BAYESIAN MODELS AND GIBBS SAMPLING STRATEGIES FOR LOCAL GRAPH ALIGNMENT AND MOTIF IDENTIFICATION IN STOCHASTIC BIOLOGICAL NETWORKS*

RUI JIANG†, TING CHEN‡, AND FENGZHU SUN§

Abstract. With increasing amounts of interaction data collected by high-throughput techniques, understanding the structure and dynamics of biological networks becomes one of the central tasks in post-genomic molecular biology. Recent studies have shown that many biological networks contain a small set of "network motifs," which are suggested to be the basic cellular information-processing units in these networks. Nevertheless, most biological networks have stochastic nature, due to the intrinsic uncertainties of biological interactions and/or experimental noises accompanying the high-throughput data. The building blocks in these networks thus also have stochastic properties. In this paper, we study the problem of identifying stochastic network motifs that are derived from families of mutually similar but not necessarily identical patterns of interactions. Motivated by existing methods for detecting sequence motifs in biopolymer sequences, we establish Bayesian models for stochastic biological networks and develop a group of Gibbs sampling strategies for finding stochastic network motifs. The methods are applied to several available transcriptional regulatory networks and protein-protein interaction networks, and several stochastic network motifs are successfully identified.

1. Introduction. With the development of modern molecular biology, it has been now widely recognized that biological functions are derived from complicated dynamic interactions of several genes and their products instead of isolated individual genes [Roberts(1998), Prill et al.(2005)]. Various biological molecules, including DNA, RNA, and protein, interact with each other to form biological networks that govern the transfer and exchange of materials, energy, and information in living cells. For example, the expression of a gene depends on the binding of transcription factors to binding sites that are located in the regulatory region of the gene [Berg et al.(2004)]. Such cooperative actions between genes and their corresponding transcription factors are encoded in transcriptional regulatory networks [Lee et al.(2002), Harbison et al.(2004)]. Molecular functions in most cases depend on protein complexes that are composed of several proteins [Berg et al.(2004)]. Such physical connections between proteins are governed by protein-protein interaction networks [Uetz et al.(2000), Ito et al.(2001)].

Many biological networks have been shown to share global statistical features,

---

such as the "hub" property that a few nodes have many more connections than most nodes in the network have [Newman(2003), Barabasi and Oltvai(2004)] and the "scale-free" property in which the fraction of nodes having k connections decays as a power law (k^γ, $2 \leq \gamma \leq 3$) [Barabasi and Oltvai(2004), Barabasi and Albert(1999)]. The modular nature of biological networks has also been recognized and utilized in the identification of protein complexes and functional modules [Spirin and Mirny(2003)]. Recently, "network motifs" have been found in a wide variety of biological networks, ranging from the regulatory network of *E.coli* to the neural network of *C.elegans* [Milo et al.(2002)]. These network motifs are patterns of interconnections occurring in networks at numbers that are significantly higher than those in randomized networks [Milo et al.(2002), Milo et al.(2004)]. In these network motifs, biological molecules collaborate with each other to form control flows that mediate the transportation of materials, the transfer of energy, and the exchange of information between molecules [Mangan et al.(2003), Shen-Orr et al.(2002)]. The research on network motifs is therefore promising in uncovering the basic information processing units in biological networks and further revealing the structural design principles of living cells.

Most contemporary studies on biological networks regard such networks as deterministic ones, in which interactions between nodes (biological molecules) are represented by the binary presence/absence status of corresponding edges between the nodes [Yeger-Lotem et al.(2004), Vazauez et al.(2004)]. Although such simplified representation of complex biological networks has demonstrated remarkable successes in the analysis of design principles of biological systems [Shen-Orr et al.(2002), Milo et al.(2002), Mangan et al.(2003), Milo et al. (2004)], the neglect of the stochastic nature of biological interactions might impair the power of such analysis and miss some meaningful results. In addition, incomplete and/or incorrect observations due to experimental resolutions, systematic errors, and random noises also introduce considerable uncertainties into the observed interactions. This situation prevails in biological networks such as protein-protein interaction networks constructed using the yeast two-hybrid (Y2H) assays [Uetz et al.(2000), Ito et al.(2001)] and transcriptional regulatory networks constructed using the chromatin immunoprecipitation (ChIP) method with microarray experiments (ChIP-chip) [Lee et al.(2002), Harbison et al.(2004)].

To take such intrinsic dynamic and experimental uncertainties into consideration, researchers have proposed to model biological networks as "stochastic networks," in which connections between biological molecules are represented by probabilities [Jiang et al.(2006), Berg and Lassig(2004), Berg and Lassig(2006)]. We have also proposed to model network motifs as "stochastic network motifs" in our previous studies [Jiang et al.(2006)], because functionally related motifs are not necessarily topologically identical, and variants in network motifs may also arise due to incomplete and/or incorrect observations [Berg and Lassig(2004)]. With this consideration, a stochastic biological network is regarded as a mixture of a family of mutually similar intercon-

nection patterns (stochastic network motifs) and a background random ensemble, and the problem of identifying the stochastic network motifs is then transferred into the estimation of statistical significant network motif patterns [Jiang et al.(2006)]. In our previous study, we have proposed an expectation maximization (EM) algorithm to estimate the stochastic network motif patterns and used a likelihood ratio test to access the statistical significance of the identified patterns [Jiang et al.(2006)].

In this paper, we formulate the network motifs identification problem from another point of view, which is analogous to the detection of sequence motifs in biopolymer (DNA or protein) sequences [Lawrence et al.(1993), Liu et al.(1995), Bailey and Elkan(1995)]. A biopolymer sequence can be regarded as (one or more) families of mutually similar subsequences embedded in a background sequence. To retrieve the subsequences from a set of biopolymer sequences, one aligns the sequences with the starting points of the subsequences. Sequence motifs can then be represented by the probabilistic patterns derived from the aligned subsequences. Similarly, a stochastic network can be thought of as coming into being by embedding families of mutually similar but not necessarily identical interconnection patterns (subgraphs) in a background random ensemble [Jiang et al.(2006)]. These subgraphs define stochastic network motifs and have different statistical properties from the random ensemble. To recover the motifs embedded in an observed network, we sample subgraphs from the network and "align" them according to their internal connection properties [Berg and Lassig(2004)]. Network motifs can then be identified and described by the stochastic patterns derived from the aligned subgraphs. Here we establish Bayesian models and develop a group of Gibbs sampling strategies to address this problem.

2. Methods.

2.1. Biological interaction networks. In this paper, we study two typical types of biological interaction networks, i.e., transcriptional regulatory networks and protein-protein interaction networks. Without considering uncertainties, a biological interaction network (also interchangeably referred to as a graph) is a collection of nodes and connections (edges) between the nodes. In transcriptional regulatory networks, nodes are genes or corresponding proteins and directed edges are the regulatory interactions between the proteins (transcription factors) and the genes. In protein-protein interaction networks, nodes are proteins and undirected edges are physical interactions between the proteins. In our study, nodes are arbitrarily labeled by numbers starting from 1. A graph $\mathcal{G}$ with N nodes is described using an *adjacency matrix* $\mathbf{A} = (a_{ij})_{N \times N}$, where $a_{ij} = 1$ if there is a directed edge pointing from node i to node j in a transcriptional regulatory network or an undirected edge connecting these two nodes in a protein-protein interaction network, and $a_{ij} = 0$, otherwise. For a certain node k, we define the in degree I_k as the number of directed edges linking to it ($I_k = \sum_{i=1}^{N} a_{ik}$), the out degree O_k as the number of directed edges starting from

it ($O_k = \sum_{j=1}^{N} a_{kj}$), and the mutual degree M_k as the number of undirected edges connecting it ($M_k = \sum_{i=1}^{N} a_{ik} = \sum_{j=1}^{N} a_{kj}$). For a graph, the in (out, mutual) degree distribution is the distribution of the in (out, mutual) degrees for all the nodes. Degree distributions can be more specifically represented as degree sequences ($\{I_k\}_{k=1}^{N}$, $\{O_k\}_{k=1}^{N}$, and $\{M_k\}_{k=1}^{N}$).

A subgraph consists of a subset of nodes and corresponding edges between the nodes 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 using an adjacency matrix $\mathbf{B} = (b_{ij})_{n \times n}$, where b_{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 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 $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$, a new adjacency matrix $\mathbf{X}_{\boldsymbol{\pi}} = (x_{ij})_{n \times n}$ corresponding to a certain isomorphic structure can be obtained by setting $x_{ij} = b_{\pi_i \pi_j}$ 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 Figure 1 (A).

A set of subgraph isomorphic structures with identical adjacency matrices defines a class of subgraph isomorphic structures. Supposing that in a certain graph $\mathcal{G}$, the number of occurrence of a specific class of subgraph isomorphic structures, l, is N_l, the probability of observing this class of subgraph isomorphic structures in the graph is then calculated as

$$\Pr(\mathbf{X}_l | \mathcal{G}) = \frac{N_l}{\sum_k N_k}$$

where $\mathbf{X}_l$ is the adjacency matrix corresponds to the class of subgraph isomorphic structures and the summation in the denominator is taken over all possible classes of subgraph isomorphic structures. This probability is also referred to as the concentration of the class of subgraph isomorphic structures [Kashtan et al.(2004)].

The intrinsic and experimental uncertainties associated with biological interactions can be represented by probabilities of having the corresponding connections in the networks. Consequently, a biological network with uncertainties can be 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 have a connection. In this paper, we would refer to biological networks in

which connections are described by probabilities as *stochastic networks*. In contrast, networks in which connections are described by the presence/absence of interactions would be referred to as *deterministic networks*. A stochastic network can be thought of as a family of mutually similar deterministic networks, in each of which edges exist independently with probabilities $\Pr(a_{ij} = 1) = \pi_{ij}$ and $\Pr(a_{ij} = 0) = 1 - \pi_{ij}$. For this reason, when talking about subgraphs in a stochastic network, we actually refer to subgraphs in the family of deterministic networks.

2.2. Stochastic network motifs and local graph alignments. A set of subgraphs with similar connectivity defines a *stochastic network motif pattern* $\mathcal{M}$, which is described using a probability matrix $\Theta = (\theta_{ij})_{n \times n}$, $0 \leq \theta_{ij} \leq 1$. θ_{ij} means the probability that a directed edge pointing from node i to node j in a transcriptional regulatory network or an undirected edge connecting them in a protein-protein interaction network. Given an n-node isomorphic structure $\mathcal{I}$ (of a certain subgraph), represented by adjacency matrix $\mathbf{X}_{\mathcal{I}} = (x_{ij})_{n \times n}$. The probability that this isomorphic structure matches the motif pattern is calculated by

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

with the assumption that edges exist independently [Jiang et al.(2006)]. The probability that a subgraph matches the motif pattern is obtained by summing up the probabilities for different isomorphic structures of the subgraph.

Given a set of n-node similar subgraphs $\{\mathcal{S}_w\}_{w=1}^{W}$ with $\mathcal{S}_w$ having P_w different isomorphic structures, we assign for each subgraph a specific isomorphic structure and obtain an *alignment* for this set of subgraphs. With an alignment available, the motif pattern which can maximize the probability of observing the set of subgraphs is $\Theta = \frac{1}{W}\sum_{w=1}^{W}\mathbf{X}_w$, where $\mathbf{X}_w$ is the adjacency matrix for the isomorphic structure assigned to subgraph $\mathcal{S}_w$. In the general case in which only a subset of the given subgraphs are similar and the alignment is unknown, we need to decide which subgraphs are similar (and thus should be included in the alignment), align the determined similar subgraphs, and then derive the motif pattern from the aligned subgraphs. We refer to this general problem as *local graph alignment* in the sense that we only intend to align a subset of the given subgraphs [Berg and Lassig(2004)]. As a demonstration, Figure 1 (B) shows the process of the local graph alignment.

2.3. The single motif model and the Bernoulli sampling strategy. A stochastic network can be thought of as coming into being by embedding (one or more) families of mutually similar subgraphs in a background random ensemble, where the families of subgraphs define motif patterns. In such a (mixture) network, each subgraph can be regarded as either coming from one of the motif patterns or from the background. To recover the embedded motif patterns, we can sample a set of

subgraphs from the network, and search for over-abundant motif patterns within the subgraphs. From the point of view of local graph alignment, the motif finding process includes two folds: (*i*) determine which of the subgraphs belong to the foreground motif patterns, and (*ii*) for each of the subgraphs classified to the foreground, determine a proper isomorphic structure so that the subgraphs can be aligned and the motif patterns can be derived. Here we assume that a stochastic network contains only one motif pattern and show how the pattern can be recovered with a Bayesian model and a Gibbs sampling strategy.

Given a network, we sample from it to obtain a set of subgraphs and permute the canonical node labels for each of them to enumerate different isomorphic structures. Let $\{\mathcal{I}_l\}_{l=1}^L$ be all the L isomorphic structures obtained in this way, $\mathbf{X}_l = (x_{ij}^l)_{n \times n}$ the adjacency matrices corresponding to $\mathcal{I}_l$, and $\mathbf{X} = \{\mathbf{X}_l\}_{l=1}^L$ the collection of all adjacency matrices. For each isomorphic structure $\mathcal{I}_l$, we introduce an indicator δ_l, where $\delta_l = 1$ if the subgraph corresponding to $\mathcal{I}_l$ comes from the foreground motif pattern (i.e., included in the alignment) and $\mathcal{I}_l$ is the isomorphic structure assigned in the alignment, and 0 otherwise. Let $\boldsymbol{\delta} = \{\delta_l\}_{l=1}^L$ be all the indicators. We assume that $\Pr(\delta_l = 1) = \varepsilon$, independently and identically distributed (iid) for $l = 1, \ldots, L$. The motif can then be recovered by determining $\boldsymbol{\delta}$, and at the same time estimating ε and the parameters associated with the motif pattern. Because for every subgraph only one of the isomorphic structures can be its alignment, there is some slight discrepancy among the δ's. However, the probability ε is in general very small; hence ignoring the discrepancy, as in the iid assumption, has little effect (the same situation exists in sequence alignment [Liu et al.(1995)]). We assume that the prior distribution for ε Beta($\boldsymbol{\mu}$), where $\boldsymbol{\mu} = (\mu_0, \mu_1)^T$ are hyper-parameters. In other words, $\Pr(\varepsilon) \propto \varepsilon^{\mu_1 - 1}(1 - \varepsilon)^{\mu_0 - 1}$.

The foreground motif pattern is described by a probability matrix $\Theta_1 = (\theta_{ij})_{n \times n}$, $0 \leq \theta_{ij} \leq 1$. We assume that the prior distribution for θ_{ij}

Beta($\boldsymbol{\alpha}_{ij}$), where $\boldsymbol{\alpha}_{ij} = (\alpha_{ij0}, \alpha_{ij1})^T$ are hyper-parameters. In other words, $\Pr(\theta_{ij}) \propto \theta_{ij}^{\alpha_{ij1} - 1}(1 - \theta_{ij})^{\alpha_{ij0} - 1}$ for $i, j = 1, \ldots, n$. The background random ensemble represents a family of randomized networks, and each of which has the same degree distributions as the given network. Let $\mathbf{I}$, $\mathbf{O}$, and $\mathbf{M}$ be the in, out, and mutual degree distribution of the given network, respectively. The background ensemble can then be characterized by $\Theta_0 = \{\mathbf{I}, \mathbf{O}, \mathbf{M}\}$. Given Θ_0, we can simulate a number of networks with the degree distributions of Θ_0. Statistical properties related to the background can then be calculated by averaging over the ensemble of the generated networks [Newman et al.(2001)].

We further introduce the following counting functions. Let $h_1(\boldsymbol{\delta}) = \sum_{l=1}^L \delta_l$ and $h_0(\boldsymbol{\delta}) = \sum_{l=1}^L (1 - \delta_l)$ be the number of isomorphic structures classified to the foreground and background, respectively. Let $h_{ij0}(\mathbf{X}, \boldsymbol{\delta}) = \sum_{l=1}^L \delta_l (1 - x_{ij}^l)$ and $h_{ij1}(\mathbf{X}, \boldsymbol{\delta}) = \sum_{l=1}^L \delta_l x_{ij}^l$ be the number of 0's and 1's for element (i, j) in adja-

cency matrices corresponding to isomorphic structures classified to the foreground, respectively. Let $\boldsymbol{h}(\boldsymbol{\delta}) = (h_0(\boldsymbol{\delta}), h_1(\boldsymbol{\delta}))^T$ and $\boldsymbol{h}_{ij}(\mathbf{X}, \boldsymbol{\delta}) = (h_{ij0}(\mathbf{X}, \boldsymbol{\delta}), h_{ij1}(\mathbf{X}, \boldsymbol{\delta}))^T$. Treating $\boldsymbol{\delta}$ as the missing data, the likelihood for the complete data $\{\mathbf{X}, \boldsymbol{\delta}\}$ given the parameters $\{\varepsilon, \Theta_0, \Theta_1\}$ is

$$\Pr(\mathbf{X}, \boldsymbol{\delta} | \varepsilon, \Theta_1, \Theta_0) = \Big[\prod_{l=1}^{L} \Pr(\mathbf{X}_l | \Theta_0)^{1-\delta_l}\Big] \times \Big[\varepsilon^{h_1(\boldsymbol{\delta})}(1-\varepsilon)^{h_0(\boldsymbol{\delta})}\Big] \times$$
$$\Big[\prod_{i=1}^{n}\prod_{j=1}^{n} (\theta_{ij})^{h_{ij1}(\mathbf{X}, \boldsymbol{\delta})}(1-\theta_{ij})^{h_{ij0}(\mathbf{X}, \boldsymbol{\delta})}\Big],$$

and

$$\Pr(\boldsymbol{\delta}, \Theta_1, \varepsilon | \mathbf{X}, \Theta_0) \propto \Pr(\mathbf{X}, \boldsymbol{\delta}, \Theta_0, \Theta_1, \varepsilon)$$
$$\propto \Pr(\mathbf{X}, \boldsymbol{\delta} | \Theta_0, \Theta_1, \varepsilon)\Pr(\Theta_1)\Pr(\varepsilon).$$

Integrating out Θ_1 and ε in $\Pr(\boldsymbol{\delta}, \Theta_1, \varepsilon | \mathbf{X}, \Theta_0)$, we obtain

$$\Pr(\boldsymbol{\delta} | \mathbf{X}, \Theta_0) \propto \Big[\prod_{l=1}^{L} \Pr(\mathbf{X}_l | \Theta_0)^{1-\delta_l}\Big] \times \Big[\frac{\Gamma(|\boldsymbol{\mu}|)}{\Gamma(\boldsymbol{\mu})}\frac{\Gamma(\boldsymbol{h}(\boldsymbol{\delta}) + \boldsymbol{\mu})}{\Gamma(|\boldsymbol{h}(\boldsymbol{\delta})| + |\boldsymbol{\mu}|)}\Big] \times$$
$$\Big[\prod_{i=1}^{n}\prod_{j=1}^{n}\frac{\Gamma(|\boldsymbol{\alpha}_{ij}|)}{\Gamma(\boldsymbol{\alpha}_{ij})}\frac{\Gamma(\boldsymbol{h}_{ij}(\mathbf{X}, \boldsymbol{\delta}) + \boldsymbol{\alpha}_{ij})}{\Gamma(|\boldsymbol{h}_{ij}(\mathbf{X}, \boldsymbol{\delta})| + |\boldsymbol{\alpha}_{ij}|)}\Big],$$

where $\Pr(\mathbf{X}_l | \Theta_0)$ is the probability of observing a class of isomorphic structures with the adjacency matrix $\mathbf{X}_l$ in the background ensemble and can be estimated in advance using the method presented in section 2.6. $\Gamma(x)$ is the *Gamma* function. For a vector $\boldsymbol{x} = (x_1, \ldots, x_K)^T$, $|\boldsymbol{x}| = \sum_{k=1}^{K} |x_k|$, $\Gamma(\boldsymbol{x}) = \prod_{k=1}^{K} \Gamma(x_k)$, and $\Gamma(|\boldsymbol{x}|) = \Gamma(\sum_{k=1}^{K} |x_k|)$.

Noticing that $\Gamma(x) = (x-1)\Gamma(x-1)$, the odds that the l-th isomorphic structure is classified to the motif pattern given the rest of the classifications can then be calculated as

$$o_l = \frac{\Pr(\delta_l = 1 | \boldsymbol{\delta}_{[-l]}, \mathbf{X}, \Theta_0)}{\Pr(\delta_l = 0 | \boldsymbol{\delta}_{[-l]}, \mathbf{X}, \Theta_0)} = \frac{1}{\Pr(\mathbf{X}_l | \Theta_0)}\Big[\frac{\hat{\varepsilon}}{1 - \hat{\varepsilon}}\Big]\Big[\prod_{i=1}^{n}\prod_{j=1}^{n} (\hat{\theta}_{ij})^{x_{ij}^l}(1 - \hat{\theta}_{ij})^{1-x_{ij}^l}\Big],$$

where

$$\hat{\varepsilon} = \frac{h_1(\boldsymbol{\delta}_{[-l]}) + \mu_1}{[h_0(\boldsymbol{\delta}_{[-l]}) + \mu_0] + [h_1(\boldsymbol{\delta}_{[-l]}) + \mu_1]},$$

and

$$\hat{\theta}_{ij} = \frac{h_{ij1}(\mathbf{X}_{[-l]}, \boldsymbol{\delta}_{[-l]}) + \alpha_{ij1}}{[h_{ij0}(\mathbf{X}_{[-l]}, \boldsymbol{\delta}_{[-l]}) + \alpha_{ij0}] + [h_{ij1}(\mathbf{X}_{[-l]}, \boldsymbol{\delta}_{[-l]}) + \alpha_{ij1}]}$$

are the posterior means of ε and θ_{ij} conditional on the observed data and the current determination of the missing data. $\mathbf{X}_{[-l]}$ is obtained by removing $\mathbf{X}_l$ from $\mathbf{X}$, and $\boldsymbol{\delta}_{[-l]}$ is obtained by removing δ_l from $\boldsymbol{\delta}$. The conditional probability $\Pr(\delta_l = 1 | \boldsymbol{\delta}_{[-l]}, \mathbf{X}, \Theta_0)$

can then be calculated as $o_l/(1 + o_l)$. This Gibbs sampler immediately suggests the following Gibbs sampling strategy, which is referred to as the *Bernoulli sampling strategy* in this paper.

1. Initialization. Set missing data $\boldsymbol{\delta} = \mathbf{0}$. Calculate initial value $\boldsymbol{h}(\boldsymbol{\delta})$, and $\boldsymbol{h}_{ij}(\mathbf{X}, \boldsymbol{\delta})$.
2. Sampling. Uniformly choose $l \in [1, \ldots, L]$. Sample a new δ_l according to $\Pr(\delta_l = 1|\boldsymbol{\delta}_{[-l]}, \mathbf{X}, \Theta_0)$.
3. Repeat 2 until convergence ($\hat{\boldsymbol{\varepsilon}}$ and $\hat{\Theta}_1$ are stable). $\qquad\square$

The counting functions should be updated in a computationally economy way. For this purpose, $\boldsymbol{h}(\boldsymbol{\delta})$ and $\boldsymbol{h}_{ij}(\mathbf{X}, \boldsymbol{\delta})$ are cached while performing the sampling. In the t-th iteration, before calculating $\Pr(\delta_l = 1|\boldsymbol{\delta}_{[-l]}, \mathbf{X}, \Theta_0)$, we perform the following updates

$$
\begin{aligned}
h_0(\boldsymbol{\delta}_{[-l]})^{(t)} &= h_0(\boldsymbol{\delta})^{(t)} - (1 - \delta_l^{(t)}), \\
h_1(\boldsymbol{\delta}_{[-l]})^{(t)} &= h_1(\boldsymbol{\delta})^{(t)} - \delta_l^{(t)}, \\
h_{ij0}(\mathbf{X}_{[-l]}, \boldsymbol{\delta}_{[-l]})^{(t)} &= h_{ij0}(\mathbf{X}, \boldsymbol{\delta})^{(t)} - \delta_l^{(t)}(1 - x_{ij}^l), \\
h_{ij1}(\mathbf{X}_{[-l]}, \boldsymbol{\delta}_{[-l]})^{(t)} &= h_{ij1}(\mathbf{X}, \boldsymbol{\delta})^{(t)} - \delta_l^{(t)} x_{ij}^l.
\end{aligned}
$$

while after the new $\delta_l^{(t+1)}$ is sampled, we perform the following updates

$$
\begin{aligned}
h_0(\boldsymbol{\delta})^{(t+1)} &= h_0(\boldsymbol{\delta}_{[-l]})^{(t)} + (1 - \delta_l^{(t+1)}), \\
h_1(\boldsymbol{\delta})^{(t+1)} &= h_1(\boldsymbol{\delta}_{[-l]})^{(t)} + \delta_l^{(t+1)}, \\
h_{ij0}(\mathbf{X}, \boldsymbol{\delta})^{(t+1)} &= h_{ij0}(\mathbf{X}_{[-l]}, \boldsymbol{\delta}_{[-l]})^{(t)} + \delta_l^{(t+1)}(1 - x_{ij}^l), \\
h_{ij1}(\mathbf{X}, \boldsymbol{\delta})^{(t+1)} &= h_{ij1}(\mathbf{X}_{[-l]}, \boldsymbol{\delta}_{[-l]})^{(t)} + \delta_l^{(t+1)} x_{ij}^l.
\end{aligned}
$$

With the above caching and updating technique, we only need to calculate 8 summations in each sampling step (multiplications can be substituted by logical operations since both δ_l and x_{ij}^l are either 0 or 1). Consequently, the Bernoulli sampling strategy can be done with high efficiency.

Hyper-parameters $\boldsymbol{\mu}$ is determined as follows. Given what is known about a network, a crude guess $\tilde{n}_1$ for the number of isomorphic structures belonging to the motif pattern is usually possible. Let $\tilde{n}_0 = L - \tilde{n}_1$. We determine the initial values of $\boldsymbol{\mu}$ by setting $\mu_m = \tilde{n}_m \tilde{L}/L$ for $m = 0, 1$, where L_0 reflects the "weight" to be put on the prior knowledge and is referred to as a "pseudo-count" (e.g. $\tilde{L} = L/50$). Similarly, within the isomorphic structures supposed to be classified to the motif pattern ($\tilde{n}_1$ of them), let $\tilde{n}_{ij1}$ be a crude guess for the number of connections from node i to node j and $\tilde{n}_{ij0} = \tilde{n}_1 - \tilde{n}_{ij1}$. We determine the initial values of $\boldsymbol{\alpha}$ by setting $\alpha_{ijk} = \tilde{n}_{ijk} \tilde{L}/L$ for $k = 0, 1$. Generally, self-connections are rare in biological networks, we thus set $\tilde{n}_{ij1} = (1 - \tau)\tilde{n}_1$ for all $i = j$, where τ is a fraction greater than 0.5 (e.g. $\tau = 0.8$). Additionally, motifs in biological networks should have enhanced

numbers of internal connections [Berg and Lassig(2004)], therefore for protein-protein interaction networks, we set $\tilde{n}_{ij1} = \tau \tilde{n}_1$ for all $i \neq j$. For regulatory networks, double-connections ($a_{ij} = a_{ji} = 1$ in adjacency matrices) are also rare, thus for each pair $(\tilde{n}_{ij1}, \tilde{n}_{ji1}), i \neq j$, we randomly set one to be $\tau \tilde{n}_1$ and the other to be $(1 - \tau)\tilde{n}_1$.

To enable the Gibbs sampler to freely maneuver in the searching space and more accurately estimate the parameters, we further apply the prior-annealing technique [Niu et al.(2002)]: in the beginning of the iteration, a set of high pseudo-counts, $\{\boldsymbol{\mu}^0, \boldsymbol{\alpha}^0\}$, that conform to the Dirichlet distribution, are used as the initial prior. As the iteration proceeds, the pseudo-counts are dwindled in a fixed rate. For example, supposing that the pseudo-counts $\boldsymbol{\mu}$ for all motif patterns are $\boldsymbol{\mu}^0 = \{\mu_0^0, \ldots, \mu_m^0\}$ and $\boldsymbol{\mu}^T = \{\mu_0^T, \ldots, \mu_m^T\}$ for the start and the end of the T-th iteration, the pseudo-counts at the t-th iteration, $\boldsymbol{\mu}^t = \{\mu_0^t, \ldots, \mu_m^t\}$, are given as

$$\mu_m^t = \mu_m^0 + \frac{t}{T}(\mu_m^T - \mu_m^0).$$

Similar formula holds for the pseudo-counts $\boldsymbol{\alpha}$.

2.4. The multiple motif model and the motif sampling strategy. We can extend the idea of the single motif model one step further to derive the multiple motif model which is capable of detecting more than one network motif simultaneously. Consider M different motif patterns, each occurring an unknown number of times, in a set of L sampled isomorphic structures. Similar to the single motif model, let the m-th motif pattern be described by parameters $\Theta_m = (\theta_{ij}^m)_{n \times n}, 0 \leq \theta_{ij}^m \leq 1$, and assume a conjugate prior Beta($\boldsymbol{\alpha}_{ij}^m$) for θ_{ij}^m, where $\boldsymbol{\alpha}_{ij}^m = (\alpha_{ij0}^m, \alpha_{ij1}^m)^T$. We also introduce the indicators $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_L)^T$, where $\delta_l = m$ if the l-th isomorphic structure belongs to the m-th motif pattern and 0 if background. Let $\Pr(\delta_l = m) = \varepsilon_m$ with $\sum_{m=0}^M \varepsilon_m = 1$ and assume a conjugate Dirichlet prior Dir($\boldsymbol{\mu}$) on ε_m ($m = 0, \ldots, M$), where $\boldsymbol{\mu} = (\mu_0, \ldots, \mu_M)^T$. The Gibbs sampler for the *motif sampling strategy* can then be derived as

$$o_l^m = \frac{\Pr(\delta_l = m | \boldsymbol{\delta}_{[-l]}, \mathbf{X}, \Theta_0)}{\Pr(\delta_l = 0 | \boldsymbol{\delta}_{[-l]}, \mathbf{X}, \Theta_0)} = \frac{1}{\Pr(\mathbf{X}_l | \boldsymbol{\Theta}_0)} \left[\frac{\hat{\varepsilon}_m}{\hat{\varepsilon}_0}\right] \left[\prod_{i=1}^n \prod_{j=1}^n (\hat{\theta}_{ij}^m)^{x_{ij}^l}(1 - \hat{\theta}_{ij}^m)^{1 - x_{ij}^l}\right],$$

and

$$p_l^m = \Pr(\delta_l = m | \boldsymbol{\delta}_{[-l]}, \mathbf{X}, \Theta_0) = \frac{o_l^m}{\sum_{m'=0}^M o_l^{m'}},$$

where o_l^m and p_l^m are the conditional odds and probability that the l-th isomorphic structure is classified to the m-th motif pattern given the rest of the classifications respectively, and $\hat{\varepsilon}_m$ and $\hat{\theta}_{ij}^m$ are the posterior means that can be calculated by the similar formula as in the previous Bernoulli sampling strategy.

2.5. The group sampling strategy. The total number of sampled isomorphic structures is huge, but the types of isomorphic structures (different adjacency matrices) in real networks are quite limited. Meanwhile, in each iteration, the chance that a certain isomorphic structure is sampled and classified to (one of) the motif pattern(s) is very small because the number of isomorphic structures is large and most of them belong to the background. These two facts motivate us to develop the following special designed Gibbs sampling strategy, which is referred to as the *group sampling strategy* in this paper.

Without loss of generality, suppose that each isomorphic structure can be classified to either the background or one of the M motif patterns, and at the same time it can be classified to one of the R isomorphic structure classes according to its adjacency matrix. Thus, there are a total of $R \times (M + 1)$ combinations (groups) of isomorphic structure classes and motif patterns. Each isomorphic structure belongs to one of the groups. Let $\mathbf{G} = \{G_u^r\}_{r=1}^{R} {}_{u=0}^{M}$ be the set of these groups and ψ_{ru} the number of isomorphic structures classified to group G_u^r. Let $\Psi = (\psi_{ru})_{R \times (M+1)}$ be all the numbers. In each iteration, the probability that one of the isomorphic structures in group G_u^r is sampled is $p_u^r = \psi_{ru}/\sum_{r'=1}^{R} \sum_{u'=0}^{M} \psi_{r'u'}$. Let $\mathbf{Y}_r = (y_{ij}^r)_{n \times n}$ be the adjacency matrix for groups G_u^r $(u = 0, \ldots, M)$ and $\mathbf{Y} = \{\mathbf{Y}_r\}_{r=1}^{R}$. Define $h_u(\Psi) = \sum_{r=1}^{R} \psi_{ru}$, $h_{ij1}^u(\mathbf{Y}, \Psi) = \sum_{r=1}^{R} \psi_{ru} y_{ij}^r$, and $h_{ij0}^u(\mathbf{Y}, \Psi) = \sum_{r=1}^{R} \psi_{ru}(1 - y_{ij}^r)$. The conditional odds that the sampled isomorphic structures in group G_u^r is classified to a certain group G_v^r is

$$o_{uv}^r = \frac{\Pr(\delta_u^r = v | \Psi_{[-(ru)]}, \mathbf{Y}, \Theta_0)}{\Pr(\delta_u^r = 0 | \Psi_{[-(ru)]}, \mathbf{Y}, \Theta_0)} = \frac{1}{\Pr(\mathbf{Y}_r | \Theta_0)} \left[\frac{\hat{\varepsilon}_v}{\hat{\varepsilon}_0}\right] \left[\prod_{i=1}^{n} \prod_{j=1}^{n} (\hat{\theta}_{ij}^v)^{y_{ij}^r} (1 - \hat{\theta}_{ij}^v)^{1 - y_{ij}^r}\right],$$

and

$$q_{uv}^r = \Pr(\delta_u^r = v | \Psi_{[-(ru)]}, \mathbf{Y}, \Theta_0) = \frac{o_{uv}^r}{\sum_{v'=0}^{M} o_{uv'}^r}.$$

δ_u^r is an indicator representing to which group the sampled isomorphic structure is classified. $\Psi_{[-(ru)]}$ is a matrix which is equal to Ψ expect for $\Psi_{[-(ru)]}(r, u) = \Psi(r, u) - 1$. $\hat{\varepsilon}_v$ and $\hat{\theta}_{ijk}^v$ are posterior means calculated as

$$\hat{\varepsilon}_v = \frac{h_v(\Psi_{[-(ru)]}) + \mu_v}{\sum_{v'=0}^{M} (h_{v'}(\Psi_{[-(ru)]}) + \mu_{v'})}$$

and

$$\hat{\theta}_{ij}^v = \frac{h_{ij1}^v(\mathbf{Y}, \Psi_{[-(ru)]}) + \alpha_{ij1}^v}{[h_{ij0}^v(\mathbf{Y}, \Psi_{[-(ru)]}) + \alpha_{ij0}^v] + [h_{ij1}^v(\mathbf{Y}, \Psi_{[-(ru)]}) + \alpha_{ij1}^v]}.$$

The probability that one of the isomorphic structures in G_u^r is sampled and classified to G_v^r is then given as $\rho_{uv}^r = p_u^r q_{uv}^r$, which can be thought of as the probability that an isomorphic structure in group G_u^r transits to group G_v^r.

Particularly, we can exclude self-transitions because they have no effect on the estimation of parameters. Let $\pi_{uv}^r = \rho_{uv}^r / (1 - \sum_{r'=1}^R \sum_{u'=0}^M \rho_{u'u'}^{r'})$ for $u \neq v$ and $\pi_{uu}^r = 0$ be the conditional probability that an isomorphic structure in group G_u^r is sampled and classified to a group G_v^r other than G_u^r. The iteration can then be performed with high efficiency by sampling a certain triple (r, u, v) according to π_{uv}^r and update Ψ by setting $\psi_{ru} \leftarrow \psi_{ru} - 1$ and $\psi_{rv} \leftarrow \psi_{rv} + 1$.

2.6. Subgraph sampling and background probability estimation. Given an N-node stochastic network represented by a probability matrix $\mathbf{P} = (\pi_{uv})_{N \times N}$, we can generate a number of J adjacency matrices $\{\mathbf{A}^j\}_{j=1}^J$ with $\Pr(a_{uv}^j = 1) = \pi_{uv}$ and $\Pr(a_{uv}^j = 0) = 1 - \pi_{uv}$. Each of these adjacency matrices then corresponds to a deterministic network. When n is small (< 5), n-node subgraphs (with canonical labeling) can be obtained from these adjacency matrices by enumerating $n \times n$ sub-matrices in each of them. After subgraphs are enumerated, isomorphic structures corresponding to subgraphs can be enumerated by permuting the canonical node labels of the subgraphs, and the enumerated isomorphic structures can be used with the Bernoulli sampling or the motif sampling strategy for motif identification. When n is large $(>= 5)$, enumerating subgraphs is prohibited in most cases. We therefore use a subgraph sampling strategy as presented in [Kashtan et al.(2004)] to estimate subgraph concentrations and multiply them by the total number of sampled subgraphs to obtain the expected number for each type of subgraph. After subgraphs are sampled, isomorphic structures corresponding to subgraphs are obtained by permuting the canonical node labels of the subgraphs, and the expected numbers of occurrences of subgraph isomorphic structures can be estimated and be used with the group sampling strategy.

$\Pr(\mathbf{X}_l | \Theta_0)$ is the probability of observing a class of isomorphic structure with the adjacency matrix $\mathbf{X}_l$ in the background ensemble and is estimated as follows. When the given network is small (e.g., containing several hundred nodes), we can randomly shuffle each generated matrix $\mathbf{A}^j$ many times while fixing the summation of each row and each column to obtain adjacency matrices corresponding to randomized networks which is uniformly drawn from the background ensemble. When the given network is large (e.g., containing several thousand nodes), the shuffling procedure is inefficient. We can apply some sampling strategy such as the Sequential Importance Sampling strategy (SIS) described in [Chen et al.(2005)] to generate adjacency matrices which correspond to randomized networks uniformly drawn from the background ensemble. Repeating the shuffling procedure or the sequential importance sampling for each matrix $\mathbf{A}^j$ many times, we obtain a number of K adjacency matrices $\{\mathbf{A}_k^j\}_{k=1}^K$ from $\mathbf{A}^j$, with each of them corresponding to a randomized network

which has the same degree distributions as the stochastic network. When n is small (< 5), the probability of observing a specific class of subgraph isomorphic structure with the adjacency matrix $\mathbf{X}_l$ in the background ensemble can be precisely calculated by enumerating subgraphs and their isomorphic structures from the ensemble of $\{\mathbf{A}_k^j\}$ ($j = 1, \ldots, J; k = 1, \ldots, K$), counting the number of occurrence of the class of isomorphic structure having the adjacency matrix $\mathbf{X}_l$, and dividing the number by the total number of isomorphic structures being enumerated. When n is large (≥ 5), enumerating subgraphs is not computationally feasible for most randomized networks. Hence, we adopt the subgraph sampling strategy [Kashtan et al.(2004)] to estimate concentrations for subgraph isomorphic structures in the randomized networks and average over the concentrations to obtain the expected concentrations for subgraph isomorphic structures in the background ensemble.

2.7. The signed rank test of significance. We use a modified Wilcoxon signed rank test as described in [Liu et al.(1995)] to assess the statistical significance of identified network motifs. Suppose that we have obtained an estimation of $\hat{\varepsilon}$ and $\hat{\Theta}_1$ by running one of the sampling strategy on a set of isomorphic structures sampled from the given network (positive data set $\mathbf{D}^+$). We then

1. Sample from randomized (control) networks (each of which has the same degree distributions as the given network) to obtain a negative data set $\mathbf{D}^-$, which contains the same number of isomorphic structures as $\mathbf{D}^+$.
2. Sample from the combined data set $\mathbf{D} = \{\mathbf{D}^+, \mathbf{D}^-\}$ (according to $\hat{\Theta}_1$ and Eq.1) to obtain a set of isomorphic structures which are most probably to be classified to the motif pattern.
3. Suppose that N_0 isomorphic structures have been identified in step (2), rank them by decreasing probabilities (i.e., the isomorphic structure with i-th largest probability is ranked as $N_0 - i + 1$), while assigning positive signs to ranks coming from $\mathbf{D}^+$ and negative signs to those from $\mathbf{D}^-$.
4. Run a Wilcoxon signed rank test (treating the signed ranks in the above step as being obtained from two paired samples) and obtain a reference p-value. □

The null hypothesis in the above test is that isomorphic structures sampled from the original and the control networks are equally likely to be classified to the motif pattern, while the alternative hypothesis is that isomorphic structures from the original network are more likely to be classified to the motif pattern. The circumstance of our test is slight different from the classical Wilcoxon test in that our test is conditional on the total number of isomorphic structures classified to the motif pattern.

3. Results.

3.1. Data sources. We studied a wide range of available transcriptional regulatory networks and protein-protein interaction networks. We downloaded the data sets

for *E.coli* and *S.cerevisiae* regulatory networks from Uri Alon's laboratory [Milo et al.(2002), Milo et al.(2004)] and the ChIP-chip data sets for the *S.cerevisiae* regulatory network from Young's laboratory [Lee et al.(2002), Harbison et al.(2004)]. We also downloaded the data sets of protein-protein interaction networks for 7 species (*E.coli*, *C.elegans*, *S.cerevisiae*(core), *H.pylori*, *M.musculus*, *D.melanogaster*, and *H.sapiens*) from DIP (Database of Interacting Proteins) [Xenarios et al.(2002), Salwinski et al.(2004)]. The details of those data sets are presented in Table 1.

3.2. Results on simulated networks. We first verify that the Gibbs sampling strategies can accurately recover the network motifs. Given a stochastic network $\mathcal{G}^{(o)}$, we use the following procedure to simulate a pseudo-network $\tilde{\mathcal{G}}^{(o)}$ with an n-node motif Θ embedded with probability λ. Suppose that G is the total number of sampled n-node subgraphs in the network and L is the corresponding number of isomorphic structures, the relation between λ and ε is simply $\lambda G = \varepsilon L$ under the iid assumption (see section 2.3).

1. Starting from the given network $\mathcal{G}^{(o)}$, generate a network $\mathcal{G}^{(r)}$, which has the same degree distributions as $\mathcal{G}^{(o)}$;
2. For the given λ, randomly choose $n_1 = [\lambda G]$ subgraphs from $\mathcal{G}^{(r)}$ and replace them with subgraphs generated according to the motif pattern Θ to obtain a pseudo-network $\tilde{\mathcal{G}}^{(o)}$, where G is the total number of n-node subgraphs in the given network $\mathcal{G}^{(o)}$. □

Node degrees in $\mathcal{G}^{(r)}$ should be fixed while subgraphs being replaced. For this purpose, we first compare the differences between the two adjacency matrices corresponding to the subgraphs before and after replacement to determine which elements should be changed from 1 to 0 and which from 0 to 1. Then we make modifications to these elements. Suppose that we want to change an element a_{st} in A (the adjacency matrix of the given graph) from 0 to 1. We scan the s-th row and the t-th column in A to find two element a_{sv} and a_{ut} equal to 1 while making sure that a_{uv} is equal to 0. Then, we modify by letting $a_{st} = 1, a_{sv} = 0, a_{ut} = 0$, and $a_{uv} = 1$. A similar method holds for changing an element from 1 to 0. When scanning elements, we should also exclude elements which have been previously modified.

We apply this method ($n = 3$) to the *E.coli* regulatory network (423 nodes, 519 edges) [Milo et al.(2002)]. Considering that $G = \binom{423}{3} \approx 1.25 \times 10^7$, we test five λ values ($1.0, 1.5, 2.0, 2.5, 3.0$ ($\times 10^{-6}$)), which are chosen based on our initial analysis of the data sets yielding estimated λ to be around 10^{-6}. For each λ, we generate 100 pseudo-networks, each of which is embedded with a randomly generated 3-node motif pattern (created by randomly assigning values in range $[0.7, 1.0]$ to elements $\theta_{ij}, i < j$ and 0 to others). For each pseudo-network, we run the group sampling strategy (with $M = 1$) for 1000 iterations and average over the last 100 iterations to obtain the estimation of $\hat{\varepsilon}$ and $\hat{\Theta}_1$. The pseudo-count $\tilde{n}_1$ is roughly set to 100, independent of

n_1 and without a bias. $\tilde{L}$ is set to $L/50$ so that the standard deviation of ε_1 is about $2/3$ of its mean. τ is chosen as 0.8 so that the standard deviation of α_{ijk} is about $1/5$ of its mean. The signed rank test is then applied to each estimated $\hat{\Theta}_1$ and a reference p-value is obtained. The quality of estimations with p-values $< 10^{-6}$ are further measured in terms of two relative error quantities

$$e_\lambda = \frac{|\hat{\varepsilon} L/G - \lambda|}{\lambda}, \text{ and } e_\Theta = \sqrt{\frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} (\frac{\theta_{ij} - \hat{\theta}_{ij}}{\theta_{ij}})^2},$$

and the results are shown in Figure 2. First, the figure shows that the sampling method can accurately estimate the parameters. For identified motifs with reference p-values $< 10^{-6}$, most relative errors are less than 10^{-2}. In other words, the embedded motif patterns are correctly recovered. Second, although the guess for $\tilde{n}_1$ is very crude, the relative errors do not change remarkably for different λ, suggesting that the sampling method is robust with $\tilde{n}_1$. We further test different pseudo-counts and find no remarkable difference for $L/25 \leq \tilde{L} \leq L/100$ and $0.7 \leq \tau < 1$.

3.3. Results on transcriptional regulatory networks. We apply both the Bernoulli and the group sampling ($M = 1$) strategies to identify 3- and 4-node network motifs in the regulatory networks of *E.coli* and *S.cerevisiae* [Milo et al.(2002), Milo et al.(2004)]. 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, the pseudo-counts are set as $\tilde{n}_1 = 100$, $\tilde{L} = L/50$, and $\tau = 0.8$. These pseudo-counts are also used for other calculations. As an example, Figure 3 shows a typical run of the Bernoulli and the group sampling ($M = 1$) strategies while identifying 3-node motifs in the *E.coli* regulatory network. Both strategies estimate $\Theta_1 = [0.0, 1.0, 1.0; 0.0, 0.0, 1.0; 0.0, 0.0, 0.0]$, which defines a *feed forward loop* motif (a transcription factor regulates another while both of them regulate a third gene) [Milo et al.(2002)]. More detailed results are presented in Table 2 (left part of the *E.coli** row). ε is estimated as 2.54×10^{-6}, and the p-value obtained using the signed rank test (2.26×10^{-6}) supports the statistical significance of the motif. For *S.cerevisiae*, both strategies identify a similar feed forward loop motif (left part of the *S.cere** row), which exists with $\varepsilon = 9.94 \times 10^{-7}$ and a reference p-value $< 10^{-8}$. Recent studies have shown that the feed forward loop serves as a sensitive delay element in regulatory networks [Milo et al.(2002), Mangan et al.(2003)]. 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 2, right). For *E.coli*, the motif exists with $\varepsilon = 9.52 \times 10^{-8}$, and is significant with a p-value $< 10^{-8}$; for *S.cerevisiae*, the motif exists with $\varepsilon = 1.40 \times 10^{-7}$, and is significant with a p-value $< 10^{-8}$. We also run the group sampling strategy with different

number of motif patterns ($M = 1, \ldots, 10$) and find the similar over-abundant motif pattern (data not shown). The deterministic bi-fan motifs (two transcription factors regulate two target genes in parallel, no interaction between the two transcription factors) have been identified previously [Milo et al.(2002), Milo et al.(2004)]. The stochastic bi-fan motifs have a novel feature (one transcription factor could also regulate the other), and reveal the existence of combinatorial transcriptional regulation in living cells [Martijn et al.(1999), Oliveira et al.(2003), Martinez-Pastor et al.(1996), Jiang et al.(2006)].

We also apply our method to the regulatory network of *S.cerevisiae* constructed using the ChIP-chip data [Lee et al.(2002), Harbison et al.(2004)]. 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 of 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% [Lee et al.(2002)]. We therefore infer that about 434 ($4,344 \times 0.1$) observed interactions are false positives, while about 858 ($4,344 \times 0.9 \times 0.18/0.82$) interactions are actually missing.

We use two method to assign interaction probabilities for protein-DNA pairs. Let π_{ij} be the probability that a transcription factor (node i) and a target gene (node j) have interaction, and p_{ij} the corresponding *p*-value provided by the data set. In the first method, we assume equal probabilities ($\pi_{ij} = 0.9$) to all the observed interactions ($p_{ij} < 0.001$) and equal probabilities ($\pi_{ij} = 858/(6,270 \times 113 - 4,344 \times 0.9) \approx 1.2 \times 10^{-3}$) to all other protein-DNA pairs ($p_{ij} \geq 0.001$). In the second method, we assume that π_{ij} and p_{ij} have the following relationship

$$\log \frac{\pi_{ij}}{1 - \pi_{ij}} = \alpha + \beta \log p_{ij}.$$

Thus protein-DNA pairs with low *p*-values would have higher interaction probabilities. An intuitive way to determine α and β is logistic regression. Because the known true interactions and known true no-interactions are limited, we determine α and β as follow.

The expected number of protein-DNA interactions in the "true" regulatory network (unknown), $\hat{M}$, can then be calculated as $\hat{M} = \sum_{i=1}^{N} \sum_{j=1}^{N} \pi_{ij}$. Let δ_{ij} be the indicator which is equal to 1 if $p_{ij} < 0.001$ and 0, otherwise. We can calculate $\hat{M}_1 = \sum_{i=1}^{N} \sum_{j=1}^{N} \pi_{ij}\delta_{ij}$, and $\hat{M}_0 = \sum_{i=1}^{N} \sum_{j=1}^{N} \pi_{ij}(1 - \delta_{ij})$, where $\hat{M}_1$ and $\hat{M}_0$ are the expected numbers of true interactions for positive observations ($p_{ij} < 0.001$) and negative observations ($p_{ij} \geq 0.001$), respectively. Let $\tilde{M}_1 = 4344 \times 0.9 \approx 3910$ and $\tilde{M}_0 = 858$ be the estimated numbers of true interactions for positive and negative observations, respectively. We determine α and β by solving the following minimization

problem

$$(\alpha^\star, \beta^\star) = \arg\min_{\alpha,\beta} \left(\frac{|\hat{M}_0 - \tilde{M}_0|}{\hat{M}_0 + \tilde{M}_0} + \frac{|\hat{M}_1 - \tilde{M}_1|}{\hat{M}_1 + \tilde{M}_1} \right),$$

where optimum value of 0 can only be obtained when $\hat{M}_0 = \tilde{M}_0$ and $\hat{M}_1 = \tilde{M}_1$. We treat this problem using a simulated annealing approach [Kirkpatrick et al.(1983)] and obtain $\alpha^\star = -13.58$, $\beta^\star = -4.51$.

The motifs identified in the stochastic network constructed using the second method are shown in Table 2. In the 3-node case, we applied the Bernoulli sampling strategy and identified a motif similar to the feed forward loop with $\varepsilon = 7.34 \times 10^{-8}$ is identified, and is significant with a reference p-value $< 10^{-8}$. In the 4-node case, a stochastic bi-fan motif is identified by the group sampling strategy with $\varepsilon = 1.06 \times 10^{-8}$, and is significant with a p-value $< 10^{-8}$. In the 5-node case, we applied the group sampling strategy to the network composed of the transcription factors only and identified two stochastic network motifs with references p-value $< 10^{-8}$. Similar motifs are found on the network constructed using the first method for probability assignment (data not shown). We notice that the identified motifs in the *S.cere* regulatory network constructed using the ChIP-chip data show more uncertainties than those constructed using human curated data, in that the estimated non-zero 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 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.

3.4. Results on protein-protein interaction networks. We also apply our method to protein-protein interaction networks for 7 species. These data sets are supposed to be reliable and thus for every interaction pair in these networks, we assign either 1 (interaction) or 0 (no-interaction) to the corresponding probability.

The motifs found by both the group sampling strategy are shown in Table 4. In the 3-node case (left part), the motifs identified for all species are the full connected *triangle*. These motifs exist with ε ranging from 10^{-8} to 10^{-6}, and are statistically significant with the reference p-values $< 10^{-8}$. As for the 4-node case (right part), we apply our method on the *E.coli*, *S.cerevisiae*, *M.musc*, and *H.sapiens* networks. For all the four species, the motifs identified are like full connected rectangles with both diagonals (one of them having low probability). These motifs exist in the networks with ε ranging from 1.39×10^{-8} to 4.98×10^{-8}. The reference p-values ($< 10^{-8}$) support the statistical significance of the motifs.

3.5. Comparison with existing methods. For comparison, we implement the widely used method for finding deterministic network motifs in deterministic networks [Milo et al.(2002), Milo et al.(2004)]. Briefly, the method 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 $Z = (N_{\text{real}} - \langle N_{\text{rand}} \rangle)/\text{std}(N_{\text{rand}})$ for regulatory networks and $\Delta = (N_{\text{real}} - \langle N_{\text{rand}} \rangle)/(N_{\text{real}} + \langle N_{\text{rand}} \rangle + \epsilon)$ for protein-protein interaction networks, where N_{real} is the occurrence number of a certain subgraph in the observed network; $\langle N_{\text{rand}} \rangle$ and $\text{std}(N_{\text{rand}})$ are the corresponding mean and standard deviation in the random ensemble; ϵ is a small positive number [Milo et al.(2004)].

According to their method, the feed forward loop and the deterministic bi-fan motifs are over-abundant in the deterministic regulatory networks for *E.coli* and *S.cerevisiae*, while the triangle and the rectangle with one or two diagonals are over-abundant in the protein-protein interaction networks for 7 species. By comparison, the feed forward loop and the stochastic bi-fan motifs are also identified by our method in the regulatory networks (see Table 1, 2) but both show uncertainties to some extent. Similarly, the triangle and the rectangle with both diagonals are identified in the protein-protein interaction networks (see Table 4) and both show some uncertainties. Furthermore, our method is also capable of finding network motifs in the stochastic regulatory network of *S.cerevisiae* and the identified motifs show more uncertainties than those in the deterministic networks (see Table 1, 2).

A closely related method for identifying network motifs in stochastic networks is introduced in [Jiang et al.(2006)], where a mixture model is used to describe stochastic networks and an expectation-maximization (EM) algorithm are utilized to determine the optimal parameters for the stochastic motif model. The experimental results presented in this paper show that the stochastic network motifs identified by our Gibbs sampling strategies are similar to what identified by the EM algorithm.

4. Conclusions and Discussion. We proposed two Bayesian models and three novelly designed Gibbs sampling strategies for identifying network motifs in stochastic biological networks and identified several stochastic network motifs 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 the network motifs (e.g. the stochastic bi-fan motif). Second, we model the 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 using the ChIP-chip

data). Third, we use a unified probabilistic model and a single statistic (ε) for different types of networks and network motifs (directed and undirected). Unlike other methods which use different statistics for different types of networks [Milo et al.(2004)], in our approach, different types of networks share the same model, which enable us to estimate and test the same statistic. Finally, with the newly designed caching technique, both the Bernoulli sampling and the group sampling strategy can run with high efficiency. These Gibbs sampling strategies with prior-annealing technique are not sensitive to the choice of pseudo-counts, which alleviates the requirement of the prior knowledge regarding the given network.

Besides the above advantages, the Bayesian model and the Gibbs sampling strategies proposed in this paper also have several advantages when compared with the mixture model and the EM algorithm proposed in [Jiang et al.(2006)]. First, the Gibbs sampling strategies simulates the posterior distribution of the parameters and estimates the parameter using the posterior mean. As a result, the method proposed in this paper is more robust than the EM algorithm, which targets in estimating the maximum likelihood parameters. Since systematic errors and experimental noises accompanying biological data are the main source for uncertainties in biological interaction networks, a more robust method is conceptually and practically preferred. Second, the Bayesian model proposed in this paper is able to deal with multiple motifs in a stochastic network, while the previous mixture model can only deal with single motif. For small motifs (i.e., 3 or 4 node), this is not a problem because the motif is simple. But when the motif is large and complicated, the advantage of multiple motif model is obvious. Third, the method proposed in this paper can deal with large scale networks by using the subgraph sampling method described in [Kashtan et al.(2004)]. With the development of high throughput techniques, existing biological interaction networks are getting complete and networks for more species will become available. Consequently, the ability of dealing with large scale networks would become crucial for future applications.

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 Bayesian framework, theoretical studies regarding the application scope of this assumption is necessary in our future study. Our stochastic 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 likelihood of observing a subgraph without a bias, but there are correlations between the likelihoods of observing a set of subgraphs. How to make corrections to this correlation is another consideration in our future research. Currently, 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 would be one of our future research focus.

REFERENCES

[Bailey and Elkan(1995)] T. L. BAILEY AND C. ELKAN. *Unsupervised learning of multiple motifs in biopolymers using expectation maximization.* Machine Learning, 21(1995), pp. 51–80.

[Barabasi and Albert(1999)] A. L. BARABASI AND R. ALBERT. *Emergence of scaling in random networks.* Science, 286(1999), pp. 509–512.

[Barabasi and Oltvai(2004)] A. L. BARABASI AND Z. N. OLTVAI. *Network biology: Understanding the cell's functional organization.* Nature Reviews Genetics, 5:1(2004), pp. 101–113.

[Berg and Lassig(2004)] J. BERG AND M. LASSIG. *Local graph alignment and motif search in biological networks.* Proc. Nat'l. Acad. Sci., 101:41(2004), pp. 14689–14694.

[Berg and Lassig(2006)] J. BERG AND M. LASSIG. *Cross-species analysis of biological networks by Bayesian alignment.* Proc. Nat'l. Acad. Sci., 103:29(2006), pp. 10967–10972.

[Berg et al.(2004)] J. BERG, S. WILLMANN, AND M. LASSIG. *Adaptive evolution of transcription factor binding sites.* BMC Evolutionary Biology, 4:1(2004), pp. 42.

[Chen et al.(2005)] Y. CHEN, P. DIACONIS, S. HOLMES, AND J. S. LIU. *Sequential monte carlo methods for statistical analysis of tables.* Journal of the American Statistical Association, 100:469(2005), pp. 109–120.

[Harbison et al.(2004)] C. T. HARBISON, D. B. GORDON, T. I. LEE, N. J. RINALDI, K. D. MACISAAC, T. W. DANFORD, N. M. HANNETT, J. TAGNE, D. B. REYNOLDS, J. YOO, E. G. JENNINGS, J. ZEITLINGER, D. K. POKHOLOK, M. KELLIS, P. A. ROLFE, K. T. TAKUSAGAWA, E. S. LANDER, D. K. GIFFORD, E. FRAENKEL, AND R. A. YOUNG. *Transcriptional regulatory code of a eukaryotic genome.* Nature, 431(2004), pp. 99–104.

[Ito et al.(2001)] T. ITO, T. CHIBA, R OZAWA, M YOSHIDA, M. HATTORI, AND Y SAKAKI. *A comprehensive two-hybrid analysis to explore the yeast protein interactome.* Proc. Nat'l. Acad. Sci., 98:8(2001), pp. 4569–4574.

[Jiang et al.(2006)] R. JIANG, Z. T. TU, T CHEN, AND F. Z. SUN. *Network motif identification in stochastic networks.* Proc. Nat'l. Acad. Sci., 103:25(2006), pp. 9404–9409.

[Kashtan et al.(2004)] N. KASHTAN, S. ITZKOVITZ, R. MILO, AND U. ALON. *Efficient sampling algorithm for stimating subgraph concentrations and detecting network motifs.* Bioinformatics, 20:11(2004), pp. 1746–1758.

[Kirkpatrick et al.(1983)] S. KIRKPATRICK, C. D. JR. GERLATT, AND M. P. VECCHI. *Optimization by simulated annealing.* Science, 220(1983), pp. 671–680.

[Lawrence et al.(1993)] C. E. LAWRENCE, S. F. ALTSCHUL, M. S. BOGUSKI, J. S. LIU, A. F. NEUWALD, AND J. C. WOOTTON. *Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment.* Science, 262(1993), pp. 208–214.

[Lee et al.(2002)] T. I. LEE, N. J. RINALDI, F. ROBERT, D. T. ODOM, Z. BAR-JOSEPH, G. K. GERBER, N. M. HANNETT, C. T. HARBISON, C. M. THOMPSON, I. SIMON, J. ZEITLINGER, E. G. JENNINGS, H. L. MURRAY, D. B. GORDON, B. REN, J. J. WYRICK, J. B. TAGNE, T. L. VOLKERT, E. FRAENKEL, D. K. GIFFORD, AND R. A YOUNG. *Transcriptional regulatory*

networks in Saccharomyces cerevisiae. Science, 298(2002), pp. 799–804.

[Liu et al.(1995)] J. S. LIU, A. F. NEUWALD, AND C. E. LAWRENCE. *Bayesian models for multiple local sequence alignment and gibbs sampling strategies.* Journal of the American Statistical Association, 90:432(1995), pp. 1156–1170.

[Mangan et al.(2003)] S. MANGAN, A. ZASLAVER, AND U. ALON. *The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks.* Journal of Molecular Biology, 334:1(2003), pp. 197–204.

[Martijn et al.(1999)] R. MARTIJN, R. VLADIMIR, G. ULRIKE, M. T. JOHAN, H. STEAFAN, A. GUS-TAV, AND R. HELMUT. *Osmotic stress-induced gene expression in Sacharomyces cerevisiae requires msn1p and the novel nuclear factor hot1p.* Molecular and Cellular Biology, 19:8(1999), pp. 5474–5485.

[Martinez-Pastor et al.(1996)] M. T. MARTINEZ-PASTOR, G. MARCHLER, C. SCHULLER, A. MARCHLER-BAUER, H. RUIS, AND F. ESTRUCH. *The Saccharomyces cerevisiae zinc finger proteins msn2p and msn4p are required for transcriptional induction through the stress response element (stre).* The EMBO Journal, 15:9(1996), pp. 2227–2235.

[Milo et al.(2002)] R. MILO, S. SHEN-ORR, S. ITZKOVITZ, N. KASHTAN, CHKLOVSKII D., AND U. ALON. *Network motifs: Simple building blocks of complex networks.* Science, 298:1(2002), pp. 824–827.

[Milo et al.(2004)] R. MILO, S. ITZKOVITZ, N. KASHTAN, R. LEVITT, S. SHEN-ORR, I. AYZENSHTAT, M. SHEFFER, AND U. ALON. *Superfamilies of evolved and designed networks.* Science, 303:1(2004), pp. 1538–1542.

[Newman(2003)] M. E. J. NEWMAN. *The structure and function of complex networks.* SIAM Review, 45:2(2003), pp. 167–256.

[Newman et al.(2001)] M. E. J. NEWMAN, S. H. STROGATZ, AND D. J. WATTS. *Random graphs with arbitrary degree distributions and their applications.* Physical Review E., 64:1(2001), pp. 026118(1) − 026118(17).

[Niu et al.(2002)] T. H. NIU, Z. S. QIN, X. XU, AND J. S. LIU. *Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms.* American Journal of Human Genetics, 70(2002), pp. 157–169.

[Oliveira et al.(2003)] E. M. M. OLIVEIRA, A. S. MARTINS, E. GARVAJAL, AND E. P. S. BON. *The role of the gata factors gln3p, nil1p, dal80p and the ure2p on asp3 regulation in Sacharomyces cerevisiae.* Yeast, 20(2003), pp. 31–37.

[Prill et al.(2005)] R. J. PRILL, P. A. IGLESIAS, AND A. LEVCHENKO. *Dynamicd properties of network motifs contribute to biological network organization.* PLoS Biology, 3:11(2005), pp. e343.

[Roberts(1998)] P. D. ROBERTS. *Classification of temporal patterns in dynamic biological networks.* Neural Computation, 10:7(1998), pp. 1831–1846.

[Salwinski et al.(2004)] L. SALWINSKI, C. S. MILLER, A. J. SMITH, F. K. PETTIT, J. U. BOWIE, AND D. EISENBERG. *The database of interacting proteins: 2004 update.* Nucleic Acids Res., 32(2004), pp. D449–451.

[Shen-Orr et al.(2002)] S. SHEN-ORR, R. MILO, S. MANGAN, AND U. ALON. *Network motifs in the transcriptional regulation network of Escherichia Coli.* Nature Genetics, 31:1(2002), pp. 64–68.

[Spirin and Mirny(2003)] V. SPIRIN AND L. A. MIRNY. *Protein complexes and functional modules in molecular networks.* Proc. Nat'l. Acad. Sci., 100:21(2003), pp. 12123–12128.

[Uetz et al.(2000)] P. UETZ, L. GIOT, G. CAGNEY, T. A. MANSFIELD, R. S. JUDSON, J. R. KNIGHT, D. LOCKSHON, V. NARAYAN, M. SRINIVASAN, P. POCHART, A. QURESHI-EMILI, Y. LI, B. GODWIN, D. GONOVER, T. KALBFLEISCH, G. VIJAYADAMODAR, M. YANG, M. JOHNSTON, S. FIELDS, AND J. M. ROTHBERG. *A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. Nature*, 403:10(2000), pp. 623–627.

[Vazauez et al.(2004)] A. VAZAUEZ, R. DOBRIN, D. SERGI, J. P. ECKMANN, Z. N. OLTVAI, AND

A. L. BARABASI. *The topological relationship between the large-scale attributes and local interaction patterns of complex networks.* Proc. Nat'l. Acad. Sci., 101:52(2004), pp. 17940–17945.

[Xenarios et al.(2002)] I. XENARIOS, L. SALWINSKI, X. J. DUAN, P. HIGNEY, S. M. KIM, AND D. EISENBERG. *Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions.* Nucl. Acids Res., 30:1(2002), pp. 303–305.

[Yeger-Lotem et al.(2004)] E. YEGER-LOTEM, S. SATTATH, N. KASHTAN, S. ITZKOVITZ, R. MILO, R. Y. PINTER, U. ALON, AND H. MARGALIT. *Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction.* Proc. Nat'l. Acad. Sci., 101:16(2004), pp. 5934–5939.
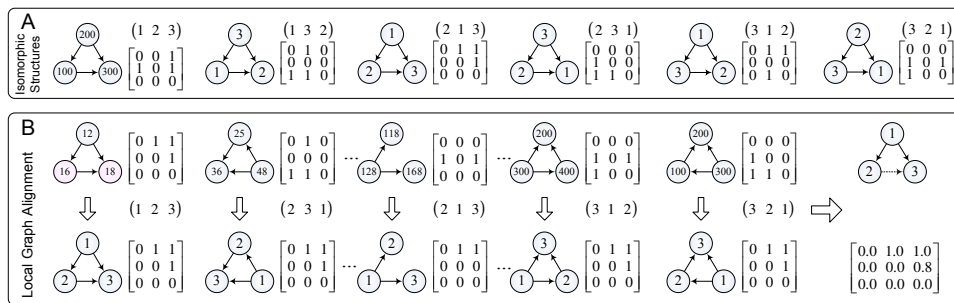
FIG. 1. *(A) A subgraph and its isomorphic structures. Left, a subgraph with three nodes (labeled by $(100, 200, 300)$ in a certain graph). The adjacency matrix is obtained by relabeling nodes as $(1, 2, 3)$, respectively. Right five columns, isomorphic structures of the subgraph. The isomorphic structures (top row) have the same connectivity as the subgraph but different adjacency matrices (middle), which can be generated by permuting the node labels (indices of the matrices), as is shown in the top. (B) An example of local graph alignment. A set of subgraphs (the top row) are sampled from a certain network and five of them are selected for alignment because they have similar connectivity. The alignment is done by assigning for each of the similar subgraphs a proper isomorphic structure (denoted by the permutation of node labels in the middle row). The stochastic motif pattern is then obtained by averaging over the adjacency matrices corresponding to aligned isomorphic structures (the right column).*
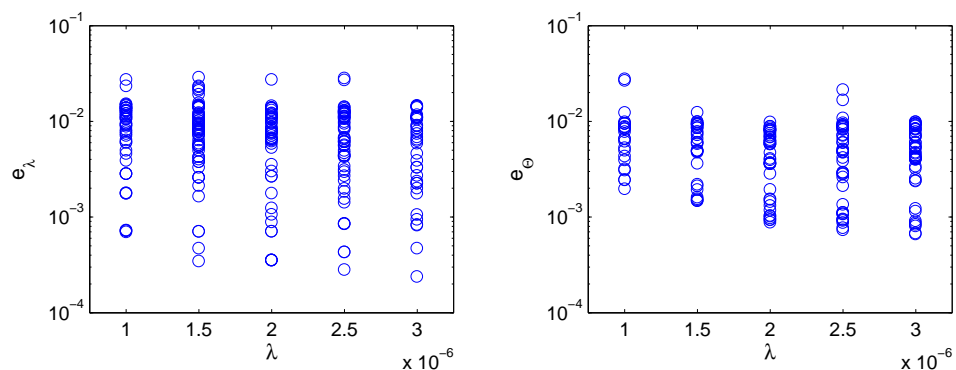
FIG. 2. *Simulation results for finding 3-node motif patterns in pseudo regulatory networks. Left, the relationship of the relative error e_λ verses λ. Right, the relationship of the relative error e_Θ versus λ.*
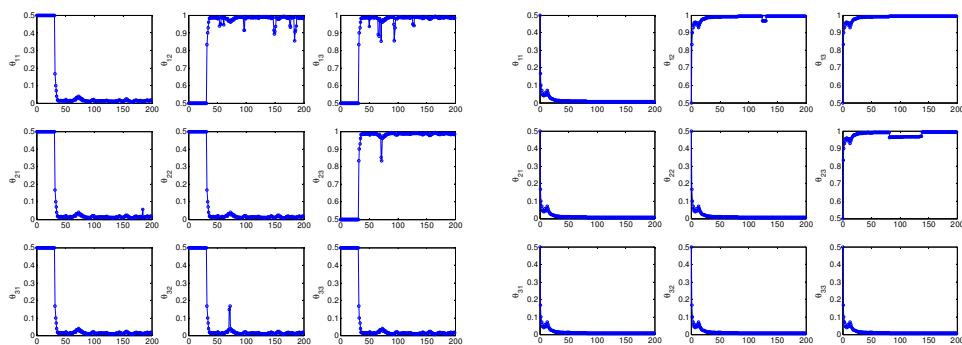
FIG. 3. *An example of running the Bernoulli and corresponding group sampling strategies to identify 3-node motifs in the* E.coli *regulatory network. The left three columns are the process of the Bernoulli sampling strategy (x-axes ($\times 1000$) are the iteration steps; y-axes are the posterior means of $\Theta_1 = (\theta_{ij})_{3\times 3}$). The right three columns are the process of the group sampling strategy with $M = 1$ (x-axes are the iteration steps; y-axes are the posterior means of Θ_1).*

TABLE 1

Details of the data studied. [1]*: Transcriptional regulatory networks.* [2]*: Protein-protein interaction networks.* [3]*: Based on human curated data.* [4]*: Based on ChIP-chip data.*

Species	#(Nodes)	#(Edges)	Average (in/out) degree
E.coli[1,3]	423	519	1.23
S.cerevisiae[1,3]	688	1,078	1.57
S.cerevisiae[1,4]	2,416	4,344	1.80
E.coli[2]	553	483	1.75
S.cerevisiae[2]	2,614	6,319	4.83
C.elegans[2]	2,638	3,970	3.01
H.pylori[2]	710	1,359	3.83
M.musculus[2]	329	274	1.67
D.melanogaster[2]	7,068	20,815	5.89
H.sapiens[2]	1,065	1,318	2.48

TABLE 2

Stochastic motifs in transcriptional regulatory networks of E.coli *and* S.cerevisiae *(based on highly reliable data from human curated databases [Milo et al.(2002), Milo et al.(2004)])*

Species	ε	Θ_1	Motif
E.coli	2.54×10^{-6}	$\begin{bmatrix} 0.00 & 1.00 & 1.00 \\ 0.00 & 0.00 & 1.00 \\ 0.00 & 0.00 & 0.00 \end{bmatrix}$	
S.cere	9.94×10^{-7}	$\begin{bmatrix} 0.00 & 1.00 & 1.00 \\ 0.00 & 0.00 & 1.00 \\ 0.00 & 0.00 & 0.00 \end{bmatrix}$	
E.coli	9.52×10^{-8}	$\begin{bmatrix} 0.00 & 0.01 & 0.96 & 0.99 \\ 0.00 & 0.00 & 0.99 & 0.92 \\ 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}$	
S.cere	1.40×10^{-7}	$\begin{bmatrix} 0.00 & 0.04 & 1.00 & 1.00 \\ 0.00 & 0.00 & 1.00 & 1.00 \\ 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}$	
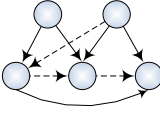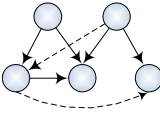
Species	ε	Θ_1	Motif
S.cere[1]	7.34×10^{-8}	$\begin{bmatrix} 0.00 & 0.99 & 0.94 \\ 0.00 & 0.00 & 0.98 \\ 0.00 & 0.00 & 0.00 \end{bmatrix}$	
S.cere[1]	1.06×10^{-8}	$\begin{bmatrix} 0.00 & 0.12 & 0.99 & 0.99 \\ 0.00 & 0.00 & 0.97 & 0.93 \\ 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}$	
S.cere[2]	6.07×10^{-5}	$\begin{bmatrix} 0.00 & 0.00 & 1.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.13 & 1.00 & 1.00 \\ 0.00 & 0.00 & 0.00 & 0.09 & 1.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.03 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}$	
S.cere[2]	5.88×10^{-5}	$\begin{bmatrix} 0.00 & 0.00 & 1.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.14 & 1.00 & 1.00 \\ 0.00 & 0.00 & 0.00 & 1.00 & 0.09 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}$	

Species	ε	Θ_1	Motif	Species	ε	Θ_1	Motif
E.coli	4.50×10^{-6}	$\begin{bmatrix} 0.00 & 0.99 & 0.99 \\ 0.99 & 0.00 & 0.99 \\ 0.99 & 0.99 & 0.00 \end{bmatrix}$		E.coli	4.38×10^{-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.cere	1.25×10^{-6}	$\begin{bmatrix} 0.00 & 1.00 & 1.00 \\ 1.00 & 0.00 & 1.00 \\ 1.00 & 1.00 & 0.00 \end{bmatrix}$		S.cere	4.86×10^{-8}	$\begin{bmatrix} 0.00 & 0.17 & 1.00 & 1.00 \\ 0.17 & 0.00 & 1.00 & 1.00 \\ 1.00 & 1.00 & 0.00 & 1.00 \\ 1.00 & 1.00 & 1.00 & 0.00 \end{bmatrix}$	
C.eleg	1.05×10^{-7}	$\begin{bmatrix} 0.00 & 1.00 & 1.00 \\ 1.00 & 0.00 & 1.00 \\ 1.00 & 1.00 & 0.00 \end{bmatrix}$					
H.pylo	1.21×10^{-6}	$\begin{bmatrix} 0.00 & 0.99 & 0.99 \\ 0.99 & 0.00 & 0.99 \\ 0.99 & 0.99 & 0.00 \end{bmatrix}$		M.musc	4.98×10^{-8}	$\begin{bmatrix} 0.00 & 0.19 & 0.99 & 0.99 \\ 0.19 & 0.00 & 0.99 & 0.99 \\ 0.99 & 0.99 & 0.00 & 0.99 \\ 0.99 & 0.99 & 0.99 & 0.00 \end{bmatrix}$	
M.musc	2.31×10^{-6}	$\begin{bmatrix} 0.00 & 0.96 & 0.96 \\ 0.96 & 0.00 & 0.96 \\ 0.96 & 0.96 & 0.00 \end{bmatrix}$					
D.mela	2.71×10^{-8}	$\begin{bmatrix} 0.00 & 1.00 & 1.00 \\ 1.00 & 0.00 & 1.00 \\ 1.00 & 1.00 & 0.00 \end{bmatrix}$		H.sapi	1.39×10^{-8}	$\begin{bmatrix} 0.00 & 0.26 & 1.00 & 1.00 \\ 0.26 & 0.00 & 1.00 & 1.00 \\ 1.00 & 1.00 & 0.00 & 1.00 \\ 1.00 & 1.00 & 1.00 & 0.00 \end{bmatrix}$	
H.sapi	1.49×10^{-6}	$\begin{bmatrix} 0.00 & 1.00 & 1.00 \\ 1.00 & 0.00 & 1.00 \\ 1.00 & 1.00 & 0.00 \end{bmatrix}$					