Supplementary Web material for Milo RE 1073478, sched. for 25 October issue

Methods

Network Databases

Transcriptional networks. The *E. coli* network was described (*S1*) and is available at www.weizmann.ac.il/mcb/UriAlon. The *S. cerevisiae* transcriptional network is based on the YPD database (*S2*). Interactions between transcription factor proteins and genes are included. Each protein complex of transcription factors is represented by a single node. This network is available at www.weizmann.ac.il/mcb/UriAlon.

Food webs. The database of seven ecosystem food webs, provided by N. Martinez, was described (*S3, S4*).

Neuronal network. We used the data (*S5*) as employed in (*S6*), with only neurons that have five or more synaptic connections. The same motifs are found also in a smaller, more stringent data set of 69 neuron classes, representing neurons with five or more synapses found in at least three of four sides of two animals studied, represented in figure 8 of (*S7*).

Electronic circuits. Electronic circuits were directly parsed from the ISCAS89 benchmark data set (*S8*), available at www.cbl.ncsu.edu/CBL_Docs/iscas89.html. The parsed networks are available at www.weizmann.ac.il/mcb/UriAlon.

World Wide Web. We used the database of (*S9*), available at www.nd.edu/~networks/database/index.html.

Internet. We used nondirected connections representing a router-level map (*S10*), provided by R. Govindan (source: www.isi.edu/~honqsuda/pub/int081099.adj.gz).

Generation of Randomized Networks

Two different algorithms were used to generate randomized networks with the same incoming and outgoing degree per node as the real network. The two algorithms gave identical results for the subgraph statistics.

Algorithm A. We employed a Markov-chain algorithm (*S11, S12*), based on starting with the real network and repeatedly swapping randomly chosen pairs of connections (X1 → Y1, X2 → Y2 is replaced by X1 → Y2, X2 → Y1) until the network is well randomized. Switching is prohibited if the either of the connections X1 → Y2 or X2 → Y1 already exist.

Algorithm B. Identical statistics were obtained by using a direct construction algorithm, modified from (*S13*). As in algorithm A, this algorithm does not allow spurious multiple connections between nodes (more than one directed connection between two nodes). Each network was presented as a connectivity matrix **M**, such that $M_{ij} = 1$ if there is a connection directed from node i to node j, and 0 otherwise. The goal is to create a randomized connectivity matrix M_{rand}, which has the same number of nonzero elements in each row and column as the corresponding row and column of the real connectivity matrix: $R_i = \sum_j M_{\text{rand}, ij} = \sum_j M_{ij}$, $C_i = \sum_i M_{\text{rand},ij} = \sum_i M_{ij}$. To generate the randomized networks, we start with an empty matrix M_{rand}. We then repeatedly randomly choose a row n according to the weights $p_i = R_i/\sum R_i$ and a column m according to the weights $q_j = R_j/\sum R_j$. If $M_{\text{rand},nm} = 0$, we set $M_{\text{rand},mn} = 1$. We then set $R_m = R_m - 1$ and $\mathrm{C}_n = \mathrm{C}_n - 1$. If the entry ($m$, n) was previously entered to the randomized matrix, that is, if $M_{\text{rand},mn} = 1$, or if $m = n$, we choose a new (m, n). This process is repeated until all $R_i = 0$ and $C_j = 0$. Rarely the algorithm can find no solution, and the process is started from scratch.

Controlling for Appearances of (*n* – 1)-Node Motifs

We generate a series of randomized network ensembles, each of which has the same (n – 1)-node subgraph count as the real network, as a null hypothesis for detecting n-node motifs. This is done to avoid assigning high significance to a structure only because of the fact that it includes a highly significant substructure.

For a null hypothesis randomized network as a basis for detecting three-node motifs, we preserve the numbers of the in- and outgoing edges for each node, as well as the number of mutual edges (X ←→ Y) for each node. This is implemented with algorithm A, treating double edges and single edges separately. A double edge is switched only with a different double edge (X1 ←→ Y1, X2 ←→ Y2 to X1 ←→ Y2, X2 ←→ Y1), and only if both (X1 and Y2) and (X2 and Y1) are unconnected by an edge in any direction. Similarly, the single directed edge switches (X1 → Y1, X2 → Y2 is replaced by X1 → Y2, X2 → Y1) are performed only if they do not form new double edges.

For a random null hypothesis network for assigning significance to the four-node subgraphs, we generate randomized networks that have the same three-node subgraph counts as the real network. This is done with a Metropolis Monte-Carlo approach (*S14*). Let $V_{\text{real},k}$, $k = 1\ldots13$ be the number of appearances of each of the 13 three-node subgraphs (Fig. 1B) in the real network and $V_{\text{rand},k}$ be the corresponding vector in the randomized network. We define an energy $E = \sum_k |V_{\text{real},k} - V_{\text{rand},k}|/(V_{\text{real},k} + V_{\text{rand},k})$. The energy E is zero only when all the three-node subgraph counts of the real and randomized graphs are equal. We start by fully randomizing the network according to algorithm A above. Then, we generate a random switch (X1 → Y1, X2 → Y2 to X1 → Y2, X2 → Y1, and similarly for double edges, as described above). If this switch lowers E, it is accepted. Otherwise, it is accepted with probability $\exp(-\Delta E/T)$, where ΔE is the difference in energy before and after the switch and T is an effective temperature. This process is repeated, with a simulated annealing regiment (*S14, S15*) to lower T slowly until a solution with $E = 0$ is obtained. This can be readily generalized to form $(n - 1)$-node null-hypothesis networks for detecting n-node motifs also for $n > 4$.

Network Motif Detection

To efficiently count all connected n-node subgraphs in a connectivity matrix **M**, the algorithm loops through all rows i. For each nonzero element (i, j), it loops through all connected elements $M_{ik} = 1$, $M_{ki} = 1$, $M_{jk} = 1$, and $M_{kj} = 1$. This is recursively repeated with elements (i, k), (k, i), (j, k), and (k, j) until an n-node subgraph is obtained. A table is formed that counts the number of appearances of each type of subgraph in the network, correcting for the fact that multiple submatrices of **M** can correspond to one isomorphic architecture owing to symmetries. This process is repeated for each of the randomized networks. The number of appearances of each type of subgraph in the random ensemble is recorded, to assess its statistical significance. The present concepts and algorithms are easily generalized to nondirected or directed graphs with several "colors" of edges and nodes, multipartite graphs, and so forth.

Criteria for Network Motif Selection

For the purposes of the present study, network motifs are subgraphs that meet the following criteria:

(i) The probability that it appears in a randomized network an equal or greater number of times than in the real network is smaller than $P = 0.01$. In the present study, P was estimated (or bounded) by using 1000 randomized networks.

(ii) The number of times it appears in the real network with distinct sets of nodes is at least $U = 4$.

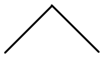
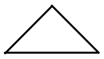
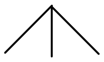
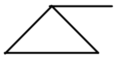
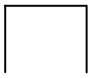
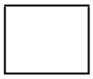
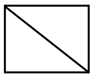
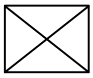
(iii) The number of appearances in the real network is significantly larger than in the randomized networks: $N_{\mathrm{real}} - N_{\mathrm{rand}} > 0.1N_{\mathrm{rand}}$. This is done to avoid detecting as motifs some common subgraphs that have only a slight difference between N_{rand} and N_{real} but have a narrow distribution in the randomized networks.

Algorithms for Nondirected Networks

Algorithm A was used, treating all edges as double edges as described above.

Network Motifs in Nondirected Networks

Table S1. Subgraphs and motifs in nondirected networks. Shown are the two types of three-node and six types of four-node nondirected subgraphs, as well as their concentration C in two networks (C is the fraction of times a given n-node subgraph occurs among the total number of occurrences of all possible n-node subgraphs). The networks are a 1843 node/2203 edge yeast protein-interaction database (*S16*) and a 228,262 node/320,147 edge database of connections between internet routers (*S10*). Motifs are indicated along with their Z score. ND, not determined because of the fact that the subgraph did not appear in the randomized network ensemble. Anti-motifs are subgraphs that satisfy the following: (i) the probability that they appear in randomized networks fewer times than in the real network is $P < 0.01$ and (ii) $N_{\mathrm{rand}} - N_{\mathrm{real}} > 0.1N_{\mathrm{rand}}$.

Pattern	**Protein interactions**	**Internet routers**
	Not a motif *C* = 0.981	Not a motif *C* = 0.977
	Motif (*Z* = 48) *C* = 0.019	Motif (*Z* = 4600) *C* = 0.023
	Motif (*Z* = 15) *C* = 0.680	Not a motif *C* = 0.931
	Anti-motif (*Z* = -19) *C* = 0.024	Motif (*Z* = 18) *C* = 0.013
	Anti-motif (*Z* = –18) *C* = 0.292	Anti-motif (*Z* = –7) *C* = 0.048
	Not a motif *C* = 0.0013	Motif (*Z* = 356) *C* = 0.004
	Anti-motif (*Z* = -4.5) *C* = 0.0019	Motif (*Z*=137) *C* = 0.002
	Not a motif *C* = 0.0004	Motif (*Z* ND) *C* = 0.0005

References and Notes

S1. S. Shen-Orr, R. Milo, S. Mangan, U. Alon, *Nature Genet.* **31**, 64 (2002).

S2. M. C. Costanzo *et al.*, *Nucleic Acids Res.* **29**, 75 (2001).

S3. R. Williams, N. Martinez, *Nature* **404**, 180 (2000).

S4. N. Martinez, *Ecol. Monogr.* **61**, 367 (1991).

S5. J. White, E. Southgate, J. Thomson, S. Brenner, *Philos. Trans. R. Soc. London Ser. B* **314** (1986).

S6. L. Amaral, A. Scala, M. Barthelemy, H. Stanley, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 11149 (2000).

S7. R. Durbin, *Cambridge University*, 1 (1987).

S8. F. Brglez, D. Bryan, K. Kozminski, *Proc. IEEE Int. Symp. Circuits Syst.* 1929 (1989).

S9. A.-L. Barabási, R. Albert, *Science* **286**, 509 (1999).

S10. R. Govindan, H. Tangmunarunkit, *Proc. IEEE Infocom 2000* (2000).

S11. R. Kannan, P. Tetali, S. Vempala, *Random Struct. Algorithms* **14**, 293 (1999).

S12. S. Maslov, K. Sneppen, *Science* **296**, 910 (2002).

S13. M. Newman, S. Strogatz, D. Watts, *Phys. Rev. E* **64**, 6118 (2001).

S14. M. Newman, G. Barkema, *Monte Carlo Methods in Statistical Physics* (Oxford Univ. Press, New York, 1999).

S15. H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, A. L. Barabasi, *Nature* **407**, 651 (2000).

S16. H. Jeong, S. Mason, A. L. Barabasi, Z. N. Oltvai, *Nature* **411**, 41 (2001).