BMC Bioinformatics

BioMed Central

Research article

Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network

Radu Dobrin[1], Qasim K Beg[1], Albert-László Barabási[2] and Zoltán N Oltvai*[1]

Address: [1]Department of Pathology, Northwestern University, Chicago, IL 60611, USA and [2]Department of Physics, University of Notre Dame, Notre Dame, IN 46556, USA

Email: Radu Dobrin - r-dobrin@northwestern.edu; Qasim K Beg - qasimbeg@northwestern.edu; Albert-László Barabási - alb@nd.edu; Zoltán N Oltvai* - zno008@northwestern.edu

* Corresponding author

Published: 30 January 2004

BMC Bioinformatics 2004, **5**:10

Received: 18 November 2003
Accepted: 30 January 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/10

Abstract

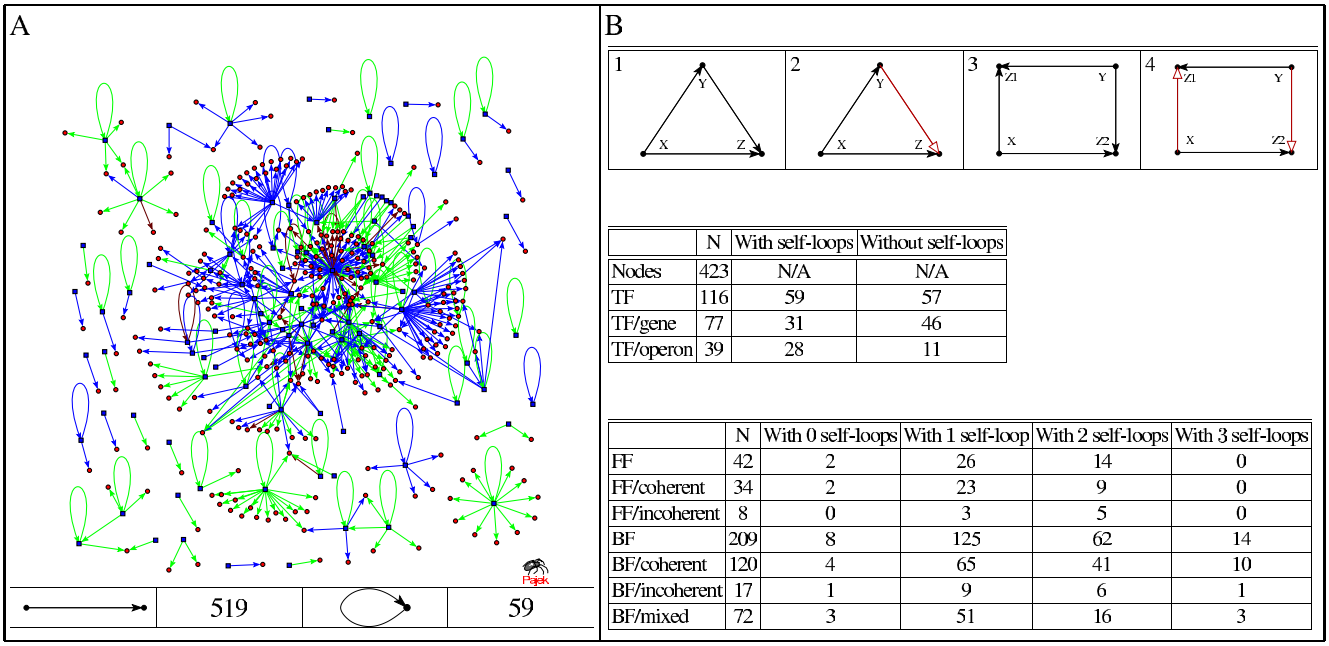
Background: Transcriptional regulation of cellular functions is carried out through a complex network of interactions among transcription factors and the promoter regions of genes and operons regulated by them.To better understand the system-level function of such networks simplification of their architecture was previously achieved by identifying the motifs present in the network, which are small, overrepresented, topologically distinct regulatory interaction patterns (subgraphs). However, the interaction of such motifs with each other, and their form of integration into the full network has not been previously examined.

Results: By studying the transcriptional regulatory network of the bacterium, *Escherichia coli*, we demonstrate that the two previously identified motif types in the network (i.e., feed-forward loops and bi-fan motifs) do not exist in isolation, but rather aggregate into homologous motif clusters that largely overlap with known biological functions. Moreover, these clusters further coalesce into a supercluster, thus establishing distinct topological hierarchies that show global statistical properties similar to the whole network. Targeted removal of motif links disintegrates the network into small, isolated clusters, while random disruptions of equal number of links do not cause such an effect.

Conclusion: Individual motifs aggregate into homologous motif clusters and a supercluster forming the backbone of the *E. coli* transcriptional regulatory network and play a central role in defining its global topological organization.

Background

Many biological functions are carried out by the integrated activity of highly interacting cellular components, referred to as functional modules [1,2]. Motifs, considered as overrepresented topological interaction patterns within complex networks, may represent the simplest building blocks of such modules [3,4]. Owing to their small size, motifs can be explicitly identified and enumerated in various cellular networks, each network being characterized by its own set of distinct motifs [3-5]. For example, triangle motifs, referred to as *feed-forward loops* in directed networks, emerge in both transcriptional regulatory and neural networks, while four node feedback loops represent characteristic motifs in electric circuits, but not in biological systems [4]. The high degree of evolutionary conservation of the motif constituents within the yeast protein interaction network [6], and the convergent evolution observed in the transcriptional regulatory network of diverse species towards the same motif types [7,8] suggest that motifs are indeed of direct biological relevance.



	N	With self-loops	Without self-loops
Nodes	423	N/A	N/A
TF	116	59	57
TF/gene	77	31	46
TF/operon	39	28	11

	N	With 0 self-loops	With 1 self-loop	With 2 self-loops	With 3 self-loops
FF	42	2	26	14	0
FF/coherent	34	2	23	9	0
FF/incoherent	8	0	3	5	0
BF	209	8	125	62	14
BF/coherent	120	4	65	41	10
BF/incoherent	17	1	9	6	1
BF/mixed	72	3	51	16	3

Figure 1

Representation and statistics of the *E. coli* transcriptional regulatory network. *a.* Graphical representation of the network. Blue diamonds represent transcription factors (TF), while the red circles denote the regulated operons. The links are color-coded according to their function: blue-activator, green-repressor, brown-activator or repressor effect. The number of the two types of elementary links, i.e., the autoregulatory loops and the directed links, is listed at the bottom of the panel. ***b.*** The top panel depicts graphical representations of (1) coherent type 1 feed-forward motif (FF), (2) incoherent type 4 FF, (3) coherent bi-fan motif (BF), and (4) incoherent BF (see Ref. [24] for nomenclature). The red links indicate repressors, while all other links denote activator links. Detailed statistics of the nodes (middle panel) and the two statistically significant motifs (bottom panel) found in the network. TF, transcription-factor; TF/gene, transcription-factor encoded as a single gene; TF/operon, transcription-factor encoded as part of an operon; BF/ mixed denotes those bi-fan motifs in which one node receives coherent input signals while the other node receives incoherent input signals.

As the molecular components of a specific motif often have interactions with nodes (i.e., the TFs and operons) outside of the motif, we need to address how the various motifs relate to these nodes and to each other. Here, we demonstrate that in the *E. coli* transcriptional regulatory network, the vast majority of motifs (feed-forward loops or bi-fan motifs) overlap, generating distinct topological units called *homologous motif clusters*. These clusters merge into a single large connected component, called *motif supercluster*, in which the specific motif clusters are no longer clearly separable. As motifs are present in all cellular networks examined to date [4], it is likely that the aggregation of motifs into motif clusters and superclusters is a general property of cellular networks.

Results

To identify successive layers of hierarchies in the local topological features of the *E. coli* transcriptional regulatory network, we first established its layout based on published data [3,9]. Following the representation of Shen-Orr et al [3], we associate the *E. coli* transcriptional regulatory network with a directed graph in which each node represents a gene or an operon encoding a transcription factor (TF) and the gene or operon regulated by the TF, while the links denote the TFs themselves. Note, that many TFs are encoded within an operon, thus the directed links represent direct transcriptional modulation from the TF to an operon, or a TF-contained operon to another operon (Fig. 1A). This representation allows us to distinguish between two different elementary links: 59 autoregulatory loops, in which a TF regulates its own expression, and 519 directed links, in which a TF regulates another TF or operon (Fig. 1A). Note, that about half of the 116 TFs have an autoregulatory loop. For those TFs that are encoded as single genes the same trend is also evident, while for the TFs that are encoded as part of an operon a significantly higher proportions possess autoregulatory loops (Fig. 1B, middle panel).

First organizational level: motifs

Motifs can be explicitly identified and enumerated in various cellular networks [3-7]. Within the *E. coli* transcriptional regulatory network we detected the two previously described motifs with uniform topology [3,7], i.e., the

feed-forward and *bi-fan* motifs (Fig. 1B, top panel). Both motifs can be further classified by the functionality of their links (activating or inhibitory). In a *coherent* feed-forward or bi-fan motif all the directed links are activating (Fig. 1B, top panel, 1 and 3), while in *incoherent* motifs one of the links inhibits the activity of its target node (Fig. 1B, top panel, 2 and 4). We find that coherent motif types are significantly more common than incoherent ones both for feed-forward and bi-fan motifs (Fig. 1B, bottom panel). We can further group the detected motifs according to the number of autoregulatory loops they possess, finding that both motifs have predominantly one or two autoregulatory loops, while no motif has an autoregulatory loop associated with each of its nodes (Fig. 1B, bottom panel).

Second organizational level: homologous motif clusters

While statistically significant motifs can be explicitly identified and enumerated, the nodes (i.e., the TFs and operons) that take part in such motifs do not exist in isolation but almost always have additional interactions with nodes outside the motif. To systematically identify such interactions, we first searched for feed-forward motifs that share at least one link and/or node with another feed-forward motif (Fig. 2A). We have also performed a similar search for bi-fan motifs that interact with each other in this manner (Fig. 2B). We find that in the *E. coli* transcriptional regulatory network the vast majority of motifs overlap generating distinct topological units that we refer to as *homologous motif clusters* (Fig. 2A,2B).

Forty-one of the 42 individual feed-forward motifs coalesce into six feed-forward motif clusters (Fig. 2A). Of these six motif clusters, three have one highly shared link, while a shared node plays a critical role in establishing the other three motif clusters (Fig. 2A). Similarly, 208 of the 209 bi-fan motifs join together into just two bi-fan motif clusters in which most of the links are shared by at least two adjacent motifs, and also among multiple motifs (Fig. 2B). The majority of links within the motif clusters are either activating or inhibitory (Figs. S1, S2, see Additional file 1), suggesting that most of the network motifs do not function in isolation but are embedded into a multi-level hierarchy of regulatory interactions. This notion is further supported by the finding that in both cases many of the topological motif clusters overlap to a large extent with known biological functions. For example, one of the feed-forward motif clusters largely overlaps with the flagella motor module, while another contains a significant number of elements responsible for regulating the aerobic/ anaerobic switch in *E. coli* (see Additional file 1 for details). While some of the motif clusters are topologically highly similar, the number of links connecting them to other network constituents is uneven. For example, the cluster encompassing most elements of the flagella motor module is relatively isolated, yet the topologically highly similar cluster overlapping the aerobic/ anaerobic switch is densely integrated with other motifs (Fig. 2C). This suggests that despite their highly similar topology, they may display qualitatively different dynamical features.

Third organizational level: motif supercluster

The homologous motif clusters are not isolated either, but are embedded into the *E. coli* transcriptional regulatory network as a whole. To understand the topological relations between different homologous motif clusters, we merged all feed-forward and bi-fan homologous motif clusters, finding that they form a single large connected component (i.e., *motif supercluster*) in which the previously identified feed-forward and bi-fan homologous motif clusters are no longer clearly separable. Indeed, we find only one feed-forward- and one bi-fan motif to be isolated from the obtained supercluster (Fig. 2C). This integration is especially evident for the feed-forward motif clusters, the vast majority of which share the same links with the bi-fan motif clusters (Fig. 2C).

The relationship of organizational levels to the global network topology

When considering the full *E. coli* transcriptional regulatory network undirected, the statistical analysis of the cumulative of its connectivity distribution demonstrates that it belongs to a class of scale-free networks [10], as previously described [3,11] (Fig. 2E), with embedded topological hierarchy [12] (Fig. 2F), and having a single connected giant component (Fig. 1A, also see Fig. S3, Additional file 1, for separate in- and out-degree distributions). To study the global relationship of motifs with the whole topological architecture of the network, we overlay the heterologous motif superclusters on the full network (Fig. 2D). It is visually evident that all the nodes from the single giant heterologous motif supercluster are part of the giant component of the full network, comprising 41.46% of its nodes and 53.53 % of its links, respectively. In fact, it appears that the heterologous motif supercluster defines the core of the connected giant component with most other nodes being connected to its nodes (Fig. 2D). Compared to the heterologous motif superstructure, the FF motif clusters use only 20.42% nodes and 21.84% links, while the BF motifs comprise 30.48% nodes and 38.32% links.

To test if the heterologous motif supercluster in fact represents the backbone of the connected giant component, we have examined the effect of removing all 250 links of the supercluster (that mimics the lack of TF binding to a promoter region) from the network [13]. Removing these 250 links (out of a total of 467) fragmented the network into 29 small, isolated subgraphs (Fig. 3A). In contrast, while the removal of 250 randomly chosen links disconnected

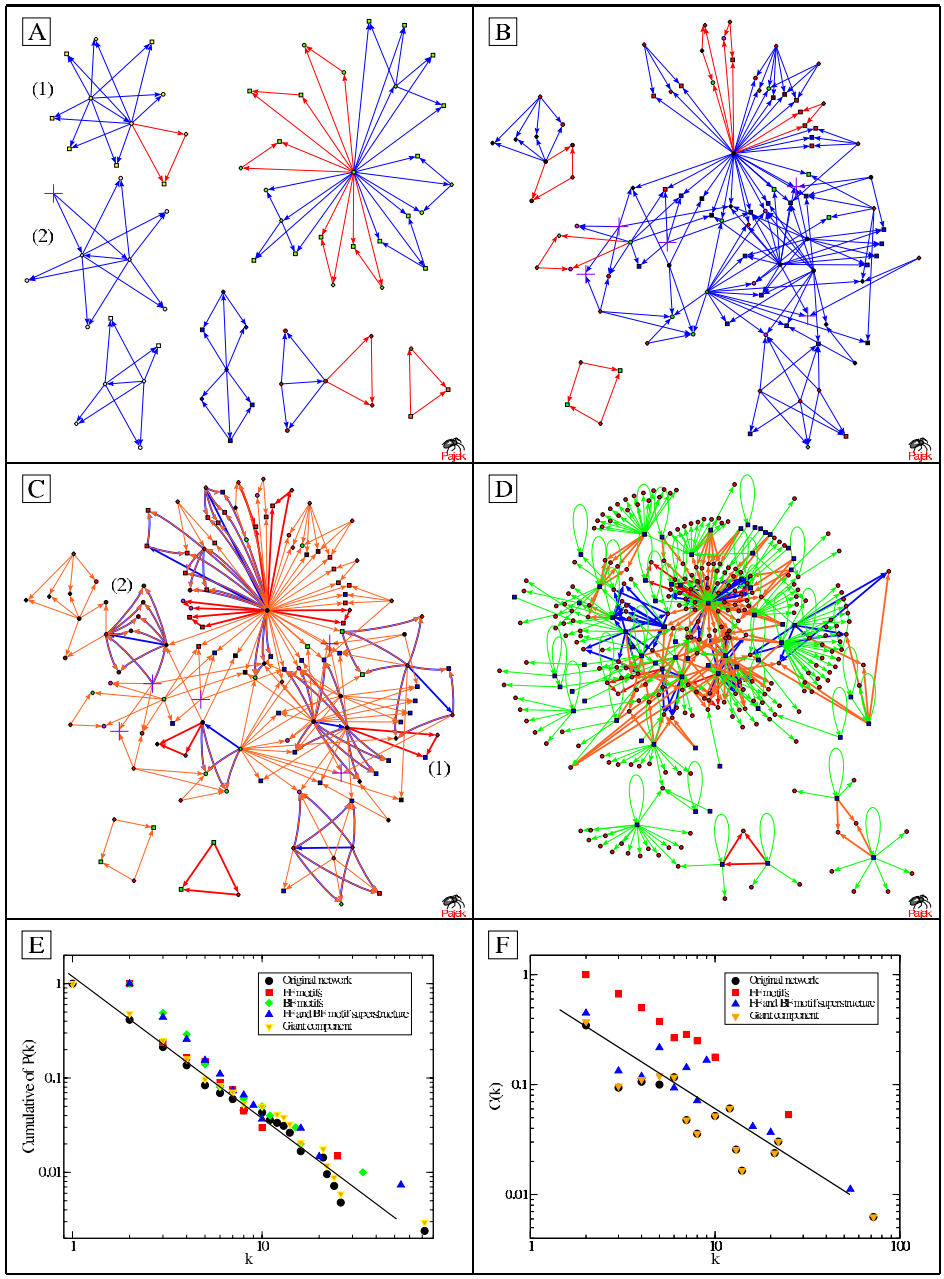


Figure 2

Motif clusters and superclusters. ***a.*** The 42 detected feed-forward loops join together forming six homologous motif clusters, and one isolated motif. ***b***. The 209 bi-fan motifs coalesce into two homologous motif clusters, and one isolated motif. In (***a,b***) motifs which share links are shown in blue, otherwise they are colored in red. The purple cross represents nodes that were originally provided in Reference 3. ***c***. The feed-forward- and bi-fan motif clusters together form a heterologous motif superstructure. The vertex coordinates are shown as in (*b*). The feed-forward motif clusters are colored as in (*a*). The thick lines mark the links involved in feed-forward motifs, while curved thick lines are those shared among links of feed-forward and bi-fan motifs. As in *a*, the flagella motor cluster (2) and the aerobic/ anaerobic switch clusters (1) are indicated. ***d.*** The connected giant component of the complete *E. coli* transcriptional regulatory network contains all components of the motif superstructures, colored as in (**c**). In panel **e**, we plot the cumulative connectivity distribution, P(k), for the original network (shown in Fig. 1a), and for the networks from panels 3*a-d*, respectively. The solid black line has an exponent $\gamma = -1.5$, and provides the best fit for the original network (black circles). The clustering distribution, C(k), as a function of connectivity for the same networks is shown in panel ***f***. The solid black line has slope $\tau = -1$, and is the best fit for all networks. The clustering coefficient of a node is a measure of its near-neighbors connectivity, and thus for the BF motifs this value is zero.

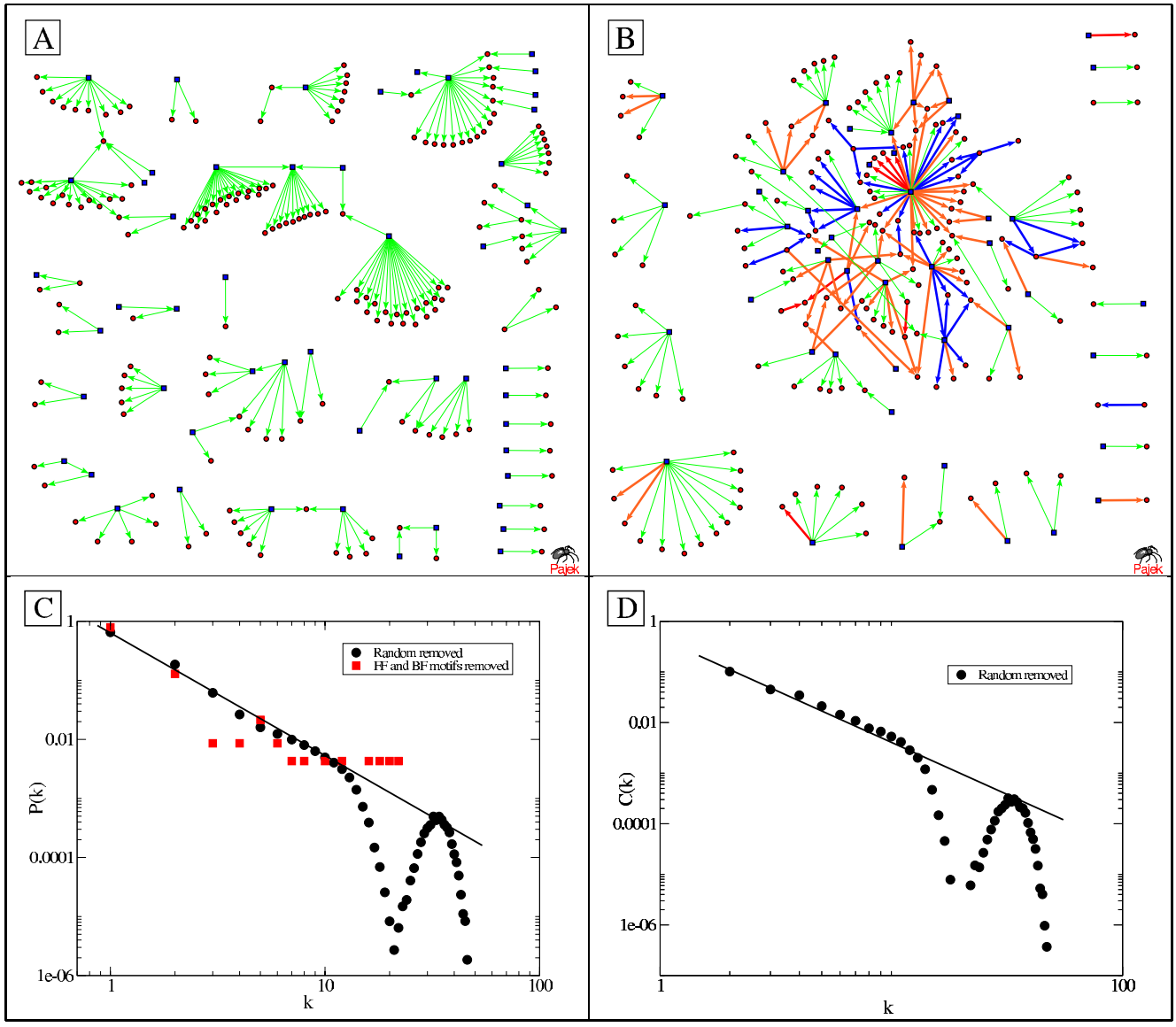


Figure 3

Network fragmentation under random- or targeted link removal. ***a.*** We have removed the links among nodes of the FF and BF motif clusters from the original network. ***b.*** The same number of links as in (*a*) are removed randomly. In order to determine qualitatively the difference between removing the links of FF and BF motif cluster and randomly removing the same number of links, we have determined (***c***) the connectivity distribution, P(k), and (***d***) the clustering distribution, C(k) of the remaining networks. The distributions for random removal of links have been averaged over 5000 different samples. The solid line in panel *c* has slope γ = -2, and is the best fit for the random link removal. The solid line in panel *d* has slope τ = -2. We do not show C(k) for the network in panel *a*, since the clustering coefficients for all nodes by definition are zero. The deviation from the solid line in panels *c*,*d*, are caused by the TF, *crp*, which represent the most connected node of the network.

the network into 16 small subgraphs, a connected giant component was retained (Fig. 3B). To quantitatively characterize the two types of reduced networks we compared the statistical features of the network following the removal of the 250 supercluster links (Fig. 3A) against 5,000 different realizations of randomized removal of the same number of links (of which one realization is shown in Fig. 3B). For networks perturbed by random link removal the cumulative of the connectivity distribution (Fig. 3C), and the scaling of C_k and k (Fig. 3D) were rela-

tively unaltered, being reminiscent to that observed for the original network (Fig. 2E,2F). However, for the network in which those links contributing to the supercluster were missing the scaling of C_k and k was completely absent (Fig. 3D). This observation quantitatively demonstrates the collapse of the network structure and its inherent topological hierarchy upon the targeted removal of the links of the motif supercluster.

Discussion

Most cellular functions are significantly influenced by the activity of transcriptional regulatory networks within living cells [14,15]. Identifying the connections [5,16,17], and decoding the organizational principles and system-level features [3,4,14,18-20] of such networks is a key challenge of post-genomic biology. Here, we extend the established motif framework [3-5] for the systematic identification of topological organizational layers within the *E. coli* transcriptional regulatory network. We show that the overwhelming majority of motifs combine to form homologous motif clusters, which further coalesce into a supercluster that serves as the backbone of the whole network. In the absence of the links that constitute the motif supercluster the network is fragmented into small isolated components.

These findings bring up a number of important questions. First, how do the identified topological features define and restrict the dynamical activity of the network? The various types of TF binding to promoter elements of genes and operons are able to establish a number of various output activities, as demonstrated both experimentally and through theoretical work, especially when small genetic circuits are considered [18-23]. However, the emergence of motif clusters and the motif superstructure suggests that the dynamical features of operon activities may be modified compared to that seen in individual motifs [24,25]. Yet, the universal presence of motifs within biological and non-biological networks, and their apparent type selection according to the function of the given network [4], strongly suggest that the dynamical range of activity is ultimately restricted by the observed topology.

From a global perspective it appears that a scale-free architecture with embedded hierarchical modularity is a general feature of cellular networks [12]. Thus, the unexpected finding that the motif supercluster represents the core determinant of the network topology suggests that similar organizational layers will also be found in other biological networks. Also, the convergent evolution of the same motif types in the transcriptional regulatory network of *E. coli* and *Saccharomyces cerevisiae* implies a sign of optimal design [7]. The identified organizational layers may thus represent an outcome of a unique balance between evolutionary processes, specific cellular function, and the dynamical range required for a robust overall functionality within the highly variable environment in which most living organisms exist.

Conclusions

This analysis demonstrates that individual motifs of the *E. coli* transcriptional regulatory network aggregate into homologous motif clusters and supercluster that are key determinants of the network's global topological organization, and which may represent distinct organizational hierarchies of transcriptional regulation. As motifs are present in all cellular networks examined to date, it is likely that the aggregation of motifs into motif clusters and superclusters is a general property of all cellular networks.

Methods

Database and motif identification

For the transcriptional regulatory network of *E. coli* we have used the data provided by Alon and coworkers [3] that is largely based on the Regulon DB [9], and was downloaded from http://www.weizmann.ac.il/mcb/UriAlon/ (version 1.1). The resulting network has 423 operons and 578 regulatory interactions, and the links defined by these interactions are directed. For detecting all n-node network motifs we used a method similar to that of Milo *et al* [4]. Briefly, the method is based on the identification of motifs by searching all rows of the adjacency matrix M [4]. For each non-zero element (i,j) representing a link, it scans through all neighbors of (i,j). This is performed recursively for all other elements $(i,k),(k,i),(k,j)$ and (j,k) until a specific n-node motif is detected. Subsequently, the detected motifs are compared to the motifs found in previous steps and eliminated if they are already in the database. 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: $\Sigma_i Mrand_{i,j} = \Sigma_i M_{ij}$. To generate the randomized networks we used a Markov-chain algorithm, as previously described [26]. Briefly, a Markov-chain algorithm is based on starting with the real network and repeatedly exchanging randomly chosen pairs of connections (X1Y1, X2Y2 is replaced by X1Y2, X2Y1) until the network is well randomized. Similar results can be obtained using the connectivity matrix M. In this case, the algorithm creates a randomized connectivity matrix with the same number of non-zeroes on each row and column, as in the original connectivity matrix. Starting with an empty matrix a random element M_{mn} is chosen and it is set

to 1 if its previous value was 0. The process is repeated until all rows and columns have the same number of nonzero elements as in the original matrix.

Characterization of global network features

We have determined the cumulative of the connectivity distribution P(k) for various subnetworks of the original regulatory network by counting all nodes with a given connectivity distribution. All networks are assumed undirected, the degree distribution k_i being defined as the total number of incoming and outgoing links for a given node *i*. It is possible to measure the in-degree and out-degree distribution, although the results will be noisy due to the relative small size of the subnetworks.

The clustering coefficient of a node gives the probability that its neighbors are connected to each other. It is defined as the number of directed links between the nearest neighbors of a node *i* divided by the total possible number of links between them. Thus, it is higher for a highly connected group of neighbors, while a loosely connected group has a clustering coefficient close to 0. It is important to notice that in a feed-forward loop (triangle) motif all nodes have a clustering coefficient $C_i = 1$, while in a bi-fan (square) motif (or in any loop with more than 4 links not reducible to lower order loops) all nodes have clustering coefficient $C_i = 0$. Since the clustering coefficient determines how connected a network is, we have disregarded the links' direction.

Authors' contributions

RD carried out all theoretical analyses, while biological analyses were carried out by QKB. ALB and ZNO provided guidance, coordinated the biological and theoretical analyses, and prepared the manuscript.

Acknowledgements

We thank G. Balázsi for discussion. Research at the University of Notre Dame and Northwestern University was supported by grants from the National Institute of Health (NIGMS) and the Department of Energy Genomes to Life Program.

References

1. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology.** *Nature* 1999, **402:**C47-52.
2. Wolf DM, Arkin AP: **Motifs, modules and games in bacteria.** *Curr Opin Microbiol* 2003, **6:**125-134.
3. Shen-Orr SS, Milo R, Mangan S, Alon U: **Network motifs in the transcriptional regulation network of Escherichia coli.** *Nat Genet* 2002, **31:**64-68.
4. Milo R, Shen-Orr SS, Itzkovitz S, Kashtan N, Alon U: **Network motifs: simple building blocks of complex networks.** *Science* 2002, **298:**824-827.
5. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA: **Transcriptional regulatory networks in Saccharomyces cerevisiae.** *Science* 2002, **298:**799-804.
6. Wuchty S, Oltvai ZN, Barabási A-L: **Evolutionary conservation of motif constituents within the yeast protein interaction network.** *Nature Genetics* 2003, **35:**176-179.
7. Conant GC, Wagner A: **Convergent evolution of gene circuits.** *Nature Genet* 2003, **34:**264-266.
8. Hinman VF, Nguyen AT, Cameron RA, Davidson EH: **Developmental gene regulatory network architecture across 500 million years of echinoderm evolution.** *Proc Natl Acad Sci U S A* 2003, **100:**13356-13361.
9. Salgado H, Santos-Zavaleta A, Gama-Castro S, Millan-Zarate D, Diaz-Peredo E, Sanchez-Solano F, Perez-Rueda E, Bonavides-Martinez C, Collado-Vides J: **RegulonDB (version 3.2): transcriptional regulation and operon organization in Escherichia coli K-12.** *Nucleic Acids Res* 2001, **29:**72-74.
10. Barabási A-L, Albert R: **Emergence of scaling in random networks.** *Science* 1999, **286:**509-512.
11. Guelzim N, Bottani S, Bourgine P, Kepes F: **Topological and causal structure of the yeast transcriptional regulatory network.** *Nat Genet* 2002, **31:**60-63.
12. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási A-L: **Hierarchical organization of modularity in metabolic networks.** *Science* 2002, **297:**1551-1555.
13. Albert R, Jeong H, Barabási A-L: **Error and attack tolerance of complex networks.** *Nature* 2000, **406:**378-382.
14. Thieffry D, Huerta AM, Perez-Rueda E, Collado-Vides J: **From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli.** *Bioessays* 1998, **20:**433-440.
15. Wyrick JJ, Young RA: **Deciphering gene expression regulatory networks.** *Curr Opin Genet Dev* 2002, **12:**130-136.
16. Simon I, Barnett J, Hannett N, Harbison CT, Rinaldi NJ, Volkert TL, Wyrick JJ, Zeitlinger J, Gifford DK, Jaakkola TS, Young RA: **Serial regulation of transcriptional regulators in the yeast cell cycle.** *Cell* 2001, **106:**697-708.
17. Zeitlinger J, Simon I, Harbison CT, Hannett NM, Volkert TL, Fink GR, Young RA: **Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling.** *Cell* 2003, **113:**395-404.
18. Rosenfeld N, Elowitz MB, Alon U: **Negative autoregulation speeds the response times of transcription networks.** *J Mol Biol* 2002, **323:**785-793.
19. Buchler NE, Gerland U, Hwa T: **On schemes of combinatorial transcription logic.** *Proc Natl Acad Sci U S A* 2003, **100:**5136-5141.
20. Yildirim N, Mackey M: **Feedback regulation in the lactose operon: a mathematical modeling study and comparison with experimental data.** *Biophys J* 2003, **84:**2841-2851.
21. Becskei A, Serrano L: **Engineering stability in gene networks by autoregulation.** *Nature* 2000, **405:**590-593.
22. Becskei A, Seraphin B, Serrano L: **Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion.** *Embo J* 2001, **20:**2528-2535.
23. Setty Y, Mayo AE, Surette MG, Alon U: **Detailed map of a cis-regulatory input function.** *Proc Natl Acad Sci U S A* 2003, **100:**7702-7707.
24. Mangan S, Alon U: **Structure and function of the feed-forward loop network motif.** *Proc Natl Acad Sci U S A* 2003, **100:**11980-11985.
25. Mangan S, Zaslaver A, Alon U: **The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks.** *J Mol Biol* 2003, **334:**197-204.
26. Maslov S, Sneppen K: **Specificity and stability in topology of protein networks.** *Science* 2002, **296:**910-913.