transcriptional programs is essential for understanding cellular responses to environmental perturbations and the molecular bases of many diseases caused by microbes.

It is now an accepted notion that transcriptional regulation can be visualized as a network consisting of TFs and their

target genes (TGs) [34,60,64]. However, at a less abstract level, transcription involves a number of *cis*-regulatory elements like promoters, TF binding sites and transcription terminators, and *trans*-acting elements like TFs and sigma factors. The interplay between the *cis* and *trans* elements provides a plethora of transcriptional programs which ultimately control the state of every gene in the cell tailored for different conditions (see Fig. 1 for a genomic view of the gene regulatory network). Although it is now becoming increasingly evident that post-transcriptional control by small non-coding RNAs plays an important role in both prokaryotic and eukaryotic organisms, here we review recent advances in the computational analysis of transcriptional regulation in microbial organisms. We first discuss progress made at the level of *trans* and *cis* element identification and finally integrate our recent understanding of microbial transcriptional regulation from a network perspective.

## 2. Evolution of *trans*-acting elements

At a *trans*-acting level, a number of protein families like sigma factors and TFs play important roles in controlling the

\* Corresponding authors. Tel.: +52 777 313 2063; fax: +52 777 317 5581.  
E-mail addresses: sarath@ccg.unam.mx (S.C. Janga), collado@ccg.unam.mx (J. Collado-Vides).

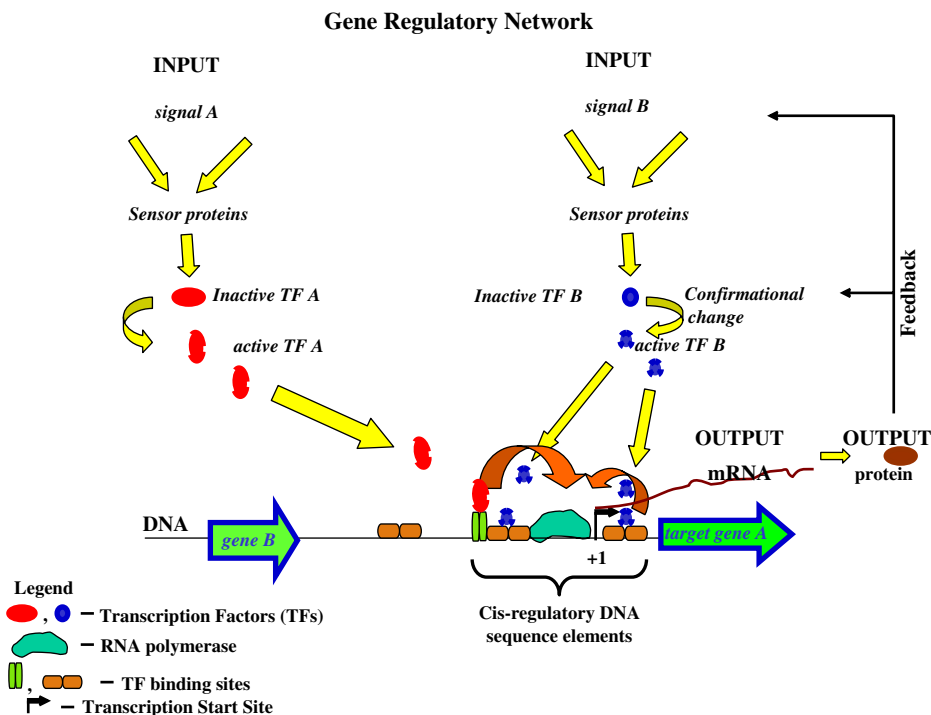


Fig. 1. Genomic view of gene regulatory network. Transcriptional regulation from a genomic perspective can be viewed as a complex interplay between *cis*-regulatory elements on the DNA and *trans*-acting factors like TFs and sigma factors. Depending on the set of input signals (extracellular or intracellular in nature), upon a cascade of events, transcription factors positively or negatively control transcription of the target gene. Often, transcription factors work in a combinatorial fashion to produce the final output. Protein products formed after transcription and translation of the gene regions are responsible for various cellular functions and ultimately feedback transcription factors at several levels to control their own transcription.

regulation of gene expression. These elements contribute significantly to rewiring the network of transcriptional interactions depending on the environmental conditions and stimuli an organism is faced with. Although our understanding of the number and repertoire of the *trans*-acting elements has greatly increased in recent years, our knowledge of the functional roles played by these proteins is far from complete.

### 2.1. Sigma factors

Sigma factors are a class of proteins forming essential dissociable subunits of prokaryotic RNA polymerase. The association of a sigma factor with core RNA polymerase provides the basis for transcriptional initiation and is an important step in the process of transcription. Sigma factors provide promoter recognition specificity to the polymerase and contribute to DNA strand separation, after which they dissociate from the RNA polymerase core enzyme following transcription initiation. The substitution of one sigma factor for another can redirect the polymerase to a different set of genes, which would otherwise be transcriptionally silent, thereby determining the transcriptional response of a group of genes.

The number of sigma factors encoded in bacterial genomes is highly variable. Although the number of sigma factors generally increases with genome size, environmental bacteria and microorganisms that have developed differentiation programs like sporulation tend to have a higher number of sigma factors

than most obligate pathogens. It is possible that the number of sigma factor genes is correlated with the diversity of lifestyles encountered by a bacterium. For instance, among the different mycobacterial species, *Mycobacterium leprae* has the lowest number of sigma factors and this seems to correlate with the fact that this organism has adapted to being an obligate pathogen, unlike other organisms of this phyla [13,58].

Sigma factors can be classified into two structurally unrelated and phylogenetically distinct families: the  $\sigma^{70}$  and  $\sigma^{54}$  families [23]. While  $\sigma^{54}$  family members are relatively rare,  $\sigma^{70}$  family members are found in all bacterial genomes. The  $\sigma^{70}$  factors typically consist of up to four conserved regions and are further classified into four different groups on the basis of their structure and physiological roles [23]. Structurally,  $\sigma^{70}$  family factors have four major regions, with the highest levels of conservation in regions 2 and 4. Subregions within region 2 are known to be involved in promoter melting (region 2.3) and  $-10$  sequence recognition (region 2.4), while the well conserved subregion 4.2 of region 4 is involved in  $-35$  recognition [54]. Within the  $\sigma^{70}$  family of sigma factors is a large, phylogenetically distinct subfamily called the extracytoplasmic function (ECF) factor, which typically contains only regions 2 and 4 of the  $\sigma^{70}$  family. These sigma factors are responsible for regulating a wide range of functions, all involved in sensing and reacting to conditions in the membrane, periplasm or extracellular environment [26]. Many bacteria contain multiple ECF factors and they generally outnumber

all other types of sigma factors combined. ECF factors are often co-transcribed with one or more negative regulators [26].

Although no sequence conservation exists between  $\sigma^{54}$  and  $\sigma^{70}$  family members, both types bind to core RNA polymerase. However, the holoenzyme formed with the  $\sigma^{54}$  class has different properties from those of the  $\sigma^{70}$  holoenzyme. For instance, all  $\sigma^{54}$  species require a separate activator protein along with the core RNA polymerase to form an open promoter complex, and promoter structures recognized by  $\sigma^{54}$ -RNAP differ from those recognized by  $\sigma^{70}$ -RNAP. The  $\sigma^{54}$  promoters generally are highly conserved, short sequences that are located at positions  $-24$  and  $-12$  upstream of the transcription start site, whereas  $\sigma^{70}$  promoter sites are typically located at  $-35$  and  $-10$  upstream [9].

## 2.2. DNA binding TFs as regulators of transcriptional control

Regulation of gene expression in an organism is predominantly controlled by DNA binding TFs. They form one of the largest protein groups in most genomes. TFs are proteins which are needed to activate or repress the transcription of a gene or operon. Most TFs form dimers and bind to the *cis*-regulatory elements on the DNA to control transcription initiation in bacterial genomes [8]. The fraction of TFs in bacterial genomes typically scales as the square of the total gene number of a genome [2,65], with the maximum number of TFs observed in *Streptomyces coelicolor* among the publicly available completely sequenced genomes [7].

TFs can be classified as activators, repressors or dual regulators depending on their mode of action on a particular promoter [8,55]. An activator stimulates the expression of its target gene by acting on a promoter to stimulate RNA polymerase. Activation is known to typically occur by binding of TFs upstream of the transcription start site and often upstream of the  $-35$  promoter element [8,40,52]. For negative control of transcription, TFs act as repressors by binding to DNA to prevent RNA polymerase from initiating transcription. Repression normally occurs when TFs bind downstream of the transcription start site, causing DNA looping, or in between the  $-35$  and  $-10$  elements of the promoter, thereby blocking RNA polymerase by steric hindrance [8,40]. Computational analyses suggest that repressors are dominant in both *Escherichia coli* and *Bacillus subtilis*, and are more likely to co-evolve with their target genes in closely related genomes [27,52,55].

DNA binding regions of prokaryotic TFs can be assigned to a number of families based on sequence and structural homologies [39,55]. TF families classified based on structural domains are three-fold—the helix-turn-helix, the winged helix and the beta ribbon [30]—with the most abundant among TFs being the classical helix-turn-helix domain [2]. It has been proposed that about 75% of the TFs in *E. coli* are formed as a result of duplication and that TFs evolve faster than their respective TGs across genomes [36,41].

Global transcription regulators have been defined as those TFs that have the ability to: regulate large number of genes

belonging to diverse functional classes, control a complex regulatory cascade by both directly and indirectly effect expression of various cellular pathways and act on target promoters that use different sigma factors [46]. Based on this, seven global regulators in *E. coli* have been proposed, which control more than 50% of the genes in the entire transcriptional regulatory network. More recent studies used connectivity (see below) of the TF as a simplified measure to assess the global nature of a TF [36,41].

## 3. Evolution of *cis*-regulatory elements

In recent years, due to accumulation of genome sequences of multiple strains of a single organism and those of phylogenetically close species, it has become possible to address a number of questions related to conservation of regulatory elements. The availability of genome sequences not only provides us with evolutionary insights into conservation of *cis*-regulatory elements like promoter regions, TF binding sites and terminator signals across organisms, but also enables us to predict them using a variety of comparative genome analysis techniques.

### 3.1. Promoter regions

Transcription initiation in bacteria requires that RNA polymerase (RNAP) recognize and bind specific DNA sequences upstream of transcription units called promoters. The recognition of promoter sequences by RNAP occurs when it associates with a small protein, known as sigma ( $\sigma$ ) factor. The primary or housekeeping sigma factor in *E. coli* is encoded by the *rpoD* gene and is known as  $\sigma^{70}$  [16]. A bacterial promoter is defined as the segment of DNA that enables a gene or set of genes to be transcribed and is located immediately proximal (6–8 bp) to the transcription start site. Although there are several other condition-specific sigma factors besides the housekeeping ones, the most frequently studied, with extensive experimentally characterized information, remains  $\sigma^{70}$ . In fact, *E. coli* has six other sigma factors which are encoded by genes *rpoN* ( $\sigma^{54}$ ), *rpoS* ( $\sigma^{38}$ ), *rpoH* ( $\sigma^{32}$ ), *rpoF* ( $\sigma^{28}$ ), *rpoE* ( $\sigma^{24}$ ) and *fecI* ( $\sigma^{19}$ ). The canonical model of the  $\sigma^{70}$  DNA promoter is characterized by two hexamers centered around positions  $-35$  and  $-10$  from the transcription start site and separated by 15–21 bp, with consensus sequences TTGACA and TATAAT, respectively. Although there is a direct relationship between promoter strength and similarity to the consensus sequence, a typical *E. coli*  $\sigma^{70}$  promoter sequence contains two mismatches within both the  $-35$  and  $-10$  hexanucleotide elements [49]. In fact, variations of over three deviations from the consensus have been reported in  $\sigma^{70}$ -dependent promoters from various studies. These variations can generate considerable differences in promoter efficiency during the transcription initiation reaction. All these factors make identification of functional promoter sequences a notoriously difficult task even in well studied model systems like *E. coli*, irrespective of the approach adopted [19,28,32,66]. In fact, Huerta and Collado-Vides [28] show that several

functional promoters are significantly different from the consensus and often occur in regulatory regions as dense overlapping signals. Further studies established that promoter densities are indeed different in the coding and regulatory regions of most bacterial genomes [28,29,31], following a regional rule which can distinguish organization of different adjacent gene pairs in bacterial genomes [50]. In contrast, certain genomes with significant size reduction were found not to show this tendency, which was attributed to a process of genome degradation resulting from the decreased efficiency of purifying selection in highly structured small populations [29]. Interestingly, several of these genomes which deviate from this tendency were found to be intracellular parasites which exhibit severe reduction not only in their genome sizes, but also a disproportionate reduction in the number of TFs. These observations also suggest that the differential distribution of promoter-like signals between regulatory and non-regulatory regions detected in large bacterial genomes might confer a fitness advantage to these organisms in their natural habitats.

### 3.2. TF binding sites

TFs recognize the TGs, whose transcription they control, due to the presence of the binding sites in the promoter regions. Typically a TF, upon binding to the promoter regions of their target genes or transcription units, can control the expression of the genes positively or negatively. While repressor sites which can inhibit the transcription of genes are known to occur downstream of the transcription start site, activators generally attach to DNA upstream of the start site [22,40,52]. In *E. coli*, there is an enrichment for factors which act as transcriptional repressors and hence the majority of genes in the transcriptional network are negatively regulated [39,55].

Two general computational approaches have emerged for inferring TF binding sites in promoter regions: (i) analysis of co-regulated sets of genes; and (ii) phylogenetic footprinting of the upstream regions of orthologous genes in closely related genomes, under the notion that selective pressure would sustain regulatory elements over the background non-coding DNA among organisms at short evolutionary distances [48,56]. Both methods aim to identify statistically significant patterns which are conserved in the background of the remaining aligned intergenic regions. Since the identification of a putative set of co-regulated genes from genome sequences alone is not straightforward, the majority of computational approaches for inferring regulatory motifs use upstream regions of orthologous genes from phylogenetically close organisms as the seed set of sequences. However, Wang and Stormo used the conserved regions of orthologous sequences in multiple sequence alignments, and then compared profiles of non-orthologous sequences (genes with in a given organism) to generate sets of co-regulated genes [67], while others used co-expressed genes as a seed set to improve motif detection by incorporating phylogenetic conservation [51]. Another recent approach took into account the phylogenetic relationship between the species, in order to distinguish conservation due

to the occurrence of functional sites from spurious conservation, which is due to evolutionary proximity, and they developed a Gibbs sampling algorithm for motif prediction from phylogenetic conservation [61].

Once regulatory elements are identified, they can either be compared with already known binding profiles for TFs, or subjected to experimental analyses to prove that they are functional and/or to determine binding factors. Some recent approaches also adopted the use of binding profiles obtained first using cross-species data and then generating genome-specific models through recursive training to attain higher specificity for identifying binding sites [20], or exploited the fact that most transcription factors bind to DNA as spaced dimers [35,53], while others used the idea that TFs often bind cooperatively to their targets; hence, statistically overrepresented motif co-occurrence patterns can help identify novel TF-TG associations [10]. Another work attempted to integrate a variety of properties like proximity of TFs to their TGs, similarity in the binding properties of TFs which belong to the same family and phylogenetic correlation to develop a system for inferring regulatory interactions on a genomic scale [62].

### 3.3. Transcriptional termination

Transcription termination typically involves the release of the mRNA transcript and RNA polymerase from the template strand at the end of transcription. Proper termination is essential for bacteria, as the regions between transcription units are generally rather small. In bacterial genomes, termination generally occurs either spontaneously, which is termed intrinsic or Rho-independent termination, or involves the use of a set of *trans*-acting factors in conjunction with the *cis*-acting elements, referred to as Rho-dependent termination. Although Rho-dependent termination has been the focus of several experimental studies, our knowledge of defined rules which can be used to identify Rho-dependent termination signals from genome sequences is rather limited, believed to be due to the complex interplay of several auxiliary elements which occur at a specific Rho site depending on the local context of the terminator [12]. The only conserved element common to Rho-dependent terminators appears to be richness in cytosine residues. On the other hand, Rho-independent termination usually occurs due to the presence of a hairpin loop structure followed by a stretch of thymine residues. It is accepted that most Rho-independent terminators can be identified from sequences due to their dyad symmetry and poly-T tail [14,18].

About half of the transcription terminators which are experimentally characterized in *E. coli* are Rho-dependent [59]. However, unlike the case in *E. coli*, the Rho protein is dispensable in *B. subtilis*, suggesting a limited role for Rho-dependent termination in the latter. In fact, recent work demonstrated that more than 90% of the termination signals in *B. subtilis* and other Firmicutes are Rho-independent in nature [15]. Another group developed an efficient algorithm for rapid determination of Rho-independent terminators and demonstrated that outside the Firmicutes division, Rho-independent termination is also

found to be dominant in the *Neisseria*, *Vibrio* and *Pasteurellaceae* genera [33].

#### 4. The link between *cis*- and *trans*-acting elements and the notion of transcriptional networks

An important notion that is emerging in post-genomic biology is that cellular components can be visualized as a network of interactions between different molecules like proteins, DNA and metabolites [5]. This has led to the application of network theory to biological problems, particularly in understanding the regulation of gene expression [60,64]. In transcriptional networks typically *trans*-acting elements like TFs and sigma factors form one set of nodes and their target genes, of which they control the activity, form the other set of nodes. The links between them which have directionality from the *trans*-acting elements to their target genes, controlled by their *cis*-regulatory elements, form a complex and directional network of interactions (see Fig. 2 for a network view of transcriptional regulation).

##### 4.1. Network structure

One of the most important and obvious pieces of information that can be obtained is the distribution of connectivity, i.e. how many connections a node has and how many nodes have a particular number of connections. In the case of transcriptional networks, these parameters actually have two sides, as incoming and outgoing connections must be considered separately. The incoming connectivity is the number of transcription factors regulating a target gene, which gives a sense of the combinatorial effect of gene regulation. The fraction of target genes with a given incoming connectivity was observed to follow an exponential distribution in both *E. coli* and *Saccharomyces cerevisiae* [24,64]. The exponential behavior indicates that most target genes are regulated by a similar number of factors and apparently reflects limits in the size of multiprotein complexes that can be bound near the promoter, as well as by the amount of DNA sequences in upstream regions of genes. On the other hand, outgoing connectivity, which is the number of target genes regulated by each transcription factor, was found to be distributed according to a power law, contrary to incoming connectivity distribution. This is indicative of a hub-containing network structure, in which a select set of transcription factors participate in the regulation of a disproportionately large number of target genes.

At a local level, in transcriptional networks, certain subnetworks appear more often than expected by chance and have been referred to as motifs, analogous to sequence motifs which occur repeatedly in sequences. Motifs were originally described in an *E. coli* transcriptional regulatory network, but were subsequently found in yeast and other organisms [1,60]. Three network motifs were found to predominantly occur in most transcriptional networks: (1) a feed-forward loop (FFL), in which a transcription factor regulates expression of another transcription factor which, in turn, regulates a gene

that is also regulated by the first transcription factor; (2) a single-input module (SIM), in which a single transcription factor regulates several genes, usually also referred to as a simple regulon [25]; (3) dense overlapping regulons (DORs) in which several TFs regulate overlapping sets of genes; these groups are also called a complex regulon. FFL appears to be the most abundant motif among the best studied transcriptional networks. FFLs have been further classified into eight motif subtypes (see Fig. 3) and two of them, namely coherent type-1 and incoherent type-1 FFL, appear to be much more predominant than others [1,43]. The former was shown to act as a sign-sensitive delay element and a persistence detector, while the latter was demonstrated to function as a pulse generator and response accelerator [44,45]. Although motifs form overrepresented subgraphs in the entire network of transcriptional regulation, they do not appear independently but rather integrate to form superstructures or modules that carry out a common biological function by sharing some of their edges [17,57].

##### 4.2. Network conservation

With the availability of documented information on transcriptional regulation, based on experimental evidence in model organisms for a significant fraction of the TFs, it has now become possible to address questions on the evolution of the structure and components of regulatory systems across bacterial genomes [42,59]. From the perspective of the evolution of transcriptional networks in a given organism, it was proposed that duplication of TFs and their TGs could have given rise to a significant proportion of the currently known regulatory networks in both *E. coli* and *S. cerevisiae* [63]. However, from a cross-genome perspective, it was found that TFs are poorly conserved across genomes in comparison to their TGs, and hence are likely to evolve faster than their TGs [36,41]. It was also found that global regulators are not more conserved than general TFs, suggesting a possible scenario for rapid evolution of gene regulatory mechanisms across bacteria. The latter group also showed that regulatory interactions within a network motif do not show any preference for evolving together and that organisms with a similar lifestyle are likely to preserve equivalent regulatory interactions and network motifs.

##### 4.3. Network dynamics

Despite several studies which focus on regulatory networks at a static level, it should be noted that the regulatory network of an organism is highly dynamic and different sections of the network could be active under different conditions [21]. In fact, it has been shown in yeast, by integrating expression and regulatory interaction data, that the regulatory subnetworks for different conditions vary significantly [37]. In particular, it was demonstrated that in multistage processes, like the cell cycle or sporulation, there are extensive variations in the regulatory networks. In an attempt to extend this work to bacterial systems, another study systematically identified

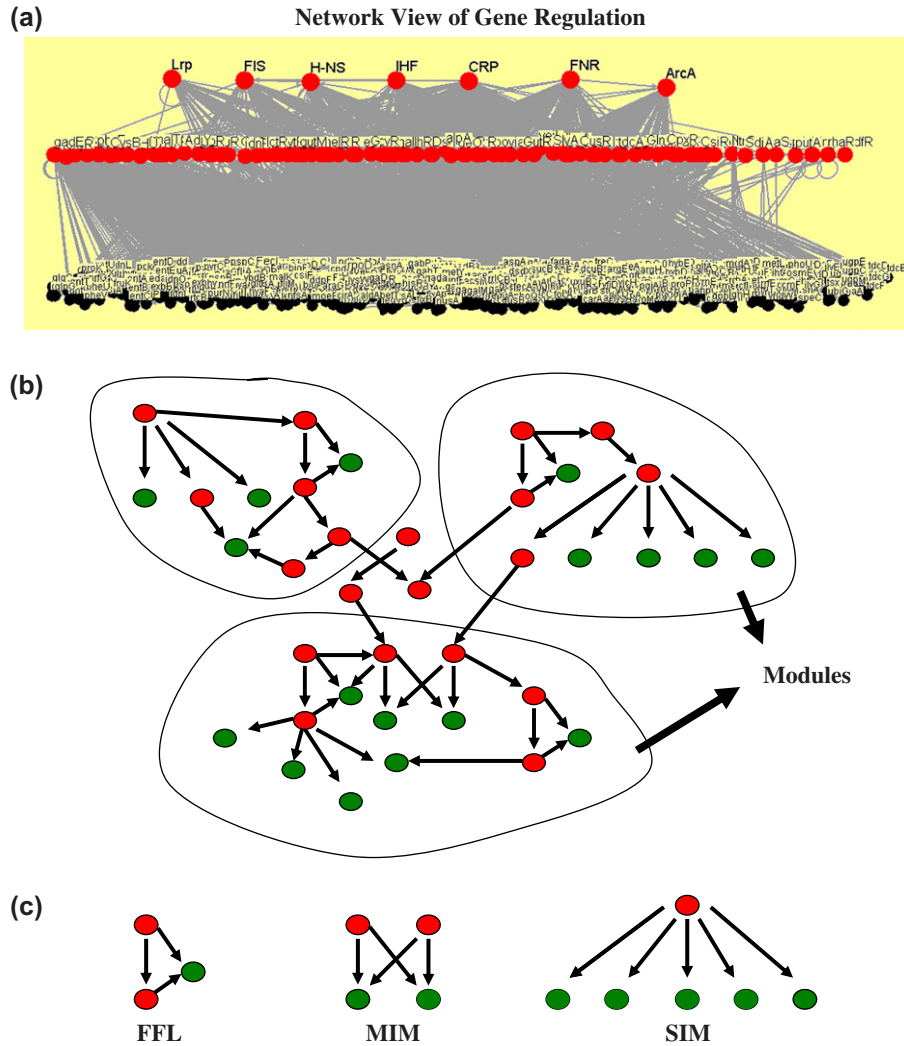


Fig. 2. Network view of transcription regulation. Transcriptional regulatory interactions at a genomic level can be visualized as a network between TFs (shown in red) and target genes (shown in black/green). (a) The transcriptional regulatory network is a multi-layer hierarchical modular structure without feedback regulation at the transcription level [38,68], with the global regulators at the top of this layout and local TFs at the bottom, regulating a few genes. (b) Modules are inter-connected clusters which divide the network of transcriptional interactions into subnetworks. Modules have been identified using a variety of approaches [17,38,57] and have been found to be semi-independent in nature. Modules are in turn formed by one or more different types of network motifs. (c) Motifs are patterns of interconnections which are overrepresented in transcriptional networks. Known transcriptional regulatory networks were found to have feed forward loops (FFLs), multiple input modules (MIMs) and SIMs, with each kind of motif playing a different role [1].

topological units called origons under the notion that different subnetworks from the completely known transcriptional regulatory network could be active under different experimental conditions depending on the environmental signals sensed by

the sensor TFs [4]. In another recent work using a static regulatory network of *E. coli*, the authors classified the complete set of TFs in the currently known regulatory network as those sensing endogenous or exogenous signals. Curiously enough,

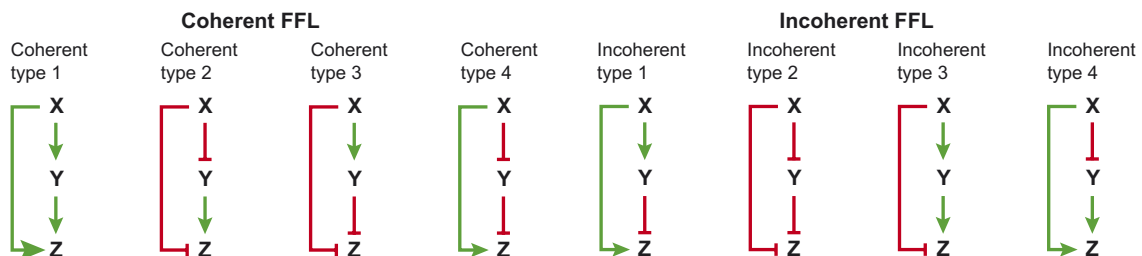


Fig. 3. Different subtypes of feed forward loop motifs. Directed arrow represents positive regulation and is shown in green, while negative transcriptional control is shown in red.

global regulators often correspond to those sensing internal signals, and TFs sensing internal signals were found to direct the activity of the regulatory network in *E. coli* [47].

## 5. Conclusions and perspectives

Although our understanding of the design principles of complex transcriptional regulatory networks is far from complete, we are beginning to design biological circuits and predict their behavior. Progress in sequencing technologies, high throughput experimental techniques like chromatin immunoprecipitation, and advances such as noise filters [6] and oscillators that can combine repressor functionality with that of a two-component system [3], together with improvements in measuring cellular quantities at high resolution [69], should not only enable us to design synthetic circuits for maneuvering bacterial systems [11], but also enable us to address several fundamentally unanswered questions in the years to come.

## Acknowledgements

We thank Arthur Wuster for critically reading and providing comments on a previous version of this manuscript. We also thank Agustino Martinez-Antonio for providing assistance in the generation of Fig. 2. This work was partially supported by NIH grant RO1 GM 071962.

## References

- [1] Alon, U. (2007) Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* 8, 450–461.
- [2] Aravind, L., Anantharaman, V., Balaji, S., Babu, M.M., Iyer, L.M. (2005) The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol. Rev.* 29, 231–262.
- [3] Atkinson, M.R., Savageau, M.A., Myers, J.T., Ninfa, A.J. (2003) Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in *Escherichia coli*. *Cell* 113, 597–607.
- [4] Balazsi, G., Barabasi, A.L., Oltvai, Z.N. (2005) Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* 102, 7841–7846.
- [5] Barabasi, A.L., Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113.
- [6] Becskei, A., Serrano, L. (2000) Engineering stability in gene networks by autoregulation. *Nature* 405, 590–593.
- [7] Bentley, S.D., Chater, K.F., Cerdeno-Tarraga, A.M., Challis, G.L., Thomson, N.R., James, K.D., Harris, D.E., Quail, M.A., Kieser, H., Harper, D., et al. (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* 417, 141–147.
- [8] Browning, D.F., Busby, S.J. (2004) The regulation of bacterial transcription initiation. *Nat. Rev. Microbiol.* 2, 57–65.
- [9] Buck, M., Gallegos, M.T., Studholme, D.J., Guo, Y., Gralla, J.D. (2000) The bacterial enhancer-dependent sigma(54) (sigma(N)) transcription factor. *J. Bacteriol.* 182, 4129–4136.
- [10] Bulyk, M.L., McGuire, A.M., Masuda, N., Church, G.M. (2004) A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in *Escherichia coli*. *Genome Res.* 14, 201–208.
- [11] Chin, J.W. (2006) Modular approaches to expanding the functions of living matter. *Nat. Chem. Biol.* 2, 304–311.
- [12] Ciampi, M.S. (2006) Rho-dependent terminators and transcription termination. *Microbiology* 152, 2515–2528.
- [13] Cole, S.T., Eiglmeier, K., Parkhill, J., James, K.D., Thomson, N.R., Wheeler, P.R., Honore, N., Garnier, T., Churcher, C., Harris, D., et al. (2001) Massive gene decay in the leprosy bacillus. *Nature* 409, 1007–1011.
- [14] d'Aubenton Carafa, Y., Brody, E., Thermes, C. (1990) Prediction of rho-independent *Escherichia coli* transcription terminators. A statistical analysis of their RNA stem-loop structures. *J. Mol. Biol.* 216, 835–858.
- [15] de Hoon, M.J., Makita, Y., Nakai, K., Miyano, S. (2005) Prediction of transcriptional terminators in *Bacillus subtilis* and related species. *PLoS Comput. Biol.* 1, e25.
- [16] deHaseth, P.L., Nilsen, T.W. (2004) Molecular biology. When a part is as good as the whole. *Science* 303, 1307–1308.
- [17] Dobrin, R., Beg, Q.K., Barabasi, A.L., Oltvai, Z.N. (2004) Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network. *BMC Bioinformatics* 5, 10.
- [18] Ermolaeva, M.D., Khalak, H.G., White, O., Smith, H.O., Salzberg, S.L. (2000) Prediction of transcription terminators in bacterial genomes. *J. Mol. Biol.* 301, 27–33.
- [19] Eskin, E., Keich, U., Gelfand, M.S., Pevzner, P.A. (2003) Genome-wide analysis of bacterial promoter regions. *Pac. Symp. Biocomput* 29–40.
- [20] Espinosa, V., Gonzalez, A.D., Vasconcelos, A.T., Huerta, A.M., Collado-Vides, J. (2005) Comparative studies of transcriptional regulation mechanisms in a group of eight gamma-proteobacterial genomes. *J. Mol. Biol.* 354, 184–199.
- [21] Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J., Gardner, T.S. (2007) Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* 5, e8.
- [22] Gralla, J.D., Collado-Vides, J. (1996) Organization and function of transcription regulatory elements. In F.C. Neidhardt, R. Curtis III, J. Ingraham, E.C.C. Lin, K.B. Low, B. Magasanik, W. Reznikoff, M. Schaechter, H.E. Umbarger, & M. Riley (Eds.), *Cellular and molecular biology: Escherichia coli and Salmonella* (pp. 1232–1245). Washington, DC: American Society for Microbiology.
- [23] Gruber, T.M., Gross, C.A. (2003) Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu. Rev. Microbiol.* 57, 441–466.
- [24] Guelzim, N., Bottani, S., Bourgine, P., Kepes, F. (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nat. Genet.* 31, 60–63.
- [25] Gutierrez-Rios, R.M., Rosenblueth, D.A., Loza, J.A., Huerta, A.M., Glasner, J.D., Blattner, F.R., Collado-Vides, J. (2003) Regulatory network of *Escherichia coli*: consistency between literature knowledge and microarray profiles. *Genome Res.* 13, 2435–2443.
- [26] Helmann, J.D. (2002) The extracytoplasmic function (ECF) sigma factors. *Adv. Microb. Physiol.* 46, 47–110.
- [27] Hershberg, R., Margalit, H. (2006) Co-evolution of transcription factors and their targets depends on mode of regulation. *Genome Biol.* 7, R62.
- [28] Huerta, A.M., Collado-Vides, J. (2003) Sigma70 promoters in *Escherichia coli*: specific transcription in dense regions of overlapping promoter-like signals. *J. Mol. Biol.* 333, 261–278.
- [29] Huerta, A.M., Francino, M.P., Morett, E., Collado-Vides, J. (2006) Selection for unequal densities of sigma70 promoter-like signals in different regions of large bacterial genomes. *PLoS Genet.* 2, e185.
- [30] Huffman, J.L., Brennan, R.G. (2002) Prokaryotic transcription regulators: more than just the helix-turn-helix motif. *Curr. Opin. Struct. Biol.* 12, 98–106.
- [31] Janga, S.C., Lamboy, W.F., Huerta, A.M., Moreno-Hagelsieb, G. (2006) The distinctive signatures of promoter regions and operon junctions across prokaryotes. *Nucleic Acids Res.* 34, 3980–3987.
- [32] Kanhere, A., Bansal, M. (2005) A novel method for prokaryotic promoter prediction based on DNA stability. *BMC Bioinformatics* 6, 1.
- [33] Kingsford, C.L., Ayanbule, K., Salzberg, S.L. (2007) Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol.* 8, R22.
- [34] Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., et al. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298, 799–804.
- [35] Li, H., Rhodius, V., Gross, C., Siggia, E.D. (2002) Identification of the binding sites of regulatory proteins in bacterial genomes. *Proc. Natl. Acad. Sci. U.S.A.* 99, 11772–11777.

- [36] Lozada-Chavez, I., Janga, S.C., Collado-Vides, J. (2006) Bacterial regulatory networks are extremely flexible in evolution. *Nucleic Acids Res.* 34, 3434–3445.
- [37] Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A., Gerstein, M. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* 431, 308–312.
- [38] Ma, H.W., Buer, J., Zeng, A.P. (2004) Hierarchical structure and modules in the *Escherichia coli* transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinformatics* 5, 199.
- [39] Madan Babu, M., Teichmann, S.A. (2003) Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res.* 31, 1234–1244.
- [40] Madan Babu, M., Teichmann, S.A. (2003) Functional determinants of transcription factors in *Escherichia coli*: protein families and binding sites. *Trends Genet.* 19, 75–79.
- [41] Madan Babu, M., Teichmann, S.A., Aravind, L. (2006) Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J. Mol. Biol.* 358, 614–633.
- [42] Makita, Y., Nakao, M., Ogasawara, N., Nakai, K. (2004) DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic Acids Res.* 32, D75–D77.
- [43] Mangan, S., Alon, U. (2003) Structure and function of the feed-forward loop network motif. *Proc. Natl. Acad. Sci. U.S.A.* 100, 11980–11985.
- [44] Mangan, S., Itzkovitz, S., Zaslaver, A., Alon, U. (2006) The incoherent feed-forward loop accelerates the response-time of the gal system of *Escherichia coli*. *J. Mol. Biol.* 356, 1073–1081.
- [45] Mangan, S., Zaslaver, A., Alon, U. (2003) The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *J. Mol. Biol.* 334, 197–204.
- [46] Martinez-Antonio, A., Collado-Vides, J. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr. Opin. Microbiol.* 6, 482–489.
- [47] Martinez-Antonio, A., Janga, S.C., Salgado, H., Collado-Vides, J. (2006) Internal-sensing machinery directs the activity of the regulatory network in *Escherichia coli*. *Trends Microbiol.* 14, 22–27.
- [48] McCue, L., Thompson, W., Carmack, C., Ryan, M.P., Liu, J.S., Derbyshire, V., Lawrence, C.E. (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.* 29, 774–782.
- [49] Mitchell, J.E., Zheng, D., Busby, S.J., Minchin, S.D. (2003) Identification and analysis of ‘extended-10’ promoters in *Escherichia coli*. *Nucleic Acids Res.* 31, 4689–4695.
- [50] Mitchison, G. (2005) The regional rule for bacterial base composition. *Trends Genet.* 21, 440–443.
- [51] Monsieurs, P., Thijs, G., Fadda, A.A., De Keersmaecker, S.C., Vanderleyden, J., De Moor, B., Marchal, K. (2006) More robust detection of motifs in coexpressed genes by using phylogenetic information. *BMC Bioinformatics* 7, 160.
- [52] Moreno-Campuzano, S., Janga, S.C., Perez-Rueda, E. (2006) Identification and analysis of DNA binding transcription factors in *Bacillus subtilis* and other Firmicutes—a genomic approach. *BMC Genomics* 7, 147.
- [53] Mwangi, M.M., Siggia, E.D. (2003) Genome wide identification of regulatory motifs in *Bacillus subtilis*. *BMC Bioinformatics* 4, 18.
- [54] Paget, M.S., Helmann, J.D. (2003) The sigma70 family of sigma factors. *Genome Biol.* 4, 203.
- [55] Perez-Rueda, E., Collado-Vides, J. (2000) The repertoire of DNA binding transcriptional regulators in *Escherichia coli* K-12. *Nucleic Acids Res.* 28, 1838–1847.
- [56] Rajewsky, N., Succi, N.D., Zapotocky, M., Siggia, E.D. (2002) The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons. *Genome Res.* 12, 298–308.
- [57] Resendis-Antonio, O., Freyre-Gonzalez, J.A., Menchaca-Mendez, R., Gutierrez-Rios, R.M., Martinez-Antonio, A., Avila-Sanchez, C., Collado-Vides, J. (2005) Modular analysis of the transcriptional regulatory network of *E. coli*. *Trends Genet.* 21, 16–20.
- [58] Rodrigue, S., Proveddi, R., Jacques, P.E., Gaudreau, L., Manganelli, R. (2006) The sigma factors of *Mycobacterium tuberculosis*. *FEMS Microbiol. Rev.* 30, 926–941.
- [59] 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–D397.
- [60] Shen-Orr, S.S., Milo, R., Mangan, S., Alon, U. (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* 31, 64–68.
- [61] Siddharthan, R., Siggia, E.D., van Nimwegen, E. (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.* 1, e67.
- [62] Tan, K., McCue, L.A., Stormo, G.D. (2005) Making connections between novel transcription factors and their DNA motifs. *Genome Res.* 15, 312–320.
- [63] Teichmann, S.A., Babu, M.M. (2004) Gene regulatory network growth by duplication. *Nat. Genet.* 36, 492–496.
- [64] Thieffry, D., Huerta, A.M., Perez-Rueda, E., Collado-Vides, J. (1998) From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. *Bioessays* 20, 433–440.
- [65] van Nimwegen, E. (2003) Scaling laws in the functional content of genomes. *Trends Genet.* 19, 479–484.
- [66] Wang, H., Benham, C.J. (2006) Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress. *BMC Bioinformatics* 7, 248.
- [67] Wang, T., Stormo, G.D. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* 19, 2369–2380.
- [68] Yu, H., Gerstein, M. (2006) Genomic analysis of the hierarchical structure of regulatory networks. *Proc. Natl. Acad. Sci. U.S.A.* 103, 14724–14731.
- [69] Zaslaver, A., Bren, A., Ronen, M., Itzkovitz, S., Kikoin, I., Shavit, S., Liebermeister, W., Surette, M.G., Alon, U. (2006) A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli*. *Nat. Methods* 3, 623–628.