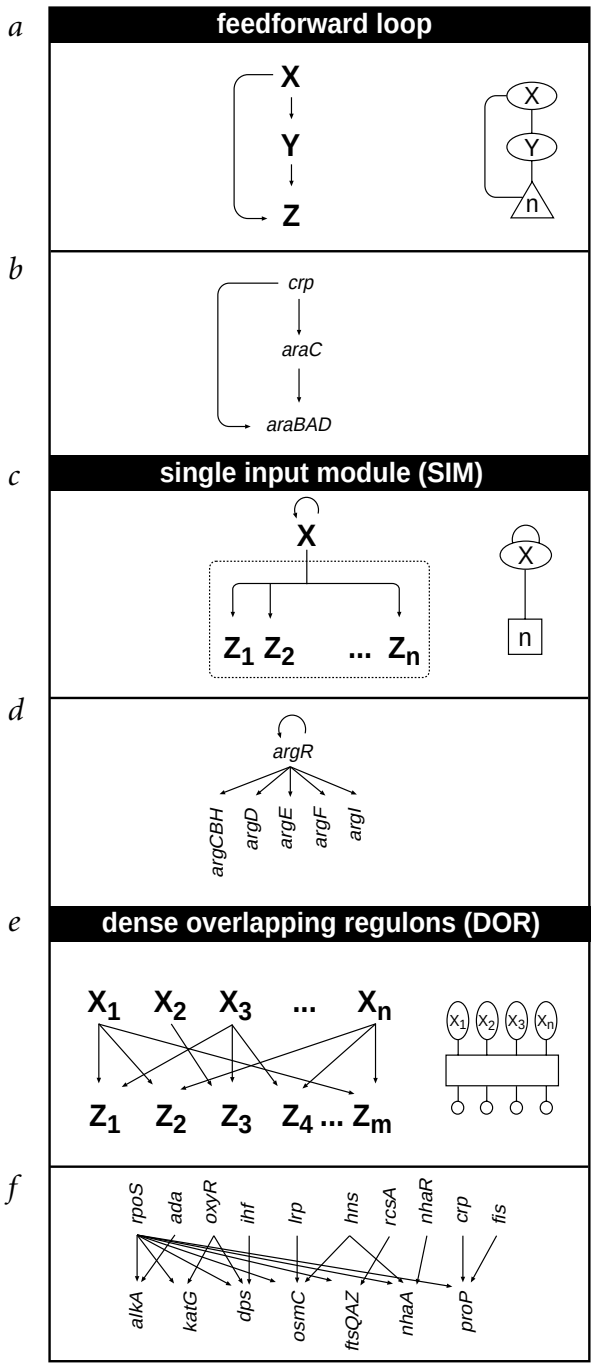
Network motifs in the transcriptional regulation network of *Escherichia coli*

Shai S. Shen-Orr[1], Ron Milo[2], Shmoolik Mangan[1] & Uri Alon[1,2]

Published online: 22 April 2002, DOI: 10.1038/ng881



Little is known about the design principles[1–10] of transcriptional regulation networks that control gene expression in cells. Recent advances in data collection and analysis[2,11,12], however, are generating unprecedented amounts of information about gene regulation networks. To understand these complex wiring diagrams[1–10,13], we sought to break down such networks into basic building blocks[2]. We generalize the notion of motifs, widely used for sequence analysis, to the level of networks. We define 'network motifs' as patterns of interconnections that recur in many different parts of a network at frequencies much higher than those found in randomized networks. We applied new algorithms for systematically detecting network motifs to one of the best-characterized regulation networks, that of direct transcriptional interactions in *Escherichia coli*[3,6]. We find that much of the network is composed of repeated appearances of three highly significant motifs. Each network motif has a specific function in determining gene expression, such as generating temporal expression programs and governing the responses to fluctuating external signals. The motif structure also allows an easily interpretable view of the entire known transcriptional network of the organism. This approach may help define the basic computational elements of other biological networks.

We compiled a data set of direct transcriptional interactions between transcription factors and the operons they regulate (an operon is a group of contiguous genes that are transcribed into a single mRNA molecule). This database contains 577 interactions and 424 operons (involving 116 transcription factors); it was formed on the basis of on an existing database (RegulonDB)[3,14]. We enhanced RegulonDB by an extensive literature search, adding 35 new transcription factors, including alternative σ-factors (subunits of RNA polymerase that confer recognition of specific promoter sequences). The data set consists of established interactions in which a transcription factor directly binds a regulatory site.

The transcriptional network can be represented as a directed graph, in which each node represents an operon and edges represent direct transcriptional interactions. Each edge is directed

Fig. 1 Network motifs found in the *E. coli* transcriptional regulation network. Symbols representing the motifs are also shown. ***a***, Feedforward loop: a transcription factor X regulates a second transcription factor Y, and both jointly regulate one or more operons $Z_1...Z_n$. ***b***, Example of a feedforward loop (L-arabinose utilization). ***c***, SIM motif: a single transcription factor, X, regulates a set of operons $Z_1...Z_n$. X is usually autoregulatory. All regulations are of the same sign. No other transcription factor regulates the operons. ***d***, Example of a SIM system (arginine biosynthesis). ***e***, DOR motif: a set of operons $Z_1...Z_m$ are each regulated by a combination of a set of input transcription factors, $X_1...X_n$. DORs are defined by an algorithm that detects dense regions of connections, with a high ratio of connections to transcription factors. ***f***, Example of a DOR (stationary phase response).

[1]*Department of Molecular Cell Biology,* [2]*Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel. Correspondence should be addressed to U.A. (e-mail: urialon@wisemail.weizmann.ac.il).*

Table 1 • Statistics of occurrence of various structures in the real and randomized networks

Structure	Appearances in real network	Appearances in randomized network (mean ± s.d.)	*P* value
Coherent feedforward loop	34	4.4 ± 3	$P < 0.001$
Incoherent feedforward loop	6	2.5 ± 2	$P \sim 0.03$
Operons controlled by SIM (>13 operons)	68	28 ± 7	$P < 0.01$
Pairs of operons regulated by same two transcription factors	203	57 ± 14	$P < 0.001$
Nodes that participate in cycles*	0	0.18 ± 0.6	$P \sim 0.8$

*Cycles include all loops greater than size 1 (autoregulation). *P* value for cycles is the probability of networks with no loops.

from an operon that encodes a transcription factor to an operon that is regulated by that transcription factor. We scanned the network with algorithms aimed at detecting recurring patterns (see Methods). We evaluated the statistical significance of the network motifs by comparison with randomized networks having the same characteristics as the real *E. coli* network. The probability that a randomized network had an equal or greater number of each of the motifs than the *E. coli* network was determined by enumerating the motifs found in 1,000 randomized networks.

The first motif, termed 'feedforward loop', is defined by a transcription factor X that regulates a second transcription factor Y, such that both X and Y jointly regulate an operon Z (Fig. 1*a*). We term X the 'general transcription factor', Y the 'specific transcription factor', and Z the 'effector operon(s)'. For example, this motif occurs in the L-arabinose utilization system (Fig. 1*b*)[15]. Here, Crp is the general transcription factor and AraC the specific transcription factor. This motif characterizes 40 effector operons in 22 different systems in the network database, with 10 different general transcription factors.

A feedforward loop motif is 'coherent' if the direct effect of the general transcription factor on the effector operons has the same sign (negative or positive) as its net indirect effect through the specific transcription factor. For example, if X and Y both positively regulate Z, and X positively regulates Y, the feedforward loop is coherent. If, on the other hand, X represses Y, then the motif is incoherent. We find that most (85%) of the feedforward loop motifs are coherent (Table 1). Feedforward loops are stylized structures that occur much more frequently in the *E. coli* network than in randomized networks (Table 1, $P < 0.001$).

The second motif, termed single-input module (SIM), is defined by a set of operons that are controlled by a single transcription factor (Fig. 1*c*). All of the operons are under control of the same sign (all positive or all negative) and have no additional transcriptional regulation. The transcription factors controlling SIM motifs are usually autoregulatory (70%, mostly autorepression), in contrast to only 50% of the transcription factors in the complete data set. An example is the arginine biosynthesis pathway, where the transcription factor ArgR uniquely controls five operons that encode arginine biosynthesis genes (Fig. 1*d*). Other amino-acid biosynthesis systems also correspond to this motif. The SIM motif appears in 24 systems in the database (including only systems with three or more operons). Large SIMs occur infrequently in randomized networks (Table 1, $P < 0.01$), because there is a low probability that a large number of operons controlled by a single transcription factor will have no other transcriptional inputs.

The third motif, termed 'dense overlapping regulons' (DOR), is a layer of overlapping interactions between operons and a group of input transcription factors (Fig. 1*e*) that is much more dense than corresponding structures in randomized networks. We find that the sets of genes regulated by different transcription factors in *E. coli* are much more overlapping than expected at random. This can be quantified by the frequency of pairs of genes regulated by the same two transcription factors (Table 1). This does not result, however, in a homogenous mesh of dense interconnections; instead, the network contains several loosely connected, internally dense regions of combinatorial interactions (DORs). As these regions are somewhat overlapping, different criteria can yield slightly different groupings.

We used a clustering approach to define DORs. This algorithm detects locally dense regions in the network with a high ratio of connections to transcription factors (see Methods). This defines six DORs. The operons in each DOR share common biological functions. Typically, every output operon is controlled by a different combination of input transcription factors. In rare cases, termed 'multi-input modules', several operons in a DOR are regulated by precisely the same combination of transcription factors with identical regulation signs. An example of a DOR is the set of operons regulated by RpoS upon entry into stationary phase (Fig. 1*f*)[16]. Different combinations of additional transcription factors, including transcription factors that respond to various stresses and nutrient limitations, control each of these operons. To fully understand the computation performed by each DOR requires a knowledge of the regulatory logic that controls how multiple inputs are integrated at each promoter[17]. A number of DORs as large and dense as in the real *E. coli* network occurs very rarely in randomized networks ($P \sim 0.001$). We note that different clustering rules can give rise to slightly different separations of operons into DORs. The significant finding is that these dense regions of overlapping interactions exist and that they seem to partition the operons into biologically meaningful combinatorial regulation clusters.

The fact that the network motifs appear at frequencies much higher than expected at random suggests that they may have specific functions in the information processing performed by the network. One clue to their possible function is provided by common themes of the systems in which they appear. Additional insight may be gained by mathematical analysis of their dynamics. The feedforward loop motif often occurs where an external signal causes a rapid response of many systems (such repression of sugar utilization systems in response to glucose, shift to anaerobic metabolism). The abundance of coherent feedforward loops, as opposed to incoherent ones, suggests a functional design (Table 1).

Mathematical analysis suggests that the coherent feedforward loop can act as a circuit that rejects transient activation signals from the general transcription factor and responds only to persistent signals, while allowing a rapid system shutdown. This can occur when X and Y act in an 'AND-gate'–like manner to control operon Z (Fig. 2*a*), as is the case in the *araBAD* operon in the arabinose feedforward loop (Fig. 1*b*)[15]. When X is activated, the signal is transmitted to the output Z by two pathways, a direct one

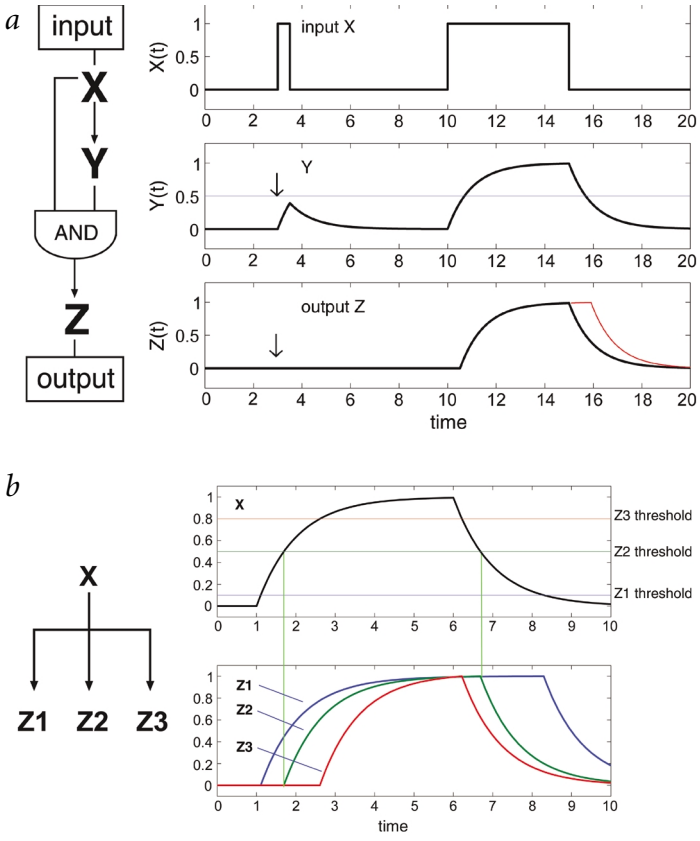


Fig. 2 Dynamic features of the coherent feedforward loop and SIM motifs. ***a***, Consider a coherent feedforward loop circuit with an 'AND-gate'–like control of the output operon Z. This circuit can reject rapid variations in the activity of the input X, and respond only to persistent activation profiles. This is because Y needs to integrate the input X over time to pass the activation threshold for Z (thin line). A similar rejection of rapid fluctuations can be achieved by a cascade, X→Y→Z; however, the cascade has a slower shut-down than the feedforward loop (thin red line in the Z dynamics panel). ***b***, Dynamics of the SIM motif. This motif can show a temporal program of expression according to a hierarchy of activation thresholds of the genes. When the activity of X, the master activator, rises and falls with time, the genes with the lowest threshold are activated earliest and deactivated latest. Time is in units of protein lifetimes, or of cell cycles in the case of long-lived proteins.

from X and a delayed one through Y. If the activation of X is transient, Y cannot reach the level needed to significantly activate Z, and the input signal is not transduced through the circuit. Only when X signals for a long enough time so that Y levels can build up will Z be activated (Fig. 2*a*). Once X is deactivated, Z shuts down rapidly. This kind of behavior can be useful for making decisions based on fluctuating external signals.

The SIM motif is found in systems of genes that function stochiometrically to form a protein assembly (such as flagella) or a metabolic pathway (such as amino-acid biosynthesis). In these cases, it is useful that the activities of the operons are determined by a single transcription factor, so that their proportions at steady state can be fixed. In addition, mathematical analysis suggests that SIMs can show a detailed temporal program of expression resulting from differences in the activation thresholds of the different genes (Fig. 2*b*). Built into this design is a pattern in which the first gene activated is the last one to be deactivated. Such temporal ordering can be useful in processes that require several stages to complete. This type of mechanism may explain the experimentally observed temporal program in the expression of flagella biosynthesis genes[18].

The motifs allow a representation of the *E. coli* transcriptional network (Fig. 3) in a compact, modular form (for an image of the full network, see Web Fig. A online). By using symbols to represent the different motifs (Fig. 1), the network is broken down to its basic building blocks. A single layer of DORs connects most of the transcription factors to their effector operons. Feedforward loops and SIMs often occur at the outputs of these DORs. The DORs are interconnected by the global transcription factors, which typically control many genes in one DOR and few genes in several DORs. An important step in visualizing the network was to allow each global transcription factor to appear multiple times, whenever it is an input to a structure. This reduces the complexity of the interconnections while preserving all the information. There are few long cascades[3], usually involving σ-factors, such as cascades of depth 5 in the flagella and nitrogen systems. Over 70% of the operons are connected to the DORs; the rest of the operons are in small disjoint systems. Most disjoint systems have only 1 to 3 operons. The remaining disjoint systems have up to 25 operons and show many SIMs and feedforward loops. A notable feature of the overall organization is the large degree of overlap within DORs between the short cascades that control most operons. The layer of DORs may therefore represent the core of the computation carried out by the transcriptional network.

Cycles such as feedback loops are an important feature of regulatory networks. Transcriptional feedback loops occur in various organisms, such as the genetic switch in λ-phage[5]. In the *E. coli* data set, there are no examples of feedback loops of direct transcriptional interactions, except for auto-regulatory loops[3]. However, the absence of feedback loops is not statistically significant, as over 80% of the randomized networks also have no feedback loops (Table 1). The many regulatory feedbacks loops in the organism are carried out at the post-transcriptional level.

We considered only transcription interactions specifically manifested by transcription factors that bind regulatory sites[3,14]. This transcriptional network can be thought of as the 'slow' part of the cellular regulation network (time scale of minutes). An additional layer of faster interactions, which include interactions between proteins (often subsecond timescale), contributes to the full regulatory behavior and will probably introduce additional network motifs. Characterization of additional transcriptional interactions may change the present motif assignment for specific systems. However, our conclusions regarding the high frequencies of feedforward loops, SIMs and overlapping regulation compared with randomized networks are insensitive to the addition or removal of interactions from the data set. These features are still highly significant, even when 25% of the connections in the *E. coli* network are removed or rearranged at random.

The concept of homology between genes based on sequence motifs has been crucial for understanding the function of uncharacterized genes. Likewise, the notion of similarity between connectivity patterns in networks, based on network motifs, may be helpful in gaining insight into the dynamic behavior of newly identified gene circuits. The present analysis may serve as a guideline for experimental study of the functions of the motifs. It would be useful to determine whether the network motifs found in *E. coli* can characterize the transcriptional networks of other cell types. In higher eukaryotes, for example, there will be many more regulators affecting each gene, and additional types of circuits may be found. The findings presented here also raise the possibility that motifs can be defined in other biological networks[7], such as signal transduction, metabolic[19] and neuron connectivity networks.

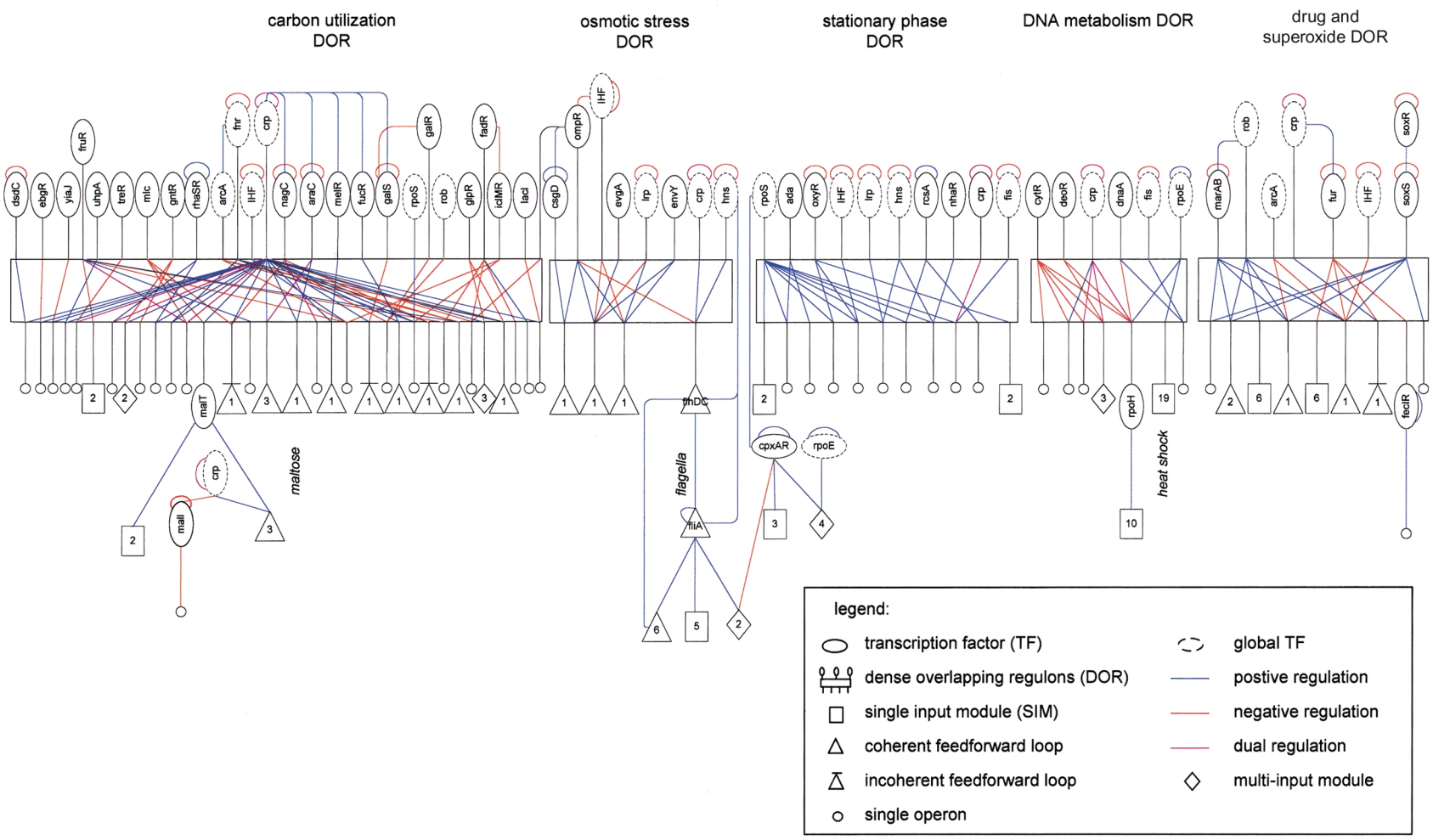


Fig. 3 Part of the network of direct transcriptional interactions in the *E. coli* data set, represented using network motifs. Nodes represent operons, and lines represent transcriptional regulation, directed so that the regulating transcription factor is above the regulated operons. Network motifs are represented by their corresponding symbols (Fig. 1). The DORs are named according to the common function of their output operons. Each transcription factor appears in only a single subgraph, except for transcription factors regulating more than ten operons ('global transcription factors'), which can appear in several subgraphs. For an image of the entire network, see Web Fig. A online.

Methods

Transcriptional interaction database. Data from RegulonDB (v. 3.2, XML format) included 81 transcription factors, with 624 interactions between transcription factors and sites. For this study, we unified interactions with several promoters for the same operon, as well as interactions of a transcription factor with several binding sites in the same promoter region. Unified interactions of different signs (negative/positive) were registered as 'dual'. We did not include interactions of unknown type or those based solely on microarray data. This reduced the effective number of interactions in RegulonDB to 390. We extended RegulonDB data by adding 35 new transcription factors, including alternative σ-factors, and 187 new interactions that we collected through a literature search. In most cases, the new interactions added were supported in the literature both by *in vivo* genetic experiments and by *in vitro* DNA binding data. Most (58%) of the interactions are positive, owing largely to the addition of the alternative σ-factors as transcription factors. Of the 58 autoregulatory interactions (50% of all transcription factors), a majority are autorepressors (70%). The distribution of the number of transcription factors controlling an operon is compact (exponential), whereas the distribution of the number of operons regulated by a transcription factor is long-tailed[10] with an average of approximately 5.

Algorithms for detecting network motifs. The transcriptional network was represented as a connectivity matrix, M, such that $M_{ij} = 1$ if operon j encodes a transcription factor that transcriptionally regulates operon i, and $M_{ij} = 0$ otherwise. We scanned all $n \times n$ submatrices of M, generated by choosing n nodes that lie in a connected graph, for $n = 3$ and $n = 4$. Submatrices were enumerated efficiently by recursively searching for nonzero elements (i,j) and then scanning row i and column j for nonzero elements. The P value for the submatrices representing each type of connected subgraph was evaluated by comparing the number of times they appeared in the real network to the number of times they appeared in the randomized ensemble. For $n = 3$, the only significant motif is the feedforward loop. For $n = 4$, only the overlapping regulation motif, where two operons are regulated by the same two transcription factors (Table 1), was found to be significant. To detect SIMs and multi-input modules, we searched for identical rows of M.

DOR detection. We used an algorithm for detecting dense regions of interactions in the network. All operons regulated by two or more transcription factors were considered. We defined a (nonmetric) distance measure between operons k and j, based on the number of transcription factors regulating both operons: $d(k,j) = 1/(1+ (\Sigma_n f_n M_{k,n} M_{j,n})^2)$, where $f_n = 1/2$ for global transcription factors (transcription factors that regulate more than ten operons); otherwise, $f_n = 1$. Using this distance measure, the operons were clustered with a standard average-linkage algorithm[20]. DORs corresponded to clusters with more than $C = 10$ connections, with a ratio of connections to transcription factors greater than $R = 2$ and a splitting distance[18] larger than the mean splitting distance. Finally, all additional operons (those regulated by a single transcription factor), which are regulated by transcription factors participating in a single DOR, were included in that DOR.

Generation of randomized networks. For a stringent comparison to randomized networks, we generated networks with precisely the same number of operons, interactions, transcription factors and number of incoming and outgoing edges for each node as in the real *E. coli* network. The corresponding randomized connectivity matrices, *Mrand*, have the same number of nonzero elements in each row and column as the corresponding row and column of the real connectivity matrix M; that is: $\Sigma_i Mrand_{ij} = \Sigma_i M_{ij}$, $\Sigma_j Mrand_{ij} = \Sigma_j M_{ij}$. We used a previously described algorithm[13] to generate the randomized networks. Briefly, the proper number of incoming and outgoing edge 'stubs' is assigned to each node. Pairs of in/out edge stubs are randomly chosen and joined, generating a directed graph. We obtained identical results using a Markov-chain algorithm[21], based on starting with the real network and repeatedly swapping randomly chosen pairs of connections ($X \rightarrow Y1$, $X2 \rightarrow Y2$ is replaced by $X1 \rightarrow Y2$, $X2 \rightarrow Y1$) until the network is well randomized. We verified that this yields networks with precisely the same $F(p,q)$, or the number of nodes with p incoming and q outgoing nodes, as the real network.

Mathematical model of network motif dynamics. We used Boolean kinetics[4]. The SIM (Fig. 2*b*) was described by $dZi/dt = F(X,Ti) - aZi$, where Zi and $i = 1,2,3$ are the protein concentrations; the activation thresholds are $T1 = 0.1$, $T2 = 0.5$, $T3 = 0.8$; the cell-cycle time (or lifetime for rapidly degradable proteins) is $a = 1$; and $F(X,T) = 0$ if $X < T$ and 1 if $X \geq T$. The feed-forward loop (Fig. 2*a*) was described by $dY/dt = F(X,Ty) - aY$, $dZ/dt = F(X,Ty)F(Y,Tz) - aZ$, with $Ty = Tz = 0.5$, $a = 1$. The cascade in Fig. 2*a* corresponds to $dY/dt = F(X,Ty) - aY$, $dZ/dt = F(Y,Tz) - aZ$. The dynamics are qualitatively similar if other sigmoidal forms for F are used instead of Boolean kinetics (such as $F(X,T) = X/(T+X)$).

Data availability. The data set is available at http://www.weizmann.ac.il/mcb/UriAlon.

Note: Supplementary information is available on the Nature Genetics website.

Acknowledgments

We thank J. Collado-Vides and the RegulonDB team for making their invaluable database available. We thank A. Arkin, H.C. Berg, J. Doyle, M. Elowitz, S. Leibler, S. Quake, J. Shapiro, M.G. Surette, B. Shilo, E. Winfree and all members of our lab for discussions. This work was supported by the Israel Science Foundation and the Minerva Foundation.

Competing interests statement

The authors declare that they have no competing financial interests.

Received 4 October 2001; accepted 15 March 2002.

1. Bray, D. Protein molecules as computational elements in living cells. *Nature* **376**, 307–312 (1995).
2. Hartwell, L.H., Hopfield, J.J., Leibler, S. & Murray, A.W. From molecular to modular cell biology. *Nature* **402**, C47–52 (1999).
3. Thieffry, D., Huerta, A.M., Perez-Rueda, E. & Collado-Vides, J. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. *Bioessays* **20**, 433–440 (1998).
4. McAdams, H.H. & Arkin, A. Simulation of prokaryotic genetic circuits. *Annu. Rev. Biophys. Biomol. Struct.* **27**, 199–224 (1998).
5. McAdams, H.H. & Shapiro, L. Circuit simulation of genetic networks. *Science* **269**, 650–656 (1995).
6. Savageau, M. & Neidhart, F.C. Regulation beyond the operon. in *Escherichia coli and Salmonella: Cellular and Molecular Biology* (ed. Neidhart, F.C.) 1310–1324 (American Society for Microbiology, Washington D.C., 1996).
7. Strogatz, S.H. Exploring complex networks. *Nature* **410**, 268–276 (2001).
8. Rao, C.V. & Arkin, A.P. Control motifs for intracellular regulatory networks. *Annu. Rev. Biomed. Eng.* **3**, 391–419 (2001).
9. Kauffman, S.A. Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* **22**, 437–467 (1969).
10. Barabasi, A.L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
11. Hughes, J.D., Estep, P.W., Tavazoie, S. & Church, G.M. Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* **296**, 1205–1214 (2000).
12. Hartemink, A.J., Gifford, D.K., Jaakkola, T.S. & Young, R.A. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomput.*, 422–433 (2001).
13. Newman, M.E., Strogatz, S.H. & Watts, D.J. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**, 026118 (2001).
14. Salgado, H. *et al*. RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.* **29**, 72–74 (2001).
15. Schleif, R. Regulation of the L-arabinose operon of *Escherichia coli*. *Trends Genet.* **16**, 559–565 (2000).
16. Hengge-Aronis, R. Survival of hunger and stress: the role of rpoS in early stationary phase gene regulation in *E. coli*. *Cell* **72**, 165–168 (1993).
17. Yuh, C.H., Bolouri, H. & Davidson, E.H. Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* **279**, 1896–1902 (1998).
18. Kalir, S. *et al*. Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. *Science* **292**, 2080–2083 (2001).
19. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. & Barabasi, A.L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
20. Duda, R.O. & Hart, P.E. *Pattern Classification and Scene Analysis* (Wiley, New York, 1973).
21. Kannan, R., Tetali, P. & Vempala, S. Simple Markov-chain algorithms for generating bipartite graphs and tournaments. *Random Structures and Algorithms* **14**, 293–308 (1999).