Network Motifs: Simple Building Blocks of Complex Networks

R. Milo,1 S. Shen-Orr,1 S. Itzkovitz,1 N. Kashtan,1 D. Chklovskii,2 U. Alon1*

Complex networks are studied across many fields of science. To uncover their structural design principles, we defined “network motifs,” patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks. We found such motifs in networks from biochemistry, neurobiology, ecology, and engineering. The motifs shared by ecological food webs were distinct from the motifs shared by the genetic networks of Escherichia coli and Saccharomyces cerevisiae or from those found in the World Wide Web. Similar motifs were found in networks that perform information processing, even though they describe elements as different as biomolecules within a cell and synaptic connections between neurons in Caenorhabditis elegans. Motifs may thus define universal classes of networks. This approach may uncover the basic building blocks of most networks.

Many of the complex networks that occur in nature have been shown to share global statistical features (1–10). These include the “small world” property (1–9) of short paths between any two nodes and highly clustered connections. In addition, in many natural networks, there are a few nodes with many more connections than the average node has. In these types of networks, termed “scale-free networks” (4, 6), the fraction of nodes having k edges, p(k), decays as a power law p(k) ~ k−γ (where γ is often between 2 and 3). To go beyond these global features would require an understanding of the basic structural elements particular to each class of networks (9). To do this, we developed an algorithm for detecting network motifs: recurring, significant patterns of interconnections. A detailed application to a gene regulation network has been presented (11). Related methods were used to test hypotheses on social networks (12, 13). Here we generalize this approach to virtually any type of connectivity graph and find the striking appearance of

Supporting Online Material
www.sciencemag.org/cgi/content/full/298/5594/821/DC1
Figs. S1 to S3
Tables S1 and S2
22 July 2002; accepted 9 September 2002

Fig. 1. (A) Examples of interactions represented by directed edges between nodes in some of the networks used for the present study. These networks go from the scale of biomolecules (transcription factor protein X binds regulatory DNA regions of a gene to regulate the production rate of protein Y), through cells (neuron X is synaptically connected to neuron Y), to organisms (X feeds on Y). (B) All 13 types of three-node connected subgraphs.
motifs in networks representing a broad range of natural phenomena.

We started with networks where the interactions between nodes are represented by directed edges (Fig. 1A). Each network was scanned for all possible \( n \)-node subgraphs (in the present study, \( n = 3 \) and \( 4 \)), and the number of occurrences of each subgraph was recorded. Each network contains numerous types of \( n \)-node subgraphs (Fig. 1B). To focus on those that are likely to be important, we compared the real network to suitably randomized networks (12–16) and only selected patterns appearing in the real network at numbers significantly higher than those in the randomized networks (Fig. 2).

For a stringent comparison, we used randomized networks that have the same single-node characteristics as does the real network. Each node in the randomized networks has the same number of incoming and outgoing edges as the corresponding node has in the real network. The comparison to this randomized ensemble accounts for patterns that appear only because of the single-node characteristics of the network (e.g., the presence of nodes with a large number of edges). Furthermore, the randomized networks used to calculate the significance of \( n \)-node subgraphs were generated to preserve the same number of appearances of all \((n-1)\)-node subgraphs as in the real network (17, 18). This ensures that a high significance was not assigned to a pattern only because it has a highly significant subpattern. The “network motifs” are those patterns for which the probability \( P \) of appearing in a randomized network an equal or greater number of times than in the real network is lower than a cutoff value (here \( P = 0.01 \)).

Patterns that are functionally important but not statistically significant could exist, which would be missed by our approach.

We applied the algorithm to several networks from biochemistry (transcriptional gene regulation), ecology (food webs), neurobiology (neuron connectivity), and engineering (electronic circuits, World Wide Web). The network motifs found are shown in Table 1. Transcription networks are biochemical networks responsible for regulating the expression of genes in cells (11, 19). These are directed graphs, in which the nodes represent genes (Fig. 1A). Edges are directed from a gene that encodes for a transcription factor protein to a gene transcriptionally regulated by that transcription factor. We analyzed the two best characterized transcriptional regulation networks, corresponding to organisms from different kingdoms: a eukaryote (the yeast *Sacccharomyces cerevisiae* (20)) and a bacterium (*Escherichia coli*) (11, 19). The two transcription networks show the same motifs: a three-node motif termed “feedforward loop” (11) and a four-node motif termed “bi-fan.” These motifs appear numerous times in each network (Table 1), in nonhomologous gene systems that perform diverse biological functions. The number of times they appear is more than 10 standard deviations greater than their mean number of appearances in randomized networks. Only these subgraphs, of the 13 possible different three-node subgraphs (Fig. 1B) and 199 different four-node subgraphs, are significant and are therefore considered network motifs. Many other three- and four-node subgraphs recur throughout the networks, but at numbers that are less than the mean plus 2 standard deviations of their appearance in randomized networks.

We next applied the algorithm to ecosystem food webs (21, 22), in which nodes represent groups of species. Edges are directed from a node representing a predator to the node representing its prey. We analyzed data collected by different groups at seven distinct ecosystems (22), including both aquatic and terrestrial habitats. Each of the food webs displayed one or two three-node network motifs and one to five four-node network motifs. One can define the “consensus motifs” as the motifs shared by networks of a given type. Five of the seven food webs shared one three-node motif and all seven shared one four-node motif (Table 1). In contrast to the three-node motif (termed “three chain”), the three-node feedforward loop was underrepresented in the food webs. This suggests that direct interactions between species at a separation of two layers [as in the case of omnivores (23)] are selected against. The biparallel motif indicates that two species that are prey of the same predator both tend to share the same prey. Both network motifs may thus represent general tendencies of food webs (21, 22).

We next studied the neuronal connectivity network of the nematode *Caenorhabditis elegans* (24). Nodes represent neurons (or neuron
classes), and edges represent synaptic connections between the neurons. We found the feedforward loop motif in agreement with anatomical observations of triangular connectivity structures (24). The four-node motifs include the bi-fan and the bi-parallel (Table 1). Two of these motifs (feedforward loop and bi-fan) were also found in the transcriptional gene regulation networks. This similarity in motifs may point to a fundamental similarity in the design constraints of the two types of networks. Both network functions carry information from sensory components (sensory neurons/transcription factors regulated by biochemical signals) to effectors (motor neurons/structural genes). The feedforward loop motif common to both types of networks may play a functional role in information processing. One possible function of this circuit is to activate output only if the input signal is persistent and to allow a rapid deactivation when the input goes off (17). Indeed, many of the input nodes in the neural feedforward loops are sensory neurons, which may require this type of information processing to reject transient input fluctuations that are inherent in a variable or noisy environment.

We also studied several technological networks. We analyzed the ISCAS89 benchmark set of sequential logic electronic circuits (7, 25). The nodes in these circuits represent logic gates and flip-flops. These nodes are linked by directed edges. We found that the motifs separate the circuits into classes that correspond to the circuit’s functional description. In Table 1, we present two classes, consisting of five forward-logic chips and three digital fractional multipliers. The digital fractional multipliers share three motifs, including three- and four-node feedback loops. The forward logic chips share the feedforward loop, bi-fan, and bi-parallel motifs, which are similar to the motifs found in the genetic and neuronal information-processing networks. We found a different set of motifs in a network of directed hyperlinks between World Wide Web pages within a single domain (4). The World Wide Web motifs may reflect a design aimed at short paths between related pages. Application of our approach to nondirected networks shows distinct sets of motifs in networks of protein interactions and Internet router connections (18).

None of the network motifs shared by the food webs matched the motifs found in the gene regulation networks or the World Wide Web. Only one of the food web consensus motifs also appeared in the neuronal network. Different motif sets were found in electronic circuits with different functions. This suggests that motifs can define broad classes of networks, each with specific types of elementary structures. The motifs reflect the underlying processes that generated each type of network; for example, food webs evolve to allow a flow of energy from the bottom to the top of food chains, whereas genes regulation and neuron networks evolve to process information. Information processing seems to give rise to significantly different structures than does energy flow.

We further characterized the statistical significance of the motifs as a function of network size, by considering pieces of various sizes (subnetworks) of the full network. The concentration of motifs in the subnetworks is about the same as that in the full network (Fig. 3). In contrast, the concentration of the corresponding subgraphs in the randomized versions of the subnetworks decreases sharply with size. In analogy with statistical physics, the number of appearances of each motif in the real networks

**Table 1.** Network motifs found in biological and technological networks. The numbers of nodes and edges for each network are shown. For each motif, the numbers of appearances in the real network (N_{real}) and in the randomized networks (N_{rand} ± SD, all values rounded) (17, 18) are shown. The P value of all motifs is P < 0.01, as determined by comparison to 1000 randomized networks (100 in the case of the World Wide Web). As a qualitative measure of statistical significance, the Z score = (N_{real} - N_{rand})/SD is shown. NS, not significant. Shown are motifs that occur at least U = 4 times with completely different sets of nodes. The networks are as follows (18): transcription interactions between regulatory proteins and genes in the bacterium E. coli (11) and the yeast S. cerevisiae (20); synaptic connections between neurons in C. elegans, including neurons connected by at least five synapses (24); trophic interactions in ecological food webs (22), representing pelagic and benthic species (Little Rock Lake), birds, fishes, invertebrates (Ythan Estuary), primarily larger fishes (Chesapeake Bay), lizards (St. Martin Island), primarily invertebrates (Skipwith Pond), pelagic lake species (Bridge Brook Lake), and diverse desert taxa (Coachella Valley); electronic sequential logic circuits parsed from the ISCAS89 benchmark set (7, 25), where nodes represent logic gates and flip-flops (presented are all five possible scans of forward-logic chips and three digital fractional multipliers in the benchmark set); and World Wide Web hyperlinks between Web pages in a single domain (4) (only three-node motifs are shown), e, multiplied by the power of 10 (e.g., 1.46e6 = 1.46 × 10^6).

<table>
<thead>
<tr>
<th>Network</th>
<th>Nodes</th>
<th>Edges</th>
<th>N_{real}</th>
<th>N_{rand} ± SD</th>
<th>Z score</th>
<th>N_{real}</th>
<th>N_{rand} ± SD</th>
<th>Z score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene regulation (transcription)</td>
<td>Feed-forward loop</td>
<td>Bi-fan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E. coli</td>
<td>424</td>
<td>519</td>
<td>70</td>
<td>11 ± 4</td>
<td>14</td>
<td>1812</td>
<td>300 ± 40</td>
<td>13</td>
</tr>
<tr>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>W</td>
<td>V</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>W</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>685</td>
<td>1052</td>
<td>70</td>
<td>11 ± 4</td>
<td>14</td>
<td>1812</td>
<td>300 ± 40</td>
<td>13</td>
</tr>
<tr>
<td>Neurons</td>
<td>Feed-forward loop</td>
<td>Bi-parallel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>W</td>
<td>V</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>W</td>
</tr>
<tr>
<td>Food webs</td>
<td>Feed-forward loop</td>
<td>Bi-parallel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>W</td>
<td>V</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>W</td>
</tr>
<tr>
<td>Electronic circuits (forward logic chips)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>W</td>
<td>V</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>W</td>
</tr>
<tr>
<td>Electronic circuits (digital fractional multipliers)</td>
<td>Bi-fan</td>
<td>Four-node feedback loop</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>W</td>
<td>V</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>W</td>
</tr>
<tr>
<td>World Wide Web</td>
<td>Feedback with two mutual dyads</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>W</td>
<td>V</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>W</td>
</tr>
<tr>
<td>nd.edu</td>
<td>3.25</td>
<td>0.729</td>
<td>1.46e6</td>
<td>5.6</td>
<td>3e3</td>
<td>9e3</td>
<td>1.8</td>
<td>1.2</td>
</tr>
</tbody>
</table>

*Has additional four-node motif: (X → Z, W; Y → Z, W; Z → W), N_{real} = 150, N_{rand} = 85 ± 15, Z = 4. †Has additional four-node motif: (X → Y, Z; Y → Z, W), N_{real} = 204, N_{rand} = 80 ± 20, Z = 6. The three-node pattern (X → Y, Z; Y → Z, W; Z → Y) also occurs significantly more than at random. It is not a motif by the present definition because it does not appear with completely distinct sets of nodes more than U = 4 times. ‡Has additional four-node motif: (X → Y, Z; Y → Z, W; Z → X), N_{real} = 914, N_{rand} = 500 ± 70, Z = 6. §Has two additional three-node motifs: (X → Y, Z; Y → Z, W; Z → Y), N_{real} = 3e5, N_{rand} = 1.46e3 ± 6e1, Z = 6000, and (X → Y, Z → Y), N_{real} = 5e5, N_{rand} = 9e4 ± 1.5e3, Z = 250.
Progression of Vertebrate Limb Development Through SHH-Mediated Counteraction of GLI3

Pascal te Welscher,1 Aimée Zuniga,1 Sanne Kuijper,2 Thij Drenth,1 Hans J. Goedemans,1 Frits Meijlink,2 Rolf Zeller1*

Distal limb development and specification of digit identities in tetrapods are under the control of a mesenchymal organizer called the polarizing region. Sonic Hedgehog (SHH) is the morphogenetic signal produced by the polarizing region in the posterior limb bud. Ectopic anterior SHH signaling induces digit duplications and has been suspected as a major cause underlying congenital malformations that result in digit polydactyly. Here, we report that the polydactyly of GLI3-deficient mice arises independently of SHH signaling. Disruption of one or both GLI3 alleles in mouse embryos lacking Shh progressively restores limb distal development and digit formation. Our genetic analysis indicates that SHH signaling counteracts GLI3-mediated repression of key regulator genes, cell survival, and distal progression of limb bud development.

The Hedgehog (Hh) signaling pathway controls many key developmental processes during animal embryogenesis (1). In Drosophila embryos, all known functions of Hh signaling are mediated by the transcriptional effector Cubitus interruptus (Ci) (2). Several homologs of Hh and Ci have been identified in higher vertebrates. In particular, Sonic Hedgehog (SHH) and the Ci homolog GLI3 are required for vertebrate limb development (3–6). GLI3 acts first during the initiation of limb bud development and before the activation of SHH signaling in posterior restriction of the basic helix-loop-helix transcription factor dHAND. dHAND in turn prevents GLI3 expression from spreading posteriorly (Fig. 1A, panel 1) (7). In addition, GLI3 restricts the SHH-independent early expression of 3’HoxD genes and Gremlin to the posterior mesenchyme (8). Subsequently, dHAND functions in the activation of Shh expression (9). Limb bud morphogenesis is then controlled by reciprocal interactions of two signaling centers (Fig. 1A, panel 2): the polarizing region, an instructive organizer located in the posterior limb bud mesenchyme, and the apical ectodermal ridge (AEI), SHH signaling by the polarizing region in combination

References and Notes
17. The randomized networks used for detecting three-rod motifs preserve the numbers of incoming, outgoing, and double edges with both incoming and outgoing arrows for each node. The randomized networks used for detecting four-rod motifs preserve the above characteristics as well as the numbers of all 13 three-rod subgraphs as in the real network. Algorithms for constructing these randomized network ensembles are described (16). Additional information is available at www.weizmann.ac.il/mcb/UriAlon.
18. Methods are available as supporting material on Science Online.
26. In Erdos-Renyi randomized networks with a fixed connectivity (2), the concentration of a subgraph with n nodes and k edges scales with network size as C ∼ n2k−6−1, thus C ∼ 1/3 for the feedforward loop of Fig. 3 where n ∼ k = 3. The sole exception in Table 1 in which C should not vanish at large size is the three-chain pattern in food webs where n = 3 and k = 2.
28. We thank S. Maslov and K. Sneppen for valuable discussions. We thank J. Collado-Vides, N. Martinez, R. Gorduan, R. Durbin, L. Amaral, R. Cancho, S. Maslov, and K. Sneppen for kindly providing data, as well as D. Alon, E. Domany, M. Elowitz, I. Kanter, O. Hobart, M. Naor, D. Mukamel, A. Murray, S. Quake, R. Raz, M. Reigl, M. Surette, K. Sneppen, P. Steinberg, E. Winfree, and members of our lab for comments. We thank Caltech and the Aspen Center for Physics for their hospitality during part of this work. We acknowledge support from the Israel Science Foundation, the Human Frontier Science Program, and the Minerva Foundation.

Supporting Online Material
www.sciencemag.org/cgi/content/full/298/5594/824/DC1
Methods
Table S1

1 May 2002; accepted 10 September 2002

*To whom correspondence should be addressed. E-mail: R.Zeller@bio.uu.nl