Network motif identification in stochastic networks

Rui Jiang, Zhidong Tu, Ting Chen[†], and Fengzhu Sun[†]

Molecular and Computational Biology Program, University of Southern California, Los Angeles, CA 90089

Edited by Philip P. Green, University of Washington School of Medicine, Seattle, WA, and approved May 8, 2006 (received for review September 8, 2005)

Network motifs have been identified in a wide range of networks across many scientific disciplines and are suggested to be the basic building blocks of most complex networks. Nonetheless, many networks come with intrinsic and/or experimental uncertainties and should be treated as stochastic networks. The building blocks in these networks thus may also have stochastic properties. In this article, we study stochastic network motifs derived from families of mutually similar but not necessarily identical patterns of interconnections. We establish a finite mixture model for stochastic networks and develop an expectation-maximization algorithm for identifying stochastic network motifs. We apply this approach to the transcriptional regulatory networks of *Escherichia coli* and *Saccharomyces cerevisiae*, as well as the protein–protein interaction networks of seven species, and identify several stochastic network motifs that are consistent with current biological knowledge.

expectation-maximization algorithm | mixture model | transcriptional regulatory network | protein–protein interaction network

Networks are ubiquitous and abundant in many scientific fields. Examples include the World Wide Web, electronic circuits, social networks, biological interaction networks, etc. Many networks have been shown to share global statistical features (1, 2), such as the "small world" property of short paths between any two nodes and highly clustered connections (3, 4). It has also been shown that many networks are "scale-free" networks, in which the node degrees follow a power-law distribution (5, 6). Recent studies have shown that many networks contain a small set of "network motifs," that is, patterns of interconnections occurring in networks at numbers that are significantly higher than those in randomized networks that are uniformly drawn from the networks with the same degree distributions as the original networks (7, 8). These network motifs may define universal classes of networks in that similar motifs have been found in a wide variety of networks, ranging from the World Wide Web to the electronic circuits, from the transcriptional regulatory networks of *Escherichia coli* to the neural network of *Caenorhabditis elegans*. The research on network motifs is therefore promising in uncovering the basic building blocks of most complex networks.

In most studies (7–10), both the networks and the motifs are deterministic in that connections between nodes are represented by the presence/absence of edges. Nevertheless, many natural networks have intrinsic uncertainties. For example, in a living cell, DNA binding proteins are believed to be in an equilibrium between the bound and unbound states, thus introducing uncertainties in protein–DNA interactions. Similar circumstance holds for protein–protein interactions, which are crucial to cellular functions both in assembling protein machinery and in signaling cascades. Consequently, both transcriptional regulatory networks and protein–protein interaction networks have intrinsic uncertainties. Additionally, incomplete and/or incorrect observations due to experimental resolutions, systematic errors, and random noises also introduce considerable uncertainties into the observed networks. This situation prevails in biological interaction networks constructed by using data collected by high-throughput techniques such as the yeast two-hybrid assays (11, 12) and the chromatin immunoprecipitation (ChIP) method (13, 14). With the intrinsic and experimental uncertainties, it is more suitable to associate connections between nodes with probabilities. Consequently, a network with N nodes can be described by a probability matrix $\mathbf{P} = (\pi_{ij})_{N\times N}$, $0 \leq \pi_{ij} \leq 1$. π_{ij} is the probability that two nodes i and j are connected. In this article, we refer to networks in which connections are described by probabilities as *stochastic networks*, and networks in which connections are described by the presence/absence of edges as *deterministic networks*. A stochastic network can be thought of as a family of similar deterministic networks.

Network motifs also have uncertainties. For example, the evolution of regulatory pathways in cells is itself a stochastic process, and some protein–DNA interactions can change without affecting the functionality of the pathways. Functionally related network motifs are therefore not necessarily topologically identical. Topological variations in network motifs may also arise because of incomplete and/or incorrect observations resulting from experimental noises. Hence, it is more appropriate to model network motifs as stochastic patterns and discuss the network motif identification problem in the circumstance of stochastic networks.

A stochastic network can be thought of as coming into being by embedding a family of mutually similar interconnection patterns (subgraphs) in a background random ensemble with a probability λ. The set of patterns defines a foreground stochastic network motif and is described by a probability matrix $\Theta_1 = (\theta_{ij})_{n\times n}$, $0 \leq \theta_{ij} \leq 1$. θ_{ij} is the probability that node i and node j are connected. The background ensemble is characterized by the degree distributions of the given stochastic network (Θ_0). With such a mixture model, the network motif can be recovered by fitting the stochastic network with a foreground motif and a suitable background ensemble. This process requires estimating the occurrence probability of the motif (λ), the parameters associated with the motif (Θ_1), and the statistical properties associated with the background ensemble. In this article, we develop a pseudo-maximum likelihood framework and an expectation-maximization (EM) algorithm to estimate these parameters. The statistical significance of the identified motif is quantified by a p value estimated by a pseudo-likelihood ratio test approach.

Results

Data Sources. We studied a wide range of available biological networks, including transcriptional regulatory networks and protein–protein interaction networks. In regulatory networks, nodes are genes or corresponding proteins, and directed interactions are the regulatory relationship between the proteins (transcription factors) and the genes. In protein–protein interaction networks, nodes are proteins, and undirected connections are physical interactions between the proteins. We downloaded the data sets of *E. coli* and *Saccharomyces cerevisiae* regulatory networks from Uri Alon's laboratory (7, 8) and the data sets of protein–protein interaction networks for seven species [*E. coli*, *S. cerevisiae* (core), *C. elegans*, *Helicobacter pylori*, *Mus musculus*, *Drosophila melanogaster*, and *Homo sapiens*] from the Database of Interacting Proteins (15, 16). These data sets come from human curated databases and

Conflict of interest statement: No conflicts declared.

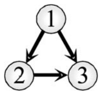
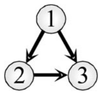
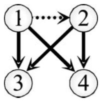
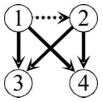
This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: ChIP, chromatin immunoprecipitation; EM, expectation-maximization.

[†] To whom correspondence may be addressed at: MCB 201, 1050 Childs Way, Los Angeles, CA 90089. E-mail: tingchen@usc.edu or fsun@usc.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0507841103

Table 1. Stochastic motifs in highly reliable regulatory networks

Species	LR	λ	Θ_1	Motif
E. coli	6.92×10^{1}	2.75×10^{-6}	$\begin{bmatrix} 0.00 & 1.00 & 1.00 \\ 0.00 & 0.00 & 1.00 \\ 0.00 & 0.00 & 0.00 \end{bmatrix}$	
S. cerevisiae	1.10×10^{2}	1.06×10^{-6}	$\begin{bmatrix} 0.00 & 1.00 & 1.00 \\ 0.00 & 0.00 & 1.00 \\ 0.00 & 0.00 & 0.00 \end{bmatrix}$	
E. coli	2.07×10^{2}	1.30×10^{-7}	$\begin{bmatrix} 0.00 & 0.01 & 0.97 & 0.98 \\ 0.00 & 0.00 & 0.99 & 0.99 \\ 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}$	
S. cerevisiae	3.23×10^{3}	1.80×10^{-7}	$\begin{bmatrix} 0.00 & 0.04 & 1.00 & 1.00 \\ 0.00 & 0.00 & 1.00 & 1.00 \\ 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}$	

LR, log pseudo-likelihood ratio.

are supposed to be reliable. We also downloaded the ChIP-chip data sets for the *S. cerevisiae* regulatory network from Young's laboratory (13, 14). This data set is less reliable, and each protein–DNA interaction pair is assigned a *p* value, indicating the confidence of the interaction. The details of these data sets are presented in the supporting information, which is published on the PNAS web site.

Results on Simulated Networks. We use a simulation method to verify that our approach can recover motifs embedded in stochastic networks. The results show that the EM algorithm can accurately estimate the parameters (λ and Θ_1), while the pseudo-likelihood test can successfully reject the null hypothesis even for small λ values ($\approx 10^{-6}$). The details of the method and the results are presented in the supporting information.

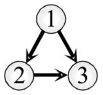
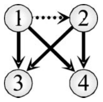
Results on Transcriptional Regulatory Networks. We apply our method to find 3- and 4-node network motifs in the regulatory networks of *E. coli* and *S. cerevisiae* (7, 8). Because these networks are highly reliable, we assign $\pi_{ij} = 1$ for any interaction pair (i, j) in the data set and $\pi_{ij} = 0$, otherwise.

In the 3-node case, we identify *feed forward loop* motifs (a transcription factor regulates another, while they both regulate a third target gene) for both species (Table 1, upper). For *E. coli*, the motif exists with $\lambda = 2.75 \times 10^{-6}$ and is significant with a *p* value of $<10^{-4}$; for *S. cerevisiae*, the motif exists with $\lambda = 1.06 \times 10^{-6}$ and is significant with a *p* value of $<10^{-4}$. Recent studies have shown that the feed forward loop serves as a sensitive delay element in regulatory networks (7, 17). It can speed up the response time of the target gene's expression following stimulus steps in one direction (e.g., off to on) but not in the other direction (on to off).

In the 4-node case, we identify *stochastic bi-fan* motifs for both species (Table 1, lower). For *E. coli*, the motif exists with $\lambda = 1.30 \times 10^{-7}$ and is significant with a *p* value of $<10^{-3}$; for *S. cerevisiae*, the motif exists with $\lambda = 1.80 \times 10^{-7}$ and is significant with a *p* value of $<10^{-4}$. The deterministic bi-fan motif (two transcription factors regulate two target genes in parallel; no interaction between the two transcription factors) has been identified previously (7, 8). The newly identified stochastic bi-fan motif has a novel feature (one transcription factor could also regulate the other) and provides us a more general presentation of the combinatorial transcriptional regulation in living cells.

We also apply our approach to the *S. cerevisiae* regulatory network constructed by using the ChIP-chip data (13, 14). The data set contains genome-wide protein–DNA interaction analysis of 113 transcription factors and 6,270 target genes. Each probed interaction is assigned a *p* value, indicating the confidence of the interaction. At the recommended *p* value threshold (0.001), the observed network contains 2,416 nodes and 4,344 edges with a false positive rate of 10% and a false negative rate of 18% (13). We therefore infer that ≈434 (4,344 × 0.1) observed interactions are false positives, while ≈858 (4,344 × 0.9 × 0.18/0.82) interactions are actually missing. We use two methods as presented in the supporting information to assign interaction probabilities for protein–DNA pairs and apply our approach to identify network motifs in the constructed stochastic networks. The results are shown in Table 2. In the 3-node case (upper portion of Table 2), a motif similar to the feed forward loop is identified ($\lambda = 7.49 \times 10^{-8}$; *p* value $<10^{-4}$). In the 4-node case, a stochastic bi-fan is identified ($\lambda = 1.15 \times 10^{-8}$; *p* value $<10^{-4}$). We notice that the identified motifs in Table 2 show more uncertainties than those in Table 1, in that the estimated nonzero probabilities (see Θ_1) in the former is less close to either 1 (always having interaction) or 0 (never having interaction). On the one hand, the identification of similar motifs in both the highly reliable and noisy networks further validates that our approach can overcome the effects of experimental noises to identify the intrinsic

Table 2. Stochastic motifs in less reliable regulatory networks (based on ChIP-chip data)

Species	LR	λ	Θ_1	Motif
S. cerevisiae	1.67×10^{2}	7.49×10^{-8}	$\begin{bmatrix} 0.00 & 0.98 & 0.98 \\ 0.00 & 0.00 & 0.96 \\ 0.00 & 0.00 & 0.00 \end{bmatrix}$	
S. cerevisiae	3.27×10^{4}	1.15×10^{-8}	$\begin{bmatrix} 0.00 & 0.13 & 0.97 & 0.97 \\ 0.00 & 0.00 & 0.98 & 0.98 \\ 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}$	

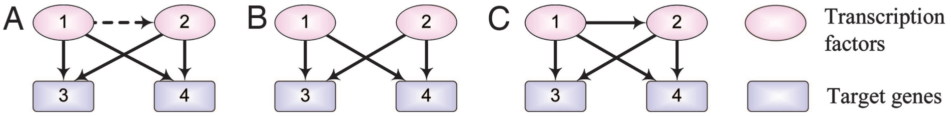


Fig. 1. The stochastic bi-fan motif and two types of related subgraphs. (*A*) The stochastic bi-fan motif identified in the *S. cerevisiae* regulatory network. (*B* and *C*) Two types of subgraphs that can match the stochastic bi-fan motif with nonzero probabilities.

building blocks of the networks. On the other hand, the observation of more uncertainties involved in the motifs in the less reliable network reveals that building blocks in stochastic networks do share stochastic properties of the networks.

Biological Evidences of the Stochastic Bi-Fan Motif. Given an identified stochastic motif, the expected number of instances for the motif can be calculated as the product of the estimated λ and the total number of sampled subgraphs. For a particular subgraph, the probability that it matches the motif can be calculated by using the method presented in *Stochastic Network Motifs*. Multiplying the above two quantities, we obtain the expected number of motif instances for the specific type of subgraph (N_m). On the other hand, the expected number of observing the same type of subgraph in the sampled subgraphs (N_o) can be obtained before the running of the EM algorithm (see *Estimation of Background Probabilities*). With these two expected numbers, the probability that the particular subgraph is a motif instance can be estimated as $p_m = N_m/N_o$, and the probability that the subgraph is observed by chance can be estimated as $p_c = 1 - N_m/N_o$. For example, given the stochastic bi-fan motif identified in the *S. cerevisiae* regulatory network (7) (see Table 1 and Fig. 1*A*), we expect to see $1.8 \times 10^{-7} \times \binom{688}{4} \approx$ 1,666 instances, with 1,599 (1,666 × 0.96) being the subgraph B in Fig. 1 and the rest 67 (1,666 × 0.04) being the subgraph C. Meanwhile, 1,843 instances for the former and 157 for the later are expected to be observed in the original network. Therefore, we estimate that for instances of the subgraph B, 1,599 out of 1,843 (86.8%) are true motifs and the rest, 244 (13.2%), are possibly observed by chance. Similarly, for the instances of the subgraph C, 67 out of 157 (42.7%) are true motifs and the rest, 90 (57.3%), are likely observed by chance.

The stochastic bi-fan motifs (see Tables 1 and 2) reveal the existence of combinatorial transcriptional regulation in *S. cerevisiae*. For example, in the highly reliable network (7), in 496 out of the 1,843 instances of the subgraph B in Fig. 1, the two transcription factors are MSN1 and MSN2 (Fig. 2*A*), and these instances are suggested to relate to the high-osmolarity glycerol (HOG) response pathway (18, 19). In the high external osmolarity conditions, the transcription factor MSN2 (together with a redundant transcription factor MSN4) can activate many osmolarity stress responsive genes (such as CTT1, HSP12, etc.). However, the HOG pathway-mediated transcriptional activation can still be found in the absence of MSN2, suggesting that MSN2 is not the only one responsible for the expression of high-osmolarity stress responsive (HOSR) genes (18). Recent studies show that the transcription factor MSN1 can also activate the HOSR genes when MSN2 is not present (19). In other words, MSN1 acts as a "backup" of MSN2 in the activation of HOSR genes. This mechanism, as illustrated in Fig. 2*C*, keeps the HOSR genes being activated even when MSN2 is not functional. As for other instances of the subgraph B in Fig. 1, we discover that in 703 instances, the two transcription factors are STA2 and TDH1; in 171 instances, the two transcription factors are SKI8 and XBP1; and in the other 473 instances, the two transcription factors show variations from instances to instances. The combinatorial transcriptional regulatory mechanisms for these groups of instances are still not clear.

In 105 of the 157 instances of the subgraph C in Fig. 1, the first transcription factor is GLN3 and the second is GAL80 (see Fig. 2*B*), and these instances are suggested to relate to the nitrogen regulation process in poor nitrogen medium (20). The transcription factor GLN3 is the major regulator in this process, and its expression is induced in the presence of poor nitrogen sources. GLN3 can activate another transcription factor DAL80 and other nitrogen-sensitive (NS) genes (such as DAL1–5). DAL80, however, can repress the expression of NS genes. Therefore, when the cell is experiencing poor nitrogen conditions, GLN3 is induced and GAL80 is up-regulated. The expression of other NS genes, however, is determined by the competitive binding of GLN3 and DAL80. This suggests a "balancing" mechanism in which the NS gene's expression level would be modulated by a balance between the GLN3 activator and the DAL80 repressor (20). This mechanism, as illustrated in Fig. 2*D*, helps to keep the expression of the NS genes from varying too much, even in the presence of poor nitrogen sources. In the other 52 instances of the subgraph C in Fig. 1, the two transcription factors vary from instances to instances, and their regulatory mechanisms are still not clear.

In summary, the stochastic bi-fan motif reveals two kinds of combinatorial transcriptional regulation in *S. cerevisiae*. In the instances with literature support, the two transcription factors work together either in a complementary way or in a competitive way to regulate the target genes, which in general share similar properties (e.g., responsive to the same stress or sensitive to the same condition).

Results on Protein–Protein Interaction Networks. We also apply our method to the protein–protein interaction networks for the seven species listed above in *Data Sources*. For every interaction pair in these networks, we assign either 1 (interaction) or 0 (no interaction) to the corresponding probability, because these data sets come from human curated database and are supposed to be reliable. The

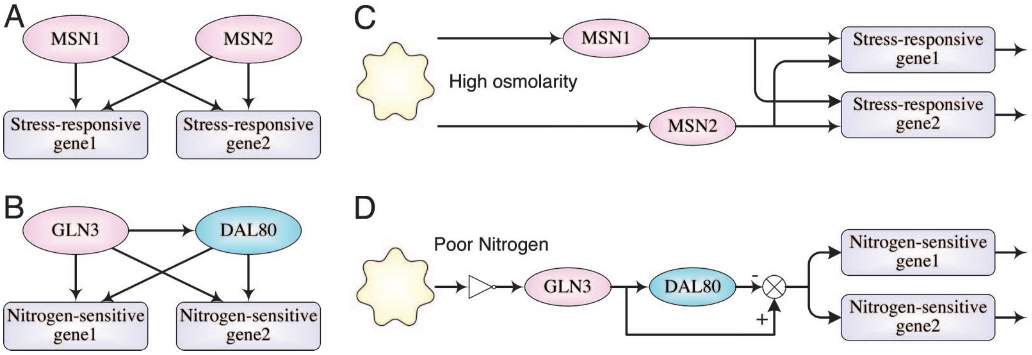
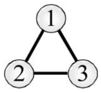
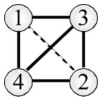
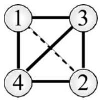


Fig. 2. Two sets of instances of the stochastic bi-fan motif in the *S. cerevisiae* regulatory network (7). (*A*) Instances related to the high-osmolarity glycerol (HOG) response pathway. (*B*) Instances related to the nitrogen regulation process. (*C*) The "backup" mechanism in the HOG pathway. (*D*) The "balancing" mechanism in the nitrogen regulation process.

Table 3. Stochastic motifs in protein–protein interaction networks

Species	LR	λ	Θ_1	Motif
E. coli	2.17×10^2	4.58×10^{-6}	$\begin{bmatrix} 0.00 & 1.00 & 1.00 \\ 1.00 & 0.00 & 1.00 \\ 1.00 & 1.00 & 0.00 \end{bmatrix}$	
S. cerevisiae	2.36×10^4	1.26×10^{-6}		
C. elegans	1.51×10^3	1.06×10^{-7}		
H. pylori	3.24×10^2	1.26×10^{-6}		
M. musculus	8.23×10^3	2.38×10^{-6}		
D. melanogaster	6.06×10^3	2.72×10^{-8}		
H. sapiens	1.80×10^5	1.52×10^{-6}		
E. coli	2.43×10^4	4.48×10^{-8}	$\begin{bmatrix} 0.00 & 0.24 & 1.00 & 1.00 \\ 0.24 & 0.00 & 1.00 & 1.00 \\ 1.00 & 1.00 & 0.00 & 1.00 \\ 1.00 & 1.00 & 1.00 & 0.00 \end{bmatrix}$	
S. cerevisiae	1.92×10^6	4.96×10^{-8}	$\begin{bmatrix} 0.00 & 0.18 & 1.00 & 1.00 \\ 0.18 & 0.00 & 1.00 & 1.00 \\ 1.00 & 1.00 & 0.00 & 1.00 \\ 1.00 & 1.00 & 1.00 & 0.00 \end{bmatrix}$	

identified motifs are shown in Table 3. In the 3-node case (upper portion of Table 3), the motifs identified for all species are the full connected *triangle*. The log pseudo-likelihood ratios ($>10^2$) and the p values ($<10^{-4}$) support the statistical significance of these motifs. As for the 4-node case (lower portion of Table 3), we apply our method to the *E. coli* and *S. cerevisiae* networks and identify motifs that are similar to fully connected rectangles with both diagonals (one of them having low probability) for both species. The log pseudo-likelihood ratios ($>10^4$) and the p values ($<10^{-4}$) support the statistical significance of these motifs.

Comparison with Existing Methods. The widely used method for finding deterministic network motifs in deterministic networks (7, 8) counts the occurrence number of subgraphs in the observed network, estimates the corresponding mean and standard deviation in the background ensemble, and calculates statistics to indicate the significance of the subgraphs. The statistics are $Z = (N_{\mathrm{real}} - \langle N_{\mathrm{rand}} \rangle)/\mathrm{std}(N_{\mathrm{rand}})$ for regulatory networks and $\Delta = (N_{\mathrm{real}} - \langle N_{\mathrm{rand}} \rangle)/(N_{\mathrm{real}} + \langle N_{\mathrm{rand}} \rangle + \varepsilon)$ for protein–protein interaction networks, where N_{real} is the occurrence number of a certain subgraph in the observed network, $\langle N_{\mathrm{rand}} \rangle$ and $\mathrm{std}(N_{\mathrm{rand}})$ are the corresponding mean and standard deviation in the random ensemble, and ε is a predefined positive number (8).

According to this method, the feed forward loop and the deterministic bi-fan are overabundant in the deterministic regulatory networks; the triangle and the rectangle with one or two diagonals are overabundant in the protein–protein interaction networks. By comparison, both the feed forward loop and the bi-fan are identified by our approach, but all show uncertainties (Table 1). The triangle and the rectangle with both diagonals are identified in the protein–protein interaction networks, both showing uncertainties (Table 3). Furthermore, our method is capable of identifying network motifs in stochastic networks, and the identified network motifs show more uncertainties (Table 2).

Conclusions and Discussion

We propose a mixture model and an EM algorithm for identifying network motifs in stochastic networks and identify several stochastic network motifs with supporting biological evidences in a wide range of biological interaction networks. Our approach has several advantages.

First, our approach is based on a probabilistic network motif model, which takes the intrinsic uncertainties of the network building blocks into consideration. Consequently, our approach can capture the stochastic properties of network motifs (e.g., the stochastic bi-fan motif).

Second, we model networks using probability matrices. Therefore, the intrinsic uncertainties and/or experimental noises can be quantified by the probabilities of connections in the networks. As a result, our approach is capable of finding stochastic network motifs in stochastic networks (e.g., the stochastic network motifs in the yeast regulatory network constructed by using the ChIP-chip data).

Third, we use a unified probabilistic model and a single statistic (λ) for different types of networks and network motifs (directed and undirected). Unlike other methods which use different statistics for different types of networks (8), in our approach, different types of networks share the same model, which enable us to estimate and test the same statistic (λ).

The EM algorithm converges very fast but might be trapped into some local optima instead of the true maximum likelihood configuration. Currently, we run the EM algorithm several times with different starting points and choose the solution with the maximum pseudo-likelihood. We find that our approach is quite robust. A majority of repeats, although starting from different initial values, eventually converge to the same motif (see the supporting information for more details).

Another limitation of our approach is that stochastic network motifs in our approach have fixed number of nodes. How to generalize our model to deal with motifs with variable number of nodes is one of our major considerations. Also, our approach is currently limited to identify small motifs ($n < 5$), as are existing methods (7, 8). The main difficulty is how to correctly and efficiently estimate the statistical properties for the background ensemble. Although approximation methods (21) are available, they are not general enough (e.g., networks should be very sparse; the constraint of fixing the occurrence of a certain kind of subgraph is not considered). As for the simulation methods, the quality of the estimation depends on the number of randomized networks simulated and the method to uniformly draw randomized networks from the background ensemble. To obtain high-confidence estimation, it is preferred to increase the number of simulated networks and the randomization steps for the simulation, and this would be very time consuming. In our current study, we resort to the parallel computation technique to partially overcome this difficulty. Recently, a highly efficient sequential importance sampling method (22) became available and is promising to be extended to efficiently and accurately estimate the background statistical properties for networks with various constraints. The application of sequential importance sampling to the network motif identification problem would be another consideration.

Our stochastic network notion assumes that the presence and absence of connections are independent events. Although this assumption works well in our current research with the pseudo-maximum likelihood framework, theoretical studies regarding the application scope of this assumption are necessary. Our stochastic

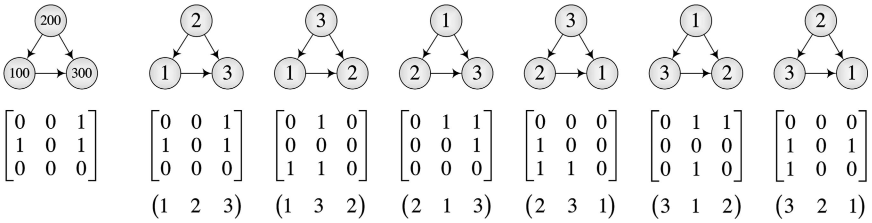


Fig. 3. A subgraph and its isomorphic structures. (*Left*) A 3-node subgraph labeled by (100, 200, 300) in a certain graph. The adjacency matrix is obtained by relabeling nodes as (1, 2, 3), respectively. (*Right*) Isomorphic structures of the subgraph. The isomorphic structures (top) have the same connectivity as the subgraph but different adjacency matrices (middle), which is generated by permuting the node labels (indices of the matrices), as shown at the bottom.

motif model assumes that edges exist in subgraphs independently. Although the presence/absence of an edge does not affect the existence of other edges in the same subgraph, it does affect those in other subgraphs. Therefore, our approach determines the probability of observing a subgraph without a bias, but there are correlations between the probabilities of observing a set of subgraphs. How to make corrections to this correlation is another consideration.

Methods

Graphs and Subgraphs. Without considering uncertainties, a network (also interchangeably referred to as a graph) is a collection of nodes and edges. In this article, nodes are labeled by numbers starting from 1; edges can be directed or undirected. A graph with N nodes is described by using an *adjacency matrix* $\mathbf{A} = (a_{ij})_{N\times N}$. For directed graphs, $a_{ij} = 1$ if there is a directed edge pointing from node i to node j, and 0 otherwise. For undirected graphs, $a_{ij} = 1$ if an undirected edge connects node i and node j, and 0 otherwise. Adjacency matrices are symmetric for undirected graphs but not necessarily so for directed graphs. For a node k, we define the in degree I_k as the number of directed edges linking to it, the out degree O_k as the number of directed edges starting from it, and the mutual degree M_k as the number of undirected edges connecting it.

A subgraph consists of a subset of nodes and corresponding edges in a graph. Intuitively, we can relabel nodes in the subgraph by numbers starting from 1 while keeping the order of the labels as in the original graph. With this "canonical" relabeling, a subgraph $\mathcal{S}$ with n nodes can be described by using an adjacency matrix $\mathbf{X}_{\mathcal{S}} = (x_{ij})_{n\times n}$, where x_{ij} is either 0 or 1 and is equal to the corresponding element in the adjacency matrix of the graph. Relabeling methods other than the canonical one may result in isomorphic structures for the same subgraph. These isomorphic structures have identical connectivity, describe the same subgraph by using different adjacency matrices, and can be mapped to each other by permuting their node labels. Given the adjacency matrix corresponding to the canonical labeling and a permutation of the canonical labels represented by an n-tuple $\mathbf{m} = (m_1, \ldots, m_n)$, a new adjacency matrix $\mathbf{X}_{\mathbf{m}} = (x_{ij}^{\mathbf{m}})_{n\times n}$ corresponding to a certain isomorphic structure can be obtained by setting $x_{ij}^{\mathbf{m}} = x_{m_i m_j}$ for $i, j = 1, \ldots, n$. Adjacency matrices for isomorphic structures of a subgraph can then be obtained by applying the above method on all permutations of the canonical labels (enumerating the permutations is possible when n is small). For an n-node subgraph, there are a total of $n!$ different permutations of the canonical labels and correspondingly $n!$ isomorphic structures. Note that some of the isomorphic structures may be identical. The relationship of a subgraph and its isomorphic structures is illustrated in Fig. 3.

When considering the intrinsic and experimental uncertainties associated with the networks, each pair of nodes in the network can be assigned a probability, indicating the chance of having a connection between the pair of nodes. In this case, a network with N nodes is described by a probability matrix $\mathbf{P} = (\pi_{ij})_{N\times N}$, $0 \le \pi_{ij} \le 1$. π_{ij} is the probability that two nodes i and j are connected. We refer to networks in which connections are described by probabilities as *stochastic networks* and to networks in which connections are described by the presence/absence of edges as *deterministic networks*. A stochastic network can be intuitively thought of as a family of mutually similar deterministic networks, in each of which edges exist independently with probabilities $\Pr(a_{ij} = 1) = \pi_{ij}$. For this reason, when talking about subgraphs in a stochastic network, we actually refer to subgraphs in the family of deterministic networks.

Stochastic Network Motifs. A set of subgraph isomorphic structures with similar adjacency matrices defines a *network motif pattern* $\mathcal{Q}$, which is described by a probability matrix $\Theta_1 = (\theta_{ij})_{n\times n}$, $0 \le \theta_{ij} \le 1$. θ_{ij} means the probability that node i and j are connected. Motif patterns are stochastic in that connections between nodes are represented by probabilities.

Given an n-node subgraph isomorphic structure $\mathcal{P}$ described by an adjacency matrix $\mathbf{X}_{\mathcal{P}} = (x_{ij})_{n\times n}$ and an n-node motif pattern $\mathcal{Q}$ described by a probability matrix Θ_1, the probability that the subgraph isomorphic structure matches the motif pattern is calculated by

$$\Pr(\mathcal{P}|\mathcal{Q}) = \Pr(\mathbf{X}_{\mathcal{P}}|\Theta_1) = \prod_{i=1}^{n}\prod_{j=1}^{n}\theta_{ij}^{x_{ij}}(1-\theta_{ij})^{1-x_{ij}},$$

with the assumption that the presence and absence of edges are independent events. The probability that a subgraph matches a motif pattern is calculated by summing up all of the probabilities for different isomorphic structures of the subgraph.

Given a set of subgraph isomorphic structures $\{\mathcal{P}_p\}_{p=1}^{P}$ with $\mathcal{P}_p$ represented by adjacency matrix $\mathbf{X}_p$, the motif pattern which can maximize the probability of observing the set of subgraph isomorphic structures is calculated as $\Theta_1 = \sum_{p=1}^{P}\mathbf{X}_p/P$. For a set of subgraphs, the motif pattern is derived by first determining for each subgraph an appropriate isomorphic structure and then averaging over the adjacency matrices corresponding to the determined isomorphic structures.

A Mixture Model for Stochastic Networks. With the above concepts, the network motif identification problem is described as searching for statistically significant motif patterns in stochastic networks. For a stochastic network, we first sample over the network to obtain a set of subgraphs, and then identify significant motif patterns from the sampled subgraphs. A stochastic network can be regarded as a mixture of a foreground motif pattern embedded in a background random ensemble with a probability λ. In such a mixture network, each subgraph can be thought of as either coming from the foreground with probability λ or from the background with probability $1 - \lambda$. The foreground model represents the motif pattern and is described by a probability matrix Θ_1. The background model represents the suitable random ensemble and is characterized by a three-tuple $\Theta_0 = \{\mathbf{I}, \mathbf{O}, \mathbf{M}\}$, where $\mathbf{I}$, $\mathbf{O}$, and $\mathbf{M}$ are the distributions of the in, out, and mutual degrees for the stochastic network, respectively. Given Θ_0, we can simulate a number of randomized networks with the degree distributions of Θ_0. Statistical properties related to the background ensemble are then estimated by averaging over the ensemble of networks generated this way (23). With background statistics ready, λ and the foreground motif pattern are estimated by using the following EM algorithm, and the significance of the motif pattern is evaluated by testing the hypothesis $\lambda > 0$.

The EM Algorithm for Parameter Estimation. For a set of W subgraphs $\{S_w\}_{w=1}^{W}$ sampled from the given stochastic network, each may come from the foreground model $\mathcal{M}_1$ with probability $\lambda_1 = \lambda$ or from the background model $\mathcal{M}_0$ with probability $\lambda_0 = 1 - \lambda$. For an individual subgraph $\mathcal{S}_w$, let P_w be the number of different isomorphic structures and $\mathbf{X}_p^w = (x_{ij}^{wp})_{n\times n}$ $(p = 1, \ldots, P_w)$ the corresponding adjacency matrices. We use random variables Z_h^w to indicate the model from which $\mathcal{S}_w$ comes. $Z_h^w = 1$ if $\mathcal{S}_w$ comes from the model $\mathcal{M}_h$ and 0, otherwise. $Z_0^w + Z_1^w = 1$ and $\Pr(Z_h^w = 1) = \lambda_h$ $(h = 0, 1)$. We use another set of random variables Y_{hp}^w to indicate which isomorphic structure (with adjacency matrix $\mathbf{X}_p^w$) should be used to derive the motif pattern. $Y_{hp}^w = 1$ if the pth isomorphic structure is used for subgraph $\mathcal{S}_w$ in the hth model and 0, otherwise. $\Sigma_{p=1}^{P_w} Y_{hp}^w = 1$. Let $\mathbf{X} = \{\mathbf{X}_p^w\}$ be all of the adjacency matrices for the sampled subgraphs, $\mathbf{Z} = \{Z_h^w\}$ and $\mathbf{Y} = \{Y_{hp}^w\}$ all of the indicators, and $\mathbf{\Theta} = \{\lambda, \Theta_0, \Theta_1\}$ all of the parameters. $\mathbf{X}$ is the observation and known. $\mathbf{Y}$ and $\mathbf{Z}$ are unknown and should be treated as missing data. The pseudo-likelihood function for the complete data $\{\mathbf{X}, \mathbf{Y}, \mathbf{Z}\}$ is given as

$$L(\mathbf{\Theta}|\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \prod_{w=1}^{W} \prod_{h=0}^{1} \left[\lambda_h \prod_{p=1}^{P_w} \Pr(\mathbf{X}_p^w|\Theta_h)^{Y_{hp}^w} \right]^{Z_h^w}. \qquad [1]$$

We use an EM algorithm to iteratively estimate the parameters that can maximize the pseudo-likelihood. In the E-step, we calculate the expectation of the log pseudo-likelihood, conditional on the observation $\mathbf{X}$ and the current estimation $\hat{\mathbf{\Theta}}$. In the M-step, we maximize the expectation of the log pseudo-likelihood to update $\hat{\mathbf{\Theta}}$. Repeating the E-step and M-step many times, the algorithm will converge to an (possibly local) optimum estimation of parameters $\mathbf{\Theta}^*$. The detailed derivations of the pseudo-likelihood function and the EM algorithm are presented in the supporting information.

Estimation of Background Probabilities. Given an N-node stochastic network represented by a probability matrix $\mathbf{P} = (\pi_{ij})_{N\times N}$, n-node subgraphs (with canonical labeling) are sampled by generating a number of K adjacency matrices $\{\mathbf{A}^k\}_{k=1}^{K}$ [with $\Pr(a_{ij}^k = 1) = \pi_{ij}$ and $\Pr(a_{ij}^k = 0) = 1 - \pi_{ij}$] and then enumerating $n\times n$ submatrices in each of them. Isomorphic structures corresponding to a certain subgraph are obtained by permuting the canonical node labels of the subgraph.

$\Pr(\mathbf{X}_p^w|\Theta_0)$ is the probability of observing a class of subgraph isomorphic structures with the adjacency matrix $\mathbf{X}_p^w$ in the background ensemble and is estimated as follows. For each generated $\mathbf{A}^k$, we randomly shuffle it many times while fixing the summation of each row and each column to obtain an adjacency matrix corresponding to a randomized network that is uniformly drawn from the background ensemble. When we repeat the random shuffling process many times, we obtain a number of L adjacency matrices $\{\mathbf{A}_l^k\}_{l=1}^{L}$ from $\mathbf{A}^k$, with each of them corresponding to a randomized network that has the same degree distributions as the stochastic network. Subgraphs are then sampled from the ensemble of $\{\mathbf{A}_l^k\}$ $(k = 1, \ldots, K; l = 1, \ldots, L)$, and their isomorphic structures are enumerated. Let $\hat{N}_q$ be the number of subgraph isomorphic structures with adjacency matrix $\mathbf{X}_p^w$ and $\hat{N}$ the total number of subgraph isomorphic structures enumerated. $\Pr(\mathbf{X}_p^w|\Theta_0)$ is estimated as $\hat{N}_p/\hat{N}$. We generally use $K = 1$, $L = 1{,}000$ for highly reliable networks (π_{ij} is either 0 or 1), and $K = 50$, $L = 100$ for networks with noises $(0 \leq \pi_{ij} \leq 1)$.

In the case of $n \geq 4$, the number of subgraph isomorphic structures in every $n' < n$ node isomorphic structure class should be fixed while shuffling each $\mathbf{A}^k$, to avoid assigning high significance to an n-node motif pattern only because of the fact that it includes a highly significant smaller motif (with $n' < n$ node). Besides the existing simulated annealing strategy (7), we propose a newly designed and more practical Monte Carlo simulation method for this purpose, as presented in the supporting information.

Pseudo-Likelihood Ratio Tests for Statistical Significance Assessment. We need to test the hypothesis $H_0 : \lambda = 0$ versus $H_1 : \lambda > 0$. That is, whether the mixture model can explain the sampled subgraphs significantly better than the null model of the random ensemble. An intuitive test statistic is the likelihood ratio $-2\log(L_0/L_1)$, where L_1 and L_0 are the pseudo-likelihoods of the observed subgraphs under the alternative and null models, respectively. L_1 can be calculated by using the EM algorithm described above, and L_0 can be easily calculated by using $\lambda = 0$. Distribution of this statistic is not clear, though, because the subgraphs are not independent, making the likelihood defined in Eq. **1** a pseudo one, and the parameter spaces for the alternative and the null hypothesis are not open sets, making the χ^2 approximation invalid. Therefore, we use a simulation method as described below to evaluate the significance. This approach does not depend on any of the two limitations.

1. Run the EM algorithm over the given stochastic network to obtain the best estimation of parameters $\mathbf{\Theta}^*$ and the expected maximum log pseudo-likelihood ratio LR*.
2. Generate a number of K networks with the same degree distributions as the given network. Run the EM algorithm over each of them to obtain the maximum log pseudo-likelihood ratio $\widehat{LR}_1, \ldots, \widehat{LR}_K$.
3. Count the number of times $\widehat{LR}_k \geq \mathrm{LR}^*$ for $k = 1, \ldots, K$. The p value is then approximated by $p = \hat{K}/K$.

We thank two anonymous referees for suggested revisions that significantly improved the presentation of this article. This work was partly supported by National Institutes of Health/National Science Foundation Joint Mathematical Biology Initiative DMS-0241102.

1. Newman, M. E. J. (2003) *SIAM Rev.* **45,** 167–256.
2. Barabasi, A. L. & Oltvai, Z. N. (2004) *Nat. Rev. Genet.* **5,** 101–113.
3. Milgram, S. (1967) *Psychol. Today* **2,** 60–67.
4. Watts, D. & Strogatz, S. (1998) *Nature* **393,** 440–442.
5. Barabasi, A. L. & Albert, R. (1999) *Science* **286,** 509–512.
6. Barabasi, A. L. & Bonab, E. (2003) *Sci. Am.* **288,** 60–69.
7. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. (2002) *Science* **298,** 824–827.
8. Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M. & Alon, U. (2004) *Science* **303,** 1538–1542.
9. Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R. Y., Alon, U. & Margalit, H. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5934–5939.
10. Vazauez, A., Dobrin, R., Sergi, D., Eckmann, J. P., Oltvai, Z. N. & Barabasi, A. L. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 17940–17945.
11. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., *et al.* (2000) *Nature* **403,** 623–627.
12. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 4569–4574.
13. Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., *et al.* (2002) *Science* **298,** 799–804.
14. Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J., Reynolds, D. B., Yoo, J., *et al.* (2004) *Nature* **431,** 99–104.
15. Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M. & Eisenberg, D. (2002) *Nucleic Acids Res.* **30,** 303–305.
16. Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U. & Eisenberg, D. (2004) *Nucleic Acids Res.* **32,** D449–D451.
17. Mangan, S., Zaslaver, A. & Alon, U. (2003) *J. Mol. Biol.* **334,** 197–204.
18. Martinez-Pastor, M. T., Marchler, G., Schuller, C., Marchler-Bauer, A., Ruis, H. & Estruch, F. (1996) *EMBO J.* **15,** 2227–2235.
19. Martijn, R., Vladimir, R., Ulrike, G., Johan, M. T., Steafan, H., Gustav, A. & Helmut, R. (1999) *Mol. Cell. Biol.* **19,** 5474–5485.
20. Oliveira, E. M. M., Martins, A. S., Garvajal, E. & Bon, E. P. S. (2003) *Yeast* **20,** 31–37.
21. Itzkovitz, S., Milo, R., Kashtan, N., Ziv, G. & Alon, U. (2003) *Phys. Rev. E* **68,** 026127.
22. Chen, Y., Diaconis, P., Holmes, S. & Liu, J. S. (2005) *J. Am. Stat. Assoc.* **100,** 109–120.
23. Newman, M. E. J., Strogatz, S. H. & Watts, D. J. (2001) *Phys. Rev. E* **64,** 026118.