Article

Network motif identification and structure detection with exponential random graph models

Munni Begum[1], **Jay Bagga**[1], **Ann Blakey**[1], **Sudipta Saha**[2]
[1]Ball State University, Muncie, IN 47306, USA
[2]University of South Carolina, Columbia, SC, USA
E-mail: mbegum@bsu.edu

Received 22 August 2014; *Accepted 25 September 2014*; *Published online 1 December 2014*

Abstract

Local regulatory motifs are identified in the transcription regulatory network of the most studied model organism *Escherichia coli* (*E. coli*) through graphical models. Network motifs are small structures in a network that appear more frequently than expected by chance alone. We apply social network methodologies such as p^* models, also known as Exponential Random Graph Models (ERGMs), to identify statistically significant network motifs. In particular, we generate directed graphical models that can be applied to study interaction networks in a broad range of databases. The Markov Chain Monte Carlo (MCMC) computational algorithms are implemented to obtain the estimates of model parameters to the corresponding network statistics. A variety of ERGMs are fitted to identify statistically significant network motifs in transcription regulatory networks of *E. coli*. A total of nine ERGMs are fitted to study the transcription factor - transcription factor interactions and eleven ERGMs are fitted for the transcription factor-operon interactions. For both of these interaction networks, *arc* (a directed edge in a directed network) and *k-istar* (or incoming star structures), for values of *k* between 2 and 10, are found to be statistically significant local structures or network motifs. The goodness of fit statistics are provided to determine the quality of these models.

Keywords biological networks; network motifs; transcriptional regulatory network; graphical models; exponential random graph models; Markov Chain Monte Carlo algorithms.

Network Biology
ISSN 2220-8879
URL: http://www.iaees.org/publications/journals/nb/online-version.asp
RSS: http://www.iaees.org/publications/journals/nb/rss.xml
E-mail: networkbiology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

Biological functions depend on complex interactions among the cell's numerous constituents such as protein, DNA, RNA and other small molecules. Thus, for biologists it is important to assess interactions among molecules at different levels of hierarchy. In particular, there is a high degree of interest in identifying interactions at the gene-gene, gene-protein, and metabolic levels. High-throughput assays that probe cells and sub-cellular systems at the genome scale can measure molecular interaction networks and their components at

each of these key levels. These may include specific gene sequences, mRNA transcription products, protein-protein interactions, protein-DNA interactions, as well as other interactions of interest (Friedman, 2004). With the advancement of high-throughput data collection techniques such as microarrays and next generation sequencing, it is now possible to investigate the status of the interactions of a particular cell's components. Thus scientists can assess complex molecular interactions by implementing proper computational methodologies.

Molecular interactions that are of the most interest include transcription regulatory networks, protein-protein interaction, and metabolic interaction. Although one can study each of these interactions separately, none of these networks functions independently but instead they form a series of interdependent networks (Barabasi and Oltvai, 2004). This large complex global system of cellular networks thus determines the major characteristics of a cell and the functions of its sub-cellular regions. In order to better understand local features of a complex global network one needs to study each interaction network separately. For example, Costanzo et al. (2010) examined the *Saccharomyces cerevisiae* cellular networks by connecting pairs of genes with similar profiles using Pearson correlation coefficients to form global interaction networks. Biological networks are characterized as functions of local network features (Saul and Filkov, 2007) using a family of statistical models from the social network methodologies, such as Exponential Random Graph Models (ERGMs). The ERGMs provide a flexible principle to study global network structure as a function of prominent 'local features'.

Gene expression is a fundamental process to the survival of an organism, both prokaryotic and eukaryotic, as well as the propagation of the various classes of viruses. An overview of the transcription regulatory networks of *E. coli* from basic biology to our current understanding on a global scale has been previously well-described by Martínez-Antonio (2011). In particular, core processes of the central dogma of biology involve DNA transcription, multi-subunit enzymes, several classes of RNA, and specialized proteins known as transcription factors, or TFs. Together these elements produce what is known as the *transcriptome* of the cell. Overall, the steps appear to be simple: induction of transcription or release from repression of transcription, initiation of transcription, elongation or synthesis of the RNA species, and termination of the RNA species. But in reality, each step involves levels of complexity that we are just beginning to understand. Some genes within the genome may require only one to two TFs to activate the transcriptional process at a promoter site, while others can require ten or more TFs for overall regulation and attenuation, as seen in the *RegulonDB* database (RegulonDB Release 7.4, 2012) (RegulonDB, 2012).Production of each of these polypedtides that function as a TF is, in itself, a highly regulated process. Thus, determination of the key points of regulation within the smaller networks that give rise to the major cascades of metabolic activities within the cell will prove to be invaluable to our understanding of the system as a whole.

The examination of transcription regulating network motifs has been attempted with various algorithms since the late 1990s. These analyses have attempted to examine gene expression from the perspective of circuitry (Thieffry et al., 1998), and smaller network motifs within a global network (Shen-Orr et al., 2002). Biological context of the data allows for refinement of the analysis of the inter-relationship of specific features within smaller local network features. These features will have both hierarchical and evolutionary implications when evaluated for their regulatory roles with a biological system (Balaji et al., 2007) and (Martnez-Antonio, 2011). Therefore, utilizing the most updated version of the *RegulonDB* as of the time of analysis (currently v7.5, soon to be v8.0) (RegulonDB, 2012) has allowed for access to the most detailed information assembled on the *E. coli* transcriptional regulatory network (Gama-Castro et al., 2011). While it remains important to consider the globality of the regulatory networks, closer examination of the unique features of the smaller motifs has provided an insight towards mechanisms of control both in vivo and in vitro.

2 Materials and Methods

2.1 Data

We consider the *E. coli* K-12 transcriptional regulatory network interactions using *RegulonDB* (RegulonDB, 2012). This database contains information about the organization of *operons*, genes organized into a single transcriptional unit, and the composition of operons into *transcriptional units* among numerous other information units within a transcriptional network. In *RegulonDB* database, an *operon* is defined as "a set of one or several genes and their associated regulatory elements, which are transcribed as a single unit". However, an additional criterion is that one particular gene cannot belong to more than one *operon* (RegulonDB, 2012). A *transcription unit* is thus defined as a set of one or more genes within an operon transcribed as a set through the utilization of a single promoter. The database also provides terminology regarding transcriptional units known as a *regulon*. A regulon in its simplest form involves the regulation of a group of genes regulated by a single *regulator*, hereafter referred to as a transcription factor but is known to exit in a complex form involving two or more regulating transcription factors or *regulators*. It is the information provided by each of the *simple regulons, complex regulons*, and *strict complex regulons* that differentiates and identifies the unique local features of those networks.

Saul and Filkov (2007) implemented ERGMs to a number of biological networks including the transcription regulatory network of *Escherichia coli* (*E. coli*) (Shen-Orr et al., 2002). The transcription regulatory network of *E. coli* is updated regularly in the repository *RegulonDB*. We selected this particular network and studied the structure of the interactions closely due to several reasons. First, all regulatory networks are the most important biological network due to their role in gene expression. In the previous works on this network, direction of the regulation was not addressed properly. We considered two types of regulatory networks in *E.coli*: 1) regulation between transcription factors namely transcription factor - transcription factor (TF-TF) interactions, and 2) regulation between transcription factors and the *operons* that contain TFs. The second network is referred to as the transcription factor - operon (TF-Operon) interaction network. In both cases, we generated directed exponential random graph models and identified prominent local features. Our results are comparable with those obtained by Saul and Filkov (2007) with the additional advantage that our approach also addresses the regulatory interactions and places them within their proper biological context.

The networks for the TF-TF interactions and TF-Operon interactions we observed are presented in Fig. 1. The network on the left panel (Fig. 1(a)) is for the TF-TF interactions and the network on the right panel (Fig. 1(b)) is for the TF-Operon interactions. We implement ERGMs to both types of interaction networks in order to identify statistically significant network motifs that can be used to represent these observed networks and compare the sets in terms of similarity and uniqueness. In the method section below, we briefly discuss the ERGMs and associated computational algorithms.

2.2 Method

Biological networks have been investigated using several network models such as the Erdos-Renyi model (Erdos and Renyi, 1960), the geometric random network model, exponential random graph models (ERGM), and graphical models (Begum et al., 2012; Zhang, 2011, 2012). In particular, the Erdos-Renyi and the geometric random network models were used in the study of graphlets in *Saccharomyces cerevisiae* protein-protein interaction (PPI) networks (Przulj et al., 2004), and exponential random graph models have been employed to study biological databases such as *RegulonDB* (RegulonDB, 2012). The ERGMs have also been used to study large social networks (Goodreau, 2007; Robins et al., 2007). In order to study two specific transcription regulatory networks of *E. coli* we generate a directed ERGM and identify the statistically significant network statistics as prominent 'local features'.

The ERGM represents a general and flexible methodology for modeling interactions among a number of

actors in a complex network. This methodology originated and had been implemented widely in the literature of social networks. ERGMs generalize the Markov random graph models (Frank and Strauss, 1986), and edge and dyadic independence models. Briefly we discuss the ERGM, also known as the p^*model and the associated computational algorithms in the following subsections.

2.2.1 The p* model

The p^*model is a more general model that includes the Markov random graph models and the dyadic independence models also known as $p_1$model as special cases. In order to specify a p^* model we follow the notations of Wasserman and Pattison (Wasserman and Pattison, 1996). Let X_{ij}^{+}denote an adjacent matrix where a tie from $i \to j$ is forced to be present. That is $X_{ij}^{+} = \{X_{kl}, \text{with } X_{ij} = 1\}$. X_{ij}^{-} denotes an adjacency matrix where a tie from $i \to j$ is forced to be absent. That is $X_{ij}^{-} = \{X_{kl}, with\ X_{ij} = 0\}$. And finally, X_{ij}^{c} denotes an adjacency matrix with complement relation for the tie from $i \to j$. That is $X_{ij}^{c} = \{X_{ki}, with(k,l) \neq (i,j)\}$.

The general log-linear form of p^* model is expressed as

$$P(\boldsymbol{X} = \boldsymbol{x}) = \frac{\exp(\boldsymbol{\theta}'\boldsymbol{z}(\boldsymbol{x}))}{\kappa(\boldsymbol{\theta})} \tag{1}$$

here $\boldsymbol{\theta}$ is a vector of model parameters, $\boldsymbol{z}(\boldsymbol{x})$ is a vector of network statistics, and $\kappa(.)$ is a normalizing constant which is hard to compute for moderate to large networks. In order to ease the estimation process of the model parameters, the log-linear model form of the p^*model can be re-expressed as a logit model. A logit model is a special case of generalized linear model where log odds of a binary variable is expressed as linear combination of several explanatory variables. The p^*model in (1) can be converted to a logistic regression model by considering the set of binary random variables $\{X_{ij}\}$,where $X_{ij} = 1$ implying a tie from *i* to *j* (Strauss and Ikeda, 1990). With the new notations, the log-linear model in (1) can be expressed as a logit model as in (4).

$$P\left(X_{ij} = 1 \middle| \boldsymbol{X}_{ij}^{c}\right) = \frac{P(\boldsymbol{X} = \boldsymbol{x}_{ij}^{+})}{P\left(\boldsymbol{X} = \boldsymbol{x}_{ij}^{+}\right) + P(\boldsymbol{X} = \boldsymbol{x}_{ij}^{-})} \tag{2}$$

$$P\left(X_{ij} = 0 \middle| \boldsymbol{X}_{ij}^{c}\right) = \frac{P(\boldsymbol{X} = \boldsymbol{x}_{ij}^{-})}{P\left(\boldsymbol{X} = \boldsymbol{x}_{ij}^{+}\right) + P(\boldsymbol{X} = \boldsymbol{x}_{ij}^{-})} \tag{3}$$

Using expression in (1) and taking the ratio of (2) and (3) one can write

$$\frac{P\left(X_{ij} = 1 \middle| \boldsymbol{X}_{ij}^{c}\right)}{P\left(X_{ij} = 0 \middle| \boldsymbol{X}_{ij}^{c}\right)} = \exp\left\{\boldsymbol{\theta}'\left[\mathbf{z}(\mathbf{x}_{\mathrm{ij}}^{+}) - \mathbf{z}(\mathbf{x}_{\mathrm{ij}}^{-})\right]\right\}$$

$$\log\left\{\frac{P\left(X_{ij} = 1 | \mathbf{X}_{ij}^{c}\right)}{P\left(X_{ij} = 0 | \mathbf{X}_{ij}^{c}\right)}\right\} = \omega_{ij} = \boldsymbol{\theta}'\left[\mathbf{z}(\mathbf{x}_{\mathrm{ij}}^{+}) - \mathbf{z}(\mathbf{x}_{\mathrm{ij}}^{-})\right]$$

$$\omega_{ij} = \boldsymbol{\theta}'\boldsymbol{\delta}\left(x_{ij}\right) \tag{4}$$

Here $\boldsymbol{\delta}\left(x_{ij}\right)$ is the vector of difference statistics obtained from the network statistics $\mathbf{z}(.)$ when the variable X_{ij}changes from 1 to 0. The model in (4) is referred to as the *logit* p^* model for single binary relation (Wasserman and Pattison, 1996). One can work with either the log-linear form of p^*model given in (1) or the logit form given in (4). However for a sparse or complete network with lack of interactions and strong interactions respectively, model in (1) is preferable.

2.2.2 Computational algorithms

As observed by Snijders et al. (2006) and Goodreau (2007), models with dyadic independence are good candidates for logit form of p^*models and should employ the method of Maximum Pseudo Likelihood Estimation (MPLE) for parameter estimation. Whereas the models with dyadic dependence should be expressed as log-linear form as in equation (1). These more general ERGM models do not have closed form

expression as the normalizing constant involves sums or integration over a large number of variables (nodes). Thus it is impossible to apply the method of maximum likelihood estimation (MLE) for estimating the parameters of these models. However, the Monte Carlo approximation of the MLE is often used for such models as in Geyer (1991), Geyer and Thompson (1992), and Saul and Filkov (2007). The Monte Carlo approximation of the MLE using Markov chains is known as the Markov chain Monte Carlo MLE (MCMC-MLE). We adopt the notations of Geyer (Geyer, 1991) in order to describe the basics of the MCMC-MLE and MPLE.

We write the ERGM model in equation (1) with respect to a generic probability measure μ (discrete or continuous) as follows:

$$f_\theta(x) = \frac{1}{\kappa(\theta)} h_\theta(x) \tag{5}$$

where $h_\theta(x) = \exp[\boldsymbol{\theta}' \boldsymbol{z}(\boldsymbol{x})]$ and $\kappa(\theta) = \int h_\theta(x) d\mu(x)$. The integral in (5) is analytically intractable. The Markov chain Monte Carlo proceeds as providing a sample $X_1, X_2, \ldots$ from any ϕ in the parameter space which can be used to estimate the log-likelihood ratio for an observation x (Geyer, 1991). Here the log-likelihood ratio is written as

$$l(\theta) = log \frac{f_\theta(x)}{f_\phi(x)} = \log \frac{h_\theta(x)}{h_\phi(x)} - \log \frac{\kappa(\theta)}{\kappa(\phi)} \tag{6}$$

Note that, because the ratio of normalizing constant can be expressed as

$\frac{\kappa(\theta)}{\kappa(\phi)} = E_\theta \left[\frac{h_\theta(x)}{h_\phi(x)}\right]$,

the log-likelihood ratio in expression (6) can be approximated by replacing the ratio of normalizing constant $\frac{\kappa(\theta)}{\kappa(\phi)}$ by its Monte Carlo estimate

$\left(\frac{1}{n}\sum_{i=1}^{n} \frac{h_\theta(X_i)}{h_\phi(X_i)}\right)$.

Finally an MCMC-MLE of θ is obtained by maximizing the approximate likelihood. The MCMC-ML estimation algorithm is implemented to the software package *statnet* under the statistical computational environment **R**. We use these two packages *statnet* and *ergm* to fit the exponential random model given in (1). The MPLE method proceeds as maximizing the pseudo-likelihood which is the product of $P(X_{ij} = 1 \,|\boldsymbol{X}_{ij}^c)$ in equation (2) for all the nodes (i, j), $i \neq j$. The MPLE of θ is obtained by maximizing the log pseudo-likelihood (Besag, 1975). Thus for the logit model derived from (2) and (3) and given in (4), finding the MPLE of θ is equivalent to fitting a logistic regression model and hence obtaining the parameter estimates. One can use the software packages *statnet* and *ergm* to obtain MPLE. As discussed earlier if dyadic independence assumption is not valid MPLE may produce poor estimates and the method of MCMC-MLE should be implemented instead.

3 Results

The transcription regulation in *E. coli* is studied with two specific interaction networks, namely the TF-TF interaction and the TF-Operon interaction. The observed networks for TF-TF interaction and TF-Operon interaction are presented in Fig. 1. There are 175 vertices and 387 edges in the TF-TF interaction network (Fig. 1(a)). There are 898 vertices and 1702 edges in the observed TF-Operon interaction network (Fig. 1(b)). Both of these are directed networks with loops. However, self looping is excluded from these networks as the ERGM methodology is unable to handle loops in modeling interaction networks. Each vertex in Fig. 1(a) represents a transcription factor (TF) and an *arc* between two TFs represents a tie. In a directed network an

edge between two vertices is termed as an *arc*. An edge from one TF to another TF represents that the first TF regulates the second TF. Each vertex in Fig. 1(b) represents an operon or TF and an *arc* between two operons and/or TFs represents regulation of an operon by another operon or TF. In the TF-TF interaction network we see that there are two big clusters and several small clusters whereas in TF-Operon interaction we see that there is one big cluster and fewer smaller clusters. We derive statistically significant network statistics for both networks as discussed in the following subsections.

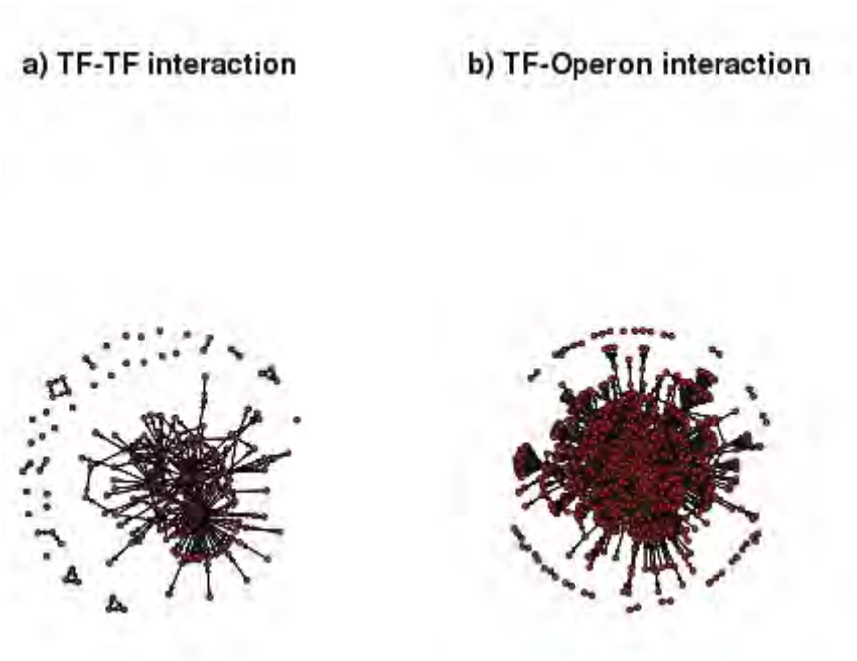

Fig. 1 Observed TF-TF and TF-Operon interaction networks in *E. coli.* a) TF-TF interaction with 175 TFs and 387 edges denoting interactions. b) TF-Operon interaction with 898 Operons and 1702 edges denoting interactions.

3.1 TF-TF interactions

We fit nine exponential random graph models (ERGMs) (Models 1-9 in Tables 1(a) and 1(b)) with varying network statistics for the TF-TF interaction network model. The network statistics include *arc* and *k-star*, where a *k-star* is defined to be a node *N* and a set of *k* different nodes $O_1, \dots, O_k$ such that the ties (N, O_i) exist for $i = 1, \dots, k$. For a directed network, as in the regularity network of *E. coli*, *k-star* statistics are denoted as *k-istar* and *k-ostar* in order to emphasize the direction as incoming *k-star* and outgoing k-star respectively. Models 1-4 include individual network statistics, such as arc and *2-istar, 3-istar, 4-istar*. Models 5-9 include multiple network statistics. A variety of other models with network statistics *triangle, istar* (with higher *k*-values), and *ostar* with various *k*-values are also fitted for the TF-TF interaction network of *E. coli*. The results from the models for which convergence criteria are met are presented in Tables 1(a) and 1(b). Table 1(a) contains all the network statistics for models 1-6 but part of the network statistics (*arc, 2-istar,...4-istar*) for models 7-9. Rest of the network statistics (*5-istar, 9-istar, and 10-istar*) of models 7-9 are presented in Table 1(b). The *k*-values of *istar* and *ostar* statistics are selected based on the actual counts from the observed network.

The estimates of the parameters (θ-scalar for models 1-4, and vector for models 5-9) and their standard errors are presented in Tables 1(a) and 1(b). For example, model 1 considers only arc as the network statistic of interest. An estimate of the coefficient of *arc* statistic -4.74 implies that the log-odds of a regulation is -4.74 $\times$ change in the number of ties which is just -4.74. Thus the corresponding probability of a TF regulating another

TF when there is an *arc* in the network can be estimated as 0.0087. These probabilities for TF-TF and TF-Operon interaction models in the presence of individual network statistics only are presented in Table 3. Models 5-9 include multiple network statistics in order to facilitate conditional impact of one statistic, holding the effects of the rest fixed in the model. For example, model 5 considers three network statistics, *arc*, *2-istar* and *3-istar*. Then the conditional log-odds of a regulation between two transcription factors is $-5.35\times$ change in the number of ties + 0.34 $\times$ change in the number of *2-istars* + (-.025)×change in the number of *3-istar*.

Table 1a Estimates and standard errors of ERGM parameters for TF-TF interactions: models 1-6 and parts of models 7-9

	Arc		2-istar		3-istar		4-istar	
	Est	Sterr	Est	Sterr	Est	Sterr	Est	Sterr
Model 1	-4.74	0.06						
Model 2			-2.01	0.0544				
Model 3					-0.9758	0.03389		
Model 4							-0.5274	0.0204
Model 5	-5.35	0.19	0.34	0.09	-0.025	0.017		
Model 6	-5.22	0.4	0.18	0.3	0.067	0.135	-0.0215	0.029
Model 7	-5.26	0.52	0.26	0.34	-0.018	0.254	0.029	0.181
Model 8	-5.26	1.16	0.26	1.47	-0.017	1.302	0.03	0.737
Model 9	-5.35	1.76	0.51	2.92	-0.48	3.637	0.505	2.958

Table 1b Estimates and standard errors of ERGM parameters for TF-TF interactions: rest of the network statistics of models 7-9.

	5-istar		9-istar		10-istar	
	Est	Sterr	Est	Sterr	Est	Sterr
Model 7	-0.014	0.057				
Model 8	-0.015	0.211	0.0004	0.022		
Model 9	-0.25	1.255	0.38	1.517	-0.69	2.665

In the absence of *2-istar* and *3-istar*thelog-odds of regulation is -5.35, in the presence of a *2-istar* but no *3-istar* the log-odds of regulation is -5.01, and in the presence of two *2-istar* and two *3-istar*, the log-odds is -4.72 and so on. The corresponding probabilities for regulation with multiple network statistics are calculated in Table 4 for both TF-TF and TF-Operon interaction networks.

3.2 TF- Operon interaction

We fit eleven ERGMs (Models 1-11 in Tables 2(a) and 2(b)) with varying network statistics to TF-Operon interaction network of *E. coli*. As in TF-TF interaction network we include *arc* and *k-istar* as the network statistics. Models 1-4 include individual network statistics, such as *arc* and *2-istar, 3-istar, 4-istar*. Models 5-11 include multiple network statistics. Tables 2(a) and 2(b) present results from the models for which convergence criteria are met. The estimates of the model parameters along with their standard errors for models 1-6 are presented in Table 2(a). Tables 2(a) and 2(b) jointly contain the parameter estimates and their standard errors for models 7-11. The parameter estimates from TF-Operon are unstable compared to those from TF-TF network, as the standard errors of the estimates are large. Nonetheless, the interpretation of the estimates is

similar as in TF-TF interaction model. The estimates of the parameters in individual network statistics models represent log-odds of regulation and those in the multiple network statistics models represent conditional log-odds. We present probabilities of regulation for individual network statistics models in Table 3 for both TF-TF and TF-Operon interaction networks. Table 4 presents the probabilities of regulation for multiple network statistics models for both TF-TF and TF-Operon interaction networks. From these results we see that the probability of regulation of one TF by another TF or of one operon by another operon is higher in models with higher *istar* network statistics.

Table 2a Estimates and standard errors of ERGM parameters for TF-Operon interactions: Models 1-6 and parts of models 7-11

	Arc		2-istar		3-istar		4-istar	
	Est	Sterr	Est	Sterr	Est	Sterr	Est	Sterr
Model 1	-6.237	0.0252						
Model 2			-2.88	0.029				
Model 3					-3.84	0.000009		
Model 4							-1.19	0.017
Model 5	-6.46	2789	0.01	3.94	0	0.0003		
Model 6	-6.24	0.1989	-0.17	0.154	0.17	0.0732	-0.04	0.017
Model 7	-5.86	0.00006	-0.08		0.09		-0.05	
Model 8	-5.79	1056.6	-0.83	895.5	1.06	577.48	-0.75	265.9
Model 9	-5.835	364.01	-0.95	533.1	1.05	638.657	-0.65	553.4
Model 10	-5.813	37.902	-1.05	76.68	1.55	140.589	-1.7	205.8
Model 11	-5.817	130.8	-1	392.2	1.47	1002	-1.65	1979

Table 2b Estimates and standard errors of ERGM parameters for TF-Operon interactions: rest of the network statistics of models 7-11.

	5-istar		6-istar		7-istar		8-istar		9-istar	
	Est	Sterr	Est	Sterr	Est	Sterr	Est	Sterr	Est	Sterr
Model 7	0.015									
Model 8	0.337	79.74	-0.07	12.03						
Model 9	0.227	327.25	-0.01	122.6	-0.02	22.767				
Model 10	1.636	230.35	-1.27	188	0.682	101.23	-0.187	27.59		
Model 11	1.668	2966	-1.38	3353	0.08	2745	-2.359	1480	8.68	403.3

Table 3 Estimated probabilities for regulation in individual network statistics models.

	TF-TF interaction		TF-Operon interaction	
	Log-odds	Prob.	Log-odds	Prob.
Model 1	-4.74	0.00866	-6.237	0.0019519
Model 2	-2	0.1192	-2.88	0.05345114
Model 3	-0.976	0.27369	-3.84	0.02104135
Model 4	-0.5271	0.37119	-1.19	0.23325894

Table 4 Estimated probabilities for regulation in multiple network statistics models.

	Model: arc +2-istar+3-istar; varying stars as (0,0), (0,1), (1,0) and (2,2)							
	None, None		None, One		One, None		Two, Two	
	Log-odds	Prob.	Log-odds	Prob.	Log-odds	Prob.	Log-odds	Prob.
TF-TF	-5.35	0.0047	-5.375	0.0046	-5.06	0.0063	-4.72	0.0088
TF-Operon	-6.46	0.0016	-6.46	0.0016	-6.45	0.0016	-6.44	0.0016

It is to be noted that a similar set of network statistics fit both TF-TF and TF-Operon interaction networks. However, model fitting suffers from the convergence problems while we include both incoming and outgoing star structures in the model along with other visible structures such as triangles.

3.3 Goodness of fit and Ostar models

It is important to investigate the performance of the exponential random graph models (ERGMs) in terms of how well these models fit the observed network. The parameters of an ERGM are estimated by the approximated maximum likelihood method. Although a maximum likelihood estimator of $\boldsymbol{\theta}$ may provide the best possible model among a particular class of models defined in equation (1) for a particular choice of a set of network statistics $\boldsymbol{z}(\boldsymbol{x})$, it does not necessarily provide a good model in a practical sense (Hunter et al., 2008). We present two goodness of fit statistics: Akaiki Information Criterion (AIC) and Bayesian Information Criterion (BIC) in Table 5. These are generated by the MCMC-MLE method of fitting of the ERGMs for TF-TF and TF-Operon interaction networks. AIC and BIC measure the relative goodness of fit of a statistical model.The smaller the values of these statistics the better the fit of model to observed data. Model 6 (*arc + 2- istar + 3- istar + 4- istar*) and model 9 (*arc + 2- istar + ... +5-istar + 9- istar + 10-istar*) fit the observed TF - TF interaction network well in terms of the AIC and BIC criteria as shown in Table 5.

Table 5 Goodness of fit statistics for TF-TF and TF-Operon interaction models.

	TF-TF interaction		TF-Operon interaction	
	AIC	BIC	AIC	BIC
Model 1	3025.1	3033.4	22771	22783
Model 2	4437.9	4446.2	29378	29390
Model 3	6041.4	6049.7	43430	43441
Model 4	7436.3	7444.7	48153	48165
Model 5	2966	2991	22704	22739
Model 6	2964	2997.3	22668	22714
Model 7	2968.3	3010	76718	76776
Model 8	2970.2	3020.1	22787	22857
Model 9	2968.4	3026.6	22697	22779
Model 10			22627	22720
Model 11			22775	22879
Model 12			81305	81421

In order to visualize how well these fits are to the observed the network, we present the simulated networks by using models 6 and 9 in Fig. 2. Comparing the observed TF-TF network in Fig. 1(a) with those in Fig. 2(a) and 2(b), we see the network structure is approximately similar. However, the clusters are not visible in the networks generated by the best two models, in terms of AIC and BIC, model 6 and model 9 respectively. Inclusion of other visible structures, that we were not able to do due to the convergence issues, may improve the closeness of the observed and simulated networks.

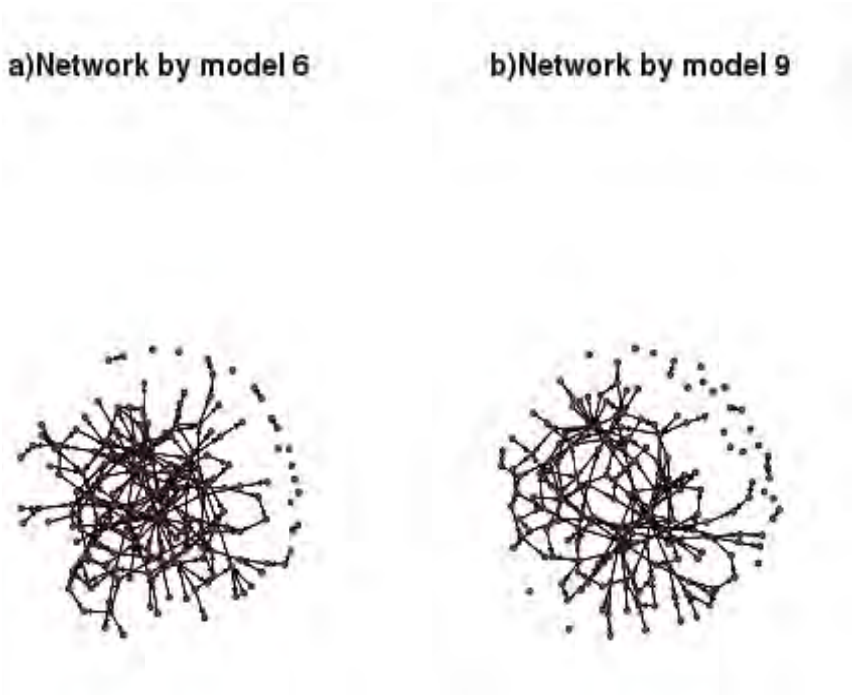

Fig. 2 Simulated networks by models 6 and 9 for TF-TF interaction in *E. coil*. a) Interaction simulated by network statistics an *arc*, *2-istar, 3-istar*, and *4-istar*. b) Interaction simulated by network statistics: *arc*, *2-istar*, …., *5-istar, 9-istar* and *10-istar*.

The goodness of fit is followed by the diagnostics plots for models 6 and 9 as shown in Fig. 3. The top plot in Fig. 3 reveals that with the network statistics *arc*, *2- istar, 3-istar* and *4-istar* the ERGM captures the geodesic distance distribution well. The geodesic distance distribution is the proportion of pairs of nodes (TFs in this case), whose shortest connecting path is of length *k*, where $k = 1, 2, ...$If the pairs of nodes are not connected in this manner we define $k = \infty$. The statistic geodesic distance is not in the ERGM. But, we want to compare if the value of this statistic observed in the original network mimic the distribution of values we get in simulated networks from our fitted model. For example, if the selected network statistics (*arc, 2-istar, 3-istar* and *4-istar*) fit the observed network well then the solid line in Fig. 3 will mimic the boxplots closely, which is the case in both plots in Fig. 3. The solid line represents the observed statistics for the TF-TF interaction network of *E. coli* and the boxplots summarize the statistics for the simulated networks using model 6. The bottom plot in Fig. 3 shows even better fit by the model 9.

Model 6 (*arc + 2-istar + 3- istar + 4-istar*) and model 10 *(arc + 2-istar + … + 8-istar*) fit the observed TF-Operon interaction network well in terms of the AIC and BIC criteria as shown in Table 5. The goodness of fit diagnostics plots presented in Fig. 5 shows reasonably good fit to the observed TF-Operon interaction network by model 6 and 10. For instance, the top plot in Fig. 5 demonstrates that with the network statistics *arc, 2-istar, 3-istar* and*4-istar* the ERGM captures the minimum geodesic distance distribution well. The solid line represents the observed statistics for the TF-Operon interaction network of *E. coli* and the boxplots summarize the statistics for the simulated networks using model 6. The bottom plot in Fig. 5 shows similar fit

by the model 10. In Fig. 4, we present the simulated networks by using models 6 and 10 to visualize the goodness of these fits.

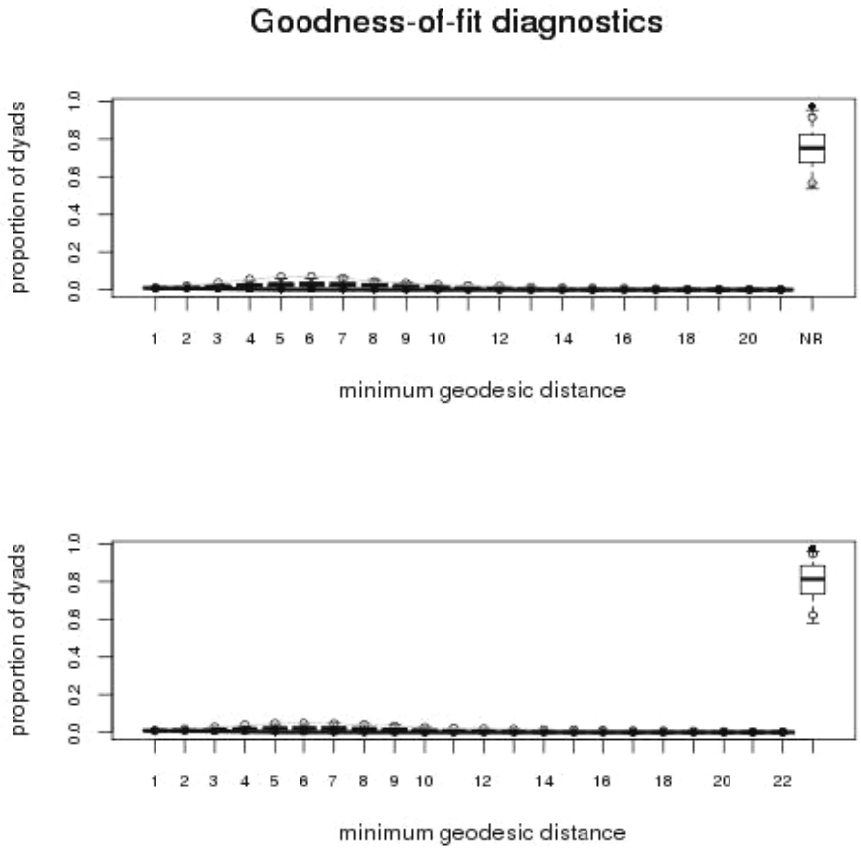

Fig. 3 Goodness of Fit (GOF) plots for models 6 and 9 for TF-TF interaction in *E. coli*. The top panel represents the GOF plot for model 6 and the bottom panel represents the GOF plot for model 9.

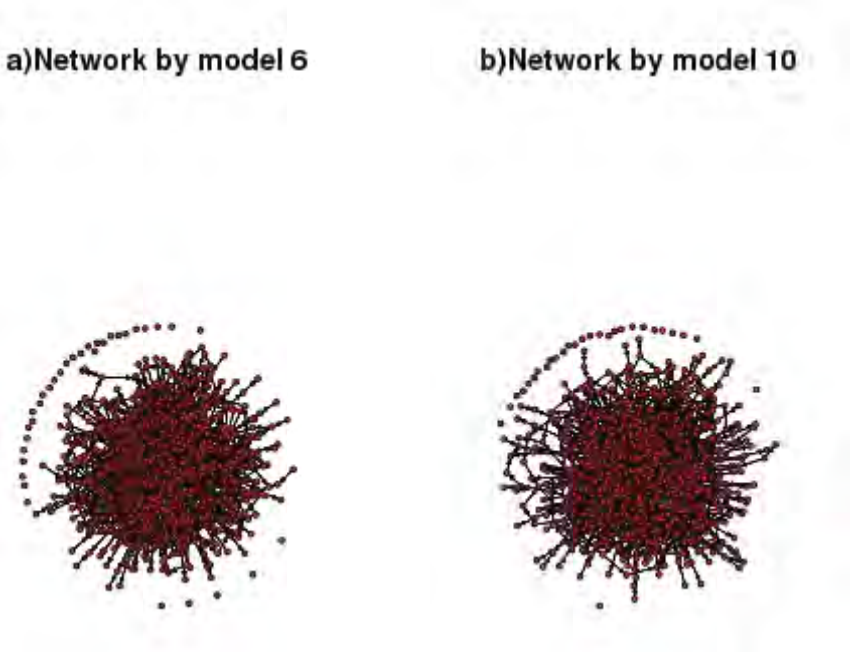

Fig. 4 Simulated network by models 6 and 10 for TF-Operon interaction in *E.coli*. a) Interaction simulated by network statistics: *arc, 2-istar, 3-istar,* and *4-istar*. b) Interaction simulated by network statistics: an *arc, 2-istar,…., 9-istar*.

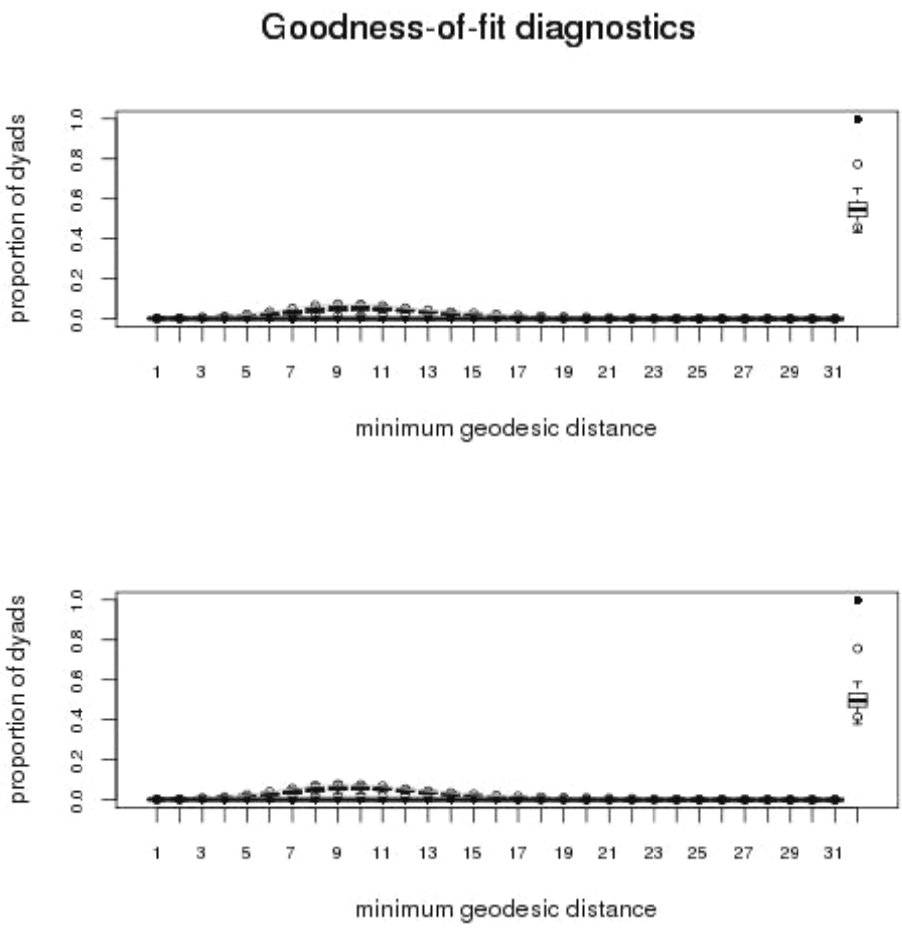

Fig. 5 Goodness of fit plots for models 6 and 10 for TF-Operon interaction in *E. coli.* The top panel represents the GOF plot for model 6 and the bottom panel represents the GOF plot for model 10.

All the models discussed so far include *arc* and *k-istar* network statistics for different *k* values as determined from the observed networks. To address the directed regulatory network of *E. coli*, we also include *k-ostar* network statistics to the ERGMs. In Fig. 6, we present simulated networks with *k-ostar* statistics for TF-TF and TF-Operon interaction networks. Fig. 6(a) is a simulated network for TF-TF interaction by the model that includes *arc, 2-ostar, 3-ostar, 4-ostar*, and *6-ostar* as the network statistics. Fig. 6(b) is a simulated network for TF-Operon interaction by the model with the network statistics *arc*, *2-ostar, …, 6-ostar*, and *10-ostar*. The goodness of fit diagnostics plots in Fig. 7 show that the ERGM with the network statistics *arc*, *2-ostar …, 6-ostar*, and *10-ostar* for TF-Operon interaction network (top panel) provide better fit to the observed TF-Operon interaction network compared to that of TF-TF interaction network (bottom panel).

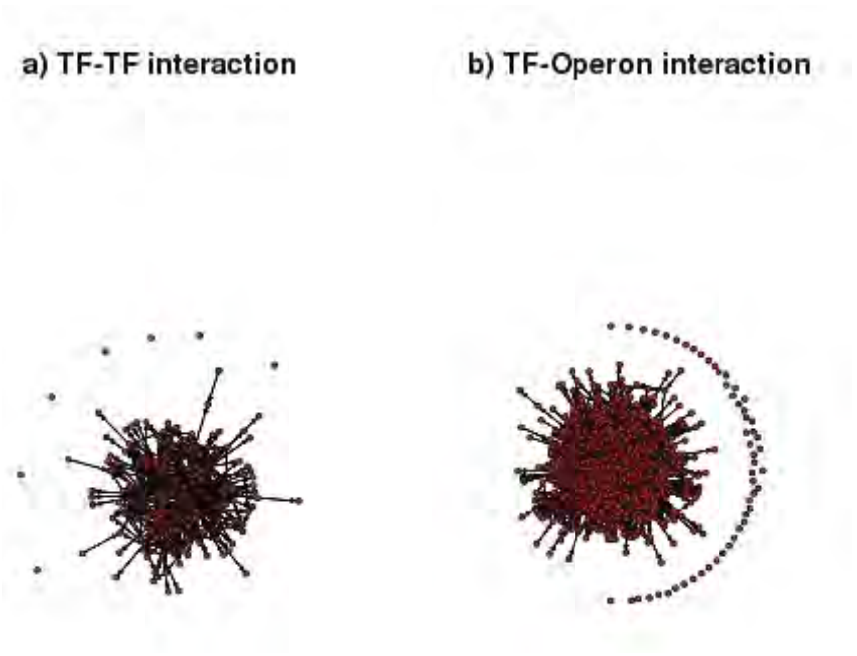

Fig. 6 Simulated networks with *k-istar* statistics for TF-TF and TF-Operon interaction in *E. coli.* a) Interaction simulated by network statistics: *arc, 2-ostar, 3-ostar, 4-ostar* and *6-ostar*. b) Interaction simulated by network statistics: an *arc, 2-ostar,…, 6-ostar* and *10-ostar*.

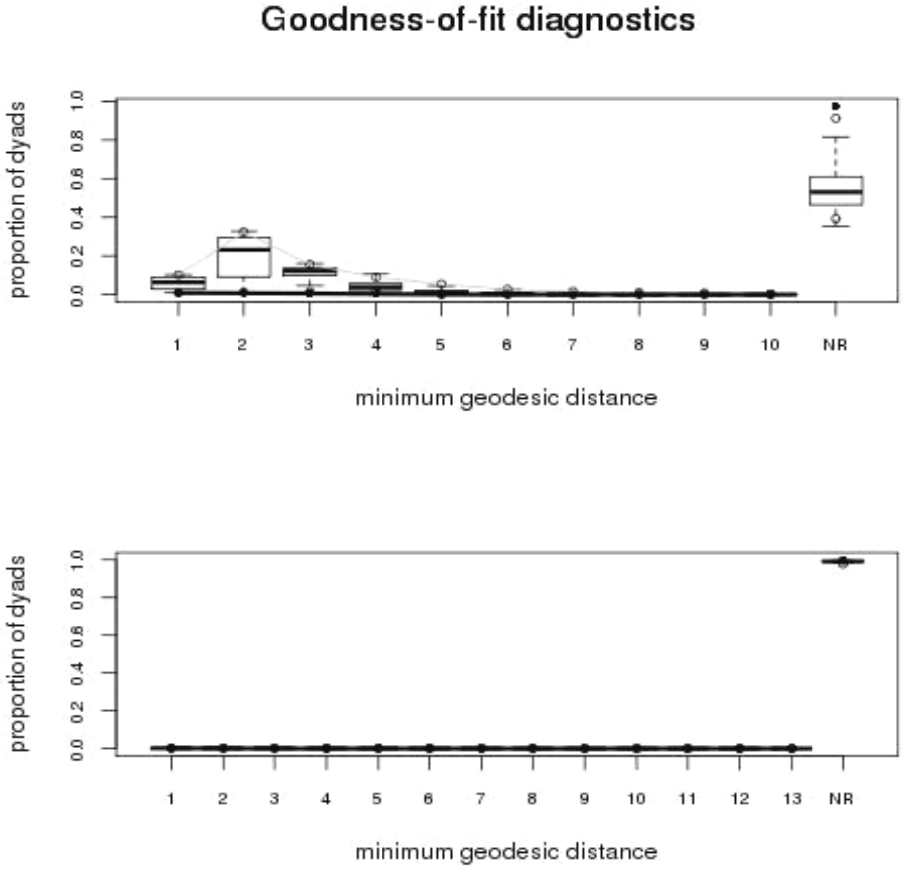

Fig. 7 Goodness of fit for TF-TF and TF-Operon interaction models with *k-ostar* statistics. The top panel represents the GOF plot for TF-TF interaction network and the bottom panel represents the GOF plot for TF-Operon interaction network.

4 Discussion

We explored a variety of Exponential Random Graph Models (ERGMs) to identify statistically significant network motifs or small sub-networks in transcription regulatory networks of *E. coli*. The regulatory networks were obtained from the *RegulonDB* database (Release 7.4, 2012). Since the process of regulation is directed, we extended Saul and Filkov (2007) principles of implementation of ERGMs to the transcription regulatory networks with network statistics appropriate for directed networks. A total of nine TF-TF ERGMs and eleven TF-Operon ERGMs were fitted with network statistics found in the observed network. The performance of each model was examined using goodness of fit statistics to determine how well the ERGMs fit the observed biological network interactions. The results show that ERGMs with multiple *k-istar* network statistics and an arc term fit both the TF-TF and TF-Operon interaction networks well. Thus for TF-TF and TF-Operon interaction networks, *arc* and *k-istar*, $2 \leq k \leq 10$, can be considered as statistically significant network motifs. Although *k-ostar*, $2 \leq k \leq 10$, statistics do well in fitting these interaction networks individually with the arc term, inclusion of both *k-istar* and *k-ostar* statistics leads to convergence problems in the estimation of model parameters. The ERGMs displayed a better fit for the observed TF-Operon networks with *k-ostar* statistics. This fit also aligns better with the biological context of how these networks function in the living organism.

Although ERGMs provide a simple and flexible principle of statistical modeling for regularity networks, there are several issues with computational algorithms in fitting the ERGMs. The first and foremost is the convergence problem in MCMC MLE method in the presence of multiple network statistics, such as *ostar*, *triangle*, *istar* and other higher order structure of subnetworks. We plan to address these issues of convergence problems and the possibility of extending the list of network motifs in our future work. Another possible expansion may include exploring if the similar set of network motifs can be used to describe the transcription regulatory networks in other prokaryotic as well as higher eukaryotic model organisms.

References

Balaji S, Babu MM, Aravind L. 2007. Interplay between network structures, regulatory modes and sensing mechanisms of transcription factors in the transcriptional regulatory network of *E. coli*. Journal of Molecular Biology, 372: 1108-1122

Barabasi AL, Oltvai ZN. 2004. Network biology: Understanding the cell's functional organization. Nature Reviews Genetics, 5: 101-113

Begum M, Bagga J, Blakey CA. 2012. Graphical modeling for high dimensional data, Journal of Modern Applied Statistical Methods, 11: 457-468

Besag J. 1975. Statistical Analysis of non-lattice data. Statistician, 24: 179-195

Costanzo M, Baryshnikova A, Bellay J, et al. 2010. The genetic landscape of the cell. Science, 327: 425-431

Erdˇos P, Rˇenyi A. 1960. On the evolution of random graphs. Publications of the Mathematical Institute of the Hungarian Academy of Science, 5: 17-61

Frank O, Strauss D. 1986.Markov graphs. Journal of the American Statistical Association, 81: 832-842

Friedman N. 2004. Inferring cellular networks using probabilistic graphical models. Science, 303: 799-805

Gama-Castro S, Salgado H, Peralta-Gil M, et al. 2011. RegulonDB version7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units). Nucleic Acids Research, 39: 98-105

Geyer CJ. 1991. Markov chain Monte Carlo maximum likelihood. In: Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface, 156-163

Geyer CJ, Thompson EA. 1992. Constrained Monte Carlo maximum likelihood for dependent data. Journal of

the Royal Statistical Society Series B, 54: 657-699

Goodreau SM. 2007. Advances in exponential random graph models applied to a large social network. Social Networks, 26: 231-248

Holland PW, Leinhardt S. 1981. An exponential family of probability distributions for directed graphs. Journal of the American Statistical Association, 76: 33-50

Hunter DR, Handcock MS, Butts CT, et al. 2008. ergm: A package to fit, simulate and diagnose exponential-family models for networks. Journal of Statistical Software, 24: 1-29

Martnez-Antonio A. 2011. *Escherichia coli* transcriptional regulatory network. Network Biology, 1: 21-33

Pr˘zulj N, Corneil DG, Jurisica I. 2004. Modeling interactome: scale-free or geometric? Bioinformatics, 20: 3508-3515

RegulonDB Release 7.4. March 2012. http://regulondb.ccg.unam.mx/index.jsp

Robins G, Pattison P, Kalish Y, Lusher D. 2007. An introduction to exponential random graph (p*) models for social networks. Social Networks, 29: 173-191

Saul ZM, Filkov V. 2007. Exploring biological network structure using exponential random graph models. Bioinformatics, 23: 2604-2611

Shen-Orr SS, Milo R, Mangan S, et al. 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. Nature Genetics, 31: 64-68

Snijders TAB, Pattison PE, Robins GL, Handcock MS. 2006. New specifications for exponential random graph models. Sociological Methodology, 36: 99-153

Strauss D, Ikeda M. 1990. Pseudo-likelihood estimation for social networks, Journal of the American Statistical Association, 85: 204-212

Thieffry D, Huerta AM, Perez-Rueda E, Collado-Vides J. 1998. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. Bioessays, 20: 433-440

Wasserman S, Pattison P. 1996. Logit models and logistic regressions for social networks: i. an introduction to markov graphs and p∗. Psychometrika, 6: 401-425

Zhang WJ. 2011. Constructing ecological interaction networks by correlation analysis: hints from community sampling. Network Biology, 1(2): 81-98

Zhang WJ. 2012. Computational Ecology: Graphs, Networks and Agent-based Modeling. World Scientific, Singapore