View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Collection Of Biostatistics Research Archive

University of Pennsylvania
UPenn Biostatistics Working Papers

Year 2010 *Paper* 34

Bayesian Methods for Network-Structured Genomics Data

Stefano Monni* Hongzhe Li†

*Cornell, stm2013@med.cornell.edu

†University of Pennsylvania, hongzhe@mail.med.upenn.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/upennbiostat/art34

Copyright ©2010 by the authors.

Bayesian Methods for Network-Structured Genomics Data

Stefano Monni and Hongzhe Li

Abstract

Graphs and networks are common ways of depicting information. In biology, many different processes are represented by graphs, such as regulatory networks, metabolic pathways and protein-protein interaction networks. This information provides useful supplement to the standard numerical genomic data such as microarray gene expression data. Effectively utilizing such an information can lead to a better identification of biologically relevant genomic features in the context of our prior biological knowledge. In this paper, we present a Bayesian variable selection procedure for network-structured covariates for both Gaussian linear and probit models. The key of our approach is the introduction of a Markov random field prior for the indicator variables that describe which covariates should be included in the model and the use of the Wolff algorithm for Markov Chain Monte Carlo inference. We illustrate the proposed procedure with simulations and with an analysis of genomic data. Finally, we present some other areas of genomics research where novel Bayesian approaches may play important roles.

Bayesian Methods for Network-Structured Genomics Data

Stefano Monni and Hongzhe Li

Department of Public Health, Weill Cornell Medical College

402 East 67th Street, New York, NY 10065-6304

and

Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104.

A Chapter in *"Frontier of Statistical Decision Making and Bayesian Analysis - In honor of James O. Berger"*, Co-Editors: Ming-Hui Chen, Dipak K. Dey, Peter Mueller, Dongchu Sun and Keying Ye.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Abstract

Graphs and networks are common ways of depicting information. In biology, many different processes are represented by graphs, such as regulatory networks, metabolic pathways and protein-protein interaction networks. This information provides useful supplement to the standard numerical genomic data such as microarray gene expression data. Effectively utilizing such an information can lead to a better identification of biologically relevant genomic features in the context of our prior biological knowledge. In this paper, we present a Bayesian variable selection procedure for network-structured covariates for both Gaussian linear and probit models. The key of our approach is the introduction of a Markov random field prior for the indicator variables that describe which covariates should be included in the model and the use of the Wolff algorithm for Markov Chain Monte Carlo inference. We illustrate the proposed procedure with simulations and with an analysis of genomic data. Finally, we present some other areas of genomics research where novel Bayesian approaches may play important roles.

1 Introduction

One of the main problems in biological research is the identification of genetic variants such as single nucleotide polymorphisms (SNPs) or gene expression levels that are responsible for a clinical phenotype such as disease status. The problem can in general be formulated as a variable selection problem for regression models. To deal with high-dimensionality, many statistical methods have been developed, including Lasso (Tibshirani, 1995) and its many extensions such as fused lasso (Tibshirani *et al.*, 2005), adaptive lasso (Zou, 2006), group lasso (Yuan and Lin, 2006), SCAD (Fan and Li, 2001), the Elastic net (Zou and Hastie, 2005), LARS (Efron *et al.*, 2004), and the Dantzig selector (Candes and Tao, 2007). These methods are mainly based on the idea of regularization. Alternatively, variable selection has also been developed and extensively studied in a Bayesian framework, especially for linear or generalized linear models (George, 2000; George and McCulloch, 1993, 1997). Hans *et al.* (2007) developed shotgun stochastic search in regression with many predictors in order

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/upennbiostat/art34

the make the Bayesian variable selection procedures applicable and feasible to the analysis of genomic data. Bayesian formulations of some regularized procedures are also available: a Bayesian Lasso, for example, has been developed recently in (Park and Casella, 2008). Many of these methods have also been employed to analyze genomic data, especially microarray gene expression data in order to identify the genes that are related to a certain clinical or biological outcome.

One limitation of all these popular approaches is that often the methods are developed purely from computational or algorithmic points without utilizing any prior biological knowledge or information and thus important structures of the data may be ignored. For many complex diseases, especially for cancers, a wealth of biological knowledge (*e.g* pathway information) is available as a result of many years of intensive biomedical research. This large body of information is now primarily stored in databases on different aspects of biological systems. Some well-known pathway databases include KEGG, Reactome (www.reactome.org), BioCarta (www.biocarta.com) and BioCyc (www.biocyc.org). Of particular interest are gene regulatory pathways that provide regulatory relationships between genes or gene products. These pathways are often interconnected and form a web of networks, which can then be combined and represented as a graph, the vertices of which are genes or gene products and the edges representations of inter-gene regulatory relationships of some kind. This information is a useful supplement to the standard numerical data collected from an experiment. Incorporating the information from these graphs into a data analysis is a non-trivial task, which is generating increasing interest. In genome-wide association studies, the SNPs are often in linkage disequilibrium (LD) and are therefore dependent. Li *et al.* (2009) introduced the idea of weighted LD graphs based on the pair-wise r^2 statistics between the SNPs. The problem we encounter is that the predictors are constrained on a graph and the challenge we face is to incorporate these constraints in the regression analysis. Motivated by a Gaussian Markov random field prior on the regression coefficients, Li and Li (2008) proposed a network-constrained regularization procedure to incorporate the network-structure information into the analysis, and demonstrated gain in sensitivity in identifying the relevant genes. In the Bayesian context, Li and Zhang (2008) proposed a variable selection for Gaussian linear models with structured covariates using an Ising prior and a Gibbs sampling. Tai and

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Pan (2009) put forward a similar approach using several different Markov random field priors. In this paper we consider a Bayesian variable selection method that takes account of the fact that the covariates are measured on a graph for both linear Gaussian and probit models. Because prior distributions model our *a priori* knowledge of the data, the network structure is introduced in a very natural way at the level of prior probabilities. We consider here an Ising prior, as in Li and Zhang (2008). An Ising model was also used for network-based analysis in Wei and Li (2007). In addition, we implement an MCMC sampler for estimating the posterior probabilities that a variable is selected that is based on the Wolff algorithm (Wolff, 1989). This algorithm was introduced to eliminate the critical slowing down of local updating schemes in Ising models, and is extremely natural in this problem, as we hope will be clear. The paper is organized as follows. In section 2, we formulate the problem in the context of Bayesian variable selection and describe the models, the prior probability distributions, and the algorithm used for inference. In section 3, we report the results of some applications of the method to simulated data sets and to a real data set. Finally, we make some comments and present some discussions.

2 Bayesian Variable Selection with a Markov Random Field Prior

From a statistics view-point, we are interested in the problem of Bayesian variable selection in the case in which the data enjoy a graphical representation. Namely, variables have pairwise relations, which are represented as edges in a graph whose nodes represent the variables. We assume the network to be simple and undirected, *i.e.* that the relations are among pair of distinct variables and are symmetric (if the variable i is related to j, then j is related to i). If one is able to assess the relative strength of the pair-wise interactions, one can furnish the edges with a quantitative label (a weight) that measures such strengths. When such an assessment is not possible, the only information an edge encodes is the existence of the interactions. Both situations are possible and will be taken into account in our model.

To fix notation, let $X = (X_1, \ldots, X_p)$ be the vector of p-covariates and Y the binary

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/upennbiostat/art34

or continuous outcome. Each variable is measured on N samples. We denote by $\mathbf{Y} = (y_1, \ldots, y_N)^T$ the vector of responses, by $\mathbf{X} = (x_{ij})$ the $N \times p$ matrix of covariate values, and by $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$, with the super-script T being transposition, the i-th row of the covariate matrix, that is, the values of the covariates for the i-th sample. Finally, we let (G_{ij}) be the adjacency matrix of the network. For unweighted networks

$$G_{ij} = \begin{cases} 1 & \text{if } X_i \text{ and } X_j \text{ are related} \quad i, j = 1, \ldots, p \\ 0 & \text{otherwise.} \end{cases}$$

The assumption that the network is simple and undirected is tantamount to (G_{ij}) being symmetric and having zeros along the diagonal.

In our approach the network structure will be taken into consideration in the choice of prior distributions and in the Markov chain used for the inference. The rest of the formalism is quite common, and will be sketched here to make the paper self-contained. We first describe the models used to relate the outcome Y to the covariates when Y is binary or continuous. We then detail the inferential strategy.

2.1 Likelihood and the Prior Distributions

Binary outcomes can be modeled in many ways. Here, we consider a probit model. This choice allows us to write marginalized quantities in a manageable form. In this model, the responses are assumed to be independent samples of Bernoulli distributions

$$Y_i | \boldsymbol{\beta}, \mathbf{X} \sim \mathsf{Bernoulli}(\mu_i) \quad i = 1, \ldots, N. \tag{1}$$

The probability μ_i of success ($y_i = 1$) is related to a linear combination of the covariates (linear predictor) by the following relation:

$$\mu_i = \Phi(\boldsymbol{x}_i^T \boldsymbol{\beta}),$$

where Φ is the cumulative distribution function of the standard normal distribution. Alternatively, if the outcome is continuous, we consider instead a Gaussian linear model

$$Y_i = y_i | \boldsymbol{\beta}, \mathbf{X}, \sigma^2 \sim \mathcal{N}(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2 I). \tag{2}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

We assume that some predictors have negligible coefficients β, which will then be considered zero. Each model will thus be labeled by a vector of latent binary variables $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$, with each component γ_i being 1 (0) if the corresponding variable X_i is (not) present in the model: namely, $\gamma_i = 0$ if and only if $\beta_i = 0$. Accordingly, we denote by $p_\gamma = \sum_i \gamma_i$ the number of variables, by $\mathbf{X}_{\boldsymbol{\gamma}}$ the matrix $N \times p_\gamma$ obtained from $\mathbf{X}$ by removing any column i such that $\gamma_i = 0$. We leave explicit the intercept β_0 in the linear predictor so that $\mathbf{X}_{\boldsymbol{\gamma}}$ will in fact be the $N \times (1 + p_\gamma)$ matrix of covariates, with the first column being a vector of 1, with an abuse of notation.

The expressions (1) and (2) are the likelihood functions for our models, which share the parameters $\boldsymbol{\beta}$. The normal model has the additional parameter σ^2, the residual variance.

We now specify the prior distributions. For the regression coefficients, we choose the commonly used prior

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}} | \boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{m}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}), \tag{3}$$

where $\boldsymbol{m}_{\boldsymbol{\gamma}}$ is the mean and $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$ is covariance matrix. The prior distribution of the parameters $\boldsymbol{\gamma}$ will instead be non-standard. Indeed, the γ_i are generally chosen independent, *e.g.* samples from a multivariate Bernoulli distribution, with probabilities $w_i = P(\gamma_i = 1)$ either predetermined or (usually Beta) random samples. Here, instead, we do not make the latter assumption as we want to take account of the network structure. This is the first difference with the usually proposed Bayesian variable selection models. Namely, we want a probability measure that enjoys the Markov property, that is, we assume the conditional probability that a variable i is in the model to depend only on its neighbors. In addition, we impose the stronger requirement that the probability that a variable is selected be greater if its neighbors are also selected. These conditions are satisfied by the following distribution

$$\pi(\boldsymbol{\gamma}|\mathbf{J}) \propto \mathrm{e}^{\sum_{i<j} G_{ij} \delta(\gamma_i, \gamma_j) J_{ij}} \cdot \rho^{-\sum_i \gamma_i} \tag{4}$$

where δ is the Kronecker delta, ρ and $J_{ij} \geq 0$ are non-negative real numbers. The omitted normalization constant is the sum over all $\boldsymbol{\gamma}$ configurations. One may have recognized the form of the prior (4) as defining an Ising model. The parameter ρ is chosen greater than one so as to penalize large models. As for the interaction terms J_{ij}, the simplest model is that with all set equal to a constant J_0. If the network is a weighted network, J_{ij} can be chosen

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/upennbiostat/art34

equal to the weights. A particular interesting case is that in which the correlation structure of the covariates is used to define $\mathbf{J}$: $J_{ij} \propto |\mathrm{Corr}(X_i, X_j)|$. With this choice, variables that are linked in the network are a priori forced to be simultaneously inside the model (or outside the model) with a probability that is higher for variables that are more highly correlated. It would also be interesting to consider the case where J_{ij} are random samples from a distribution $\pi(\mathbf{J})$ (that is, a random Ising model). There are some computational difficulties associated with this situation. For example, the dependence of the normalization constant of the prior (4) on J makes it difficult to find a prior distribution that leads to a conditional distribution completely available in its analytic form.

2.2 Posterior distributions

Once the likelihood and the prior distributions are specified, we can apply the Bayes' formula to obtain the posterior probability. Since the main goal of our analysis is to determine which variables enter the model, we can do away with the sampling of the regression coefficients, and average over them to compute marginalized posterior probabilities. For continuous responses, we use the following values for the parameters of the prior distribution (3) of the regression coefficients

$$\boldsymbol{m}_{\boldsymbol{\gamma}} = \mathbf{0}, \quad \boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = \tau\sigma^2(X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}} + \lambda I_{\boldsymbol{\gamma}})^{-1}, \tag{5}$$

and assume the variance σ^2 to be a random variable distributed according to the law

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2}.$$

The prior for the regression coefficients with parameters (5) reduces to that of Smith and Kohn (1996) when $\lambda = 0$, which is related to Zellner's g-prior (1986). We fix $\tau = N$. For possible implications of the values of τ in model selection, we refer the reader to Chipman *et al.* (2001). The constant λ in the covariance matrix is introduced so that $L_{\boldsymbol{\gamma}}$ can be computed even when the number of selected variables $p_{\boldsymbol{\gamma}}$ is larger than the sample size N.

With these choices, the posterior distribution is

$$p(\boldsymbol{\gamma}, \mathbf{J}|\mathbf{Y}) \propto p(\mathbf{Y}|\boldsymbol{\gamma}, \mathbf{J}, \mathbf{X})\pi(\boldsymbol{\gamma}|\mathbf{J}),$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

where

$$\begin{aligned} p(\mathbf{Y}|\boldsymbol{\gamma},\mathbf{J},\mathbf{X}) &= \int \mathrm{d}\sigma^2\mathrm{d}\boldsymbol{\beta}\, p(\mathbf{Y}|\boldsymbol{\gamma},\boldsymbol{\beta},\mathbf{J},\mathbf{X},\sigma^2)\pi(\sigma^2)\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}) \\ &\propto (Y^T L_{\boldsymbol{\gamma}} Y)^{-n/2}\frac{1}{\sqrt{\det R_{\boldsymbol{\gamma}}}}, \end{aligned}$$

with $L_{\boldsymbol{\gamma}}$ and $R_{\boldsymbol{\gamma}}$ being the matrices

$$L_{\boldsymbol{\gamma}} = I - \frac{\tau}{\tau+1}X_{\boldsymbol{\gamma}}\left(X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}} + \frac{\lambda}{\tau+1}I\right)^{-1}X_{\boldsymbol{\gamma}}^T,$$

$$R_{\boldsymbol{\gamma}} = I + \tau X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}}\left(X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}} + \lambda I\right)^{-1}.$$

For the binary case, we follow Albert and Chib (1993) and introduce N Gaussian latent variables Z_i $i = 1,\ldots,N$ in terms of which the responses $\mathbf{Y}$ are recovered via the relation: $Y_i = \mathcal{I}(Z_i > 0)$, with $\mathcal{I}$ being the indicator function. As a consequence, one now needs to consider the joint posterior probability of the parameters of the model and of the latent variables, which is

$$p(\mathbf{Z},\boldsymbol{\beta},\boldsymbol{\gamma},\mathbf{J}|\mathbf{Y}) \propto \prod_{i=1}^{N} f(Y_i|Z_i) f(Z_i|\boldsymbol{\beta}_{\boldsymbol{\gamma}},\boldsymbol{\gamma})\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma}|\mathbf{J})$$

with

$$f(Z_i|\boldsymbol{\beta}_{\boldsymbol{\gamma}},\boldsymbol{\gamma}) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(Z_i - \mathbf{x}_{i\boldsymbol{\gamma}}^T\boldsymbol{\beta})^2}, \tag{6}$$

and

$$f(Y_i|Z_i) = P(Y_i = y_i|Z_i) = \mathcal{I}(z_i > 0)\delta(y_i,1) + \mathcal{I}(z_i \le 0)\delta(y_i,0).$$

The marginalized joint distribution of $\boldsymbol{\gamma}$ and $\mathbf{Z}$

$$p(\mathbf{Z},\boldsymbol{\gamma},\mathbf{J}|\mathbf{Y}) = p(\mathbf{Z}|\boldsymbol{\gamma},\mathbf{J},\mathbf{Y})\pi(\boldsymbol{\gamma}|\mathbf{J})$$

is expressed in term of the marginalized distribution of $\mathbf{Z}$, which we now compute. Choosing the values (5), with $\sigma^2 = 1$, for the parameters of the prior distribution (3) of the regression coefficients, we have

$$\begin{aligned} p(\mathbf{Z}|\boldsymbol{\gamma},\mathbf{Y},\mathbf{J}) &\propto \int \prod_{i=1}^{N} f(Y_i|Z_i) f(Z_i|\boldsymbol{\beta}_{\boldsymbol{\gamma}},\boldsymbol{\gamma})\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})\mathrm{d}\boldsymbol{\beta}_{\boldsymbol{\gamma}} \\ &\propto \frac{e^{-\frac{1}{2}Z^T L_{\boldsymbol{\gamma}} Z}}{\sqrt{\det R_{\boldsymbol{\gamma}}}} \prod_{i=1}^{N}\left(\mathcal{I}(z_i > 0)\delta(y_i,1) + \mathcal{I}(z_i \le 0)\delta(y_i,0)\right), \end{aligned} \tag{7}$$

with the matrices $L_{\boldsymbol{\gamma}}$ and $R_{\boldsymbol{\gamma}}$ as above.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/upennbiostat/art34

2.3 Markov chain Monte Carlo Inference

We have determined the marginalized posterior probabilities up to a normalizing constant, which can not be computed. To deal with this problem and identify high-probability models, we consider a Metropolis algorithm for $\boldsymbol{\gamma}$. In the binary case, we additionally draw Z from the conditional distribution which can be read out from (7). It is in the Metropolis algorithm where we make use of the network structure. Namely, we apply the algorithm devised by Wolff (1989). We randomly select a variable, i, and construct a cluster of nodes $Cl(i)$ around it iteratively and stochastically. Each neighbor j of each node k in the cluster is added to the cluster with probability $p_{kj} = G_{kj}\delta(\gamma_k, \gamma_j)\lambda_{kj}$. The cluster $Cl(i)$ initially contains only the vertex i and is iteratively grown until no neighbor is available to be added to the cluster. $Cl(i)$ is therefore composed of nodes that have all the same gamma values as i. Each proposed move is $\boldsymbol{\gamma} \to \boldsymbol{\gamma}'$ with

$$\boldsymbol{\gamma}'_k = \begin{cases} \boldsymbol{\gamma}'_k + 1 \mod 2 & \text{if } k \in \text{Cl(i)} \\ \boldsymbol{\gamma}'_k & \text{otherwise.} \end{cases}$$

It is clear that if the randomly chosen variable i has no neighbors, it is the only one that is added to (if $\gamma_i = 0$) or removed from (if $\gamma_i = 0$) the present model to obtain the proposed model. In our implementation, we alternate a proposal to add variables to the model with a proposal to remove variables from it. The proposed configuration $\boldsymbol{\gamma}'$ is accepted with probability $F(z) = \min\{1, z\}$ where

$$z = \frac{p(\mathbf{Z}|\boldsymbol{\gamma}', \mathbf{J}, \mathbf{Y})}{p(\mathbf{Z}|\boldsymbol{\gamma}, \mathbf{J}, \mathbf{Y})} \cdot \rho^{-\sum_i(\gamma'_i - \gamma_i)} \tag{8}$$

for discrete $\mathbf{Y}$, and

$$z = \frac{p(\mathbf{Y}|\boldsymbol{\gamma}', \mathbf{J}, \mathbf{X})}{p(\mathbf{Y}|\boldsymbol{\gamma}, \mathbf{J}, \mathbf{X})} \cdot \rho^{-\sum_i(\gamma'_i - \gamma_i)} \tag{9}$$

for continuous $\mathbf{Y}$. For the above relations (8, 9) to hold true, one must choose the proposal probability $\lambda_{ij} = 1 - \exp(-J_{ij})$ because of equation (4) and the detailed balance condition. For vanishing values of J_{ij} the algorithm reduces to a single-variable updating, as in this case the network is effectively a collection of isolated vertices. Larger values of J_{ij} favor larger clusters, and for sufficiently large values, variables in the same connected component of the network will have the same values of γ. The parameter ρ instead discourages large

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

models. Thus, the choice of J and the choice of ρ affect the realizations of the model. When the network is very complex, it may seem preferable to stop the construction of the cluster $Cl(i)$ about the randomly selected variable i to include only its neighbors up to some distance. For example, one can add to the cluster only the nearest neighbors j of i with probability $\lambda_{ij} = 1 - \exp(-J_{ij})$ without iterating this procedure any further. In this case, the equations (8) and (9) do not hold true anymore. Indeed, the first factor of the prior (4) would give a contribution to the ratios z that would only be partially canceled by the kernel of the proposed move. Ideally, and more naturally, the same goal could be reached by modelling J with a distribution that decreases rapidly with the distance. The advantage of this collective updating algorithm over single updating algorithms is that very few steps are generally necessary to go from one configuration to an independent one. The Wolff algorithm can be viewed as a one-cluster variant of the Swendsen-Wang algorithm (1987), which was applied in variable selection in Nott and Green (2004), and has the advantage of being more easily implemented.

3 Numerical Examples

We present in this section some applications of the method to simulated and real data.

3.1 Simulated regulatory network

We have considered a simulation with $p = 399$ covariates, one continuous outcome Y and the network represented in Figure 1. The rectangular nodes represent the variables entering the simulated model, i.e., the variables that are related to the response. The regression coefficients were assigned values according to two different schemes. The assignment was completely random in a first set of simulations, with values drawn uniformly in the interval $\mathcal{I} = [-2, -0.5] \cup [0.5, 2]$. In a second set of simulations, the values were chosen in the same interval $\mathcal{I}$, but constrained in such a way that the top node in each group of defining variables had larger values than the rest. For both simulations, the variables X were drawn from a multivariate normal distribution with variance-covariance matrix $Cor(X_i, X_j) = 0.3^{|i-j|} + G_{ij}0.2^{|i-j|}$. This choice of variance-covariance matrix insures that neighboring variables are

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/upennbiostat/art34

a bit more correlated than non-neighboring variables, although the added correlation may be very small. The outcomes were sampled from a normal distribution centered about the linear predictor and with variance σ^2 such that the noise-to-signal ratio (NSR) for the data had values 0.1, 0.3, 0.4, 0.5, 1. We present the results of runs carried out using one of the data sets simulated in one of the two simulation schemes. Similar conclusions are valid for the other data. As one can expect, the lower the noise-to-signal ratio, the easier it is to select the true model. In the easiest case $NSR = 0.1$, all variables were selected but two (variable 300 and variable 302) and two false positives were identified using as a criterion that the posterior probability that a variable i is in the model, $P(\gamma_i = 1)$, is greater than 0.5. Increasing the latter value to 0.75, the two false positives disappear. The same results are obtained in this case if the network structure is ignored, *viz.* if one considers a network of isolated vertices. For $NSR = 1$, which is the hardest case, only one variable is selected (variable 302) when no network structure is used, while the network helps select few other variables, but at the same time some false positives. This is a pattern that was verified for other values of the noise-to-signal ratio as well: employing the network structure detects more variables of the true model at the expense of introducing some false positives. Table 1 summarizes the results for $NSR = 0.4$ and $NSR = 0.3$. Figure 2 shows the plots of the true positive rate versus the false positive rate for four values of NSR, which give further illustration of the advantage of employing the network structure. We observed that in general the areas under the ROC curves are higher when the network structures are utilized in the prior distribution and in the MCMC inferences.

3.2 Simulation based on a KEGG regulatory network

We also considered a data set with a discrete outcome and a more complicated network, with $p = 400$ nodes, which is represented, with the exception of some isolated nodes, in Figure 3. This network is a subset of a real KEGG network (Kanehisa and Goto, 2002) that was used in Wei and Li (2007; 2008). We sampled the coupling J for each edge from an exponential

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

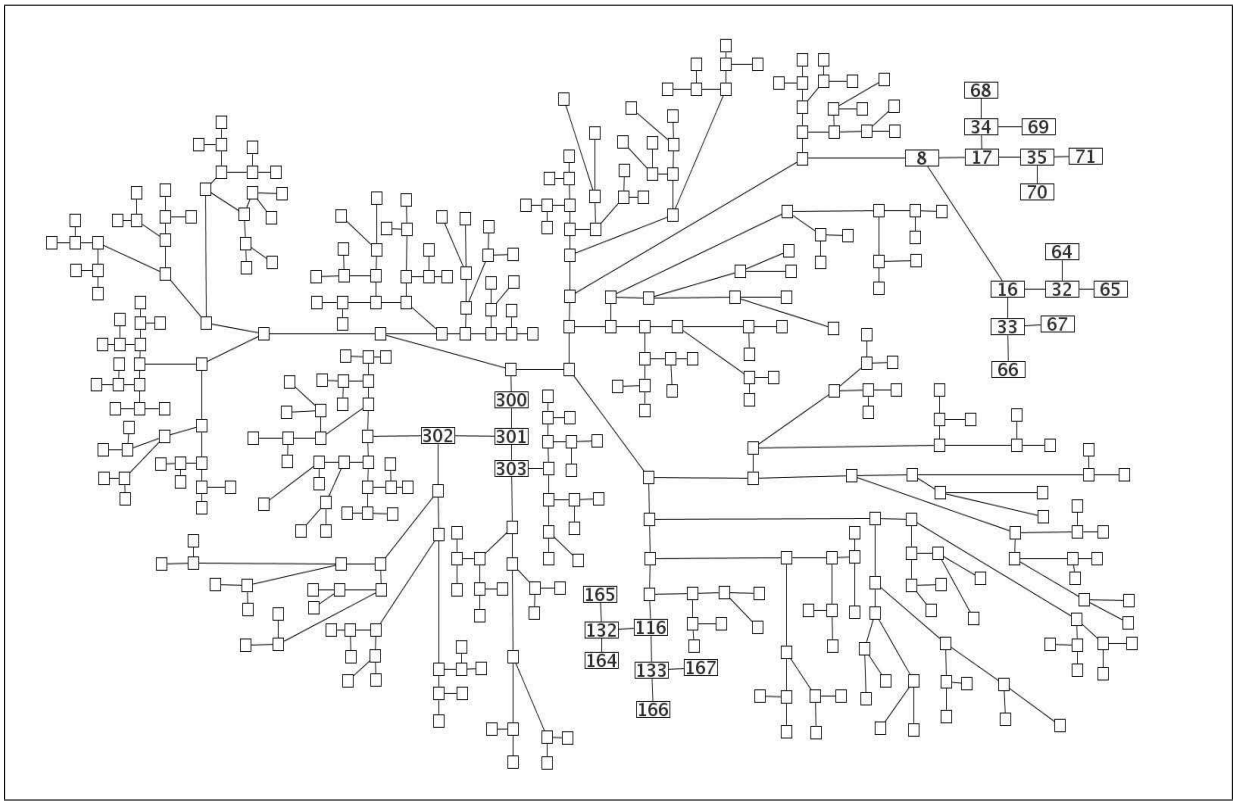


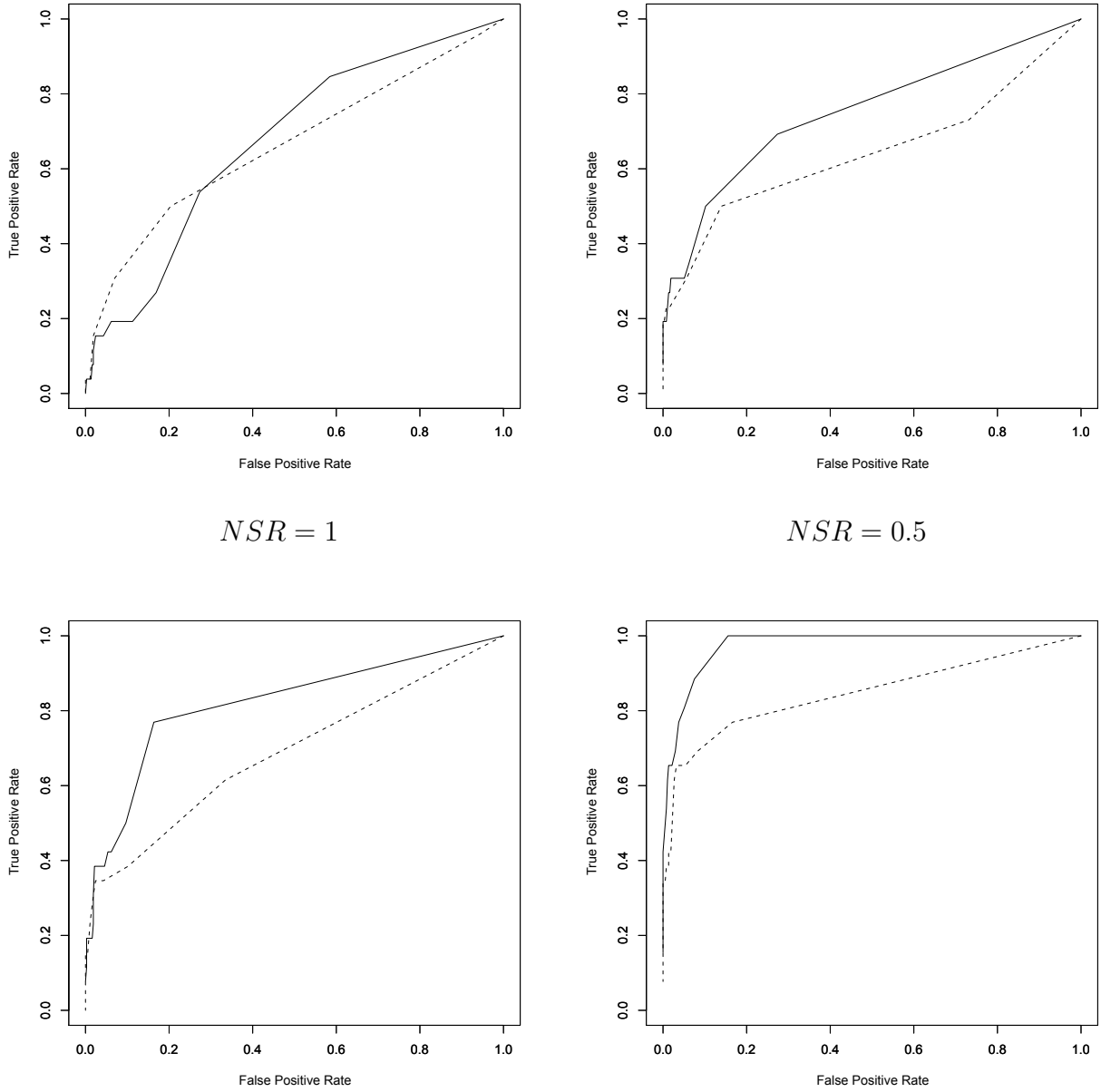
Figure 1: *Tree-structured network used for the first simulated data sets. Rectangular nodes represent the relevant variables*

Table 1: *Results of simulations using the network represented in Figure 1 or without using the network structure for two different noise-to-signal ratios (NSR). The true model consists of variables* $8, 16, 17, 32-35, 64-71, 116, 132, 133, 164-167, 300-303$. *The variables listed have a posterior probability of being present in the model greater than or equal to 0.5.*

	NSR = 0.4	NSR= 0.3
with network	8, 16, 17, 32, 33, 116	8, 16, 32, 116, 132
	132, 301, 303	164, 300, 301, 302
	(356)	(90,131,138, 168,261,298)
without network	8, 16, 17, 32, 116	8, 16, 132, 301
	132, 301 (122, 322)	(83)

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/upennbiostat/art34



$NSR = 1$ $NSR = 0.5$

$NSR = 0.3$ $NSR = 0.1$

Figure 2: *Regulatory network (tree) example. Results of simulations using the network structure or without using the network structure: true positive rates vs. false positive rates for the simulated data set for different values of the noise-to-signal ratio (NSR), with the solid lines represent results using the network structure and dashed lines are results without using the network structures.*

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

distribution (**??**), and, starting from a random assignment of values, we decided if a variable was in the model or not using the conditional distributions obtained from (4)

$$P(\gamma_i = 1|\cdots) = K \cdot e^{\sum_{j\neq i} G_{ij}\delta(\gamma_j,1)J_{ij}}\rho^{-1} \quad (10)$$

$$P(\gamma_i = 0|\cdots) = K \cdot e^{\sum_{j\neq i} G_{ij}\delta(\gamma_j,0)J_{ij}} \quad (11)$$

with

$$K^{-1} = e^{\sum_{j\neq i} G_{ij}\delta(\gamma_j,1)J_{ij}}\rho^{-1} + e^{\sum_{j\neq 1} G_{ij}\delta(\gamma_j,0)J_{ij}}$$

The variables selected are depicted as black nodes in Figure 3. We then sampled X_{ik} from a normal distribution with variance-covariance matrix $Cov(X_i, X_j) = G_{ij}/2, i \neq j$ for $k = 1, \ldots, N = 200$, and β, from a uniform distribution in the interval $[-5, -2] \cup [2, 5]$, rather than take them from a multi-variate distribution (see eq. (3)). Finally, for each i, we drew Z_i from (6) and took as Y_i its sign.

The variables identified by the algorithm are the square and rectangular nodes in Figure 3. The two shapes refer to two values of the posterior probabilities used as the criteria to identifying the selected variables, with the rectangles referring to a posterior probability of at least 0.5 and the squares of at least 0.4. We note that all the variables in the models that are isolated nodes have been omitted from Figure 3. Three of these variables enter the simulated models, two of which were correctly identified with a posterior probability greater than 0.5. Other runs gave similar results, with some variations in the variables selected in each true cluster.

3.3 Application to real data

Aging of human brain is one of the most complex biological processes. It is cause of cognitive decline in the elderly and a major risk factor in age-based degenerative diseases such as Alzheimer's. For this reason, uncovering the genetic underpinning of brain-aging has become the focus of recent research. Indeed, there have been a number of efforts to collect genetic data from brain tissue of individuals of different ages. In particular, Lu *et al.* (2004) gathered the transcriptional profiling of the human frontal cortex from 30 persons of age ranging from 26 to 106, using the Affymetrix HG-U95Av2 oligonucleotide arrays. In this

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/upennbiostat/art34

section, we present the results of an analysis carried out using these data. Specifically, we set out to identify which genes and which pathways were related to brain aging. To do this, we supplemented Lu's data with pathways information acquired from the KEGG data bases. We first constructed a (non-connected) 1668-node network by combining 33 KEGG regulatory pathways (Kanehisa and Goto, 2002), and then considered only those genes on the U95Av2 chip and those nodes in the network that overlapped and for which data were available on the entire cohort of 30 patients ($N = 30$). This resulted in $p = 1302$ genes and a network with 5258 edges. Our method could have also been applied to the entire genes on the U95Av2 chip by treating those genes as additional isolated nodes for which no pathways information was available. We log-transformed, centered and standardized the data. As responses we used the logarithm (in base 10) of the age. For this analysis, we fixed $J_{ij} = J \cdot |Corr(X_i, X_j)|$, so as to favor highly correlated variables that are connected in the network to be jointly in the model. The constants J and ρ were chosen so as to allow very high acceptance rates and reasonable model size. We have considered variables that have a posterior probability of being in the model greater than 0.5: $P(\gamma_i) \geq 0.5$. With this criterion, 44 variables were selected. Figure 4 depicts the subnetwork composed of vertices among this set, except for isolated vertices. There are a few interesting observations from these identified subnetworks. First, we identified a small subnetwork with 4 genes including Somatostatin gene (SST) and its receptors (SSTR4 and SSTR5) and another gene cortistatin (CORT) that also binds to the same receptors as SST. Somatostatin is an important regulator of endocrine and nervous system function (Yacubova and Komuro, 2002). Because its levels change with age, it is likely that age-related changes are affected or affect SST (Reed *et al.*, 1999). A role for SST in Alzheimer's disease has also been proposed (Saito *et al.*, 2005). Another interesting pair of genes, the complement component 1 inhibitor gene (SERPINGS) and the the complement component 1 (C1R), was also reported to be related to aging related phenotypes. For example, Ennis *et al.* (2008) identified an association between the SERPING1 gene and age-related macular degeneration using a two-stage case-control study. The selenium transport protein, selenoprotein P (SELP), and its ligand (SELPLG), are essential for neuronal survival and function and were reported to be associated with Alzheimer's pathology in human cortex (Bellinger *et al.*, 2008).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

4 Discussion and Future Direction

Motivated by the application of incorporating prior pathway and network structure information into the analysis of genomic data, we have considered Bayesian variable selection for both linear Gaussian models and probit models when the covariates are measured on a graph. In our approach, a flexible Markov random field prior that takes account of the graph structure is employed and a Markov chain sampler based on the Wolff algorithm is used. Our simulations indicate that incorporating the graph structure can lead to increased sensitivity in identifying the relevant variables. The algorithm performs better for continuous than for binary outcomes, as in the latter case sampling of the Gaussian latent variables $\mathbf{Z}$ is required. This paper focuses on how to utilize the prior genetic pathway and network information in the analysis of genomic data in order to obtain a more interpretable list of genes that are associated with the genotypes. An equally important topic is how to construct these pathways and networks. One area of intensive research in the last several years has been on estimating sparse Gaussian graphical models based on gene expression data (Gui and Li, 2007; Peng *et al.*, 2009). Although such models built from gene expression data can provide some information on how genes are related at the expression level, they hardly correspond to any of the real biological networks. The future will likely see more research on how to build meaningful biological networks by integrating various types of genomic data. This leads to great challenges due to both the complexity of the real biological networks and the high-dimensionality of the genomic data. Again, utilizing the prior network information in the framework of Bayesian analysis can lead to better network inference (Mukherjee and Speed, 2008). Alternative to the Gaussian graphical models, Bayesian networks provide more detailed information on causal relationship among genes based on various types of genomic data. However, the computation is even more challenging given the fact that a very large model space has to be explored and novel MCMC methods are required (Ellis and Wong, 2008). Finally, as more and more biological networks are accumulated, statistical methods for analysis of these large graphs are also needed. Some interesting problems include the identification of network modules and network motifs. Here as well, Bayesian approaches seem to provide important solutions to these problems (Monni and Li, 2007; Berg and Lassig,

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/upennbiostat/art34

2004, 2006).

Acknowledgments

This research was supported by NIH grants R01ES009911 and R01CA127334. We thank Professor Ming-Hui Chen and other editors for inviting us to contribute a chapter to this book in honor of Professor James O. Berger.

References

Albert, J.H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88, 669-679.

Bellinger, F.P., He, Q., Bellinger, M., Lin, Y., Raman, A.V., White, L.R. and Berry, M. (2008) Association of selenoprotein P with Alzheimers pathology in human cortex. *Journal of Alzheimers Disease*, 15: 465-472.

Berg, J. and and Lassig, M. (2004) Local graph alignment and motif search in biological networks. *Proceedings of the National Academy of Sciences*, 101(41), 14689-14694.

Berg, J. and and Lassig, M. (2006) Cross-species analysis of biological networks by Bayesian alignment. *Proceedings of the National Academy of Sciences*, 103 (29), 10967-10972.

Candes, E. and Tao T. (2007) The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35: 2313-2351.

Chipman, H., George, E.I., McCulloch, R.E. (2001) The practical implementation of Bayesian model selection, in *Model Selection*, P. Lahiri editor, IMS Lecture Notes, Volume 38, 67-134, Beachwood, OH

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression. *Annals of Statistics*, **32**, 407499.

Ellis, B., and Wong, W.H. (2008) Learning Causal Bayesian Network Structures From Experimental Data. *Journal of the American Statistical Association*, 103, 778-789.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Ennis, S., Jomary, C., Mullins, R., Cree, A., Chen, X., MacLeod, A., Jones, S., Collins, A., Stone, E., and Lotery, A. (2008) Association between the SERPING1 gene and age-related macular degeneration: a two-stage case-control study. *The Lancet*, 372: 1828-1834.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348-1360.

George, E.I. (2000) The variable selection problem. *Journal of the American Statistical Association*, 95, 13041308.

George, E.I. and McCulloch, R.E. (1993) Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88, 881-889.

George, E.I. and McCulloch, R.E. (1997) Approaches for Bayesian variable selection. *Statistica Sinica*, 7, 339-374.

Hans, C. Dobra, A., and West, M. (2007) Shotgun stochastic search in regression with many predictors. *Journal of the American Statistical Association*, 102: 507-516.

Kanehisa, M. and Goto, S. (2002) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28, 27-30.

Li, C., Li, H. (2008) Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24, 1175-1182.

Li, F. and Zhang, NR (2008) Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Technical report, Stanford University.*

Li, H. and Gui, J. (2006) Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics*, 7: 302-317.

Li, H., Wei, Z. and Maris, J. (2009): A hidden Markov model approach for genome-wide association studies. *Biostatistics*, in press.

Lu, T., Pan, Y., Kao, S.-Y., Li, C., Kohane, I., Chan, J. and Yankner, B.A. (2004) Gene regulation and DNA damage in the aging human brain. *Nature*, 429, 883-891.

Nott, D.J., Green, P.J. (2004) Bayesian variable selection and the Swendsen-Wang algorithm.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/upennbiostat/art34

Journal of Computational and Graphical Statistics, 13: 141-157.

Monni, S. and Li, H. (2008) Vertex clustering of graphs using reversible jump MCMC. *Journal of Computational and Graphical Statistics*, 17(2): 388-409.

Mukherjee, S. and Speed, T.P. (2008) Network inference using informative priors.*Proceedings of the National Academy of Sciences*, 105(38): 14313-14318.

Park, T. and Casella, G. (2008) The Bayesian Lasso. *Journal of the American Statistical Association*, 103: 681- 686.

Reed, D.K., Korytko, A.I., Hipkin, R.W., Wehrenberg, W.B., Schonbrunn, A., Cuttler, L. (1999) Pituitary somatostatin receptor (sst)1-5 expression during rat development: age-dependent expression of sst2. *Endocrinology*, 140:4739-4744

Peng, J., Wang, P., Zhou, N. and Zhu, J. (2009) Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, in press.

Swendsen, R. H., and Wang, J. (1987): Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letter*, 58(2):8688.

Saito, T., Iwata, N., Tsubuki, S., Takaki, Y., Takano, J., Huang, S.M., Suemoto, T., Higuchi, M., Saido, T.C. (2005) Somatostatin regulates brain amyloid beta-peptide A-beta-42 through modulation of proteolytic degradation. *Nature Medicine*, 11:434-439.

Smith, M., and Kohn, R. (1996) Non-parametric regression using Bayesian variable selection. *Journal of Econometrics* **75** 317-344

Tai, F. and Pan, W. (2009) Bayesian variable selection in regression with networked predictors. *Manuscript.*

Tibshirani, R.J. (1995) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, **58**, 267-288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005) Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society B*, **67**, 91-108.

Wei, Z. and Li, H (2007) A Markov Random Field Model for Network-based Analysis of Genomic Data. *Bioinformatics*, 23: 1537-1544.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Wei, Z. and Li, H.(2008) A Hidden Spatial-temporal Markov Random Field Model for Network-based Analysis of Time Course Gene Expression Data. *Annals of Applied Statistics*, 2(1), 408-429.

Wolff, U. (1989) Collective Monte Carlo Updating for Spin Systems, *Physical Review Letters*, 62: 361,

Yacubova, E. and Komuro, H. (2002) Stage-specific control of neuronal migration by somatostatin *Nature*, 415:77-81.

Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, 68: 49-67.

Zellner, A. (1986) On assessing prior distributions and Bayesian regression analysis with g-prior distribution. In *Bayesian Inference and Decision Techniques. Essays in Honor of Bruno De Finetti*, Goel, P.K. and Zellner, A. Editors, pp. 233-243, North Holland, Amsterdam

Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418-1429.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67, 301-320.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/upennbiostat/art34

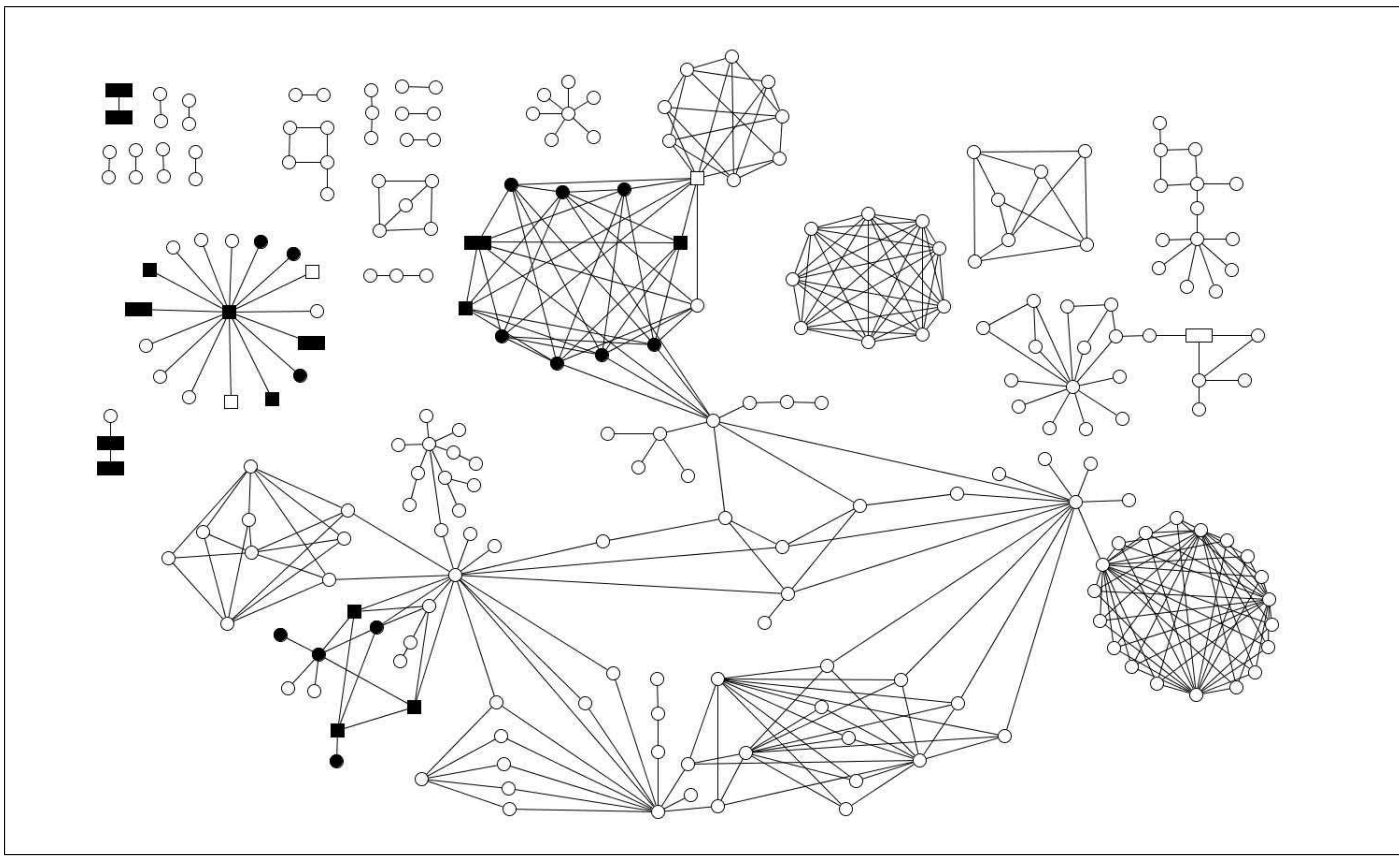


Figure 3: *A subset of the KEGG network used for the second simulated data set. The black nodes are the variables of the simulated models. Variables inferred with a posterior probability of 0.5 or greater are represented as rectangles and those that have a posterior probability of 0.4 or greater as squares. Isolated nodes are not represented in the pictures. Three isolated nodes enter the simulated model and two were inferred with a posterior probability greater than 0.5.*

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

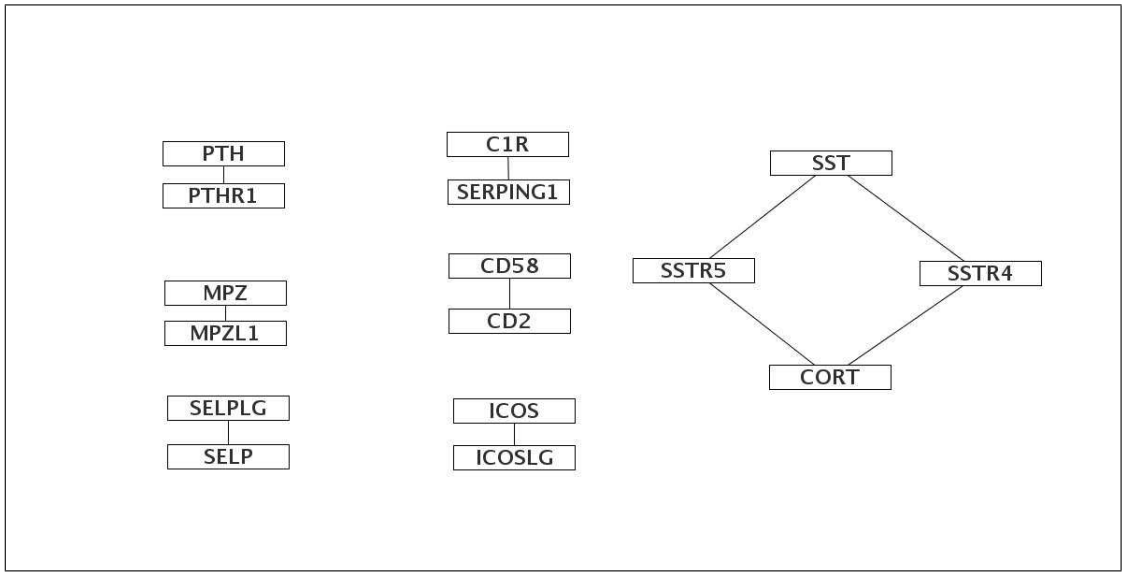


Figure 4: *A subnetwork of the KEGG network obtained by considering only vertices that represent variables inferred with a posterior probability of 0.5 in the real data analysis. Isolated nodes are not represented in the picture.*

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/upennbiostat/art34