BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 19 2007, pages 2604–2611
doi:10.1093/bioinformatics/btm370

Systems biology

Exploring biological network structure using exponential random graph models

Zachary M. Saul* and Vladimir Filkov

Department of Computer Science, University of California, Davis, 1 Shields Avenue, Davis, CA 95616, USA

Received on November 27, 2006; revised on June 6, 2007; accepted on July 11, 2007

Advance Access publication July 20, 2007

Associate Editor: Trey Ideker

ABSTRACT

Motivation: The functioning of biological networks depends in large part on their complex underlying structure. When studying their systemic nature many modeling approaches focus on identifying simple, but prominent, structural components, as such components are easier to understand, and, once identified, can be used as building blocks to succinctly describe the network.

Results: In recent social network studies, *exponential random graph models* have been used extensively to model global social network structure as a function of their 'local features'. Starting from those studies, we describe the exponential random graph models and demonstrate their utility in modeling the architecture of biological networks as a function of the prominence of local features. We argue that the flexibility, in terms of the number of available local feature choices, and scalability, in terms of the network sizes, make this approach ideal for statistical modeling of biological networks. We illustrate the modeling on both genetic and metabolic networks and provide a novel way of classifying biological networks based on the prevalence of their local features.

Contact: saul@cs.ucdavis.edu

1 INTRODUCTION

The goal of much of systems biology is to understand the functioning of biological systems which, in large part, depends on their complex underlying structure. Summarizing a biological system into a network representation lets us study the complex structure via the interactions among its components and the simple recurring patterns, or features, which they form. Thus, when studying the systemic nature of biological networks many modeling approaches focus on simple, but prominent, structural features, as they are easier to understand than the global networks and, once identified, can be used as building blocks to succinctly describe the network.

One class of approaches, statistical network modeling, has recently gained visibility in the systems biology community, and a number of methods and models have been proposed as frameworks for investigating large biological networks (Barabási and Albert, 1999; Milo *et al.*, 2002; Pržulj *et al.*, 2004). In those studies, features like node degree distribution and small connected subgraphs (graphlets), have been demonstrated to capture well some facets of biological network structure, but tools that allow us to systematically study these and other local features, as well as the ways they collaborate to form the network architecture, are needed.

Outside of biology, statistical network modeling has a long history in the social and economic networks literature. For example, the concept of *network motifs*, small subgraphs that appear in a graph more often than expected due to chance (Milo *et al.*, 2002), were studied under the name *triad census* in 1970 (Holland and Leinhardt, 1970). Because biological networks are much larger than social networks, application of social network models has not historically been possible. However, recent advances both in understanding of the behavior of these models and in the availability of multiprocessor technology have made some application to biological networks feasible and should continue to make further application possible.

Scaling up from recent social network modeling efforts, this article discusses modeling biological networks using a family of statistical models called *exponential random graph (ERG) models*, also known as *p* models*. ERG models provide a tool to further our understanding of the network-scale interactions in biological systems. We are particularly interested in studying the way that a network's global structure (and function) depend on its local structure. How does one use an understanding of local notions such as protein–protein interaction, synthetic lethality or even node degree to understand the more global notion of the function of a network system? In this article, we make the following contributions:

- we introduce exponential random graph models for biological network exploration;
- we discuss the process of modeling biological networks using ERG models, including the choice of explanatory variables and fitting methods;
- we illustrate the modeling on genetic networks, metabolic networks and power-law random networks;
- we provide a novel way of classifying biological networks based on the prevalence of their local features.

Many models currently used for biological networks are descriptive, and simply specify a feature of a graph. For example, power-law networks (sometimes called scale-free) are

*To whom correspondence should be addressed.

described as networks with a node degree distribution governed by a power law (Barabaási and Albert, 1999). Other biological network models specify a procedure for creating networks. Erdös–Rényi random graphs are created by considering each pair of nodes in a given node set as a potential edge. For each potential edge, a fair *n*-sided die is cast, if the die comes up above a given threshold, the edge is included. Otherwise, it is not. An exponential random graph model takes a different, more general, approach.

An ERG model models the probability distribution function (pdf) for a given class of graphs. Given an observed graph and a set of explanatory variables on that graph the pdf is estimated. The pdf provides a concise summary of the class of graphs to which the observed graph belongs, i.e. the pdf can be used to calculate the probability that any given graph is drawn from the same distribution as the observed graph. The advantage of this approach is that it is very general and scalable as the architecture of the graph is represented by locally determined explanatory variables, and the choice of explanatory variables is quite flexible and can be easily amended.

The rest of this article is organized as follows. In Section 2, we discuss the theory of exponential random graph models and how to fit them. Section 3 contains the description of the network data sets that we used to evaluate the ERG modeling. In Section 4, we discuss the models that we fitted and goodness of fit measurement for ERG models. Section 5 covers the results of our explorations and our experiences with both maximum pseudo-likelihood estimation (MPLE) and Markov chain Monte Carlo maximum likelihood estimation (MCMC MLE). Finally, Section 6 summarizes our conclusions, presenting the benefits of exponential random graph models for biological networks.

2 EXPONENTIAL RANDOM GRAPH MODELS

We wish to model the probability distribution of networks explained by a given set of explanatory variables (or local patterns). Any function from the observed graph to the real numbers can act as an explanatory variable. As with all models, the variables to be included in an exponential random graph model are determined by the modeler based on what features of the graph under study are thought to be pertinent. An example, non-exhaustive, set of explanatory variables is given in Table 1.

Let $\boldsymbol{X}$ be a random variable representing the matrix form of a biological network from a particular class of networks. To model this class of networks, we need to estimate the probability distribution function (PDF) for $\boldsymbol{X}$, $P(\boldsymbol{X}=\boldsymbol{x})$. That is, if this PDF were known, we would know the probability that an observed graph, $\boldsymbol{x}$, is of the type of graph that our random variable $\boldsymbol{X}$ represents. Unfortunately, the probability distribution function of $\boldsymbol{X}$ is unknown. Therefore, we cannot directly calculate $P(\boldsymbol{X}=\boldsymbol{x})$.

However, we can model $P(\boldsymbol{X}=\boldsymbol{x})$ with a log linear model. To do so, we need first to identify a vector of explanatory variables, $\boldsymbol{z}(\boldsymbol{x})=(z_1(\boldsymbol{x}), z_2(\boldsymbol{x}), \ldots, z_r(\boldsymbol{x}))$. These explanatory variables can be any graph statistic (e.g. number of triangle subgraphs) or any node statistic (e.g. molecular weight of molecule), but each explanatory variable should be a function of the observed data. To model the pdf of $\boldsymbol{X}$, we postulate that there exists $\boldsymbol{\theta}=(\theta_1, \theta_2, \ldots, \theta_r)$ such that:

$$\log(P(\boldsymbol{X}=x)) \propto \theta_1 z_1(\boldsymbol{x})+\theta_2 z_2(\boldsymbol{x})+, \ldots, +\theta_r \boldsymbol{z_r}(\boldsymbol{x}) \quad (1)$$

$$\propto \boldsymbol{\theta}^{\boldsymbol{T}}\boldsymbol{z}(\boldsymbol{x}) \quad (2)$$

Exponentiating both sides and adding a normalizing constant, $\kappa(\boldsymbol{\theta})$, to assure that the probabilities will sum to one, we get the following model:

$$P(\boldsymbol{X}=\boldsymbol{x})=\frac{e^{\boldsymbol{\theta}^T}\boldsymbol{z}(\boldsymbol{x})}{\kappa(\boldsymbol{\theta})} \quad (3)$$

This model is the standard log linear probability model that is used in a wide range of fields from the social sciences to biology (Infante-Rivard *et al.*, 2006; Kaplan, 2004).

Depending on which two of $\boldsymbol{x}, \boldsymbol{\theta}$ and $\boldsymbol{z}(\boldsymbol{x})$ are known, the third can be estimated or solved for. In practice, $\boldsymbol{z}(\boldsymbol{x})$ is a starting point and we are typically interested in the other two quantities. For a given $\boldsymbol{\theta}$ and statistics $\boldsymbol{z}(\boldsymbol{x})$, one can simulate networks drawn from the probability distribution $P(\boldsymbol{X}=\boldsymbol{x})$. The values $\boldsymbol{\theta}$ can be thought of as weights for the various variable values with stronger weights indicating that a variable more strongly determines the properties of the network distribution.

On the other hand, having observed a data matrix $\boldsymbol{x}$ and explanatory variables $\boldsymbol{z}(\boldsymbol{x})$, one is interested in fitting, or estimating, the model parameters $\boldsymbol{\theta}$ to the observed data, thereby characterizing the network $\boldsymbol{x}$ in terms of the relative importance of the explanatory variables in determining the response variable (Anderson *et al.*, 1999).

In this article we are interested in the latter, the model, or parameter fitting part, i.e. we would like to estimate $\boldsymbol{\theta}$, the vector of model parameters. Standard maximum likelihood estimation of the parameters are difficult to apply in this case, because the function for the normalizing constant $\kappa(\boldsymbol{\theta})$ is not known a priori. However, calculating $\kappa(\boldsymbol{\theta})$ can be avoided by approximating the probabilities based on differences in the $\boldsymbol{z}(\boldsymbol{x})$ statistics. There are two methods commonly used in the statistics and social networks communities to estimate the maximum likelihood fit to exponential random graph models, Markov chain Monte Carlo maximum likelihood estimation and maximum pseudo-likelihood estimation. They can also be used for network simulation. We describe them briefly next, and note their respective strengths and weaknesses and the types of networks to which they can be applied.

2.1 MCMC MLE fitting

Markov chain Monte Carlo maximum likelihood estimation (MCMC MLE) refers to a family of methods based on the Newton–Raphson algorithm for maximum likelihood estimation. Let $\boldsymbol{\mu}(\boldsymbol{\theta})$ be the vector of expected values of the explanatory variables and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ be the covariance matrix under a given parameter vector $\boldsymbol{\theta}$. Then, the standard Newton–Rhapson algorithm with iteration step $\hat{\boldsymbol{\theta}}^{(n+1)} = \hat{\boldsymbol{\theta}}^{(n)} - \left(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}^{(n)})\right)^{-1}\left(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}^{(n)}) - \boldsymbol{z}(\boldsymbol{x})\right)$ would normally be used to find the maximum likelihood estimate of $\boldsymbol{\theta}$. However, this is not feasible for exponential random graph models because $\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ are not known. MCMC likelihood estimation gets around this problem by estimating $\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. It does so by simulating the distribution of graphs given $\boldsymbol{\theta}$ and estimating $\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ based on a sample from the simulated distribution. The simulation is typically achieved through a standard MCMC process using either Gibbs sampling or the Metropolis–Hastings algorithm (Snijders, 2002).

There are a number of software packages available to fit exponential random graph models using MCMC MLE. These include the **statnet** package for the R statistical computing environment (Handcock *et al.*, 2003) and the SIENA Software (Snijders *et al.*, 2006).

However, ERGM fitting using MCMC MLE methods is only lately becoming possible, because of both computational and degeneracy problems. Computationally, networks with approximately 500 nodes are the largest that can usually be fitted on current hardware, although sparse thousand node networks have recently been fitted successfully (Goodreau, 2007). Biological networks typically have thousands (regulatory networks) if not tens of thousands of nodes (protein–protein interaction networks) and many of them will soon come within reach of the current technology, although for multiple model fitting of such networks parallel machines are recommended.

More fundamentally, though, some connected sets of fitted parameters correspond to degenerate networks (e.g. graphs with almost no edges or nearly complete graphs) and the MCMC MLE methods exhibit convergence problems when encountering such neighborhoods, yielding networks that do not resemble the original data. Important recent work by Snijders *et al.* (2006) has dealt with this problem of degeneracy and has suggested practical approaches for avoiding it. In a nutshell, whenever strong transitive relationships are suspected in a network, Snijders *et al.* suggest to use at least one of two aggregate variables: the geometrically weighted degree and the geometrically edgewise shared partners (presented in Table 1). These variables can be used alone, or together with their simpler counterparts, the node degrees and number of triangles in the network and have been used successfully to fit even large networks (Goodreau, 2007).

2.2 Maximum pseudo-likelihood estimation (MPLE)

The *logit p** model is a model related to p* (ERGM) in such a way that the maximum likelihood set of parameters of a *logit p** model is an estimate of the maximum likelihood parameters of the corresponding p* (ERG) model. The *logit p** model has no normalizing constant, $\kappa(\boldsymbol{\theta})$, thus allowing normal maximum likelihood estimation to be used to fit the *logit p** model. Estimation of the parameters of a p* network by estimating the parameters of the corresponding *logit p** model is called maximum pseudo-likelihood estimation. The following description is a synthesis of the descriptions presented in the social networks literature (Anderson *et al.*, 1999; Wasserman and Pattison, 1996).

Table 1. This table shows example explanatory variables. As can be seen from the wide variety of variables, ERG models are flexible

Variable	Description
k-Star	The number of nodes in the network with exactly k adjacent edges with unconnected end points.
Triangle	The number of 3-cycles in the network.
k-Cycle	The number of k-cycles in the network.
k-Degree	The number of nodes in the graph with degree k.
k-Edgewise shared partners	The number of edges in the network that have exactly *k* shared partners.
Geometrically weighted degree	The weighted sum of the counts of each degree, weighted by the geometric sequence, $(1-e^{-\alpha})^i$ where α is a decay parameter.
Geometrically edgewise shared partners	The weighted sum of the number of edges in the network that have exactly *i* shared partners weighted by the geometric sequence, $(1-e^{-\alpha})^i$ where α is a decay parameter.
Maximum geodesic	The length of the longest of the shortest paths between each pair of nodes.
Edge count	The number of edges in the graph.
Node count	The number of nodes in the graph.
Isolates	The number of nodes in the network with no neighbors.

Let $\boldsymbol{x}_{ij}^{+}$ refer to the matrix representation of a graph identical to the observed graph $\boldsymbol{x}$ except that in $\boldsymbol{x}_{ij}$ the edge from *i* to *j* is guaranteed to exist; let $\boldsymbol{x}_{ij}^{-}$ refer to the matrix representation of a graph identical to $\boldsymbol{x}$ except that the edge from *i* to *j* is guaranteed not to exist, and let $\boldsymbol{x}_{ij}^{c}$ represent the matrix identical to that of the observed graph with the exception that there is no entry at position (*i*,*j*) in $\boldsymbol{x}_{ij}^{c}$. This single piece of information is missing from $\boldsymbol{x}_{ij}^{c}$.

A logit is the log odds of a binary random variable. That is, for some binary random variable, Y, the logit is $\log(P(Y=1)/P(Y=0))$. The random variable, $\boldsymbol{X}$, in the p* model is not binary, but we can get around this limitation if we consider the set of binary random variables $\{X_{ij}\}$, where $X_{ij}=1$ indicates that there is an edge between nodes *i* and *j*. If we model the conditional distributions, $P(X_{ij}=1|\boldsymbol{x}_{ij}^{c})$, $P(\boldsymbol{X}=\boldsymbol{x})$ can be calculated by the Hammersly–Clifford theorem (Wasserman and Pattison, 1996).

Now, note the following:

$$P(X_{ij}=1|\boldsymbol{x}_{ij}^{c}) = \frac{P(\boldsymbol{X}=\boldsymbol{x}_{ij}^{+})}{P(\boldsymbol{X}=\boldsymbol{x}_{ij}^{+})+P(\boldsymbol{X}=\boldsymbol{x}_{ij}^{-})} \quad (4)$$

Using this probability, we can write the expression for the odds ratio of the graph with the edge linking *i* and *j* to the graph without this edge.

$$\frac{P(X_{ij}=1|\boldsymbol{x}_{ij}^{c})}{P(X_{ij}=0|\boldsymbol{x}_{ij}^{c})} = \frac{P(\boldsymbol{X}=\boldsymbol{x}_{ij}^{+})}{P(\boldsymbol{X}=\boldsymbol{x}_{ij}^{-})} \quad (5)$$

Next, substituting Equation(3), we get the following:

$$\frac{P(\boldsymbol{X}_{ij}=1|\boldsymbol{x}_{ij}^{c})}{P(\boldsymbol{X}_{ij}=0|\boldsymbol{x}_{ij}^{c})}=\frac{e^{\boldsymbol{\theta}^{T}\boldsymbol{z}(\boldsymbol{x}_{ij}^{+})}}{e^{\boldsymbol{\theta}^{T}\boldsymbol{z}(\boldsymbol{x}_{ij}^{-})}} \tag{6}$$

$$= e^{\boldsymbol{\theta}^{T}[\boldsymbol{z}(\boldsymbol{x}_{ij}^{+})-\boldsymbol{z}(\boldsymbol{x}_{ij}^{-})]} \tag{7}$$

Next, take the log of both sides to get the log odds ratio (logit) for the edge (i,j), which we will call ω_{ij}:

$$\omega_{ij}=\log\left[\frac{P(\boldsymbol{X}_{ij}=1|\boldsymbol{x}_{ij}^{c})}{P(\boldsymbol{X}_{ij}=0|\boldsymbol{x}_{ij}^{c})}\right] \tag{8}$$

$$=\boldsymbol{\theta}^{T}[\boldsymbol{z}(\boldsymbol{x}_{ij}^{+})-\boldsymbol{z}(\boldsymbol{x}_{ij}^{-})] \tag{9}$$

Defining $\boldsymbol{\delta}(i,j)=[\boldsymbol{z}(\boldsymbol{x}_{ij}^{+})-\boldsymbol{z}(\boldsymbol{x}_{ij}^{-})]$, gives a succinct statement of *logit p**:

$$\omega_{ij}=\boldsymbol{\theta}^{T}\boldsymbol{\delta}(i,j) \tag{10}$$

Thus, the logit for each pair of nodes (i,j) is the product of the model parameters and the vector of network statistics that arises when variable $\boldsymbol{X}_{ij}$ changes from 1 to 0. This last vector, $\boldsymbol{\delta}(i,j)$ is called the *difference statistics* vector.

Fitting using MPLE is computationally a much simpler task than MCMC MLE as it reduces to solving a logistic regression. In practice, 1000 node sparse networks can fit fairly quickly on modern hardware. However, although both MCMC MLE and MPLE are an approximative methods for estimating the model parameters, there are indications that in practice MPLE may do worse than MCMC MLE (Geyer and Thompson, 1992). This is especially the case for networks which have strong dyadic dependence (i.e. edges are dependent on other edges given the rest of the graph). Note that the MPLE estimate corresponds to the exact solution when no dyadic dependences exist in the graph. In practice, MPLE may be a good approximation when the dyadic dependence is weak.

3 DATA

Our goal in this article is to illustrate the fitting of different biological networks using ERG models and to note the differences between the best fitted models and parameters for each of several networks. Thereby, we can learn which variables characterize which classes of networks and possibly identify groups of networks with very similar fits and, hence, similar architectures. To that end, we evaluated exponential random graph modeling using biological networks of different origin, size and types. First, we studied two transcriptional regulation networks. The first is an updated *Escherichia coli* network (Shen Orr *et al.*, 2002) based on the well-known network available from RegulonDB (Salgado *et al.*, 2001). In this network, each node represents an operon, and an edge from one operon to another indicates that the first operon encodes the transcription factor that regulates the second. This network contains 418 nodes and 578 edges. The second transcriptional regulation that we studied is the network of TF-DNA binding for yeast (Lee *et al.*, 2002), containing 106 transcription factors and 6270 genes and 1842 edges. We used three nested versions of this network created with different edge inclusion thresholds corresponding to binding *P*-values of 0.01, 0.001 and 0.0001.

Second, we considered a collection of metabolic networks for 43 organisms introduced in an earlier work (Jeong *et al.*, 2000), coming from the WIT database. This database contains metabolic pathways that were predicted using the sequenced genomes of the several organisms. The nodes in these networks are enzymes, substrates and intermediate complexes, and the edges indicate an interaction. Of the 43 organisms, 6 are archea, 5 eukaryotes and 32 bacteria. The sizes of the networks vary from 595 nodes and 1354 edges to 2982 nodes and 7300 edges. This group of biological networks are particularly important to our task at hand of classifying networks based on structural similarity between them because of two reasons: (1) all 43 networks are fairly similar to each other as they all contain basic metabolic pathways which are fairly conserved along the evolutionary tree. Thus, we expect that the same choice of variables would provide good fits for all of them; and (2) because these networks summarize relationships between proteins and metabolites and vice-versa, they are bipartite graphs. Hence, they have no trivial transitive relationships and are likely to have low dyadic dependence, which makes them well suited for the MPL estimation method.

Finally, we generated two random power-law networks. We did so using the preferential attachment model (Barabási and Albert, 1999). This model grows a network from one edge, adding new nodes one at a time, attaching each to an existing node with probability proportional to the degree of that node. Both networks generated had 1000 nodes and approximately 3900 edges. In total, we used 49 large networks and treated all of them as undirected graphs.

4 METHODS

We fitted a number of different models to the different networks described above and investigated the relative importance of many different explanatory variables in these networks. In addition, we studied the relative merits of the two available methods to estimate the fit of an exponential random graph model.

4.1 Fitting ERGMs with MCMC MLE and MPLE

The variables, $\boldsymbol{z}(\boldsymbol{x})$, used in an ERG model can be any function from the observed graph $\boldsymbol{x}$ to the real numbers. However, as can be seen from Equation (10), the variables actually used are $\boldsymbol{\delta}(i,j)$, the vector of difference statistics. These difference statistics are the differences between the value of $\boldsymbol{z}$ when (i,j) is forced to be present in the graph and the value of $\boldsymbol{z}$ when (i,j) is forced to be absent.

To fit the ERG models with both MCMC MLE AND MPLE methods, we used the **statnet** package (Handcock *et al.*, 2003) for the R statistical computing environment. We fit several models to networks of different types and sizes using both MCMC MLE and MPLE. Although we were able to fit many of our networks using both methods, we also found that for a sizable fraction of the biological networks in our study it was computationally intractable to fit models using MCMC MLE. Based on the networks that we were able to fit (a selection of which is given in Section 5), it seems that MPL fitting is often an appropriate substitute for MCMC MLE in fitting biological networks, possibly because of their low dyadic dependence.

4.2 Explanatory variables

We illustrate the process of choosing the explanatory variables on the *E.coli* regulatory network from RegulonDB. The software **statnet** supports a number of explanatory variables for undirected networks. These include Edges, k-Star, k-Degree, k-Cycle, k-ESP, GWDegree, GWESP and Isolates. For detailed descriptions of each of these variables, see Table 1. The GWDegree and GWESP variables are used to address the degeneracy issues mentioned before in Section 2.1.

To determine the variables to use in our model, we used an iterative exploratory technique of progressively increasing the model complexity. First, we fitted single variable models consisting of each of the possible variables. Then, using the Akaike Information Criterion (AIC), the goodness of fit (gof) technique outlined below and the built-in goodness of fit method in **statnet**, we selected several pairs of variables to form models. We repeated this technique for models with 3 – 8 variables. Our results are discussed below.

4.3 Goodness of fit

Evaluating the importance of an explanatory variable or group of variables to the fit of an ERG model can be achieved by fitting the model both with and without the variables in question and comparing the goodness of fit of the two models.

By goodness of fit, we simply mean how well the model fits. To estimate how well a model fits, **statnet** has a function that simulates a sample of networks using the fitted model and, then, compares the values of several explanatory variables in the original network to the values of the same variables in the sampled networks. For the purposes of our study, we created an additional method to estimate the goodness of a fit. We compared the overlap, in terms of edges, between the observed network and our own sample of networks simulated using the fitted model.

In particular, to evaluate the importance of each explanatory variable, we fitted several nested models as discussed in Section 5.2. We used the fitted models to sample 30 networks (using **statnet**'s function simulate.ergm). Then, for every pair of nodes in the original graph, we counted the number of times that pair of nodes was adjacent in our sample of 30 graphs. Using these observed adjacency frequencies, we estimated probability of an edge ($p_{i,j}$) between each pair of nodes i and j (i.e. we normalized the observed frequencies).

We, then, created 99 nested networks, one for each $p \in \{0.99, 0.98, \ldots, 0.01\}$. In each network, we allowed edge (i,j) iff $p_{i,j} > p$. Then, interpreting each pair of nodes in each network as a prediction as to the presence or absence of an edge in the original biological network, we calculated the false positive rate and the false negative rate of these predictions. We plotted these 99 pairs of false positive and false negative rates as a receiver operating characteristic (ROC) curve (Fig. 1).

5 RESULTS AND DISCUSSION

Using an 80-processor cluster, we performed the iterative model fitting procedure described above, fitting every model to the RegulonDB *E.coli* network using MCMC MLE. The fits of a few sample models of various complexity are shown in Table 2.

We used the RegulonDB network for the iterative fitting procedure because it was one of the smallest that we considered and fitting our large, biological networks with MCMC MLE was very difficult.

Comparing the rows of Table 2 pairwise can be illustrative. For example, by comparing M1 and M4, one can see the effect of removing the k-Deg variables from the full model. This removal strongly changes the other fitted parameter values.

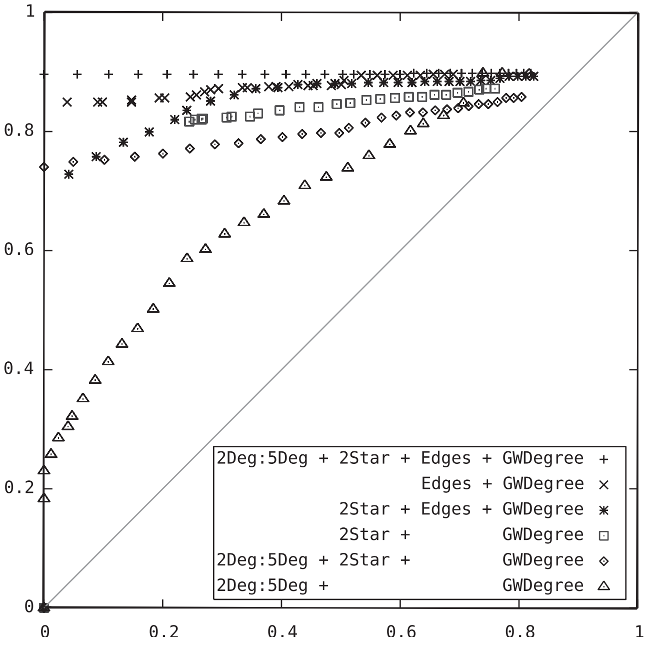

Fig. 1. We used the fitted parameters to sample several networks and then using the sampled networks as an edge predictor of the original network, and we calculated a false positive rate and a false negative rate. This plot shows the false positive rate versus the false negative rate (an ROC plot). The models that included gwdegree and 2star or edges all performed well.

Table 2. The parameter values for several sample models fitted to the RegulonDB *E.coli* network (using MCMC MLE and MPLE)

	Fitted with MCMC MLE								Fitted with MPLE							
	2-Deg	3-Deg	4-Deg	5-Deg	2-Star	Edges	GWDeg	AIC	2-Deg	3-Deg	4-Deg	5-Deg	2-Star	Edges	GWDeg	AIC
M1	−1.71	−2.22	−2.42	−3.05	0.011	10.6	−3.78	4976	−1.44	−2.08	−2.33	−3.03	0.011	10.5	−3.77	5010
M2	−1.49	−2.18	−2.48	−3.15	0.038	–	−1.08	4981	−1.51	−2.18	−2.44	−3.15	0.038	–	−1.07	5020
M3	−1.63	−2.53	−2.98	−3.53	–	–	−0.928	5262	−1.63	−2.41	−2.72	−3.53	–	–	−0.911	5300
M4	–	–	–	–	−0.079	48.6	−13.8	5410	–	–	–	–	−0.079	48.6	−13.8	5450
M5	–	–	–	–	–	19.2	−6.32	5482	–	–	–	–	–	19.2	−6.32	5520
M6	–	–	–	–	0.047	–	−1.47	5581	–	–	–	–	0.047	–	−1.45	5620
M7	–	–	–	–	–	–	−1.35	6078	–	–	–	–	–	–	−1.33	6120

Table 3. The results of fitting three models on several larger networks

Name	Nodes	Edges	Triangles	2-Star	4-Cycle	GWDegree	GWESP	AIC
Scale free 1	1000	3933	0.026	0.027	−0.022	−1.97	–	43979
Scale free 2	1000	3875	0.082	0.026	−0.031	−1.95	–	43813
Sacharomyces cerevisiae (ChIP-chip)	6270[a]	1842	136	−1.13	1.14	–	−60.8	189702
Aeropyrum pernix (WIT)	595	1354	−14.7	−0.005	0.120	−1.68	–	12179

[a]This network has 106 transcription factors and 6270 genes.

The increased weighting of GWDegree can account for the removal of the k-Deg variables because GWDegree is a summary statistic for all of the k-Deg variables. Next, compare M2 and M3; this represents the removal of the 2-Star variable from the model. In this case, the other model parameter values do not significantly change, indicating the independence of 2-Star from all the other parameters.

After we had explored several models using the RegulonDB network, we attempted to fit a few of them to other, larger networks using MCMC MLE. A few larger networks that we were able to fit can be seen in Table 3, where only the best fitting models are reported.

Comparing the rows in Table 3, shows that the parameter choice will need to vary from network to network. For example, the value for Triangles varies wildly across all of the networks. In fact, the two networks with large, anomalous values for the Triangles variable are both not likely to contain many triangles because of their structure. This example shows the importance of picking the correct parameters when attempting to fit a network.

Although some authors have reported that fitting large networks with MCMC MLE can lead to diverging parameter estimates (Goodreau, 2007), we found this not to be the case with our biological networks. For every network and model that we attempted to fit, the parameter estimates either converged or appeared to be converging to finite values.

Additionally, as mentioned above, the GWDegree variable was included in most of our models because it has been shown in earlier work to reduce degeneracy in the model fits. However, including this variable tended to slow computation and may have prevented fitting from finishing for some models on larger networks.

5.1 Comparison of MCMC MLE and MPLE

We were able to fit several of the large biological networks in our data set using MCMC MLE, but we were not able to fit all of them. At this time, MCMC MLE is probably not feasible for many large networks such as seen in bioinformatics. Each iteration in the MLE algorithm requires the simulation of a sample population of networks using MCMC, but this process is quite time and space intensive for large networks. Simply because large networks have more edges, the mixing time for Gibbs sampling is longer. So, each network in the population takes much longer to simulate, and each iteration in the MLE requires from hundreds to thousands of networks be simulated.

Although MCMC MLE provides superior fits over maximum pseudo-likelihood estimation, we would like to know if we can use MPLE as a substitute fitting method for networks where MCMC MLE fails. To answer this question, we compared the model fits to the RegulonDB network, where we were able to fit using both MCMC MLE and MPLE.

MCMC MLE and MPLE fits can be compared in Table 2. The results show that for these models, the parameter values are quite similar. Further, for all of the models and networks given in Table 3, the parameter values are also similar for the MCMC MLE and MPLE fits. Unfortunately, we were unable to include the MPLE parameter values here due to space constraints. In addition, previous researchers have shown that MCMC MLE is superior to MPLE when there is strong dyadic dependence, but there is not necessarily strong dyadic dependence in biological networks. For example, in the WIT networks, there are no triangle graphlets. So, in cases where strong dyadic dependence is not suspected, MPLE fitting can serve as a computationally feasible substitute for MCMC MLE fitting.

5.2 Goodness of fits

After fitting the nested sequence of models, we performed the goodness of fit test described in Section4.3. The results can be seen in Figure 1. We found that with biological networks, we achieved acceptable fits as long as we included the network statistic GWDegree. This is in line with similar results in social networks and biology. The degree distribution of the network appears to be a contributing factor in the overall structure. However, as can be seen in Figure 1, the various k-Degree variables are not, in themselves, enough to give the best fit.

5.3 Classifying Networks via Their Topological Profiles

Associating network topology to biological function is a major goal in systems biology. Recently, researchers have reported success characterizing networks by using network motif profiles to classify various evolved and designed networks (Milo *et al.*, 2004). Others have similarly used local network topology to infer the likeliest mechanisms of the networks' evolution or design (Middendorf *et al.*, 2005).

To demonstrate the utility of ERGMs for biological network modeling, we sought to classify all 49 networks in our data set using each network's fitted parameters as its profile. We chose

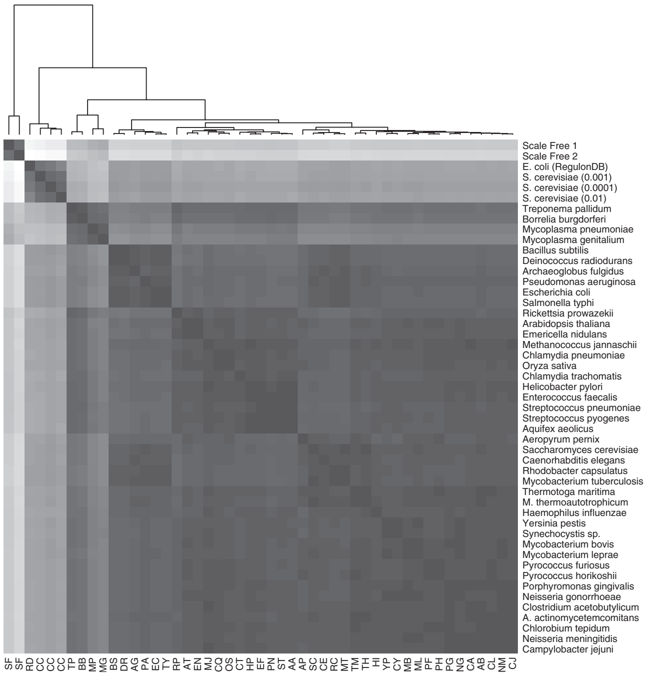

Fig. 2. Heatmap of the Euclidean distance matrix of the networks' profiles.

to use the model *model* = *4Deg* + *5Deg* + *2Star* + *GWDegree* for each of the networks and fitted the model using maximum pseudo-likelihood estimation. As mentioned before, using MPLE instead of MCMC MLE here was acceptable because the WIT networks were found to have low transitivity and likely low dyadic dependence.

We calculated the Euclidean distance matrix between the parameter profiles of all the networks and used it to cluster the set of networks using complete-linkage hierarchical clustering. The symmetric heatmap of the distance matrix is given in Figure 2, with the cluster dendrogram attached on top. The apparently strong cluster structure within it indicates that ERGM parameter profiles can be used as a means to classify these networks.

Looking at the whole heatmap, the clusters segregated strikingly well the networks by biological type or experimental origin. Namely, the two scale-free random nets were clearly separated from all the other networks. Similarly, the *E.coli* gene network clustered together with the three TF-DNA gene networks from ChIP-chip studies, while all four together were well isolated from the other clusters. Of note here is that the gene networks exhibit different architecture from the two scale-free networks. In addition all 43 WIT metabolic networks clustered closer together than to either the scale-free or gene regulation networks.

Next, it is very interesting to examine the clusters of only the 43 metabolic networks. Here, we give some very general taxonomic and functional observations which were derived by consulting bacterial resources on the Internet, in particular EBI's Karyn's Genomes website.[1] In the following, we use the two letter abbreviations for the 43 organisms from Figure 2, given on the *X* axis. First, of note is the fact that the five eukaryotic genomes AT, EN, OS, SC and CE clustered near each other, because of their very similar parameter profiles as evidenced by the dendrogram, while the six archea organisms were spread around in different clusters. Second, one of the clusters was comprised almost exclusively (13 out of 16 organisms) of non-motile organisms (the cluster at the lower right corner of the heatmap), while another cluster (fourth from the top along the diagonal) all but one of the organisms (DR) were motile. Finally, the fifth cluster from the top consists of only eukaryotes and anaerobic bacteria.

Although appealing, it is impossible to speculate further on the reasons for, or the meaning behind the functional features or phenotypes of organisms clustered together without a more detailed associative study of the parameter profiles and organisms features.

These results make it plausible that ERGM parameter profiles can be used for structural classification of different biological networks. They also raise the possibility that functional features of organisms that have a systemic level network manifestation are more likely to come up as class differentiators in such studies. Hence, classification using parameter profiles derived from ERGM models can be potentially used to identify system level functional features in biological networks.

6 CONCLUSION

In this article we have introduced exponential random graph models, a family of network models that have previously been used to study social networks, and we have demonstrated their utility in modeling biological networks. In addition, we have argued that fitting exponential random graph models to biological networks can best be achieved using pseudo-likelihood maximization. We demonstrated that topological profiles derived from biological networks by fitting ERG models can be used to classify organisms in clearly separate biological and functional groups.

There are a number of reasons that exponential random graph models should be considered for use in biology. First, the statistics underlying ERG models are more principled than seen in the previous network modeling efforts in biology. Previous efforts in biology have relied on comparing networks to simulated random networks (which depend heavily on the random model) or investigating a single network feature such as degree distribution (Albert and Barabási, 2000; Barabási and Albert, 1999; Milo *et al.*, 2002, 2003; Yeger-Lotem *et al.*, 2004). Second, exponential random graph models allow for much more flexibility than current biological network models. As seen in Table 1, the explanatory variables can be almost anything including subgraph counts, shortest path lengths, clusteredness and simple graph statistics. This flexibility allows researchers to ask and answer specific questions. For example, researchers have suggested that PPI networks are arranged in a hierarchy of modules (Han *et al.*, 2004). The hierarchy is usually identified using one or more properties of the nodes in the network. *Logit p** models provide an independent method to investigate this phenomenon. A further reason that exponential random graph models should be considered for biological networks is that they

[1] http://www.ebi.ac.uk/2can/genomes/genomes.html

provide an excellent framework for the comparison of networks. Finally, exponential random graph models provide a method to control for lower order effects by including them in the model. That is, if there is a suspected bias in the parameter of interest due to some lower level variable, the researcher can test and correct for this bias by including a parameter for the lower level variable in the model.

Conflict of Interest: none declared.

ACKNOWLEDGEMENTS

We would like to thank Anand Swaminathan for introducing us to p* models and the anonymous reviewers for their constructive comments.

REFERENCES

Albert,R. and Barabási,A.-L. (2000) Topology of evolving networks: local events and universality. *Phys. Rev. Lett.*, **85**, 5234–5237.

Anderson,C.J. *et al.* (1999) A p* primer: logit models for social networks. *Soc. Networks*, **21**, 37–66.

Barabási,A.-L. and Albert,R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

Geyer,C.J. and Thompson,E,A. (1992) Constrained monte carlo maximum likelihood for dependent data. *J. R. Stat. Soc. Ser. B*, **54**, 657–699.

Goodreau,S.M. (2007) Advances in exponential random graph (p*) models applied to a large social network. *Soc. Networks*, **29**, 231–248.

Han,J.-D.J. *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**, 88–93.

Handcock,M,S., *et al.* (2003) statnet: an r package for the statistical modeling of social networks, http://www.csde.washington.edu/statnet

Holland,P.W. and Leinhardt,S. (1970) A method for detecting structure in sociometric data. *Am. Sociol.*, **70**, 492–513.

Infante-Rivard,C. *et al.* (2006) Xenobiotic-metabolizing genes and small-for-gestational-age births: interaction with maternal smoking. *Epidemiology*, **17**, 38–46.

Jeong,H. *et al.* (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651.

Kaplan,D. (2004) *The Sage Handbook of Quantitative Methodology for the Social Sciences*. Sage Publications Inc.

Lee,T. *et al.* (2002) Transcriptional regulatory networks in *saccharomyces cerevisiae*. *Science*, **298**, 799–804.

Middendorf,M. *et al.* (2005) Inferring network mechanisms: the drosophila melanogaster protein interaction network. *Proc. Natl Acad. Sci.*, USA, **102**, 3192–3197.

Milo,R. *et al.* (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.

Milo,R. *et al.* (2003) On the uniform generation of random graphs with prescribed degree sequences. *condmat/0312028*. Retrieved from http://arxiv.org/PS_cache/ condmat/pdf/0312/0312028.pdf

Milo.R, *et al.* (2004) Superfamilies of evolved and designed networks. *Science*, **303**.

Pržulj,N. *et al.* (2004) Modeling interactome: scale-free or geometric? *Bioinformatics*, **20**, 3508–3515.

Salgado,H. *et al.* (2001) Regulondb (version 3.2): transcriptional regulation and operon organization in Escherichia coli k-12. *Nucleic Acids Res.*, **29**, 72–74.

Shen-Orr,S. *et al.* (2002) Network motifs in the transcriptional regulation network of *escherichia coli. Nat. Genet*.

Snijders,T.A.B. (2002) Markov chain monte carlo estimation of exponential random graph models. *J. Soc. Struct.*, **3**.

Snijders,T.A.B. *et al.* (2006) New specifications for exponential random graph models. *Sociol. Methodol.*

Wasserman,S. and Pattison,P. (1996) Logit models and logistic regressions for social networks: I. an introduction to markov graphs and p*. *Psychometrica*. **61**, 401–425.

Yeger-Lotem,E. *et al.* (2004). Network motifs in integrated cellular networks of transcription regulation and protein-protein interaction. *Proc Natl Acad. Sci*, USA, **101**, 5934–5939.