- 19 Kelly, D.P. and Scarpulla, R.C. (2004) Transcriptional regulatory circuits controlling mitochondrial biogenesis and function. *Genes Dev.* 18, 357–368
- 20 Li, R. *et al.* (1996) Expression of the human cytochrome c1 gene is controlled through multiple Sp1-binding sites and an initiator region. *Eur. J. Biochem.* 241, 649–656
- 21 Gulick, T. *et al.* (1994) The peroxisome proliferator-activated receptor regulates mitochondrial fatty acid oxidative enzyme gene expression. *Proc. Natl. Acad. Sci. U. S. A.* 91, 11012–11016
- 22 Arbeitman, M.N. *et al.* (2002) Gene expression during the life cycle of *Drosophila melanogaster*. *Science* 297, 2270–2275
- 23 Jegga, A.G. *et al.* (2002) Detection and visualization of compositionally similar cis-regulatory element clusters in orthologous and coordinately controlled genes. *Genome Res.* 12, 1408–1417
- 24 Hertz, G.Z. and Stormo, G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15, 563–577
- 25 Pavesi, G. *et al.* (2001) An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics* 17(Suppl. 1), S207–S214
- 26 Pavesi, G. *et al.* (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.* 32, W199–W203
- 27 Grillo, G. *et al.* (2003) PatSearch: A program for the detection of patterns and structural motifs in nucleotide sequences. *Nucleic Acids Res.* 31, 3608–3612

0168-9525/\$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2004.11.009

Modular analysis of the transcriptional regulatory network of *E. coli*

Osbaldo Resendis-Antonio, Julio A. Freyre-González, Ricardo Menchaca-Méndez, Rosa M. Gutiérrez-Ríos, Agustino Martínez-Antonio, Cristhian Ávila-Sánchez and Julio Collado-Vides

Program of Computational Genomics, Nitrogen Fixation Research Center, Universidad Nacional Autónoma de México, Ave Universidad s/n, Col Chamilpa, Cuernavaca, Morelos 62100 México

The transcriptional network of *Escherichia coli* is currently the best-understood regulatory network of a single cell. Motivated by statistical evidence, suggesting a hierarchical modular architecture in this network, we identified eight modules with well-defined physiological functions. These modules were identified by a clustering approach, using the shortest path to trace regulatory relationships across genes in the network. We report the type (feed forward and bifan) and distribution of motifs between and within modules. Feed-forward motifs tend to be embedded within modules, whereas bi-fan motifs tend to link modules, supporting the notion of a hierarchical network with defined functional modules.

There is experimental evidence suggesting that, at certain times, different fractions of the complete set of transcriptional factors (TFs) are used depending on the growth conditions within the cell [1,2]. The global topological analysis of transcriptional networks supports the notion of their organization into modules or large groups of genes that, in many cases, respond to external or internal stimuli [2,3]. A network with a scale-free property is expected to be robust to failure of individual components, and could contribute to the ability of the cell to respond to changes in environmental and evolutionary conditions [4–9]. At a finer topological scale, over-represented topological units called network motifs, consisting of three or four genes, contribute to the local dynamic behavior of transcriptional regulation [10,11]. Despite studies that are

focused on motifs and modules, little is known about how these motifs can integrate to construct modular structures and whether their components correspond to subsets of genes that regulate integrated cellular responses to external conditions.

It was recently reported that the transcriptional regulatory network of *Escherichia coli* is a scale-free network consisting of a hierarchy of modules [12]. In this article, we describe our reconstruction of eight modules (Figure 1a–d) with clearly defined physiological functions. In addition, using the fraction of feed forward (FF) and bi-fan (BF) motifs [11] that are shared between two modules, we quantified the overlap between them (Figure 2a–f). We found that the largest module that is involved in carbon sources has the greatest number of connection via motifs with other modules.

We analyzed the network of known transcriptional interactions of *E. coli* K-12 that were organized in RegulonDB [13] (http://www.cifn.unam.mx/Computational_Genomics/regulondb/). Neglecting genes without experimental evidence, indicating that they encode TFs (supplementary material online), and ignoring autoregulation, the total number of TFs that regulate the expression of other TFs is 55, all of which control the expression of 747 genes. Based on the number of genes in the genome, and the total number of estimated TFs (~320), we estimate that this fraction represents ~18% of the transcriptional network in *E. coli*. In our graphical representation, vertices represent genes and the transcriptional interactions between them are represented by edges.

Corresponding author: Collado-Vides, J. (collado@cifn.unam.mx).

Available online 25 November 2004

Figure 1. Identifying modules in *Escherichia coli*. **(a)** Clustering analysis where the intensity of each spot represents the shortest path length between pairs of genes. The intensity value can be 1.0 or 0.0, where a value of 1.0 means that two genes have a direct connection and a value of 0.0 means that one pair of genes is not connected in the network. Global regulators described in Ref. [16] are highlighted in yellow (Box 2). **(b)** This hierarchical structure shows the sets of genes that conform each module (denoted by M). **(c)** The biological function for each of the eight modules. **(d)** The color code used to annotate each gene with its corresponding functional class [15].

Identifying modules

The clustering analysis was performed using only TF-encoding genes. We obtained the shortest path length between every pair of genes (d_{ij} is the shortest path length between gene i and gene j). Next, we calculated the association function ($1/d_{ij}^2$) of these shortest path lengths [14]. This function gives a measure of the closeness among genes, amplifying close relationships and

minimizing remote distances. The data were used as input to perform a hierarchical agglomerative average-linkage clustering [2] (Figure 2a and supplementary material online). Modules identified by this algorithm and the functional classes of the genes that compose them are shown in Figure 1.

An analysis of sensitivity and specificity for each set of genes inside a module indicates that topological modules

Figure 2. The relationship between modules and motifs. **(a)** An example of a feed forward (FF) motif formed by the interaction of two different modules M_A and M_B . **(b)** The percentage of FF motifs formed by the interaction of pairs of modules (InterM) and inside (IntraM) the modules previously identified (M1–M8). The bar shows the distribution of FF in each module. **(c)** A matrix representation of the percentage of FF motifs formed by the interaction of every specific pair of modules. **(d)** An example of a bi-fan (BF) motif formed by the interaction of two different modules M_A and M_B . **(e)** and **(f)** show the same analysis as **(b)** and **(c)** respectively, but focus on the bi-fan motif. The network has a total of 325 FF motifs and 3918 BF motifs.

are composed of functionally related genes (Box 1 and Table 5 in the supplementary material online). To perform this analysis, we defined a control set using the available information in RegulonDB about the expression of the TFs and their regulated genes in a particular condition. In addition, we used information about the functional classes, described by Riley [15], as stored in RegulonDB. Then, we tested how well the modules, shown

in Figure 1a, were integrated by functionally correlated genes (Box 1).

Module 1 (M1) contains genes involved in respiration, M2 contains genes involved in stress response and M3 contains genes involved in chemotaxis, motility and biofilm formation. The module with the greatest number of TFs (23) is involved in the regulation of the different carbon sources (M4). This large module is devoted to

Box 1. Specificity and sensitivity analysis of modules

To quantify how clustered genes and modules are functional correlated, we performed a sensitivity and specificity analysis. First, we identified a set of functionally correlated groups using information from RegulonDB [13]. This information was used as the control set to test how well the hierarchical clustering method grouped genes into functionally correlated groups. Then, we perform the sensitivity and specificity analysis for each module.

Sensitivity was defined as:

$$\text{Sensitivity} = \frac{\text{pres}^+}{\text{pres}^+ + \text{pres}^-}$$

Specificity was defined as:

$$\text{Specificity} = \frac{\text{abs}^-}{\text{abs}^- + \text{abs}^+}$$

where:

pres^+ is the number of genes that belong to the module and that were clustered into that module (true positives).

pres^- is the number of genes that belong to the module but were not clustered into the module (false negatives).

abs^+ is the number of genes that do not belong to the module but were clustered into the module (false positives).

abs^- is the number of genes that do not belong to the module, and that were not clustered into that module (true negatives).

the use of preferential carbon sources and can be divided into smaller sub-modules related to alternative carbon sources. Furthermore, four smaller modules (M5, M6, M7 and M8) of interacting TFs are fully disconnected from the largest connected network component. These involve genes regulating relevant cellular responses such as sulfur assimilation, metabolism of nitrogen sources, fermentative metabolism and chromosome replication.

Global TFs (Box 2) [16] are evenly distributed within major modules. Cyclic AMP receptor protein (CRP) is positioned clearly within the module of carbon metabolism, together with factor for inversion stimulation (FIS) and leucine-responsive regulatory protein (Lrp). Aerobic respiration regulatory protein (ArcA) and Fumarate and nitrate reductase regulatory protein (FNR) belong to the respiration-response module. Histone-like protein or nucleoid-associated protein (Hns) and integration host factor (IHF) belong to the module involved in chemotaxis, motility and biofilm formation (Figure 1b,c).

Relationship between modules and regulons

We distinguish between simple regulons (genes that are regulated by only one TF) and complex regulons (genes that are regulated by the same set of two or more TFs) as defined in Ref. [17]. We define the degree of participation as the largest fraction of TFs and regulated genes of a regulon that belong to one module, comparing both strict complex regulons (SCR) and the non-strict complex

regulons (NSCR), where strict means that a regulator has the same effect (either positive or negative) on every gene in the regulon. The SCRs and NSCRs have an average degree of participation of 0.29 ± 0.16 and 0.24 ± 0.16 , respectively (supplementary material online). On average, complex regulons involve 2.42 different TFs, implying that the probability of each TF of a SCR belonging to a different module is almost 0.75. In this sense regulons are multi-functional. The interesting conclusion is that a single promoter subject to regulation by several proteins, can function as an integrator of TFs that belong to different modules, suggesting that the promoter senses the status of different physiologically relevant cellular entities or modules.

The interrelationship between motifs and modules

Network motifs are statistically significant interconnection patterns with three-to-four genes that characterize the local topology of the regulatory network [11]. It has been shown that these network motifs were not formed by gene duplication but were formed as a result of convergent evolution [18]. When analyzing their distribution within the larger modules, we found that 71% of the FF motifs are inside a single module, whereas this is true for only 35% of the BF motifs (Figure 2b,e, and supplementary material online). This is not surprising because BF motifs are complex regulons, regulating TFs that belong to distinct modules. As described in the supplementary material online, we identified the fraction of motifs that are shared between every pair of modules (Figure 2c,f).

Module M4.6 has the highest degree of overlap with other modules, via BF motifs mainly involving CRP. It also has a high degree of overlap with modules M1 (related to respiration) and M3 (related with chemotaxis, motility and biofilm formation). The connection between M4.6 and M1 makes sense because carbon metabolism is closely related to respiration through the tricarboxylic acids cycle. Modules M4.6 and M3 are highly related, being both involved in nutritional mechanisms. However, the same modules have a lower degree of overlap when they are related via FF motifs. Respiration (M1) is connected with the module of chemotaxis, motility and biofilm formation (M3) via BF motifs (Figure 2f). Certainly, *E. coli* is a facultative, anaerobic organism that uses its motility capability to change environments, seeking optimal oxygen concentrations, a mechanism known as aerotaxis [19]. The genes involved in this mechanism have been identified in *E. coli* [20,21]. These specific cases indicate that the BF motif is the main motif connecting modules, whereas FF motifs are mainly located inside modules.

Concluding remarks

This study was initially motivated by the mathematical suggestion, based on the scale-free distribution of the clustering coefficient, that transcriptional regulation is a hierarchical modular network [12]. We wanted to provide biological content and, at the same time, enhance our detailed understanding of the internal organization of this network. Physiological functions are usually performed by the activity of several proteins. Depending on the particular conditions and cellular responses, genes function coordinately to perform a given task.

Box 2. Definition of global regulators

We used an operational definition of 'global' TFs based on a collection of diagnostic criteria: (i) they regulate many genes; (ii) they regulate several genes encoding TFs; (iii) they cooperate with numerous TFs and together regulate other genes; (iv) they directly affect the expression of a variety of promoters that use different sigma factors; and (v) their regulated genes belong to different functional classes. The analysis supporting these criteria is described in detail in Ref. [16].

We know that genes in bacteria are organized in operons and regulons. More recently, it has been shown that small sets of genes form motifs that convey specific dynamical advantages to cellular responses and coordinated expression [22,23]. We show here the interplay of such motifs with larger groups of physiologically related genes or modules that reflect a higher level of organization of gene content of the cell. The hierarchical organization is demonstrated by FF motifs being confined to individual modules that are devoted to particular cellular processes. Larger modules are not strictly separated because a significant number of BF motifs overlap and, thus, interconnect. This interconnection is also reflected by complex regulons, which are mostly regulated by TFs that belong to different modules. Thus, we have begun to understand in detail the precise structure of the biological complexity of the regulatory network of a single cell.

It came as a surprise that blind computational processing of undirected shortest path distances between regulatory genes (local topological properties of the network) permitted a highly accurate reconstruction of groups of functionally related genes. This is an interesting result validated in *E. coli* and has potentially useful applications in describing the regulatory networks of others organisms.

Acknowledgements

J.F-G. is supported by Ph.D. fellowship 176341 from CONACyT-Mexico. This work is supported by grant GM 62205-03 from NIH to J.C-V. We thank Verónica Jiménez for help extracting the dataset from RegulonDB. We also acknowledge Fabiola Sánchez for her computing support. We acknowledge the Editor for helping to improve this article.

Supplementary data

Supplementary data associated with this article can be found at [doi:10.1016/j.tig.2004.11.010](https://doi.org/10.1016/j.tig.2004.11.010)

References

- Segal, E. *et al.* (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 34, 166–176
- Eisen, M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868
- Hartwell, H. *et al.* (1999) From molecular to modular cell biology. *Nature* 402, C47–C52
- Barabasi, A.L. and Albert, R. (1999) Emergence of scaling in random networks. *Science* 286, 509–512
- Ravasz, E. *et al.* (2002) Hierarchical organization of modularity in metabolic networks. *Science* 297, 1551–1555
- Albert, R. and Barabasi, A.L. (2002) Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74, 47–97
- Ravasz, E. and Barabasi, A.L. (2003) Hierarchical organization in complex networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 67, 026112(<http://link.aps.org/abstract/PRE/v67/e026112>)
- Spiro, S. and Guest, J.R. (1991) Adaptive responses to oxygen limitation in *Escherichia coli*. *Trends Biochem. Sci.* 16, 310–314
- 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
- Shen-Orr, S.S. *et al.* (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* 31, 64–68
- Milo, R. *et al.* (2002) Network motifs: simple building blocks of complex networks. *Science* 298, 824–827
- Dobrin, R. *et al.* (2004) Aggregation of topological motifs in the *Escherichia coli* regulatory network. *BMC Bioinformatics* 5, 10
- Salgado, H. *et al.* (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res.* 32, D303–D306
- Rives, A.W. and Galitski, T. (2003) Modular organization of cellular networks. *Proc. Natl. Acad. Sci. U. S. A.* 100, 1128–1133
- Riley, M. (1997) Genes and proteins of *Escherichia coli* K-12 (GenProtEC). *Nucleic Acids Res.* 25, 51–52
- Martínez-Antonio, A. and Collado-Vides, J. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr. Opin. Microbiol.* 6, 482–489
- Gutiérrez-Ríos, R.M. *et al.* (2003) Regulatory network of *Escherichia coli*: consistency between literature knowledge and microarray profiles. *Genome Res.* 13, 2435–2443
- Conant, G.C. and Wagner, A. (2003) Convergent evolution of gene circuits. *Nat. Genet.* 34, 264–266
- Adler, J. (1966) Effect of amino acids and oxygen on chemotaxis in *Escherichia coli*. *J. Bacteriol.* 92, 121–129
- Rebbapragada, A. *et al.* (1997) The Aer protein and the serine chemoreceptor Tsr independently sense intracellular energy levels and transduce oxygen, redox, and energy signals for *Escherichia coli* behavior. *Proc. Natl. Acad. Sci. U. S. A.* 94, 10541–10546
- Bibikov, S.I. *et al.* (1997) A signal transducer for aerotaxis in *Escherichia coli*. *J. Bacteriol.* 179, 4075–4079
- Rosenfeld, N. and Alon, U. (2003) Response delays and the structure of transcription networks. *J. Mol. Biol.* 329, 645–654
- Mangan, S. *et al.* (2003) The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *J. Mol. Biol.* 334, 197–204

0168-9525/\$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2004.11.010