Link Prediction in Biological Networks using Multi-Mode Exponential Random Graph Models

Ali Shojaie
Department of Biostatistics, University of Washington
F650 Health Sciences Building
Seattle, Washington
ashojaie@u.washington.edu

ABSTRACT

We propose a novel multi-mode exponential random graph model for supervised prediction of gene networks, coupled with a penalized estimation framework for improved prediction performance. The proposed framework facilitates the analysis of gene networks with multiple edge types, and provides a systematic framework for incorporating multiple sources of biological data, as well as diverse attributes regarding function and location of genes, and structure of observed networks. Results of numerical experiments indicate that the method enjoys superior performance compared to other state-of-the-art reconstruction methods.

1. INTRODUCTION

Reconstruction of genetic networks is an important and challenging problem in systems biology, for which many new technologies [21, 35] and computational methods [5, 3, 22, 23, 9] have been proposed. Despite significant progress on both technological and computational fronts, a number of challenges continue to limit the ability of reconstruction methods for providing reliable estimates of genetic networks. These challenges include the high level of noise in biological experiments, high dimensionality of the problem relative to the available sample size in typical biological experiments, inadequacy of some of the abundantly available data sources for revealing genetic interactions, and limitations of computational models for addressing the complexities of gene networks.

Gene networks represent abstract models for complex interaction mechanisms among genes and proteins. Thus, while mathematical models can e.g. predict expression levels of a small group of genes in few well-studied pathways, the nature of genetic interactions remains largely unknown. Consequently, different assumptions about genetic interactions used in unsupervised network reconstruction methods may not generalize to other interaction types or organisms. Examples of assumptions used in unsupervised reconstruction methods include the presence of linear or nonlinear correlations among pairs of connected genes [e.g. 22, 9], specific functional forms for effect of transcription factors on regulated genes [31], or similarity in mRNA expression levels of genes regulated by the same transcription factor [25]. See [2] for a review of some of the existing unsupervised methods.

Supervised network reconstruction methods, on the other hand, offer the opportunity to gain additional insight into mechanisms of genetic interaction. As more interactions among genes are experimentally discovered and/or validated, statistical learning methods can be used to both determine the nature of such interactions, and also to use this knowledge to predict new, unobserved interactions. However, existing supervised method of network reconstruction often focus on a single type of genetic interaction. For instance, the SIRENE algorithm [25] can only be applied to estimation of regulatory interactions between transcription factors and regulated genes, and requires a large set of observed interactions for each transcription factor. On the other hand, [4, 20] consider estimation of protein-protein interaction networks from sequence data, while [38, 39] focus on prediction of interactions among enzymes. Thus, the above methods cannot be applied directly to estimation of other types of interactions among genes, an issue that we will revisit shortly. On the other hand, a number of existing methods focus solely on the prediction of new edges without providing direct insight into the underlying mechanisms of genetic interactions. This is mainly due to the unprobabilistic nature of such methods, and the lack of a rigorous framework for testing the model for presence of interactions among two genes. A typical example of this limitation is reconstruction methods based on support vector machines (SVM), which despite their many desirable properties, do not directly determine the factors affecting whether two genes/proteins/enzymes would interact with each other. As we discuss in Section 2, one of the main challenges in defining probabilistic models for genetic networks is computations needed for obtaining a closed-form probability distribution for the edges of the network; addressing this challenge is one of the main motivations of the method proposed in this paper.

Gene networks are known to include at least two types of interactions, or edges: (i) undirected edges corresponding to protein-protein interactions, and (ii) directed edges corresponding to *protein-DNA* (PD) interactions among transcription factors (TFs) and target genes (TGs), as well as *protein-protein* (PP) interactions among two genes, which suggest similar patterns of expression for the connected genes. Moreover, genes in a biological pathways are more likely to interact with each other than those in other pathways.

Eleventh Workshop on Mining and Learning with Graphs. Chicago, Illinois, USA

Finally, the degree distribution of the network is heterogeneous, with few "hub" genes that are connected to many other genes.

In this work, we seek to develop models that allow for simultaneous analysis and prediction of biological networks with *multiple edge types*, while incorporating additional features including high degree of clustering and presence of hub nodes. To achieve these goals, we propose a penalized multi-mode exponential random graph model, termed MP-ERGM, for supervised prediction of gene networks from diverse data sources. This model exploits and extends the framework of exponential random graph models (ERGMs), also known as p^* models, which are widely used for analysis of social networks [37].

Another attractive feature of the proposed model is its flexibility for incorporating different sources of biological data as variables in the probability distribution. For instance, one can seamlessly integrate mRNA expression levels, gene/protein sequences and data from ChIP-Seq experiments. This is particulary important as recent evidence shows that integration of multiple sources of biological data often result in improved estimation of network structure [see e.g. 29]. In this paper, we focus on a simple setting where the genetic interactions are modeled using mRNA expression levels, as well as the pathway membership of genes; other sources of data can be included in the model through additional covariates (see Section 2.3).

2. METHODS

Throughout this paper, we denote a graph $\mathcal{G} = (V, E)$ with the node set $V, |V| = m$ and edge set E. An edge is considered to be directed if $(i, j) \in E \Rightarrow (j, i) \notin E$, whereas an undirected edge implies that $(i, j) \in E$ iff $(j, i) \in E$. Here, we assume that network information is observed on a randomly selected subset of nodes in V, denoted $V_0, |V_0| = m_0$. We denote by $\mathcal{G}_0$ the subnetwork of $\mathcal{G}$ induced by V_0. We use random variables $Y_{ij}, i, j = 1, \dots m, i \neq j$ to denote the presence or absence of an edge between nodes i and j in $\mathcal{G}$. Thus, $\mathbf{Y}$ defines the adjacency matrix of the graph $\mathcal{G}$. Note that for undirected edges $Y_{ij} = Y_{ji}$. For each pair of nodes (genes), we use random variables $Z_{ijk}, i, j = 1, \dots, m, k = 1, \dots p$ to denote the set of p corresponding attributes. Each pair of nodes in the network is often referred to as a *dyad*. In general, Z_{ijk} can include attributes calculated based on different biological measurements, gene attributes or network features. For instance, let X be the $m \times n$ matrix of gene expression values from n samples, and W denote the $m \times q$ matrix gene attributes. Then, $Z_{ijk} = h_k(\mathbf{Y}, X_i, X_j, W_i, W_j)$, where X_i and W_i denote ith rows of X and W. This notation emphasizes that the kth dyadic attribute is a function of features of the network ($\mathbf{Y}$), expression levels of the corresponding genes (X_i and X_j), as well as other gene attributes (W_i and W_j). The function h_k denotes any summary measure based on different data sources, e.g. similarity or distance based kernels.

Our goal is to develop probabilistic models to relate the values of Y_{ij} to Z_{ijk}'s in the settings where Y_{ij}'s can be of different types. Using such a model, we then predict the values of Y_{ij}'s for $(i, j) \in \mathcal{G} \backslash \mathcal{G}_0$ by estimating the parameters of the model on the training set, $\mathcal{G}_0$. To this end, we exploit (and extend) the framework of ERGMs to build supervised models for predicting network edges based on observed links in $\mathcal{G}_0$, gene attributes, including their mRNA expression levels, as well as other relevant attributes, including whether two genes are in the same biological pathway, or have similar protein sequences.

In Section 2.1, we provide a short introduction to ERGMs. Details about the proposed multi-mode penalized ERGM framework are given in Section 2.2, where we also discuss parameter estimation and inference. Section 2.3 discusses the choice of covariates, as well as implementation considerations.

2.1 The Exponential Random Graph Model

To describe the general ERGM framework, let $\mathbf{y}$ denote a realization of $\mathbf{Y}$ as the adjacency matrix of the observed network. Also, assume that $\mathbf{y} \in \mathcal{Y}$, where $\mathcal{Y}$ denotes the set of all possible adjacency matrices. In general, $\mathcal{Y}$ can be set to $\{0, 1\}^{m \times m}$, allowing e.g. for self regulatory interactions, which can be estimated from time-course gene expression data. However, for simplicity, here we assume that diagonal entries are all zero. The multi-mode ERGM framework of Section 2.2 imposes additional restriction on the set $\mathcal{Y}$.

The general ERGM framework posits an exponential family distribution for the observed network [see e.g. 16]:

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}) = \exp\left(\theta^\intercal h(\mathbf{y}, X)\right) / \phi(\theta, \mathbf{y}), \tag{1}$$

where θ is the vector of model parameters, ϕ is the normalizing factor, calculated by summing over all configurations of $\mathbf{y}$. As before, h denotes the set of (node and dyad) attributes in the model.

The model in (1) provides a flexible framework for incorporating different attributes, although it has been mainly used in social network literature to incorporate features of the observed networks. The main challenge in application of ERGMs to high dimensional networks comes from the calculation of the normalizing factor ϕ. This involves summing over all possible configurations in $\mathcal{Y}$, which can be intractable even for moderate-size networks. Therefore, a number of approximate algorithms have been proposed for estimation of model parameters in ERGMs. One such idea is the pseudo likelihood approach of [33], which we briefly discuss next.

Let $\tilde{h}(\mathbf{y}_{ij}) \equiv h(\mathbf{y}^+_{ij}) - h(\mathbf{y}^-_{ij})$ denote the vector of changes in h when y_{ij} changes from 1 to 0. Then, conditional on the rest of the dyads in the network $\mathbf{Y}^c_{ij} = \{Y_{kl}, (k, l) \neq (i, j)\}$, it follows from (1) that

$$\text{logit}\left(\mathbb{P}(Y_{ij} = 1 | \mathbf{Y}^c_{ij} = \mathbf{y}^c_{ij})\right) = \theta^\intercal \tilde{h}(\mathbf{y}_{ij}), \tag{2}$$

where for $0 < x < 1$, $\text{logit}(x) = \log(x/(1-x))$.

This reparametrization in terms of change statistics, allows for intuitive interpretation of model parameters in terms of their effect on probability of an edge from i to j. In addition, it forms the basis for the pseudo maximum likelihood (PML) estimation method of [33]. In particular, the PML estimate of parameters θ can be obtained by assuming independence among values of Y_{ij}:

$$\mathbb{P}(Y_{ij} = 1 | \mathbf{Y}^c_{ij} = \mathbf{y}^c_{ij}) = \mathbb{P}(Y_{ij} = 1).$$

However, empirical evidence indicates that the pseudo likelihood method does not provide reliable estimates of model parameters, and [36] suggest the use of Markov Chain Monte Carlo (MCMC) methods for estimation of model parameters.

Although the PML framework may not provide a reliable estimate of the model parameters, it can be shown that

for a special class of models, called *dyadic independent*, the pseudo likelihood and likelihood methods are *equivalent*, and hence calculation using (2) provides *exact* estimates of θ [see e.g. 33, for more details]. An ERGM is said to satisfy dyadic independence if change statistics $\tilde{h}(\mathbf{y}, X)$ for all dyads (i, j) can be calculated without any knowledge of values of y_{kl} for $(k, l) \neq (i, j)$, except for (j, i). Clearly, for dyadic independence models, the estimation of model parameters is significantly simplified, and can be performed for high dimensional networks.

Examples of dyadic independent and dependent ERGMs are discussed in [33, 16]. In particular, a commonly used class of dyadic dependent models includes the Markov dependence model of [11], which, as the authors show, involves parameters for counts of k-stars and triangles. In view of these theoretical developments, commonly used methods of biological network reconstruction, based on attributes of genes fall in the category of dyadic independent models. However, the equivalence of pseudo likelihood and likelihood estimates in case of dyadic independent models does not hold if the network includes directed edges. This is because each dyad in directed networks consists of two edges; therefore, calculation of $\tilde{h}(y_{ij}, X_i, X_j)$, may also depend on y_{ji} even in dyadic independence ERGMs. This implies that the PML may no longer provide a reliable estimate of parameters of traditional ERGM models, and hence cannot be used to test hypotheses about the factors that affect the presence of biological interactions. On the other hand, the more reliable MCMC methods become computationally intensive for large biological networks, and may suffer from convergence issues [14].

In the next section, we propose a class of multi-mode ERGMs, which distinguish between directed and undirected edges in gene networks. We show that these models satisfy dyadic independence for a general set of attributes. This implies that parameters of the model can be efficiently estimated even for large gene networks. Further, we propose a penalized version of the model in (2) for simultaneous parameter estimation and model selection in ERGMs.

2.2 Penalized Multi-Attribute ERGMs

We first describe a penalized version of the regular ERGM for improved parameter estimation and prediction accuracy. Penalized estimation are widely used in high dimensional settings, and different penalties have been recently proposed to achieve accurate prediction and estimation. Penalized estimation methods have also been used in the graphical models setting, for unsupervised prediction of genetic networks, both for directed graphs [31, 30, 28], as well as undirected conditional independence graphs [24, 10].

In addition to general benefits of regularized estimation methods, there are two main motivations for using penalized ERGMs to predict edges of genetic networks. First, as pointed out in [26], inclusion of terms corresponding to different network features may introduce linear dependencies (i.e. multi-collinearity) among features of the model, which can result in unreliable parameter estimation. The addition of regularization penalties will facilitate model fitting in such settings and generalizes the applicability of ERGMs. Second, with multiple sources of biological data, and additional gene and network attributes, the number of predictors in the model p can increase relative to the size of the network. Consequently, the direct application of ERGMs for prediction of gene networks may suffer from over-fitting of the model to the training data. In such cases, the addition of the regularization penalty can result in improved prediction accuracy. Note that this is especially important in our setting, where the available network information is incomplete, wherein over-fitting can further imped the performance of predictive models.

Let $J(\theta)$ denote a general penalty function. Examples of J include the l_2 or Ridge [15], and l_1 or Lasso [34] penalties. The Ridge penalty encourages similarity amongst parameters for correlated variables, while the Lasso penalty encourages sparsity (and hence model selection) by setting some of the θ coefficients to zero. Here, we use the *elastic net* penalty [40], a combination of Ridge and Lasso penalties which offers improved estimation in settings with correlated predictors.

Using the elastic net penalty, the penalized ERGM for prediction of genetic networks is obtained by solving the following optimization problem:

$$\begin{aligned} \hat{\theta} &= \underset{\theta}{\operatorname{argmax}}\, l(\theta; \mathbf{y}) - J(\theta) \qquad (3) \\ J(\theta) &= \alpha\lambda\|\theta\|_1 + (1-\alpha)\lambda\|\theta\|_2^2 \end{aligned}$$

Here, $l(\theta; \mathbf{y})$ is the log-likelihood as a function of θ, and α and λ are tuning parameters controlling the size of the penalty. Equation (1) gives the form of the log-likelihood for the general ERGM. However, as we discussed in Section 2.1, in dyadic independent ERGMs, the log-likelihood function reduces to the log pseudo likelihood, given by the sum over all observed dyads in the network:

$$l(\theta; \mathbf{y}) = \sum_{\mathbf{y}} y_{ij} \log(\pi_{ij}) + \sum_{\mathbf{y}} (1 - y_{ij}) \log(1 - \pi_{ij}), \qquad (4)$$

where $\pi_{ij} = \mathbb{P}(Y_{ij} = 1)$. In this case, our network prediction problem corresponds to a *penalized logistic regression* model, and can be efficiently solved for large networks. Here, we use the coordinate descent algorithm of [12] for estimation of generalized linear models with the elastic net panalty, which is implemented in the R-package `glmnet`.

As discussed earlier, pseudo likelihood estimation can be used to obtain exact estimates of model parameters in case of dyadic independent ERGMs. However, in presence of multiple edges types, including directed edges, pseudo likelihood estimation may not give valid estimates of model parameters. This includes the case where the model only includes node attributes, and/or simple networks attributes. Next, we introduce the class of multi-mode ERGMs, which alleviate this problem and allow for efficient parameter estimation with multiple edge types. It should be noted that when the goal is solely network prediction, the pseudo likelihood approach in (4) can be still utilized, as there is no need for calculating the normalizing factor $\phi(\mathbf{Y}, \theta)$ (in this case, only the ratio of the two probabilities is needed). However, such a model would no longer provide insight into the underlying mechanisms of genetic interactions. On the other hand, as we will show later, the additive flexibility resulting from joint modeling of multiple edge types can also result in improved reconstruction performance.

As pointed out in the Introduction, different types of associations govern the underlying interaction mechanism of genes in biological networks. The penalized multi-mode ERGM (MP-ERGM) framework described next, extends the model of Section 2.1 to allow for multiple edge types and

addresses the general problem of supervised network reconstruction in the setting of gene networks.

Assume that the network is comprised of T different edge types: $\mathcal{Y} = \mathcal{Y}_1 \bigoplus \ldots \bigoplus \mathcal{Y}_T$, where $\bigoplus$ denotes partition of the space of interactions into different interaction types. Throughout this paper, we focus on the case of $T = 2$, corresponding to PD and PP interactions in the gene networks. However, the general framework can be used to allow for other types of genetic interactions (e.g. post-transcriptional regulations through miRNA's).

The above framework can also be extended to allow for presence of multiple edge types for the same dyad. In particular, if we instead consider the union of all possible edges for each dyad, $\mathcal{Y} = \mathcal{Y}_1 \biguplus \ldots \biguplus \mathcal{Y}_T$, more than one edge types can be estimated for each dyad. Finally, before discussing the parametrization and estimation in this framework, it is worth noting that the above framework differs from the multivariate ERGM framework of [27] in that the partitioning of edge types further limits the set of possible edges for each subclass $\mathcal{Y}_t$. An issue that facilitates the estimation and can result in improved prediction as further discussed below.

Using the above decomposition, we consider the following extension of (3), which estimates different parameters for each edge type, $\hat{\theta}_t$, and can be used for edge prediction via the multi-mode ERGM:

$$\underset{\theta}{\operatorname{argmax}} \sum_{\mathbf{y}_t} y_{ij} \log(\pi_{ij}) + \sum_{\mathbf{y}_t} (1 - y_{ij}) \log(1 - \pi_{ij}) - J(\theta),$$

where $\mathbf{y}_t$ are observed dyads for edge type t, with $t = 1, 2$ corresponding to PD and PP interactions, respectively. In this framework, PD interactions are only estimated between transcription factors (TFs) and target genes (TGs). Thus, edges corresponding in $\mathcal{Y}_1$ correspond to TF to TG interactions in the network. On the other hand, TG to TG edges correspond to $\mathcal{Y}_2$, which represent undirected associations among genes. These associations could correspond to both PP interactions, as well as any other undirected associations between two genes i and j. In particular, since directed TF to TF interactions cannot be estimated from steady state expression data, in this paper we also allow for estimation of undirected edges among TFs. Finally, under this model, no TG to TF, or self-regulatory edges are estimated.

2.3 Implementation Considerations

Covariates. In general, three classes of covariates can be incorporated into the proposed MP-ERGM framework: (a) network attributes, including degrees of nodes i and j, and the sparsity of the observed network; (b) data on gene activity in the cell, including steady state, or time course gene expression levels; (c) functional genes attributes, including whether genes i or j are transcription factors, and their pathway membership. In all models considered in Section 3, we focus mainly on attributes from classes (b) and (c) above. However, we also use the sparsity of the observed network in order to propose a weighted estimation scheme for improved prediction, as discussed next.

For each dyad, we consider four expression-based attributes, namely empirical correlation matrix and its absolute value, as well as estimated partial correlation matrix and its absolute value. The estimate of the correlation and partial correlation matrices are obtained using the graphical lasso (`glasso`) algorithm [13], which results in more robust estimates in small sample settings. Further, we define two function-based attributes, namely whether genes i and j are in the same KEGG pathway, and whether (i, j) dyad correspond to a TF to TG association. The last attribute is only used in the penalized ERGM estimator (referred to as ERGM hereafter), while the remaining attributes are shared between the ERGM and MP-ERGM estimators.

Weighted Parameter Estimation. It is well known that genetic networks are sparse, in the sense that the total number of edges in the network is far less than the m^2 possible edges, and it often scales linearly with the number of genes. Therefore, the training data for the proposed framework, are highly unbalanced in terms of number positive and negative examples (i.e. the number of 1's compared to the number of 0's in $\mathbf{y}$). A number of solutions have been explored in the literature for obtaining better estimators, including subsampling from the negative examples, or over-sampling from positive ones. Here, we consider an alternative strategy, motivated by developments in the area of survey sampling [8]. Specifically, we consider an observation weight for positive labels given by $(2\sum \mathbf{y}/|\mathcal{Y}|)^{-1}$. Note that for equal number of positive and negative labels, this weighting scheme amounts to equal weights for the two classes. However, in unbalanced settings, this method results in higher weights for positive examples.

Computational Complexity. As discussed earlier, the dyadic independence in the proposed MP-ERGM framework allows for efficient estimation of parameters using a penalized logistic regression model. To determine the computational complexity of the this procedure, it suffices to note that the computational cost of the algorithm is dominated by the cost of obtaining estimates of the parameters θ which requires $O(m_0^2 p^2)$ computations. However, the MP-ERGM algorithm also requires calculation of the covariates described in Section 2.3. This involves estimation of (partial) correlation matrix of genes using the `glasso` algorithm, which requires $O(m^3)$ operations. Thus, the computational complexity of the proposed MP-ERGM algorithm is $O(\max(m^3, m_0^2 p^2))$.

3. RESULTS

We consider network of genetic interactions of yeast *Saccharomyces cerevisiae*, for which information on transcription regulatory and protein-protein interactions are available from perturbation screens and two-hybrid experiments. To evaluate the proposed MP-ERGM estimator, we compare its performance with a number of state-of-the-art methods for gene network reconstruction. The competing methods have been chosen to provide a representative sample of network reconstruction methods, and include both supervised and unsupervised methods. We also consider different evaluation methods, as well as both real and simulated data. Next, we briefly describe each of the *competing methods* considered in our analysis.

1) **COEXP**: Coexpression analysis is perhaps the simplest method for reconstructing biological networks. In this method, a network is simply built by thresholding the absolute values of pairwise correlations between genes at a threshold τ. The resulting network is by default undirected, and is well known to include spurious edges e.g. due to high correlations between genes regulated by the same transcription factor.

2) **GLASSO**: In this method, the gene network is estimated by the graph of conditional dependencies among

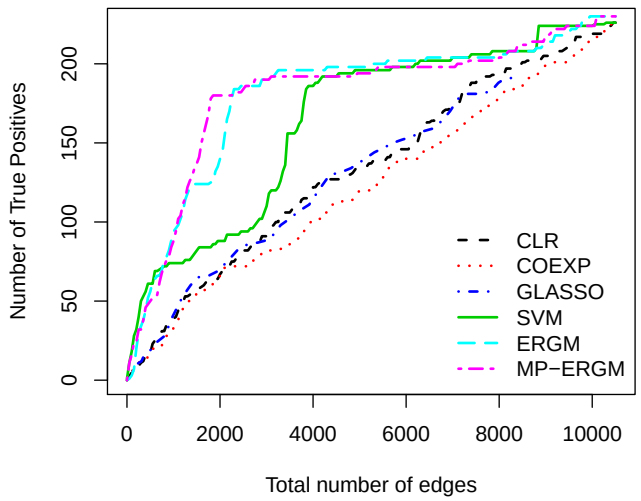

Figure 1: **Number of true positives vs. total number of edges in the graph for different reconstruction network methods.**

genes. More specifically, if the mRNA expression levels follow a joint normal distribution $X \sim N(0, \Sigma)$ (after log-transformation, and centering), the nonzero entries of Σ^{-1} define conditional dependence relationships among variables, and can be represented by an undirected graph. A number of regularization methods have been recently proposed for estimation of Σ^{-1} when the number of variables (genes) is larger than the sample size. Here, we use the *graphical lasso* algorithm as implemented in the R-package `glasso` [13].

3) **CLR**: The CLR method estimates the gene network using mutual information between pairs of genes [9]. This method is based on a background correction for the *Relevance Network* method [5], and is considered one of the best unsupervised method of network reconstruction. We use the matlab implementation of CLR, using the '`rayleigh`' method for calculating mutual information, as suggested by the authors for small to medium graphs.

4) **PCALG**: This method is based on the PC-algorithm of [17], wherein the gene network is estimated by a partially directed graph estimated based on conditional independence relations among genes. The PC-algorithm is the only method considered here that builds a partially directed network, and is used to provide an unsupervised benchmark for estimation of mixed edge types.

5) **SVM**: We develop a supervised estimation method using kernel-SVM's. This method is motivated by the proposal of [4], however, it differs from their proposal in that it uses expression data instead of sequence information. More specifically, to obtain a fair comparison, the SVM model is trained using the same set of covariates as in the penalized ERGM model (see Section 2.3), which includes mRNA expression levels, as well as pathway membership information. We use the *C-svc* method with Gaussian kernels (using $\sigma = 0.1$). We use the R-implementation in the package `kernlab` [19], which is based on the LIBSVM library [6]. The value of the penalty parameter C is chosen based on best performance in validation data among the set of values $\{100, 200, \ldots, 900, \} \cup \{1000, \ldots, 5000\}$.

3.1 Reconstruction of gene network of yeast

The network considered in our first experiment is comprised of list of documented transcription regulatory interactions between TFs and TGs in yeast obtained from YEASTRACT [1], coupled with protein-protein interactions in yeast from BioGRID [32] downloaded using the BioGRID Cytoscape plug-in.

Here, we focus on the subset of genes involved in 'metabolism' and 'cell growth' in yeast, which provide a window into changes in yeast cells at different stages of growth, and under different environmental conditions. In particular, we consider the genes in Amino Acid Metabolism, Carbohydrate Metabolism, and Cell Growth based on information from KEGG [18]. Expression levels for these genes were obtained from the data in [7] (GEO Accession No. GSE5499), which includes 270 arrays under different conditions. The resulting network included a network of 411 yeast genes with both gene expression data and network information.

Figure 1 shows the number of true positive (TP) edges for each reconstruction method as a function of the total number of edges (TE) in the estimated network. Such a comparison removes the need for choosing a cutoff parameter for determining whether an edge is present among two genes, and provides a more complete picture of performance of different methods across the range of the tuning parameters. The result includes a line for each of the methods described previously, with the exception of PCALG; the PC-algorithm implementation in the R-package `pcalg` uses the probability of false positives α as the tuning parameter, which limits the total number of edges in the network, and results in an incomplete TP-TE line.

It is worth noting that the unsupervised estimates (COEXP/GLASSO/CLR) are based solely on gene expression measurements. On the other hand, the supervised methods (SVM/ERGM/MP-ERGM) can incorporate multiple types of biological information, and are developed using the covariates discussed in Section 2.3. We split the network into training and test sets, by randomly selecting 75% of the genes and the network of interactions amongst them for training. We then evaluate the performance of each of the estimators on the remaining part of the network. Note that unsupervised methods do not require a training network. However, GLASSO and CLR are designed to remove spurious edges using information from other genes in the network. Therefore, to prevent bias, GLASSO and CLR estimates are obtained based on the complete gene expression matrix for 411 genes; we then take the subset of the network corresponding to the test genes for evaluation.

The results in Fig 1 show that, as expected, incorporating the knowledge of the network and gene attributes results in significant gain in performance for the supervised methods. Interestingly, the unsupervised methods have very similar performances, with CLR and GLASSO having a slight edge over COEXP.

On the other hand, these results indicate that the methods based on exponential random graph models, ERGM and MP-ERGM, offer significant advantages compared to other methods considered here, including the kernel SVM method. Finally, while MP-ERGM seems to have a slight edge over ERGM, the difference does not seem to be significant.

3.2 Reconstruction from simulated data

In this section, we focus on the performance of network

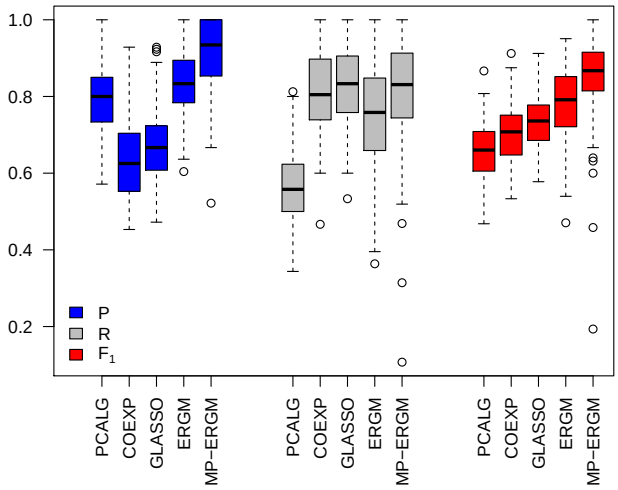

Figure 2: precision, recall, and F_1 measure for estimating the network of yeast genes using different prediction methods.

reconstruction methods, when the cutoff parameter for each method is optimally determined. For supervised classification methods, this corresponds to choosing the cutoff for probability of an edge. In unsupervised methods, on the other hand, the choice of tuning parameter is often determined based on the properties of the estimator, or the fit to the data. However, when edges in the network are partially known, such information can be used to choose the cutoff and/or tuning parameter in order to improve the performance of the method. This amounts to a bridge between supervised and unsupervised methods. Here, we consider variants of GLASSO and COEXP methods where the corresponding tuning parameter – the threshold parameter τ for COEXP and the penalty coefficient λ for GLASSO – is chosen so that the method gives an optimal performance on the available network. To evaluate the performance of the reconstruction methods, we report for each method its *precision* P and *recall* R, as well as their harmonic mean, F_1.

We consider ERGM, MP-ERGM and modified versions of GLASSO and COEXP. However, considering the equivalent performances of CLR and GLASSO, we do not include CLR in the analysis of this section. Also, considering the inferior performance of kernel SVM, and its larger computational cost, compared to other supervised methods, we do not include this method either. The last method included in this section is the PCALG estimator, which as mentioned before, provides a partially directed estimate and can hence be more directly compared to the MP-ERGM estimator. This method also provides a benchmark for comparing the effect of using partial knowledge of the network, either to train the classifier (in case of ERGM and MP-ERGM), or to choose the tuning parameter (in case of COEXP and GLASSO). Following the suggestion of the [17], we set the probability of false positive in PCALG to $\alpha = 0.01$.

To gain further insight into the performance of reconstruction methods, in this section we compare their performances using data simulated according to the given gene network. By doing this, we guarantee that the underlying network is a *gold standard*, in the sense that no missing or incorrect edges are present in the network. The availability of gold standard is one of the main challenges in evaluating the performance of network reconstruction methods, and by simulating data from the network, we can directly compare the performance of the competing methods in reconstructing the network from available data. However, a clear limitation of this approach is that in order to generate data from the network, we need to pose different assumptions on the distribution of the data, as well as effects of regulatory and protein-protein interactions.

To prevent any bias towards the proposed methods, and to see whether supervised methods offer any advantages in the settings where unsupervised methods correctly capture the data generation mechanism, we generate data from a Gaussian graphical model (GGM). In this setting, the (log) expression profiles are assumed to follow a normal distribution $N(0, \Sigma)$. To determine the joint covariance matrix Σ, we consider a model wherein the gene network corresponds to the conditional independence graph. In other words, the presence of an edge between two genes indicates that they are conditionally dependent, given all other genes. Considering the symmetry of the covariance matrix, we remove the directionality of the edges in the network, and transform the adjacency matrix A into a symmetric matrix. It then follows, from the theory of GGMs, that the adjacency matrix has the same non-zero pattern as Σ^{-1}.

For simplicity, we assign a constant value of partial correlation to all edges of the network (set here as $\rho = 0.6$). However, to obtain a well-defined probability distribution, we need Σ^{-1} to be positive definite. We achieve this by making the matrix sub-stochastic (a.k.a. diagonally dominant). Expression levels can then be generated as multivariate normal with covariance matrix Σ. We generate $n = 100$ i.i.d. samples. Note that the data generated from such a procedure matches exactly the underlying assumption of the GLASSO estimator, except for presence of directed edges in the network. Therefore, it is expected that by obtaining the optimal value of tuning parameter based on the partial knowledge of network, GLASSO would result in good reconstruction performance.

The distributions of P, R and F_1 for different network reconstruction methods are shown in Figure 2. The results are for 100 randomly selected sets of training and test networks. As in Section 3.1, it can be seen that, ERGM and MP-ERGM outperform the other methods in terms of F_1 measure. (p-values for pairwise tests between supervised and unsupervised methods based on Wilcoxon Rank Sum tests are all $< 10^{-4}$.) In this setting, GLASSO and COEXP have high recall values, which is expected as the underlying data generation mechanism obeys the multivariate normal assumption of these methods. These results also suggest that the use of partial knowledge of network in selection of tuning parameters for GLASSO and COEXP results in considerable improvements compared to the unsupervised PCALG method.

The results of above experiment further highlight the advantages of the supervised methods, even when the underlying model matches the assumptions of unsupervised methods. Clearly, this advantage can be more pronounced if the underlying model does not correspond to the assumption of unsupervised methods. These results also suggest that the joint modeling of multiple edge types in MP-ERGM can offer additional improvements in network reconstruction.

4. DISCUSSION

In this paper, we proposed a multi-mode exponential random graph model for supervised prediction of genetic interactions. The proposed model offers a systematic framework for analysis of gene networks with multiple edge types, and allows for seamless integrations of diverse sources of biological data, as well as additional information on function and location of genes in the cell. Numerical experiments indicate that this model offers improved predictive performance compared to existing models. Another appealing feature of this framework, compared to some of the existing supervised methods, is that by considering different models for multiple types of genetic interactions, and providing an exact inference framework, it offers additional insight into mechanisms of genetic interactions.

The proposed framework offers a number of possible extensions. In particular, it is of interest to further explore the capability of this model to integrate multiple data sources, include other sources of association measures and more complex network features, and extend its applicability to other types of biological networks, such as metabolic networks. The model developed in this paper implicitly assumes that the available network information are selected randomly from the set of all possible interactions. Investigating the effects of such assumptions, and developing models that allow for different sampling mechanisms remain the topic of future research.

Acknowledgement

This work was partially supported by the Grant DMS-1161565 from the National Science Foundation.

References

[1] D. Abdulrehman, P.T. Monteiro, M.C. Teixeira, N.P. Mira, A.B. Lourenço, S.C. dos Santos, T.R. Cabrito, A.P. Francisco, S.C. Madeira, R.S. Aires, et al. Yeastract: providing a programmatic access to curated transcriptional regulatory associations in saccharomyces cerevisiae through a web services interface. *Nucleic acids research*, 39(suppl 1):D136–D140, 2011.

[2] M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. Di Bernardo. How to infer gene networks from expression profiles. *Molecular systems biology*, 3(1), 2007.

[3] Z. Bar-Joseph, G.K. Gerber, T.I. Lee, N.J. Rinaldi, J.Y. Yoo, F. Robert, D.B. Gordon, E. Fraenkel, T.S. Jaakkola, R.A. Young, et al. Computational discovery of gene modules and regulatory networks. *Nature biotechnology*, 21(11):1337–1342, 2003.

[4] A. Ben-Hur and W.S. Noble. Kernel methods for predicting protein–protein interactions. *Bioinformatics*, 21(suppl 1):i38–i46, 2005.

[5] A.J. Butte, P. Tamayo, D. Slonim, T.R. Golub, and I.S. Kohane. Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences*, 97(22):12182, 2000.

[6] C.C. Chang and C.J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[7] G. Chua, Q.D. Morris, R. Sopko, M.D. Robinson, O. Ryan, E.T. Chan, B.J. Frey, B.J. Andrews, C. Boone, and T.R. Hughes. Identifying transcription factor functions and targets by phenotypic activation. *Proceedings of the National Academy of Sciences*, 103(32):12045–12050, 2006.

[8] W.G. Cochran. *Sampling techniques*. Wiley-India, 2007.

[9] J.J. Faith, B. Hayete, J.T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J.J. Collins, and T.S. Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, 5(1):e8, 2007.

[10] J. Fan, Y. Feng, and Y. Wu. Network exploration via the adaptive LASSO and SCAD penalties. *Annals of Applied Statistics*, 3(2):521–541, 2009.

[11] O. Frank and D. Strauss. Markov graphs. *Journal of the American Statistical Association*, pp. 832–842, 1986.

[12] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Technical Report, Department of Statistics Stanford University*, 2008.

[13] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432, 2008.

[14] W.R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in practice: interdisciplinary statistics*, volume 2. Chapman & Hall/CRC, 1995.

[15] A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, pp. 55–67, 1970.

[16] D.R. Hunter, M.S. Handcock, C.T. Butts, S.M. Goodreau, and M. Morris. ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, 24(3):nihpa54860, 2008.

[17] M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *The Journal of Machine Learning Research*, 8:613–636, 2007.

[18] M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[19] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis. kernlab-an s4 package for kernel methods in r. 2004.

[20] T. Kato, K. Tsuda, and K. Asai. Selective integration of multiple biological data for supervised network inference. *Bioinformatics*, 21(10):2488–2495, 2005.

[21] 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, et al. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science's STKE*, 298(5594):799, 2002.

[22] A.A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R.D. Favera, and A. Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7(Suppl 1):S7, 2006.

[23] A.A. Margolin, K. Wang, W.K. Lim, M. Kustagi, I. Nemenman, and A. Califano. Reverse engineering cellular networks. *Nature Protocols*, 1(2):662–671, 2006.

[24] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals Of Statistics*, 34(3):1436–1462, 2006.

[25] F. Mordelet and J.P. Vert. SIRENE: supervised inference of regulatory networks. *Bioinformatics*, 24(16):i76–i82, 2008.

[26] M. Morris, M.S. Handcock, and D.R. Hunter. Specification of exponential-family random graph models: terms and computational aspects. *Journal of Statistical Software*, 24(4):1548, 2008.

[27] P. Pattison and S. Wasserman. Logit models and logistic regressions for social networks: Ii. multivariate relations. *British Journal of Mathematical and Statistical Psychology*, 52(2):169–193, 1999.

[28] A. Shojaie, S. Basu, and G. Michailidis. Adaptive thresholding for reconstructing regulatory networks from time-course gene expression data. *Statistics in Biosciences*, 4(1):66–83, 2012.

[29] A. Shojaie, A. Jauhiainen, M. Kallitsis, and G. Michailidis. Inferring Regulatory Networks by Combining Perturbation Screens and Steady State Gene Expression Profiles. *PLoS One,* tentatively accepted, 2013.

[30] A. Shojaie and G. Michailidis. Discovering graphical Granger causality using the truncating lasso penalty. *Bioinformatics*, 26(18):i517, 2010.

[31] A. Shojaie and G. Michailidis. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538, 2010.

[32] C. Stark, B.J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M.S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, et al. The BioGRID interaction database: 2011 update. *Nucleic acids research*, 39(suppl 1):D698–D704, 2011.

[33] D. Strauss and M. Ikeda. Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, pp. 204–212, 1990.

[34] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Methodological*, 58(1):267–288, 1996.

[35] A.H.Y. Tong, G. Lesage, G.D. Bader, H. Ding, H. Xu, X. Xin, J. Young, G.F. Berriz, R.L. Brost, M. Chang, et al. Global mapping of the yeast genetic interaction network. *Science*, 303(5659):808, 2004.

[36] M.A.J. van Duijn, K.J. Gile, and M.S. Handcock. A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. *Social Networks*, 31(1):52–62, 2009.

[37] S. Wasserman and P. Pattison. Logit models and logistic regressions for social networks: I. an introduction to markov graphs andp. *Psychometrika*, 61(3):401–425, 1996.

[38] Y. Yamanishi, J.P. Vert, and M. Kanehisa. Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, 20(suppl 1):i363–i370, 2004.

[39] Y. Yamanishi, J.P. Vert, and M. Kanehisa. Supervised enzyme network inference from the integration of genomic data and chemical information. *Bioinformatics*, 21(suppl 1):i468–i477, 2005.

[40] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.