Universitá degli Studi di Firenze
Dipartimento di Statistica "G. Parenti"
Dottorato di Ricerca in Statistica Applicata
XXII ciclo - SECS-S/01

Bayesian Methods for Data Integration with Variable Selection: New Challenges in the Analysis of Genomic Data

Francesco Claudio Stingo

Tutor: **Prof. Giovanni M. Marchetti**
Co-tutor: **Prof. Marina Vannucci**
Co-tutor: **Prof. Elena Stanghellini**

Coordinatore: **Prof. Fabio Corradi**

Abstract

This thesis contributes to the broad field of Bayesian variable selection. Motivated by real problems in the analysis of genomic data we build new statistical models that answer important questions and help to explain underlying biological processes.

In the first project we propose a novel method that provides insights in the analysis of genomic data in situations where the target of the analysis is to find which genes (predictors) are related to a specific phenotype, the response variable. Our model considers, first, that it has been clearly recognized that genes are biological elements that affect each other and, second, that recently interest in biology has moved from the analysis of single genes to the analysis of known groups of genes, called pathways. We build upon Bayesian variable selection methods for linear/generalized linear models by adding two novel features into the model: first we are able to incorporate information on gene networks in our prior formulation, second we develop stochastic search methods that are able to identify both pathways and pathway elements involved in particular biological processes. To achieve these goals we define a new Ising-type prior on the latent indicators of genes included into the model. In addition, we define pathway scores that synthesize the activity of each pathway via partial least square techniques. We construct a Monte Carlo Markov chain scheme with a double layer of selection indicators, one for genes and one for pathways, that takes into account a set of constraints for both identifiability of the model and biological interpretation. The method we put forward represents a consistent framework for testing whether a pathway is significantly related to a phenotype without using a 2-step procedure whose statistical properties, like the significance level of 2 hypothesis testing performed in sequence, would be difficult to investigate. We use our method both with simulated data and on an application to gene

expression data with censored survival outcomes. In addition to a better understanding of the underlying molecular processes, our method also improves on prediction accuracy.

In the second project we construct a statistical procedure to infer a biological network of very high dimensionality, where microRNAs, small RNAs, are supposed to down-regulate mRNAs, also called target genes. The main goal of the model is to understand which elements are connected and which ones are not. In addition, specific biological characteristics/constraints need to be considered. From a statistical point of view, we address this problem by building a network that represents the biological regulatory system, indicating which microRNA regulates which gene. In particular, we provide a novel graphical modeling approach that includes constraints on the regression coefficients to take into account the down-regulatory effect of the network. Our approach is able to select single connections in the network, unlike previous methods in the Bayesian variable selection literature, which only allow the selection of covariates (microRNAs) that affect either all the genes or none of them. The main challenge of this project is represented by the dimensionality of the data. The network is potentially formed by more than 30,000 connections and the data are formed by 12 observations. By developing a stochastic search variable selection type of algorithm we are able to efficiently explore the space of all possible networks and to find, for each gene, which microRNAs have high posterior probability of being down-regulating the gene. To help the selection, we also propose a new prior formulation which is able to integrate different sources of data, by exploiting information from previous sequence and structure analysis. Because we integrate many sources of information, our model is also able to determine which information is consistent with the data via posterior inference on the parameters defined in our data-integration prior. The proposed method is general and can be easily applied to other types of network inference by integrating multiple data sources.

List of Figures

1.1 An example of undirected graphical model 8
1.2 The *graphical model* structure of the linear regression model. . 13

2.1 Directed acyclic graph for model – observables are represented with squares and parameters with circles 25
2.2 Proposed move types: (1.i) add a pathway and a gene (+,+); (1.ii) remove a pathway and a gene (-,-); (2.i) add a gene in an existing pathway (n,+); (2.ii) remove a gene from an existing pathway (n,-); (3.i) add a pathway without touching genes (+,n); (3.ii) remove a pathway without touching genes (-,n). . 36
2.3 Marginal posterior probabilities for pathway selection, $p(\theta_k|\boldsymbol{T}, Y)$, for the three simulated data sets. The open circles indicate the four relevant pathways. 40
2.4 Conditional posterior probabilities, $p(\gamma_j|\boldsymbol{T}, Y, I\{\sum_{k\in\mathcal{P}} \theta_k s_{kj} > 0\})$, for gene selection for the three simulated data sets. The open circles indicate the genes used to generate the outcome variable. 42
2.5 Trace plot of the number of included pathways. 45
2.6 Trace plot of the number of included genes. 45
2.7 Scatterplot of $p(\theta_k|\boldsymbol{T}, Y)$ across the two MCMC chains. 45
2.8 Microarray data: Marginal posterior probabilities for pathway selection, $p(\theta_k|\boldsymbol{T}, Y)$. The 11 pathways with largest probabilities are marked with symbols. 46

2.9 Microarray data: Conditional posterior probabilities for gene selection, $p(\gamma_j|\boldsymbol{T}, Y, I\{\sum_{k\in\mathcal{P}}\theta_k s_{kj} > 0\})$. The 102 probes with largest probability that belong to the 11 selected pathways in Figure 2.8 are marked with $\triangle$. 47
2.10 Microarray data: Graphical representation of a subset of selected pathways and genes with their corresponding islands. The genes in the islands are listed in Table 2.2. 49

3.1 Graphical representation of the miRNA regulatory network. . 61
3.2 Structure of the graphical model 63
3.3 Trace plot for number of selected arrows and for the log-posterior probability for the time invariant model 74
3.4 Selected network for the control group using a threshold of 0.8 on the posterior probability 77
3.5 Estimation of the significant β_{gm} for the time invariant model under hyperthermia condition 79
3.6 Density Kernel estimate using the not time dependent model for the control group 80

List of Tables

2.1 Conditional and marginal posterior probabilities for pathway and gene selections, $p(\gamma_j = 1|\boldsymbol{T}, Y, I\{\sum_{k\in\mathcal{P}} \theta_k s_{kj} > 0\})$ and $p(\theta_k = 1|Y, \boldsymbol{T})$. 43

2.2 Selected 102 genes divided by islands with associated pathway indices (in parenthesis). The pathway indices correspond to: 1-Purine metabolism, 2-MAPK signaling pathway, 3-Cytokine-cytokine receptor interaction, 4-Neuroactive ligand-receptor interaction, 5-Cell cycle, 6-Axon guidance, 7-Cell adhesion molecules (CAMs), 8-Complement and coagulation cascades, 9-Regulation of actin cytoskeleton, 10-Insulin signaling pathway, 11-Pathways in cancer. The threshold for pathway inclusion was set to a marginal posterior probability of 0.45. . . . 51

Contents

1 Introduction **6**
1.1 Brief Introduction to Graphical Models 6
1.2 Stochastic Search Variable Selection 10
1.3 Description of the Projects 17

2 Bayesian Models for Pathway and Gene selection **21**
2.1 Introduction 22
2.2 Model Specification 25
2.2.1 Regression on Latent Measures of Pathway Activity . . 26
2.2.2 Models for Categorical or Censored Outcomes 28
2.2.3 Prior for Regression Parameters 30
2.2.4 Priors for Pathway and Gene Selection Indicators . . . 31
2.3 Model Fitting 34
2.3.1 Marginal Posterior Probabilities 34
2.3.2 MCMC Sampling 35
2.3.3 Posterior Inference 37
2.4 Application 38
2.4.1 Simulation Studies 38
2.4.2 Application to Microarray Data 43
2.5 Discussion 50

3 Bayesian Network for Genomic Data Integration **54**
3.1 Introduction 55

3.2 Neural Tube Defects 57
3.2.1 miRNA Expression Levels 58
3.2.2 Target Prediction via Sequence and Structure Information 59
3.2.3 Target mRNA Expression Levels 59
3.3 Model 60
3.3.1 A Bayesian Network for Gene & miRNA Expression . . 60
3.3.2 Prior Model for Variable Selection 64
3.3.3 Using Association Scores in the Prior Model 65
3.3.4 Time-dependent Coefficients Model 66
3.4 Posterior Inference 68
3.5 Neural Tube Defects Application 72
3.5.1 Parameter Settings 72
3.5.2 Results 74
3.6 Conclusions 80

Chapter 1

Introduction

In this Chapter we introduce two topics, graphical models and Bayesian variable selection, which will be extensively used in the following chapters. First, in Section 1.1, we describe graphical models, both directed and undirected, focusing on Gaussian graphical models and graphical models for binary variables. In Section 1.2 we review Bayesian approaches for variable selection in linear models; we describe different specification for the hierarchical model and various approaches for posterior inference, including several Monte Carlo Markov Chain (MCMC) strategies. Finally we briefly describe the projects that will form the following chapters, by giving a general idea of the problems addressed and indicating the main features.

1.1 Brief Introduction to Graphical Models

Graphical models, see Whittaker (1990) and Cowell et al. (1999) among others, are methods and techniques that use Graph Theory to model the relationships between random variables. A graph is formed by nodes and arcs; nodes represent random variables and the lack of arcs represent conditional independence assumptions. Hence they provide a compact representation of joint probability distributions. A graph is consistent with the conditional

independence statements expressed in the distribution. However note that it is not always possible to obtain a perfect map; this means that some distributions characterized by specific conditional independencies cannot be represented using a graphical model.

Arcs can be undirected (used for symmetric dependencies) or directed (used when there is a direction in the dependence). These dependencies can come from prior knowledge or from data analysis. Undirected graphical models, also called Markov Random Field (MRF), have a simple definition of independence: e.g., two sets of nodes A and B are conditionally independent given a third set, C, if all paths between the nodes in A and B contain at least one node belonging to C.Figure 1.1 displays an example of graphical model. In a graph it is possible to describe the conditional independencies among the variables using three equivalent *Markov* properties. The Pairwise Markov property says that two variables that are not neighbors in the graph are conditional independent given all the other variables. The Local Markov property says that every variable X_i is independent of the variables not in $\text{cl}(X_i)$ conditional on the set of variables directly connected to X_i (boundary of X_i), where $\text{cl}(X_i)$, the closure of X_i, is the set of nodes formed by X_i and its boundary. The Global Markov property says that two sets of variables B and C that are not connected are independent given a third set of variables S formed by all the variables that separate B and C. Some of the conditional independencies that we can gather from figure 1.1 are:

- $X_1 \perp\!\!\!\perp X_4 | (X_2, X_3, X_5)$ - Pairwise Markov,
- $X_1 \perp\!\!\!\perp (X_4, X_5) | (X_2, X_3)$ - Local Markov,
- $(X_1, X_2) \perp\!\!\!\perp X_5 | (X_3, X_4)$ - Global Markov ($B = \{X_1, X_2\}$, $C = \{X_5\}$, $S = \{X_3, X_4\}$).

Directed graphical models, also called Bayesian Network (BN), need a specific ordering of the variables. Since we do not allow for the presence of cycles we work with directed acyclic graph (DAG). Conditional independencies in

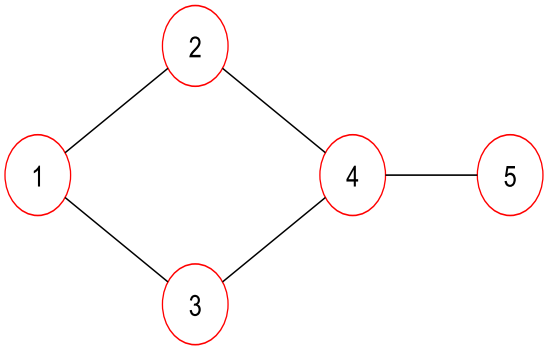

Figure 1.1: An example of undirected graphical model

a DAG depend on the ordering of the variables. When the joint distribution is a multivariate normal the model is called Graphical Gaussian model (GGM). The graph $\mathcal{G}$ and the covariance matrix Σ entirely define a GGM $\mathcal{M}$, $\mathcal{M} \equiv (\mathcal{G}, \Sigma)$. Nodes that are directly connected to node j and precede j in the ordering are called parents of j , $pa(j)$. In a Bayesian Network, X_j is independent, given its parents, of the set of all the other variables in the graph, except its parents.

In Chapter 2 a graphical model with binary random variables is used. This type of model, called Ising model, is widely used in statistics. An Ising model is a system with 2^p states, where p indicates the number of variables of which corresponding graph is a lattice system. The variables x_j, with $j = 1, \ldots, p$, can assume only two values, $x_j \in \{0, 1\}$, and their probability distribution is assumed to be:

$$P(\mathbf{x}) = \frac{\exp(\mathbf{d}'\mathbf{x} + \mathbf{x}'\mathbf{G}\mathbf{x})}{Z} \tag{1.1}$$

with $\mathbf{d} = d\mathbf{1}_p$, $\mathbf{1}_p$ the unit vector of dimension p and $\mathbf{G}$ a symmetric matrix with elements $\{g_{ij}\}$ usually set to some constant g; if nodes i and j are not connected in the graph the corresponding g_{ij} is equal to zero. Z is usually called partition function and, except for the 1-D Ising model, where each

node has a maximum of two neighbors, it is not possible to calculate its value in closed form. To understand the role of the parameters d and $\mathbf{G}$ it is helpful to look at the conditional probability distribution

$$P(x_j|x_i, i \in N_j) = \frac{\exp(x_j(d + \sum_{i \in N_j} g_{ij} x_i))}{1 + \exp(d + \sum_{i \in N_j} g_{ij} x_i)}. \tag{1.2}$$

The larger d the higher the probability assigned to configurations with most of the x_j's equal to 1. Chapter 2 explains how this parameter is linked to the concept of sparsity. The hyperparameters g_{ij}'s represent the prior belief on the strength of coupling between the pairs of neighbors (i, j). This parametrization has been recently used in the contest of variable selection by Li & Zhang (2009).

A problem faced by Li & Zhang (2009) is the phase transition, that is, the expected number of variables equal to 1 can increase massively for small increments of $\mathbf{G}$. This problem can happen because equation (1.2) can only increase as a function of the number of x_j's equal to 1. In Chapter 2 we adopt an alternative parametrization that has been used in statistics, at least, since the work of Besag (1974), see also Besag (1986), and that allows us to avoid the phase transition problem. Recently this distribution has been used by Wei & Li (2007) and Wei & Li (2008) in the context of gene expression analysis for a frequentist multiple hypothesis testing procedure; they refer to this model as a Markov Random Field (MRF). Here we assume that the global distribution on $\mathbf{x}$ is given by

$$P(\mathbf{x}) \propto \exp(\mu\ n_1 - \eta\ n_{01}), \tag{1.3}$$

where n_1 is the number of x_j's equal to 1 and n_{01} is the number of edges linking variables with different values,

$$n_1 = \sum_{j=1}^{p} x_j, \qquad n_{01} = \frac{1}{2} \sum_{i=1}^{p} \left[\sum_{j=1}^{p} r_{ij} - \left| \sum_{j=1}^{p} r_{ij}(1 - x_i) - \sum_{j=1}^{p} r_{ij} x_j \right| \right],$$

where r_{ij} is the generic element of the $p \times p$ matrix R that describes edges of the graphical model, with $r_{ij} = 1$ if x_i and x_j have a direct link in the network, and $r_{ij} = 0$ otherwise. The corresponding conditional distribution is

$$P(x_j|x_i, i \in N_j) = \frac{\exp(x_j F(x_j))}{1 + \exp(F(x_j))}, \tag{1.4}$$

where $F(x_j) = \mu + \eta \sum_{i \in N_j}(2x_i - 1)$ and N_j is the set of direct neighbors of x_j in the graph. The parameter μ controls the sparsity of the model, with higher values of η encouraging neighboring variables to take on the same x_j value. For variables with no neighbors, this distribution reduces to an independent Bernoulli with parameter $p = \exp(\mu)/[1 + \exp(\mu)]$, which is a logistic transformation of μ.

1.2 Stochastic Search Variable Selection

This Section is a brief introduction to the general theory of the Stochastic Search Variable Selection (SSVS) method introduced by George & McCulloch (1993) in the linear regression framework and subsequently adapted to other modeling settings by many other authors, see for example Brown et al. (1998*b*) and Sha et al. (2004).

This method allows to select the "best" subset of covariates from the 2^p possible models in a linear framework where Y is the dependent variable and $X_1, \ldots, X_p$ are the potential predictors. This approach has been developed to handle situations where the number of variables p is bigger, and often much bigger, than the number of observations n. In such situations most of the standard methods can not be used. Note, for example, that the matrix $X'X$ is not invertible. By imposing a hierarchical Bayes mixture prior on the regression coefficients, this procedure puts a probability distribution on the set of all possible regression models and then uses a MCMC algorithm (Gibbs sampler in the original paper and Metropolis-Hastings in most of the recent papers) to 'run' through this set. "Best" models are clearly

those with high probability. The main property of this procedure is that, in a high-dimensional framework, the MCMC can be used to obtain a sample from the posterior distribution quickly and efficiently; this avoids the overwhelming (and often mathematically or computationally impossible) burden of calculating the posterior probabilities of all subsets.

The linear regression model is

$$Y \sim N_n(X\beta, \sigma^2 I_n)$$

where Y is a $n \times 1$ vector, X is a $n \times p$ matrix and β is a $p \times 1$ vector of regression coefficients. Selecting a subset of predictors is equivalent to setting to zero the elements of β corresponding to the excluded predictors. In the milestone paper of George & McCulloch (1993), the prior on β is a mixture of two normal distributions, the first one with most of its mass concentrated about zero and the second one with its mass spread out over plausible values. Using the latent variable $\gamma_i = 0, 1$, the prior of each element of β can be expressed as

$$P(\beta_i|\gamma_i) = (1 - \gamma_i)N(0, \tau_i^2) + \gamma_i N(0, c_i^2\tau_i^2).$$

The hyperparameter τ_i is set small and c_i^2 is set large so that $N(0, \tau_i^2)$ is concentrated about zero and $N(0, c_i^2\tau_i^2)$ is diffuse. With this prior, if $\gamma_i = 0$ then β_i is so small that it could be estimated by 0, whereas if $\gamma_i = 1$ a nonzero estimate of β_i corresponds to an important predictor. In this model the β_i are independent given the vector $\gamma = (\gamma_1, \ldots, \gamma_p)$. The prior on γ is the product of p independent Bernoulli

$$P(\gamma) = \prod_{i=1}^{p} w_i^{\gamma_i}(1 - w_i)^{(1-\gamma_i)}$$

with $w_i = P(\gamma_i = 1)$. This probability can be interpreted as the prior probability that X_i should be included in the model; the set of parameters w_i regulates the sparsity of the model, defining the a priori expected number

of included regressors. The prior on σ^2 is the conjugate Inverse Gamma

$$\sigma^2 \sim InvGa(\nu/2, \nu\lambda/2)$$

and priors on β and σ^2 are treated as independent distributions.

For posterior inference, the posterior distribution of γ,

$$P(\gamma|Y) \propto P(Y|\gamma)P(\gamma),$$

allows to identify the "best" models, that is those most supported by the data and by the prior distribution. The main target of the Gibbs sampler is to generate the sequence

$$\gamma^{(0)}, \gamma^{(1)}, \gamma^{(2)}, \ldots$$

which converges in distribution to $P(\gamma|Y)$. The algorithm does not need to explore the entire distribution to find the most probable models, because many models have small posterior probability and can be ignored. This is due to the idea of 'sparsity', that relates to situations where many of the possible relations (the coefficients in our case) are so small to be practically zero or represent only noise. In other words, we think here of the 'true' model as being sparse.

The above sequence can be obtained by applying the Gibbs sampler to the complete posterior $P(\beta, \sigma^2, \gamma|Y)$; the output is the full sequence of parameter values

$$\beta^{(0)}, \sigma^{(0)}, \gamma^{(0)}, \beta^{(1)}, \sigma^{(1)}, \gamma^{(1)}, \ldots$$

a Markov chain generated by the full conditional distributions $P(\beta|\sigma^2, \gamma, Y)$, $P(\sigma^2|\beta, \gamma, Y)$ and $P(\gamma_i|\beta, \sigma^2, \gamma_{-i}, Y)$, with $\gamma_{-i} = (\gamma_1, \gamma_2, \ldots, \gamma_{i-1}, \gamma_{i+1} \ldots, \gamma_p)$. From Figure 3.2, representing the graphical structure of the model, we see that the choice of the prior distributions allows to obtain the full conditionals in closed form leading to the following simplifications:

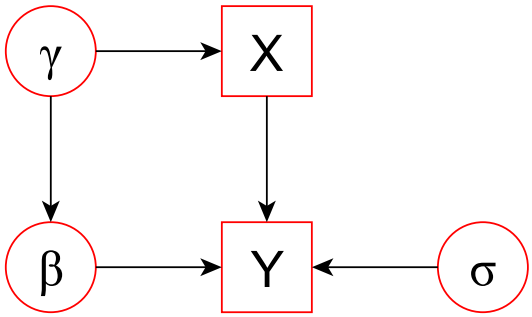

Figure 1.2: The *graphical model* structure of the linear regression model.

$$P(\sigma^2|\beta,\gamma,Y) = P(\sigma^2|\beta,Y)$$
$$P(\gamma_i|\beta,\sigma^2,\gamma_{-i},Y) = P(\gamma_i|\beta,\gamma_{-i}),$$

since the full conditional of any node depends only on the values of its parents, children and co-parents, where 'co-parents' are other parents of the children of the node (see Figure 3.2).

Smith & Kohn (1996) have extended this variable selection procedure to semiparametric additive models reframing the problem in terms of a linear model. The main feature of their procedure is the implicit introduction of a *spike and slab* prior, explicitly introduced and analyzed by George & McCulloch (1997), for the regression coefficients

$$\beta_i|\gamma_i,\sigma^2 \sim (1-\gamma_1)\delta_0 + \gamma_i N(0,\sigma^2 c x_{ii}),$$

where δ_0 is a Dirac's delta concentrated at 0 and x_{ii} is the *i-th* element of the diagonal of $(X'X)^{-1}$. They then specify a g-prior on the vector of the selected regression coefficients $\beta|\gamma,\sigma^2 \sim N(0,\sigma^2 c(X'X)^{-1})$, while the non-selected β's are excluded from the model. This setting leads to a faster computing algorithm since we are able to integrate β and σ^2 out from the

model. The only parameter sampled in the corresponding MCMC is then γ and, consequently, this algorithm needs to explore only the model space represented by $P(\gamma|Y)$. Smith & Kohn (1996) used a Gibbs sampler where at each step they sample, for $i = 1, \ldots, p$, from the full conditional distributions

$$P(\gamma_i = 1|\gamma_{-i}, Y) = \frac{1}{1+h}$$

where

$$h = \frac{1-w_i}{w_i}(c+1)^{1/2}(S(\gamma^1)/S(\gamma^0))^{n/2}$$

and $\gamma^1 = (\gamma_1, \ldots, \gamma_{i-1}, \gamma_i = 1, \gamma_{i+1}, \ldots, \gamma_p)$, $\gamma^0 = (\gamma_1, \ldots, \gamma_{i-1}, \gamma_i = 0, \gamma_{i+1}, \ldots, \gamma_p)$, $S(\gamma) = Y'Y - \frac{c}{1+c}Y'X(X'X)^{-1}X'Y$.

After obtaining a sample from the posterior distribution of γ given the data D, and the associated posterior probabilities $P(\gamma|D)$, two strategies for posterior inference are possible. The first approach uses only the model with the highest posterior probability, while the second approach, proposed by Madigan & Raftery (1994) and Madigan & York (1995), uses a set of models with high posterior probability. This set can include all the models visited in the MCMC; this procedure gives a good approximation of the marginalization over every possible model and is directly linked to an exploration algorithm of the model space. Madigan & York (1995) propose a method, in the context of model selection for discrete graphical models, to perform posterior inference for a quantity of interest Δ under model uncertainty. Because the procedure is not specific to the selection of a regressor in a linear model they do not use the parameter γ but indicate with M_k one of the possible models belonging to class of models under consideration $\mathcal{M}$. The model averaging posterior inference for Δ under model uncertainty is then performed using:

$$P(\Delta|D) = \sum_{k=1}^{K} P(\Delta|M_k, D)P(M_k|D). \tag{1.5}$$

This formula averages the posterior distribution under each of the models weighted by their posterior probabilities. To obtain the probabilities $P(M_k|D)$, at least for a significant subset of models, Madigan & York (1995) propose a procedure called Markov Chain Monte Carlo model composition (MC^3). Integrating out all the parameters and defining, implicitly, a flat prior on the model space, the MC^3 consists of a Metropolis-Hasting algorithm to explore the posterior distribution $P(M|D)$, where at each step a new model, that differs from the model selected at the previous step only for the inclusion or exclusion of an edge, is proposed.

Brown et al. (1998*a*) adapted and perfected the MC^3 algorithm in the context of linear regression. Starting from a parametrization similar to Smith & Kohn (1996), after integrating out the model parameters β and σ^2, it is possible to define an MC^3-type procedure for γ. First a value for γ is randomly chosen and then moves through a sequence of further values of γ are performed, with each step in the sequence having an element of randomness. At each point in the sequence a new candidate γ is generated by randomly modifying the current one. If the new candidate has a higher probability than the current one, then the chain will move to the new configuration γ'. If not, then the move is still possible, but now only with a certain probability. This feature explains why this algorithm is also called random search. Brown et al. (1998*a*), specifically, define two types of moves:

1. (Adding or deleting) Choose one of the p covariates at random. If the variable is currently in the model, delete the variable; if it is not currently in the model, add it to the model. Thus the new candidate γ' differs from the previous γ in one of its entries.

2. (Swapping) Choose at random one of the currently included variables and at random one of the currently excluded covariates. Exclude the previously included variable from the candidate model and include the previously excluded covariate.

By indicating with $g(\gamma) = P(\gamma)P(Y|X,\gamma)$, the new candidate model γ' is accepted with probability

$$\min\left[\frac{g(\gamma')}{g(\gamma)}, 1\right].$$

Note that the proposal distribution associated to this MCMC scheme is symmetric and does not appear in the acceptance ratio. In every model where it is possible to integrate out all the parameters but γ, this Metropolis algorithm is preferred to the Gibbs sampler because it allows a faster exploration of the space of the relevant models.

Several different MCMC schemes have been proposed to achieve a faster exploration of the posterior distribution $P(\gamma|Y)$. Recently Bottolo & Richardson (2009) have proposed and evolutionary Monte Carlo scheme combined with a parallel tempering approach that allows the algorithm to explore the model space faster by avoiding to remain stuck in local modes. Beside the different ideas combined in Bottolo & Richardson (2009), an interesting features of their algorithm is that they use an efficient Gibbs sampling. In Bayesian variable selection the Metropolis algorithm of Brown et al. (1998*a*) is usually preferred to the Gibbs sampler because, using the latter, at each iteration the algorithm has to go through all the full conditionals of every variable indicator, sampling each element of the vector γ. Therefore most of the non selected variables at the previous step will not be included into the model because completely unrelated to Y. Bottolo & Richardson (2009) have built an MCMC scheme that does not need to sample all the indicators at each step, skipping a lot of heavy computations related to non significative variables.

Brown et al. (1998*b*) have generalized the SSVS to multivariate regression model with q response variables. Defining the SSVS procedure requires the introduction of matrix variate distributions. Following Dawid (1981), the notation $Y - M \sim \mathcal{N}(\Gamma, \Sigma)$ indicates a $n \times q$ normal matrix-variate where M

indicates the mean and $\gamma_{ii}\Sigma$ and $\sigma_{jj}\Gamma$ indicate the covariance matrices of the *i-th* row and *j-th* column, respectively. The data model, with p covariates, q response variables and n independent samples can be defined as $Y - \mathbf{1}\alpha' - X\mathbf{B} \sim \mathcal{N}(I_n, \Sigma)$. Using the following priors

$$\begin{aligned} \alpha - \alpha_0 &\sim \mathcal{N}(h, \Sigma) \\ \mathbf{B} - \mathbf{B}_0 &\sim \mathcal{N}(H, \Sigma) \\ \Sigma &\sim \mathcal{IW}(\delta, Q) \end{aligned}$$

it is possible to integrate all three parameters out from the model. This feature, jointly with a QR deletion-addition algorithm for fast updating in the calculation of the marginal likelihood, leads to a very efficient Gibbs MCMC scheme for posterior inference. Note that Brown et al. (1998*b*) specify a latent $p \times 1$ vector indicator for the inclusion of the covariates, with the *j-th* element equal to 1 if the *j-th* covariate is significant for all q response variables. Consequently, it is not possible to define different sets of significant covariates for different response variables. Moreover, Brown et al. (1998*b*) use the *model averaging* idea of Madigan & York (1995) for prediction of new observations Y_f. This procedure is based on the predictive distribution $p(Y_f|Y, X_f)$ and exploits the conjugacy of the model; after integrating α, $\mathbf{B}$ and Σ out it is possible to calculate Y_f as weighted mean of the expected values of $p(Y_f|Y, X_f)$ given different configurations of γ, with the weights being the posterior probabilities of these configurations. Only the best k configurations, according to the posterior probabilities, are used for prediction.

1.3 Description of the Projects

This thesis is composed by two different projects that share some basic features:

1. They are motivated by new challenges in the analysis of genomic data;
2. They relate to Bayesian models for variable selection;

3. They are built for *p much bigger than n* problems;

4. They integrate different sources of data.

In Chapter 2, starting from the classical problem of finding genes that relate to a response variable, we build a model that takes into account both the dependence structure among the genes and the fact that genes usually work in groups, called pathways. Gene selection is important for disease diagnosis and therapeutic target selection. However, gene selection alone may not be sufficient. For example cancer drugs are increasingly designed to target specific pathways. The identification of critical genes and pathways in disease development is one of the most important tasks in the post-genomic era.

Most of the available methods make inference about pathways using a two step procedure where at the first step the significant genes are selected and, at the second step, a test statistic for the detection of the significant pathways is calculated based on the previously selected genes. Because of the two-steps nature of these procedures, it is not possible to determine the real significance level of the tests.

We propose a model that incorporates biological knowledge from pathway databases into the analysis of DNA microarrays to identify both pathways and genes related to a phenotype. In our model information on pathway membership and gene networks is used to define pathway summaries, to specify prior distributions that account for the dependence structure between genes, and to define the MCMC moves to fit the model. The group behavior of genes in pathways is summarized using the first PLS (Partial Least Squares) latent component obtained from the genes in each of the pathways included into the model. PLS allows us to extract, from every pathway, most of the information that explains the response variable. The gene network information is modeled using a Ising-type Markov Random Field prior on the binary selection indicators of the genes.

The employment of the gene network prior and the synthesis of the pathway information through PLS bring additional information into the model

that is especially useful in microarray data, where there is low sample size and large measurement error. We evaluate the performance of the method using simulated data. We also illustrate the method on a breast cancer gene expression dataset with survival outcomes. We show that our model can lead to the selection of significant genes that would have been missed otherwise. In addition we achieve better prediction results compared to models that do not treat genes as connected elements that work in groups or pathways. Chapter 2 is entirely derived from a paper written together with Yian A. Chen (Moffit Cancer Center), Mahlet G. Tadesse (Georgetown University) and Marina Vannucci (Rice University); this manuscript is a technical report (TR2010-01) of the Department of Statistics at Rice University and it has been submitted to the *Annals of Applied Statistics*.

In Chapter 3 we describe how to infer a regulatory network where genes are supposed to be regulated by microRNAs (an abundant class of small, $\sim$22 nucleotide, RNAs). The motivation of this work comes from one of the main questions in genomics: *What functional relationships exist among genes and how are they influenced by other biological elements?* This question can be appropriate in many situations. Our first problem is to understand the regulatory process of microRNAs (miRNAs) on the genes (miRNA targets). Because of the usual lack of samples we use external information that helps finding significant connections between genes and miRNAs. Specifically, we use a directed graphical model (Bayesian Network) with a predetermined ordering of the nodes based on biological considerations and use the Bayesian paradigm for inference. This model is able to answer to the baseline question of '*which miRNAs regulate which targets*' and allows us to build a fast computational procedure required in a high-dimensional framework. The challenge of the analysis is in building a fast computational procedure which is able to find sets of most probable models (that is, a *model selection* problem). We use a stochastic search variable selection type procedure adapted to a high-dimensional graphical model with prior distributions reflecting biologi-

cal information. Chapter 3 is entirely derived from a paper written together with Yian A. Chen (Moffit Cancer Center), Marina Vannucci (Rice University), Marianne Barrier (Texas A&M University) and Philip E. Mirkes (Texas A&M University); this manuscript is a technical report (TR2009-01) of the Department of Statistics at Rice University and has been invited for revision by *Annals of Applied Statistics*.

Chapter 2

Bayesian Models for Pathway and Gene selection

The vast amount of biological knowledge accumulated over the years has allowed researchers to identify various biochemical interactions and define different families of pathways. There is an increased interest in identifying pathways and pathway elements involved in particular biological processes. Drug discovery efforts, for example, are focused on identifying biomarkers as well as pathways related to a disease. We propose a Bayesian model that addresses this question by incorporating information on pathways and gene networks in the analysis of DNA microarray data. These information are used to define pathway summaries, specify prior distributions, and structure the MCMC moves to fit the model. We illustrate the method with an application to gene expression data with censored survival outcomes. In addition to identifying markers that would have been missed otherwise and improving prediction accuracy, the integration of existing biological knowledge into the analysis provides a better understanding of underlying molecular processes.

2.1 Introduction

DNA microarrays have been used successfully to identify gene expression signatures characteristic of disease subtypes Golub et al. (1999) or distinct outcomes to therapy Shipp et al. (2002). Many statistical methods have been developed to select genes for disease diagnosis, prognosis, and therapeutic targets. There is an increased consensus, however, that gene selection alone may not be sufficient. In cancer pharmacogenomics, for instance, cancer drugs are increasingly designed to target specific pathways to account for the complexity of the oncogenic process and the complex relationships between genes Downward (2006). A pathway is generally defined as an ordered series of chemical reactions in a living cell, and it can be activated or inhibited at multiple points. For example, if a gene at the top of a signaling cascade is selected as a target, it is not guaranteed that the reaction will be successfully inactivated because multiple genes downstream can still be activated or inhibited. Even if a branch of the pathway is completely blocked by inhibition or activation of multiple genes, the signal may still be relayed through an alternative branch or even through a different pathway Bild et al. (2006), Solit et al. (2006). Downward (2006) pointed out that targeting a single pathway or a few signaling pathways might not be sufficient. Thus, the focus is increasingly on identifying both relevant genes and pathways. We propose a Bayesian model that addresses this question by incorporating information on pathways and gene networks in the analysis of DNA microarray data. These information are used to define pathway summaries, specify prior distributions, and structure the MCMC moves to fit the model.

Several public and commercial databases have been developed to structure and store the vast amount of biological knowledge accumulated over the years into functionally or biochemically related groups. These databases focus on describing signaling, metabolic or regulatory pathways. Some examples include Gene Ontology (GO) The Gene Ontology Consortium (2000), the Kyoto Encyclopedia of Genes and Genomes (KEGG) Ogata et al. (1999),

MetaCyc Krieger et al. (2004), PathDB (www.ncgr.org/pathdb), Reactome KnowledgeBase Joshi-Tope et al. (2005), Invitrogen iPath (www.invitrogen.com), and Cell Signaling Technology (CST) Pathway (www.cellsignal.com). The need to integrate gene expression data with the biological knowledge accumulated in these databases is well recognized. Several software packages that query pathway information and overlay DNA microarray data on pathways have been developed. Nakao et al. (1999) implemented a visualization tool that color-codes KEGG pathway diagrams to reflect changes in their gene expression levels. GenMAPP Dahlquist et al. (2002) is another graphical tool that allows visualization of microarray data in the context of biological pathways or any other functional grouping of genes. Doniger et al. (2003) have made use of GenMAPP to view genes involved in specific GO terms. Another widely used method that relates pathways to a set of differentially expressed genes is the gene set enrichment analysis (GSEA) Subramanian et al. (2005). Given a list of genes ordered according to their correlation with a phenotype, GSEA computes an enrichment score to reflect the degree to which a pre-defined pathway is over-represented at the top or bottom of the ranked list. These procedures are useful starting points to observe gene expression changes in the context of known biological processes.

Some recent studies have gone a step further and have focused on incorporating pathway information or gene-gene network information into the analysis of gene expression data. For example, Park et al. (2007) have attempted to incorporate GO annotation to predict survival time, by first grouping genes based on their GO membership, calculating the first principal component to form a super-gene within each cluster then applying a Cox model with L_1 penalty to identify super-genes, i.e., GO terms related to the outcome. Wei & Li (2007) have considered a small set of 33 pre-selected signaling pathways and used the implied relationships among genes to infer differentially expressed genes, and Wei & Li (2008) have extended this work by including a temporal dimension. Li & Li (2008) and Pan et al. (2009) have proposed

two different procedures that use the gene-gene network to build penalties in a regression model framework for gene selection. Bayesian approaches have also been developed. Li & Zhang (2009) have incorporated the dependence structure of transcription factors in a regression model with gene expression outcomes; in their approach a network is defined based on the Hamming distance between candidate motifs and used to specify a Markov random field prior for the motif selection indicator. Telesca et al. (2008) have proposed a model for the identification of differentially expressed genes that takes into account the dependence structure among genes from available pathways while allowing for correction in the gene network topology.

These methods use the gene-pathway relationships or gene network information to identify either the important pathways or the genes. Our goal is to develop a more comprehensive method that selects both pathways and genes using a model that incorporates pathway-gene relationships and gene dependence structures. In order to identify relevant genes and pathways, latent binary vectors are introduced and updated using a two-stage Metropolis-Hastings sampling scheme. The gene networks are used to define a Markov random field prior on the gene selection indicators and to structure the Markov chain Monte Carlo (MCMC) moves. In addition, the pathway information is used to derive pathway expression measures that summarize the group behavior of genes within pathways using the first latent components obtained by applying partial least squares (PLS) regressions on the selected genes from each pathway. We apply the model to simulated and real data using the pathway structure from the KEGG database. The integration of the pathway information allows the identification of relevant predictors that would have been missed otherwise and also improves the prediction accuracy.

The paper is organized as follows. In Section 2, we discuss the model formulation and prior specification. Section 3 describes the MCMC procedure to fit the model and strategies for posterior inference. In Section 4, we evaluate the performance of the method using simulated data and illustrate

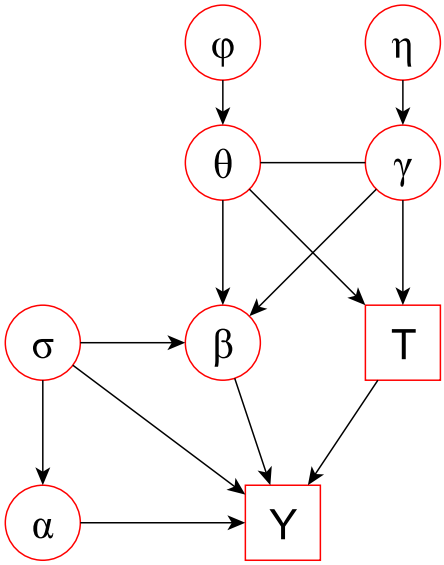

Figure 2.1: Directed acyclic graph for model – observables are represented with squares and parameters with circles

an application of the method to gene expression data with survival outcomes. We conclude with a brief discussion in Section 5.

2.2 Model Specification

In this Section, we describe how we incorporate pathway and gene network information into a Bayesian modeling framework for gene and pathway selection. Figure 3.2 shows a graphical representation of the *a priori* dependence structure among variables and parameters in the proposed model.

2.2.1 Regression on Latent Measures of Pathway Activity

Our goal is to build a model for identifying pathways related to a particular phenotype while simultaneously locating genes from these selected pathways that are involved in the biological process of interest. The data we have available for analysis consist of:

1. Y, an $n \times 1$ vector of outcomes.

2. X, an $n \times p$ matrix of gene expression levels. Without loss of generality, X is centered so that its columns sum to 0.

3. S, a $K \times p$ matrix indicating membership of genes in pathways, with elements $s_{kj} = 1$ if gene j belongs to pathway k, and $s_{kj} = 0$ otherwise.

4. R, a $p \times p$ matrix describing relationships between genes, with elements $r_{ij} = 1$ if genes i and j have a direct link in the gene network, and $r_{ij} = 0$ otherwise.

The matrices S and R are constructed using information retrieved from pathway databases, see the application in Section 2.4.2 for details.

Since the goal of the analysis is to study the association between the response variable and the pathways, we need to derive a score as a measure of "pathway expression", which summarizes the group behavior of included genes within pathways. We do this by using the latent components from a PLS regression of Y on selected subsets of genes from each pathway. A number of recent studies have, in fact, applied dimension reduction techniques to capture the group behavior of multiple genes. Pittman et al. (2004), for instance, first apply k-means clustering to identify subsets of potentially related genes, then use as regressors the first principal components obtained from applying principal component analysis (PCA) to each cluster. Bair et al. (2006) start by removing genes that have low univariate correlation with the outcome variable then apply PCA on the remaining genes to form clusters

or conceptual pathways, which are used as regressors. Similarly, Carvalho et al. (2008) use a factor model and interpret the factors as pathways. In our method, instead of attempting to infer conceptual pathways, we use the existing pathway information. We compute a pathway activity measure by applying PLS regression of Y on a subset of selected genes from the pathway. PLS has the advantage of taking into account the correlation between covariates and the response variable Y, whereas PCA focuses solely on the variability in the covariate data. The selection of a subset of gene expressions to form the PLS components is similar in spirit to the sparse PCA method proposed by Zou et al. (2006), which selects variables to be used to form the principal components.

One approach for locating genes and pathways to be included in the model consists of first deriving all possible first PLS components for pathway k using each of the $(2^{p_k} - 1)$ subsets of genes that can be formed from the $p_k = \sum_{j=1}^{p} s_{kj}$ genes belonging to the pathway (the empty set is not considered). Let $\mathbf{T}_k$ be the corresponding $n \times (2^{p_k} - 1)$ matrix of first latent components. In order to identify the genes that contribute to the summary measure of pathway k, a multinomial latent vector λ_k of size $(2^{p_k} - 1)$ is introduced with $\lambda_{kl} = 1$ if column l of $\mathbf{T}_k$ is used as score for pathway k, and $\lambda_{kl} = 0$ otherwise. In addition, a latent binary vector $\boldsymbol{\theta}$ is introduced for pathway selection. The linear regression model that relates the response variable to the selected pathways and genes is then written as

$$Y = \mathbf{1}\alpha + \sum_{r=1}^{K_\theta} \boldsymbol{T}_{r(\lambda_r)} B_{r(\lambda_r)} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}), \tag{2.1}$$

where $K_\theta = \sum_{k=1}^{K} \theta_k$ is the number of selected pathways and where the subscript (λ_r) identifies the first PLS latent component that corresponds to the non-zero element of the multinomial vector λ_r and that is used as score of pathway r, among the $(2^{p_r} - 1)$ columns of $\boldsymbol{T}_r$; $\mathbf{B}_r$ is the $(2^{p_r} - 1)$-vector of regression coefficients associated with the matrix $\boldsymbol{T}_r$ and $B_{r(\lambda_r)}$ is the non-

zero element corresponding to the selected column $\boldsymbol{T}_{r(\lambda_r)}$.

We introduce a simpler parametrization that is equivalent to model (2.1) to ease the prior specification and the MCMC implementation. Instead of defining K multinomial vectors of length $(2^{p_k} - 1)$ to locate genes relevant from each pathway, we use a single $p \times 1$ binary gene selection vector γ, where $\gamma_j = 1$ if gene j is selected, and $\gamma_j = 0$ otherwise. The linear regression model can then be defined as

$$Y = \mathbf{1}\alpha + \sum_{r=1}^{K_\theta} T_{r(\gamma)}\beta_{r(\gamma)} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}), \tag{2.2}$$

where $T_{r(\gamma)}$ corresponds to the first latent PLS component generated based on the expression levels of selected genes belonging to pathway r, that is using the X_j's corresponding to $s_{rj} = 1$ and $\gamma_j = 1$. Note that, since in (2.1) we only allow the selection of one column of $\boldsymbol{T}_k$ as the score of pathway k, the use of λ_k in (2.1) is equivalent to using γ coupled with the matrix S in (2.2).

As we explain in Sections 2.2.3 and 2.2.4, the alternative parametrization in (2.2) makes it easier to incorporate the gene network in the prior distribution for gene selection. In addition, the use of a $p \times 1$ binary vector, instead of K multinomial vectors of size $(2^{p_k} - 1)$, simplifies the implementation of an MCMC algorithm that explores the space of gene subsets and facilitates the posterior inference for gene selection.

2.2.2 Models for Categorical or Censored Outcomes

In the construction above, we have assumed a continuous response. However, our model formulation can easily be extended to handle categorical or censored outcome variables.

When Y is a categorical variable taking one of G possible values, $0, \ldots, G-1$, a probit model can be used Albert & Chib (1993), Sha et al. (2004), Kwon et al. (2007). Briefly, each outcome Y_i is associated with a vector

$(p_{i,0}, \ldots, p_{i,G-1})$, where $p_{ig} = P(Y_i = g)$ is the probability that subject i falls in the g-th category. The probabilities p_{ig} can be related to the linear predictors using a data augmentation approach. Let $\mathbf{Z}_i$ be latent data corresponding to the unobserved propensities of subject i to belong to one of the classes. When the observed outcomes Y_i correspond to nominal values, the relationship between Y_i and $\mathbf{Z}_i = (z_{i,1}, \ldots, z_{i,G-1})$ is defined as

$$Y_i = \begin{cases} 0 & \text{if } \max_{1 \le l \le G-1}\{z_{i,l}\} \le 0 \\ g & \text{if } \max_{1 \le l \le G-1}\{z_{i,l}\} > 0 \text{ and } z_{i,l} = \max_{1 \le r \le G-1}\{z_{i,l}\} \end{cases} . \quad (2.3)$$

A multivariate normal model can then be used to associate $\mathbf{Z}_i$ to the predictors

$$\mathbf{Z}_i = \alpha \mathbf{1}^T + \sum_{r=1}^{K_\theta} T_{i,r(\gamma)} \boldsymbol{\beta}_{r(\gamma)} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \Sigma), \quad i = 1, \ldots, n. \quad (2.4)$$

If the observed outcomes Y_i correspond, instead, to ordinal categories, the latent variable Z_i is defined such that

$$Y_i = g \qquad \text{if } \delta_g < Z_i \le \delta_{g+1}, \quad g = 0, \ldots, G-1, \quad (2.5)$$

where the boundaries δ_g are unknown and $-\infty = \delta_0 < \delta_1 < \ldots < \delta_{G-1} < \delta_G = \infty$. The latent variable Z_i is then associated with the predictors through the linear model

$$Z_i = \alpha + \sum_{r=1}^{K_\theta} T_{i,r(\gamma)} \beta_{r(\gamma)} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n. \quad (2.6)$$

For censored survival outcomes, an accelerated failure time (AFT) model can be used, Wei (1992), Sha et al. (2006). In this case, the observed data are $Y_i = \min(T_i, C_i)$ and $\delta_i = I\{Y_i \le C_i\}$, where T_i is the survival time for subject i, C_i is the censoring time, and δ_i is a censoring indicator. A data augmentation approach can be used and variables Z_i can be introduced such

that

$$\begin{cases} Z_i = \log(Y_i) & \text{if } \delta_i = 1 \\ Z_i > \log(Y_i) & \text{if } \delta_i = 0 \end{cases} . \tag{2.7}$$

The AFT model can then be written in terms of the latent Z_i,

$$Z_i = \alpha + \sum_{r=1}^{K_\theta} T_{i,r(\gamma)} \beta_{r(\gamma)} + \varepsilon_i, \tag{2.8}$$

where the ε_i's are independent and identically distributed random variables that may take one of several parametric forms. Sha et al. (2006) consider cases where ε_i follows a normal or a t-distribution.

2.2.3 Prior for Regression Parameters

The regression coefficient β_k in (2.2) measures the effect of the PLS latent component summarizing the effect of pathway k on the response variable. However, not all pathways are related to the phenotype and the goal is to identify the predictive ones. Bayesian methods that use mixture priors for variable selection have been thoroughly investigated in the literature, in particular for linear models, see George & McCulloch (1993, 1997) for multiple regression, Brown et al. (1998*b*) for extensions to multivariate responses and Sha et al. (2004) for probit models. A comprehensive review on special features of the selection priors and on computational aspects of the method can be found in Chipman et al. (2001). Similarly, here, we use the latent vector $\boldsymbol{\theta}$ to specify a scale mixture of a normal density and a point mass at zero for the prior on each β_k in model (2.2):

$$\beta_k | \theta_k, \sigma^2 \sim \theta_k \cdot \mathcal{N}(\beta_0, h\sigma^2) + (1 - \theta_k) \cdot \delta_0(\beta_k), \qquad k = 1, \ldots, K. \tag{2.9}$$

where $\delta_0(\beta_k)$ is a Dirac Delta function. For model (2.1), the prior specification for the regression coefficients $\boldsymbol{B}_k$ requires conditioning on θ_k as well as λ_k,

$$\begin{aligned} B_{kj}|\theta_k, \lambda_k, \sigma^2 \;\sim\; & \theta_k \cdot [\lambda_{kj} \cdot \mathcal{N}(B_{0j}, h\sigma^2) + (1-\lambda_{kj}) \cdot \delta_0(B_{kj})] \\ & + (1-\theta_k) \cdot \delta_0(B_{kj}) \end{aligned} \tag{2.10}$$

for $k = 1, \ldots, K$ and $j = 1, \ldots, (2^{p_k} - 1)$. When conditioning on γ, the posterior distributions of the selected regression coefficients induced by prior (2.9) for model (2.2) and by prior (2.10) for model (2.1) are equivalent. It is, however, easier to specify a prior that incorporates the gene dependence structure for the p-vector $\boldsymbol{\gamma}$ than it is for the K multinomial vectors $\boldsymbol{\lambda}_k$, see Section 2.2.4 below. The hyperparameter h in (2.9) regulates, together with the hyperparameters of $p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\eta)$, the amount of shrinkage in the model. We follow the guidelines provided by Sha et al. (2004) and specify h in the range of variability of the data so as to control the ratio of prior to posterior precision.

For the intercept term, α, and the variance, σ^2, we take conjugate priors

$$\begin{aligned} \alpha|\sigma^2 &\sim \mathcal{N}(\alpha_0, h_0\sigma^2) \\ \sigma^2 &\sim \text{Inv-Gamma}(\nu_0/2, \nu_0\sigma_0^2/2), \end{aligned} \tag{2.11}$$

where α_0, β_0, h_0, h, ν_0 and σ_0^2 are to be elicited.

2.2.4 Priors for Pathway and Gene Selection Indicators

In this Section we define the prior distributions for the pathway selection indicator, $\boldsymbol{\theta}$, and gene selection indicator, $\boldsymbol{\gamma}$. These priors are first defined marginally then jointly, taking into account some necessary constraints.

Each element of the latent K-vector $\boldsymbol{\theta}$ is defined as

$$\theta_k = \begin{cases} 1 & \text{if pathway } k \text{ is represented in the model} \\ 0 & \text{otherwise} \end{cases} \tag{2.12}$$

for $k = 1, \ldots, K$. We assume independent Bernoulli priors for the θ_k's,

$$p(\boldsymbol{\theta}|\varphi_k) = \prod_{k=1}^{K} \varphi_k^{\theta_k}(1-\varphi_k)^{1-\theta_k}, \tag{2.13}$$

where φ_k determines the proportion of pathways expected *a priori* in the model. A mixture prior can be further specified for φ_k to achieve a better discrimination in terms of posterior probabilities between significant and non-significant pathways by inflating $p(\theta_k = 0)$ toward 1 for the non-relevant pathways, as suggested by Carvalho et al. (2008),

$$p(\varphi_k) = \rho\delta_0(\varphi_k) + (1-\rho)\mathcal{B}(\varphi_k|a_0, b_0), \tag{2.14}$$

where $\mathcal{B}(\varphi_k|a_0, b_0)$ is a Beta density function with parameters a_0 an b_0. Since inference on φ_k is not of interest, it can be integrated out to simplify the MCMC implementation. This leads to the following marginal prior for θ

$$p(\boldsymbol{\theta}) = \prod_{k} \left[\rho \cdot (1-\theta_k) + (1-\rho) \cdot \frac{B(a_0+\theta_k, b_0+1-\theta_k)}{B(a_0, b_0)}\right], \tag{2.15}$$

where $B(\cdot,\cdot)$ is the Beta function. Prior (2.15) corresponds to a product of Bernoulli distributions with parameter $\varphi_k^* = \frac{a_0(1-\rho)}{a_0+b_0}$.

For the latent p-vector $\boldsymbol{\gamma}$, we specify a prior distribution that is able to take into account not only the pathway membership of each gene but also the biological relationships between genes within and across pathways, which are captured by the matrix R. Following Wei & Li (2007), we model these relations using a Markov random field (MRF), where genes are represented by nodes and relations between genes by edges. A MRF is a graphical model in which the distribution of a set of random variables follow Markov properties that can be described by an undirected graph. In particular, a pair of genes that are not connected are considered conditionally independent given all other genes Besag (1974). Relations on the MRF are represented by the

following probabilities

$$P(\gamma_j|\eta, \gamma_i, i \in N_j) = \frac{\exp(\gamma_j F(\gamma_j))}{1 + \exp(F(\gamma_j))}, \tag{2.16}$$

where $F(\gamma_j) = \mu + \eta \sum_{i \in N_j}(2\gamma_i - 1)$ and N_j is the set of direct neighbors of gene j in the MRF using only pathways represented in the model, *i.e.*, pathways with $\theta_k = 1$. The parameter μ controls the sparsity of the model, with higher values of η encouraging neighboring genes to take on the same γ_j value. If a gene does not have any neighbor, its prior distribution reduces to an independent Bernoulli with parameter $p = \exp(\mu)/[1 + \exp(\mu)]$, which is a logistic transformation of μ. The corresponding global distribution on the MRF is given by

$$P(\boldsymbol{\gamma}|\theta, \mu, \eta) \propto \exp(\mu\ n_1 - \eta\ n_{01}), \tag{2.17}$$

where n_1 is the number of selected genes and n_{01} is the number of edges linking genes with different values of γ_j (*i.e.*, edges linking included and non-included genes among all pathways),

$$n_1 = \sum_{j=1}^{p} \gamma_j, \qquad n_{01} = \frac{1}{2} \sum_{i=1}^{p} \left[\sum_{j=1}^{p} r_{ij} - \left| \sum_{j=1}^{p} r_{ij}(1 - \gamma_i) - \sum_{j=1}^{p} r_{ij}\gamma_j \right| \right].$$

In addition, we specify a Gamma hyperprior for η,

$$\eta \sim Gamma(c_0, d_0). \tag{2.18}$$

Constraints need to be imposed to ensure both interpretability and identifiability of the model. We essentially want to avoid the following scenarios:

1. creation of empty pathways, *i.e.*, selecting a pathway but none of its member genes;

2. creation of orphan genes, *i.e.*, selecting a gene but none of the pathways that contain it;

3. selection of identical subsets of genes by different pathways, a situation that generates identical values $T_{k(\gamma)}$ and $T_{k'(\gamma)}$ to be included in the model.

These constraints imply that some combinations of θ and γ values are not allowed. The joint prior probability for $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ taking into account these constraints is given by

$$\pi(\boldsymbol{\theta}, \boldsymbol{\gamma}|\eta) \propto \begin{cases} \prod_{k=1}^{K} \varphi_k^{*\theta_k}(1-\varphi_k^*)^{1-\theta_k} \exp(\mu n_1 - \eta n_{01}) & \text{for valid comb.,} \\ 0 & \text{for invalid comb.} \end{cases} \tag{2.19}$$

We note that specifying a prior of type (2.17) on the array $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$ from model (2.1) would in practice amount to reparameterizing the K multinomial vectors into a p-binary vector, therefore bringing us back to model (2.2). In addition, it is more straightforward to impose the constraints on the p-binary vector $\boldsymbol{\gamma}$ than it is on the K multinomial vectors λ_k.

2.3 Model Fitting

We now describe the MCMC procedure to fit the model and discuss strategies for posterior inference.

2.3.1 Marginal Posterior Probabilities

The model parameters consist of $(\alpha, \beta, \sigma^2, \boldsymbol{\gamma}, \boldsymbol{\theta}, \eta)$. The MCMC procedure can be made more efficient by integrating out some of the parameters. Here, we integrate out the regression parameters, α, β and σ^2. This leads to a multivariate t-distribution

$$f(Y|\boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \sim \mathcal{T}_{\nu_0}(\alpha_0 \mathbf{1}_n + \boldsymbol{T}_{(\theta,\gamma)}\beta_0, \sigma_0^2(\mathbf{I}_n + h_0 \mathbf{1}_n \mathbf{1}_n' + \boldsymbol{T}_{(\theta,\gamma)}\Sigma_0 \boldsymbol{T}'_{(\theta,\gamma)})), \tag{2.20}$$

with ν_0 degrees of freedom and $\mathbf{1}_n$ an n-vector of ones, and where $\Sigma_0 = h\mathbf{I}_{K_\theta}$, with $\mathbf{I}_n$ the $n \times n$ identity matrix, and $T_{(\theta,\gamma)}$ the $n \times K_\theta$ matrix derived from

the first PLS latent components for selected pathways using selected genes. In the notation (2.20) the two arguments of the t-distribution represent the mean and the scale parameter of the distribution, respectively. The posterior probability distribution of the pathway and gene selection indicators is then given by

$$f(\boldsymbol{\theta}, \boldsymbol{\gamma}, \eta | \boldsymbol{T}, Y) \propto f(Y | \boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \cdot \pi(\boldsymbol{\theta}, \boldsymbol{\gamma} | \eta) \cdot \pi(\eta). \tag{2.21}$$

2.3.2 MCMC Sampling

The MCMC steps to fit the model consist of: (I) sampling the pathway and gene selection indicators from $p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\text{rest})$; (II) sampling the MRF parameter from $p(\eta|\text{rest})$; (III) sampling additional parameters that would be introduced when fitting a probit model for categorical outcomes or an AFT model for survival outcomes.

(I) The parameters $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ are updated using a Metropolis-Hastings algorithm in a two-stage sampling scheme. The pathway-gene relationships are used to structure the moves and account for the constraints specified in Section 2.2.4. Figure 2.2 summarizes the transition moves in a flow chart. Details of the MCMC moves for updating $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ are provided in the Appendix A. Briefly, they consist of randomly choosing one of the following three move types:

1. change the inclusion status of gene and pathway by randomly choosing between adding the score of a pathway and a gene (move 1.i in Figure 2.2) or removing them both (move 1.ii);
2. change the inclusion status of gene but not pathway by randomly choosing between adding a gene (2.i) or removing a gene (2.ii);
3. change the inclusion status of pathway but not gene by randomly choosing between adding a pathway (3.i) or removing a pathway (3.ii).

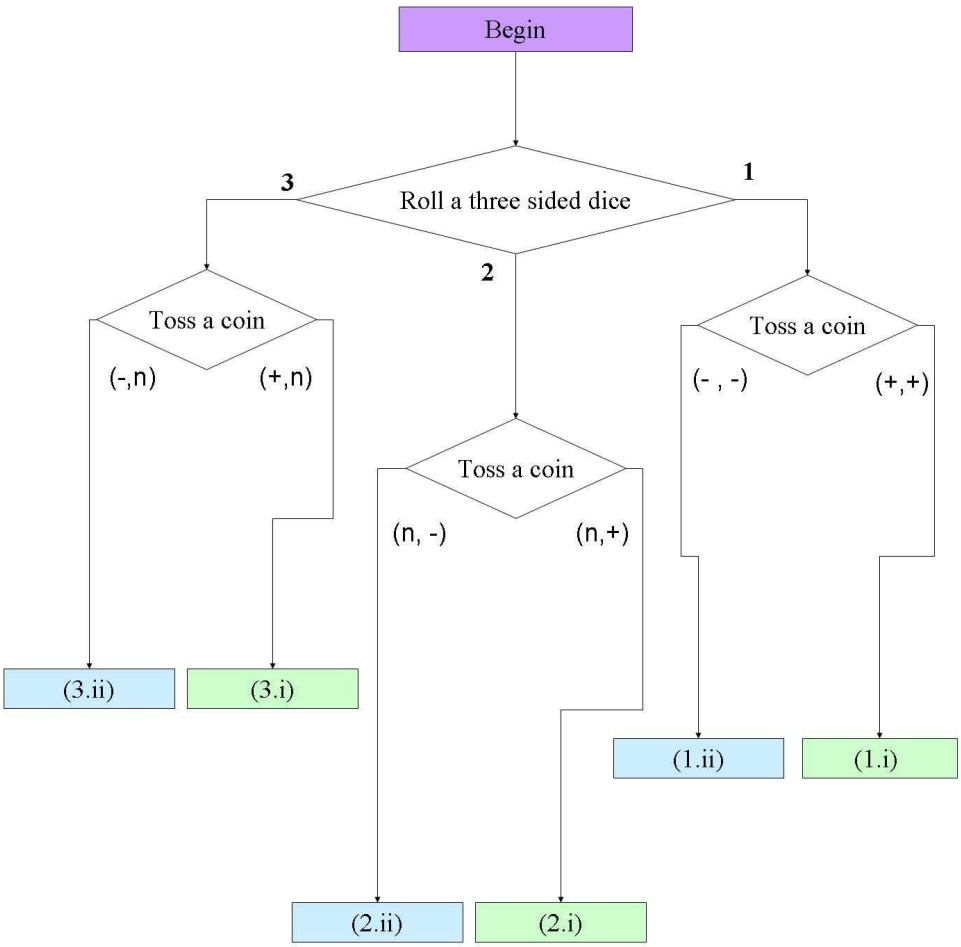

Figure 2.2: Proposed move types: (1.i) add a pathway and a gene (+,+); (1.ii) remove a pathway and a gene (-,-); (2.i) add a gene in an existing pathway (n,+); (2.ii) remove a gene from an existing pathway (n,-); (3.i) add a pathway without touching genes (+,n); (3.ii) remove a pathway without touching genes (-,n).

(II) The MRF parameter η is sampled using a Metropolis-Hastings algorithm, in which a candidate η^{new} is generated from a truncated normal distribution, with mean η^{old} and truncation at 0, which we denote by $q(\eta^{old}; \eta^{new})$. The variance of this distribution represents a tuning parameter chosen to allow a wide exploration of the parameter space and to give a good acceptance rate. Alternatively, a Gamma density could be used as proposal distribution. The acceptance probability is

$$\min\left\{\frac{f(\gamma|\eta^{new},\theta)f(\eta^{new})q(\eta^{old};\eta^{new})}{f(\gamma|\eta^{old},\theta)f(\eta^{old})q(\eta^{new};\eta^{old})},1\right\}. \tag{2.22}$$

(III) In the case of classification or survival outcomes, the augmented data Z need to be updated from their full conditionals using Gibbs sampling (see Sha et al. (2004, 2006), Kwon et al. (2007) for details on this step).

2.3.3 Posterior Inference

The MCMC procedure results in a list of visited models with included pathways indexed by $\boldsymbol{\theta}$ and selected genes indexed by $\boldsymbol{\gamma}$, and their corresponding relative posterior probabilities. Pathway selection can be based on the marginal posterior probabilities $p(\theta_k|\boldsymbol{T}, Y)$, estimated by the relative frequency of inclusion of pathway k in the models visited by the MCMC sampler. Relevant pathways can be identified as those with largest marginal posterior probabilities. Then relevant genes from these pathways can be identified based on their marginal posterior probabilities conditional on the inclusion of a pathway of interest, $p(\gamma_j|\boldsymbol{T}, Y, I\{\theta_k s_{kj} = 1\})$. An alternative inference for gene selection is to focus on a subset of pathways, $\mathcal{P}$, and consider the marginal posterior probability conditional on at least one pathway the gene belongs to being represented in the model, $p(\gamma_j|\boldsymbol{T}, Y, I\{\sum_{k\in\mathcal{P}} \theta_k s_{kj} > 0\})$.

Inference for a new set of observations, $(\boldsymbol{X}_f, Y_f)$ can be done via least squares prediction,

$$\widehat{Y}_f = \mathbf{1}_n\tilde{\alpha} + \boldsymbol{T}_{f(\theta,\gamma)}\tilde{\boldsymbol{\beta}}_{(\theta,\gamma)}, \tag{2.23}$$

where $\boldsymbol{T}_{f(\theta,\gamma)}$ is the first principal component based on selected genes from relevant pathways and

$$\tilde{\alpha} = \bar{Y}, \qquad \tilde{\boldsymbol{\beta}}_{(\theta,\gamma)} = (\boldsymbol{T}'_{(\theta,\gamma)}\boldsymbol{T}_{(\theta,\gamma)} + h^{-1}\boldsymbol{I}_{K_\theta})^{-1}\boldsymbol{T}'_{(\theta,\gamma)}Y,$$

with Y the response variable in the training and $\boldsymbol{T}_{(\theta,\gamma)}$ the scores obtained from the training data using selected pathways and genes included in the model. Note that for prediction purposes, since we do not know the future Y_f, a PLS regression cannot be fit. Therefore, we generate $T_{f(\theta,\gamma)}$ by considering the first latent component obtained by applying PCA to each selected pathway using the included genes.

In the case of categorical or censored survival outcomes, the sampled latent variables Z would be used to estimate $\widehat{Z}_f$ then the correspondence between Z and the observed outcome outlined in Section 2.2.2 can be invoked to predict Y_f (Sha et al. (2004, 2006), Kwon et al. (2007)).

2.4 Application

We assess the performance of the model on simulated data then illustrate an application to a breast cancer data using the KEGG pathway database to define the MRF.

2.4.1 Simulation Studies

We investigated the performance of the model using simulated data based on the gene-pathway relations, S, and gene network, R, of 70 pathways and 1098 genes from the KEGG database. The relevant pathways were defined by selecting 4 pathways at random. For each of the 4 selected pathways, one gene was picked at random and its direct neighbors that belong to the selected pathways were chosen. This resulted in the selection of 4 pathways and 15 genes: 7 out of 30 from the first pathway, 3 out of 35 from the second, 3 out of 105 from the third, and 2 out of 47 from the fourth pathway. Gene

expressions for $n = 100$ samples were simulated for these 15 genes using an approach similar to Li & Li (2008). This was accomplished by first creating an ordering among the 15 selected genes by changing the undirected edges in the gene networks into directed edges. The first node on the ordering, which we denote by X_{F_1}, was selected from each pathway and drawn from a standard normal distribution; note that this node has no parents. Then all child nodes directly connected only to X_{F_1} and denoted by X_{F_2} were drawn from

$$X_{F_2} \sim \mathcal{N}(X_{F_1}\rho, 1).$$

Subsequent child nodes at generation j, X_{F_j}, were drawn using all parents from

$$X_{F_j} \sim \mathcal{N}(\rho X_{pa(F_j)}\mathbf{1}_{|pa(F_j)|}, 1)$$

where $X_{pa(F_j)}$ is a matrix containing the expressions of all the $|pa(F_j)|$ parents for node j. The expression levels of the remaining 1073 genes deemed irrelevant were simulated from a standard normal density. The response variables for the $n = 100$ samples were generated from

$$Y_i = \sum_{j=1}^{15} X_{ij}\beta + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, 1), \quad i = 1, \ldots, 100.$$

For the first dataset we set $\beta = \pm 0.5$, using the same sign for genes that belong to the same pathway. For the second and third data sets we used $\beta = \pm 1$ and $\beta = \pm 1.5$, respectively. Note that the generating process is different from the model (2.2) being fit.

We report the results obtained by choosing, when possible, hyperparameters that lead to weakly informative prior distributions. A vague prior is assigned to the intercept parameter α by setting h_0 to a large value tending to ∞. For σ^2, the shape parameter can be set to $\nu_0/2 = 3$, the smallest integer such that the variance of the inverse-gamma distribution is defined, and the scale parameter $\nu_0\sigma_0^2/2$ can be chosen to yield a weakly informative prior. For the vector of regression coefficients, β_k, we set the prior mean to $\beta_0 = 0$

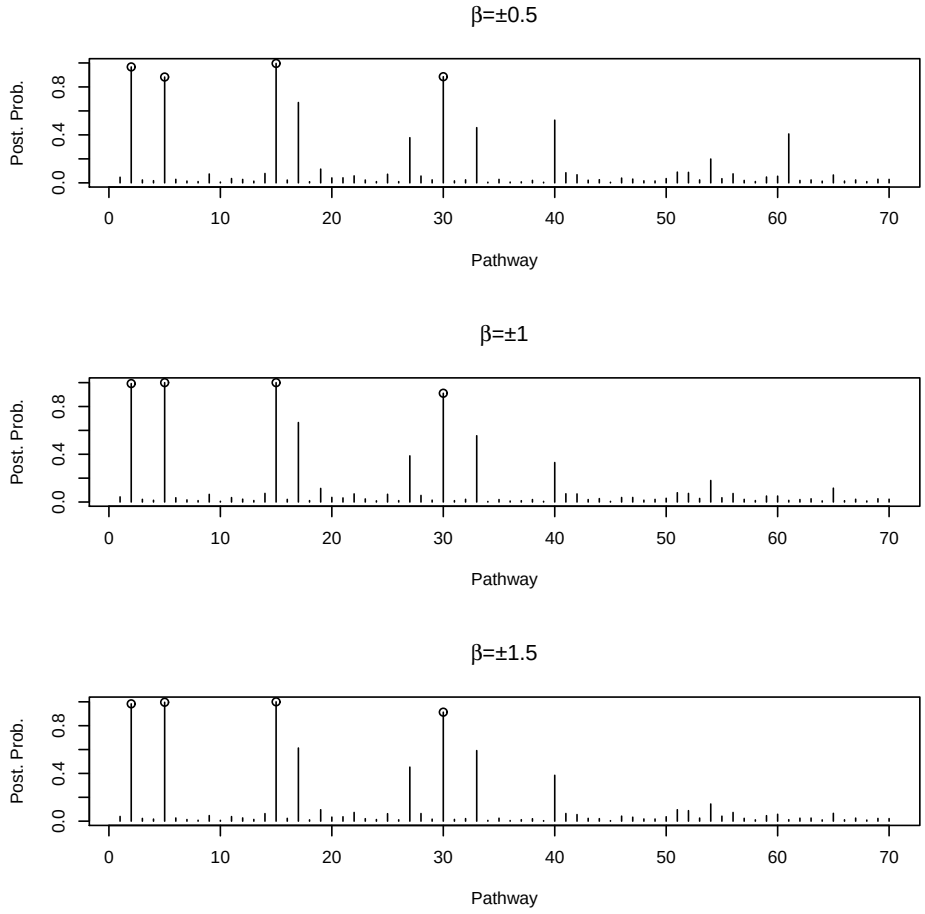

Figure 2.3: Marginal posterior probabilities for pathway selection, $p(\theta_k|\boldsymbol{T}, Y)$, for the three simulated data sets. The open circles indicate the four relevant pathways.

and h, as described in Section 2.2.3, is chosen in the range of variability of the covariates. Specifically, we set $h_0 = 10^6$, $\alpha_0 = \beta_0 = 0$, $\nu_0\sigma_0/2 = 0.5$, and $h = 0.02$. For the pathway selection indicators, θ_k, we set $\varphi_k^* = 0.01$. As for the prior at the gene level, we set $\mu = -3$, which corresponds to setting the proportion of genes expected *a priori* in the model to around 5%, and for the Gamma hyperprior on η we set $c_0 = 3.5$ and $d_0 = 0.2$ to obtain a prior distribution with most of the mass between 0.1 and 2. This controls the prior probability of selecting a gene based on how many of its neighbors are selected, as defined in (2.16).

The MCMC sampler was run for 300,000 iterations with the first 50,000 used as burn-in. We computed the marginal posterior probabilities for pathway selection, $p(\theta_k = 1|Y, \boldsymbol{T})$, and the conditional posterior probabilities for gene selection given a subset of selected pathways previously determined, $p(\gamma_j|Y, \boldsymbol{T}, I\{\sum_{k\in\mathcal{P}} \theta_k s_{kj} > 0\})$. Table 2.1 gives these posterior probabilities for the relevant pathways and genes used in simulating the response. Figure 2.3 displays the marginal posterior probabilities of inclusion for all 70 pathways and Figure 2.4 displays the conditional posterior probabilities of inclusion for all 1098 genes. The procedure successfully identified all significant pathways and genes. All four relevant pathways were selected with marginal posterior probabilities greater than 0.8. Reducing the selection threshold to a marginal posterior probability of 0.5 pulls in two false positive pathways, for all the three simulated scenarios considered. One of these is the pathway indexed 17 in Figure 2.3, which contains more than 100 genes.

A closer investigation of the MCMC output reveals that different subsets of its member genes are selected whenever it is included in the model, resulting in a high marginal posterior of inclusion for the pathway but low marginal posterior probabilities for all its member genes. The second false positive pathway appears to be selected often because it contains two or three of the relevant genes that were used to simulate the response variable and were also included in the model with high marginal posterior probabilities; all

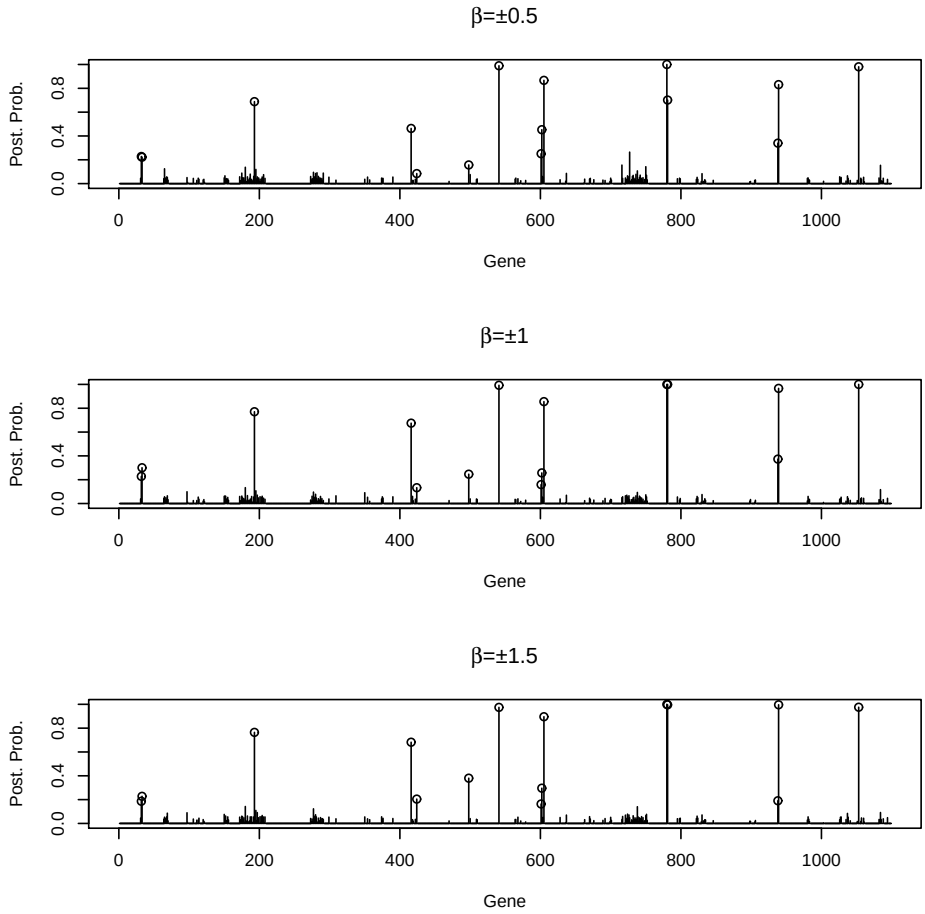

Figure 2.4: Conditional posterior probabilities, $p(\gamma_j|\boldsymbol{T}, Y, I\{\sum_{k\in\mathcal{P}} \theta_k s_{kj} > 0\})$, for gene selection for the three simulated data sets. The open circles indicate the genes used to generate the outcome variable.

Path/Gene	$\beta = \pm 0.5$	$\beta = \pm 1$	$\beta = \pm 1.5$	Path/Gene	$\beta = \pm 0.5$	$\beta = \pm 1$	$\beta = \pm 1.5$
Path1	0.967	0.992	0.983	Gene1.7	0.701	1.000	0.996
Path2	0.883	1.000	0.996	Gene2.1	0.981	1.000	0.976
Path3	0.996	1.000	1.000	Gene2.2	0.084	0.133	0.205
Path4	0.885	0.911	0.913	Gene2.3	0.156	0.246	0.380
Gene1.1	0.867	0.856	0.896	Gene3.1	0.688	0.771	0.765
Gene1.2	0.228	0.228	0.186	Gene3.2	0.340	0.373	0.190
Gene1.3	0.222	0.300	0.228	Gene3.3	0.831	0.967	0.996
Gene1.4	0.251	0.159	0.163	Gene4.1	0.463	0.675	0.682
Gene1.5	0.453	0.257	0.296	Gene4.2	0.990	0.993	0.974
Gene1.6	1.000	1.000	1.000				

Table 2.1: Conditional and marginal posterior probabilities for pathway and gene selections, $p(\gamma_j = 1|\boldsymbol{T}, Y, I\{\sum_{k\in\mathcal{P}} \theta_k s_{kj} > 0\})$ and $p(\theta_k = 1|Y, \boldsymbol{T})$.

its other member genes have very low probabilities of selection. As expected, the identification of the relevant genes is easier when the signal-to-noise ratio is higher. For the scenario with $\beta = \pm 0.5$, a marginal posterior probability cut-off of 0.5 leads to the selection of 7 relevant genes and no false positive. In order to select all 15 relevant genes, marginal posterior probabilities as low as 0.084 need to be considered and this results in the inclusion of 14 false positives. For the simulated data with $\beta = \pm 1$, on the other hand, a marginal posterior probability cut-off of 0.5 leads to the selection of 8 relevant genes with no false positive, and reducing the cut-off to 0.132 results in the inclusion of all 15 relevant genes with a single false positive.

2.4.2 Application to Microarray Data

We consider the van't Veer et al. (2002) breast cancer microarray data, available at www.rii.com/publications/2002/vantveer.htm. Gene expression measures were collected on each patient using DNA microarray with 24,481 probes. Missing expression values were imputed using a k-nearest neighbor algorithm with $k = 10$. The procedure consists of identifying the k closest genes to the one with missing expression in array j using the other $n - 1$ arrays, then imputing the missing value by the average expression level of the

k neighbors, Troyanskaya et al. (2001). We focus on the 76 sporadic lymph-node-negative patients, 33 of whom developed distant metastasis within 5 years and the remaining 43 did not; the latter are viewed as censored cases. We randomly split the patients into a training set of 38 samples and a test set of the same size. The goal is to identify a subset of pathways and genes that can predict time to distant metastasis for breast cancer patients.

The gene network and pathway information were obtained from the KEGG database. This was accomplished by mapping probes to pathways using the links between pathway node identifiers and LocusLink ID provided at ftp://ftp.genome.ad.jp/pub/kegg/pathways/hsa/hsa_gene_map.tab and ftp://ftp.genome.ad.jp/pub/kegg/pathways/map_title.tab.

Using the R package *KEGGgraph* Zhang & Wiemann (2009), we first downloaded the gene network for each pathway then merged all the networks into a single one with all the genes. A total of 196 pathways and 3,592 corresponding probes were included in the analysis. There is a many-to-many correspondence between pathways and genes – each pathway contains multiple genes and most genes are associated with several pathways.

We ran two MCMC chains with different starting numbers of included variables, 50 and 80, respectively. We used 600,000 iterations with a burn-in of 100,000 iterations. We incorporated the first latent vector of the PLS for each pathway into the analysis as described in Section 2.2.1 and set the number of pathways expected *a priori* in the model to 22. For the gene selection, we set the hyperparameter of the Markov random field $\mu = -3.5$; this parameter regulates the sparsity of the model and indicates that *a priori* at least 3% of genes are expected to be selected. As for η, the parameter that regulates the smoothness of the distribution of γ over the graph and influences the selection of neighboring genes, we set $c_0 = 3.5$ and $d_0 = 0.2$. Sensitivity analysis with different choices of these hyperparameters showed that the posterior inference is not affected for values of c_0 between 2 and 4, and values of d_0 between 0.15 and 2. For the prior of the regression

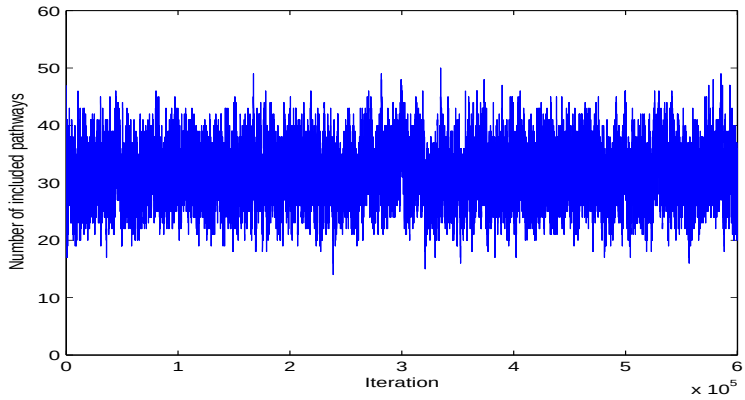

Figure 2.5: Trace plot of the number of included pathways.

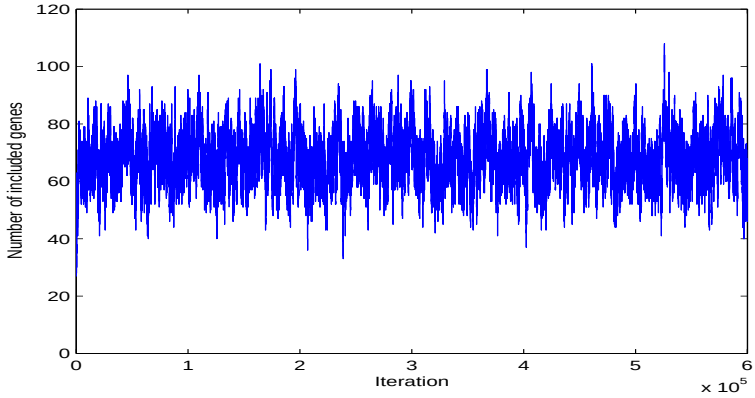

Figure 2.6: Trace plot of the number of included genes.

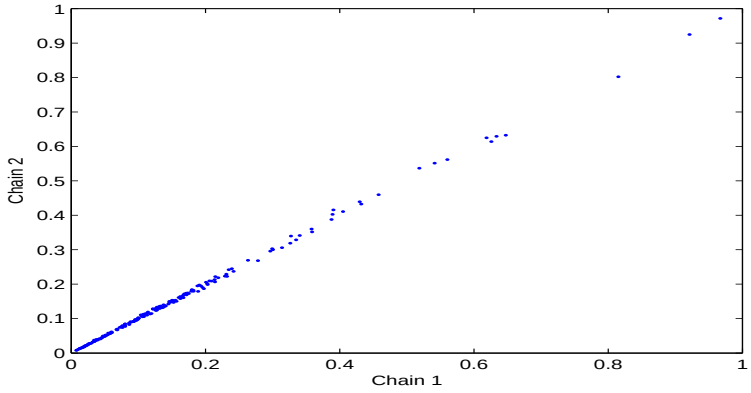

Figure 2.7: Scatterplot of $p(\theta_k|\boldsymbol{T}, Y)$ across the two MCMC chains.

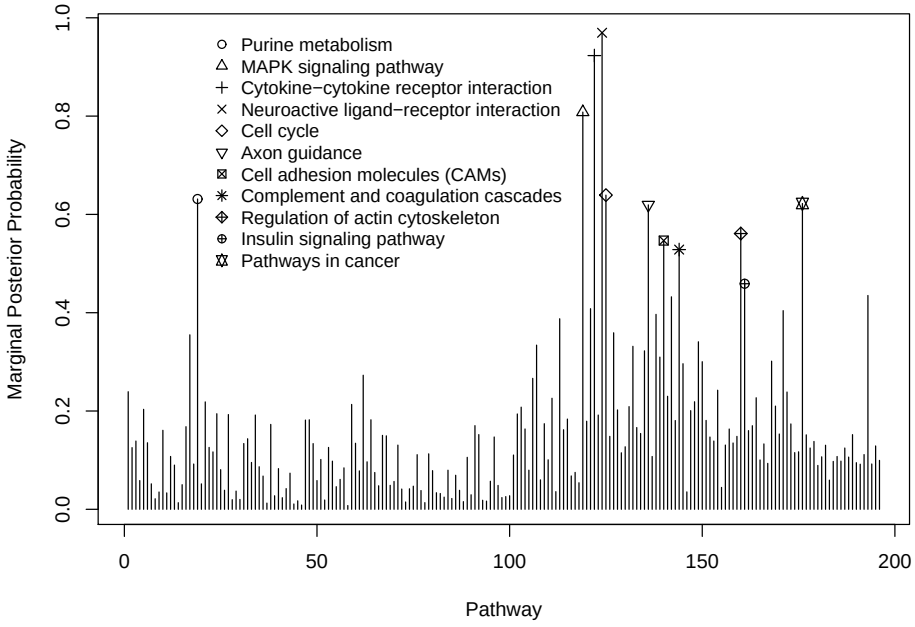

Figure 2.8: Microarray data: Marginal posterior probabilities for pathway selection, $p(\theta_k|\boldsymbol{T}, Y)$. The 11 pathways with largest probabilities are marked with symbols.

parameters, we set $\alpha_0 = \beta_0 = 0, h_0 = 10^6$ and $h = 0.1$. A vague prior was specified for σ^2 by choosing $\nu_0/2 = 3$ and $\nu_0\sigma_0^2/2 = 0.5$.

Figures 2.5 and 2.6 show the trace plots for the number of included pathways and the number of selected genes for one of the MCMC chains. The MCMC samplers mostly visited models with 20-45 pathways and 50-90 genes. To assess the agreement of the results between the two chains, we looked at the correlation between the marginal posterior probabilities for pathway selection, $p(\theta_k|\boldsymbol{T}, Y)$, and found good concordance between the two MCMC chains with a correlation coefficient of 0.9996 (Figure 2.7).

The model also shows good predictive performance. Sha et al. (2006)

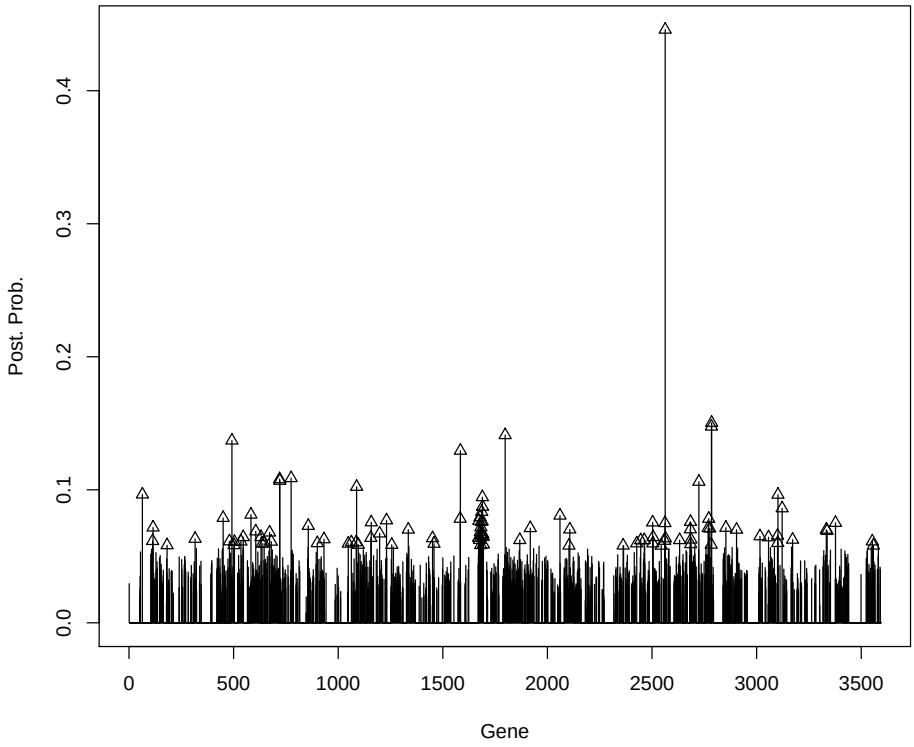

Figure 2.9: Microarray data: Conditional posterior probabilities for gene selection, $p(\gamma_j|\boldsymbol{T}, Y, I\{\sum_{k\in\mathcal{P}} \theta_k s_{kj} > 0\})$. The 102 probes with largest probability that belong to the 11 selected pathways in Figure 2.8 are marked with $\triangle$.

already analyzed these data using an AFT model with 3,839 probes as predictors and selected 11 probe sets with highest marginal probabilities. They obtained a predictive MSE($Z_f, \hat{Z}_f$) of 1.9317 in the log scale. We applied the method of Sha et al. to the 3,592 probe sets considered for analysis in this paper. Using 4 chains with 600,000 iterations and 100,000 burn-in, we selected a marginal model with 12 genes which resulted in a MSE($Z_f, \hat{Z}_f$) of 2.6735. Our model incorporating pathway information achieved a predictive MSE($Z_f, \hat{Z}_f$) of 1.3955 using 11 selected pathways and 102 probe sets with highest posterior probabilities. The selected pathways and genes are indicated in the marginal posterior probability plots displayed in Figures 2.8 and 2.9. If we increase the marginal probability thresholds to select a comparable number of probe sets as the method of Sha et al. (2006) and consider a model with 7 selected pathways and 12 genes, we obtain a MSE($Z_f, \hat{Z}_f$) of 1.5742.

The genes corresponding to the 102 selected probe sets are listed in Table 2.2 divided by islands, which correspond to sets of connected genes in the Markov random field. The islands help with the biological interpretation by locating relevant branches of pathways. A subset of the selected pathways along with the islands and included genes that fall into them are displayed in Figure 2.10. Several of the identified pathways are known to be involved in tumor formation and progression. For instance, the mitogen-activated protein kinase (MAPK) signaling pathway, which is involved in various cellular functions, including cell proliferation, differentiation and migration, has been implicated in breast cancer metastasis Lee et al. (2007), Keyse (2008). The KEGG pathway in cancers was also selected with high posterior probability. Other interesting pathways are the insulin signaling pathway, which has been linked to the development , progression, and outcome of breast cancer, and purine metabolism which is involved in nucleotide biosynthesis and affects cell cycle activity of tumor cells. In addition, several genes with known association to breast cancer were also selected. For example, several dual specificity phosphatase (DUSP) genes were selected from the MAPK signaling pathway,

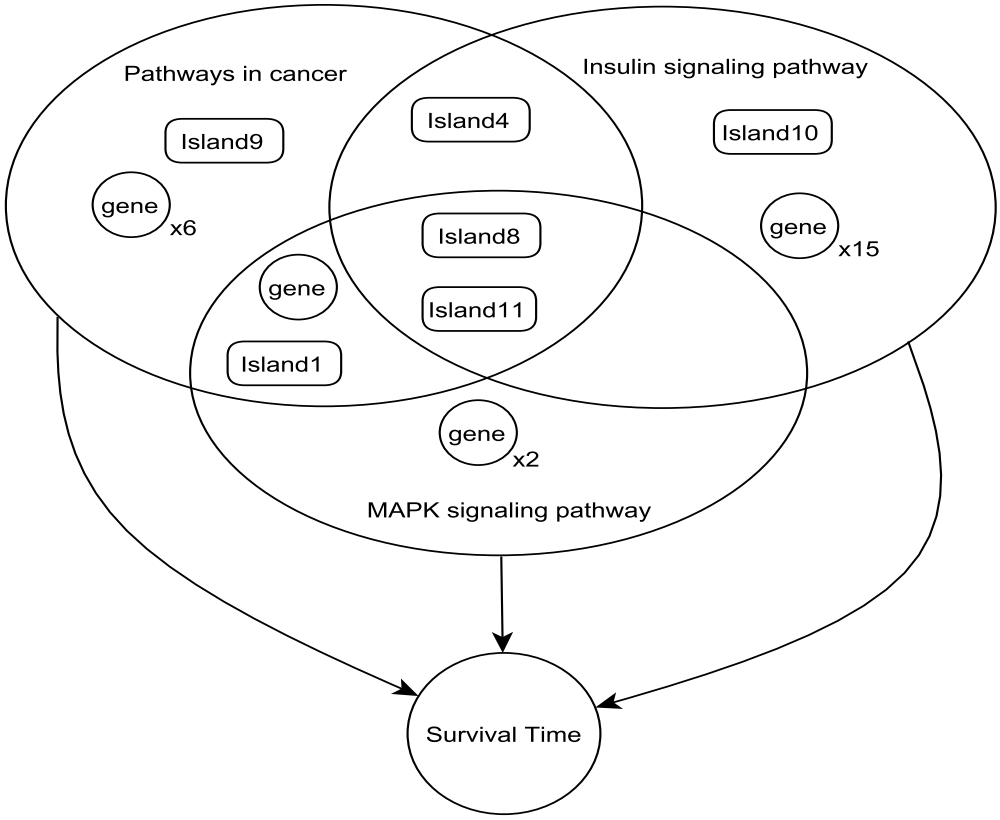

Figure 2.10: Microarray data: Graphical representation of a subset of selected pathways and genes with their corresponding islands. The genes in the islands are listed in Table 2.2.

including DUSP3 found to map in a region that contains the BRCA1 locus which confers susceptibility to breast and ovarian cancer Kamb et al. (1994). Fibroblast growth factor 1 (FGF1) and protein kinase C alpha (PKCalpha), which belong both to the MAPK pathway and the KEGG pathways in cancer, were also selected by the algorithm. The FGF family members have broad mitogenic and cell survival activities, and are involved in a variety of biological processes, including cell growth, tissue repair, tumor growth and invasion, and in particular proliferation of breast cancer cells. PKCalpha has been shown to be overexpressed in some antiestrogen resistant breast cancer cell lines and to be involved in the growth of tamoxifen resistant human breast cancer cells Frankel et al. (2007). Other known genes that were selected include Interleukin 8 (IL8) from the KEGG pathways in cancer, a useful prognostic factor in metastatic breast cancer patients Ahmed et al. (2006); ribosomal protein S6 kinase, 70kDa, polypeptide 1 (RPS6KB1) from the insulin signaling pathway, which is overexpressed in some breast cancer cell lines Kim et al. (2009); and DNA polymerase epsilon (POLE) from the purine metabolism pathway, which is involved in DNA repair and has been shown to be associated with breast cancer Zhou et al. (2008).

2.5 Discussion

We have proposed a model that incorporates biological knowledge from pathway databases into the analysis of DNA microarrays to identify pathways and genes related to a phenotype. Information on pathway membership and gene networks are used to define pathway summaries, specify prior distributions that account for the dependence structure between genes, and define the MCMC moves to fit the model. The gene network prior and the synthesis of the pathway information through PLS bring in additional information that is especially useful in microarray data, where there is low sample size and large measurement error. The performance of the method was evaluated us-

Singleton genes (no direct neighbor selected)
ACACB(10), AGTR2(4), ARPC1B(9), BIRC5(11), C4A(8), CACNB4(2), CCL13(3), CCNB2(5), CD22(7), CD28(7), CDC2(5), CFB(8), CFB(8), CR2(8), CTNNA1(11), CXCL9(3), DIAPH3(9), DOCK1(9), EPHA8(6), EPHB1(6), FBP1(10), FGF1(2, 9, 11), GMPS(1), GYS1(10), GYS2(10), IFNA7(3), IL8(3, 11), ITGA7(9), MAP4K2(2), ORC2L(5), P2RX7(4), PCK1(10), PCK2(10), PCNA(5), PFKL(10), PFKM(10), PFKP(10), PFN1(9), POLD2(1), POLD3(1), POLD4(1), POLE(1), PPARGC1A(10), PRKAR1A(10), RALBP1(11), SELP(7), SHC1(10), SHC1(10), SHC2(10), SLC2A1(11), SORBS1(10), TGFA(11), TMSB4Y(9)
Island 1
ACVR1B(2, 3, 11), ACVR1B(2, 3, 11), TGFB3(2, 3, 5, 11)
Island 2
BUB1(5), MAD2L1(5)
Island 3
C8B(8)8, C9(8)
Island 4
CALM3(10), NOS1(11)
Island 5
CD4(7), HLA-DMA(7), HLA-DMB(7), HLA-DOA(7), HLA-DOB(7), HLA-DPA1(7), HLA-DPB1(7), HLA-DQB1(7), HLA-DRA(7), HLA-DRB1(7), HLA-DRB3(7), HLA-DRB4(7), HLA-DRB5(7)
Island 6
CD8A(7), HLA-A(7), HLA-B(7), HLA-B(7), HLA-E(7) HLA-F(7), HLA-G(7)
Island 7
CLDN11(7), OCLN(7)
Island 8
DUSP3(2), DUSP4(2), MAPK10(2, 10, 11)
Island 9
DVL3(11), FZD9(11), WNT1(11), WNT2B(11)
Island 10
PDPK1(10), PRKCI(10), RPS6KB1(10)
Island 11
F11R(7), GNAI1(6), PLA2G4A(2), PRKCA(2, 11), PRKX(2, 10), PRKY(2, 10), PRKY(2, 10), PTGS2(11)

Table 2.2: Selected 102 genes divided by islands with associated pathway indices (in parenthesis). The pathway indices correspond to: 1-Purine metabolism, 2-MAPK signaling pathway, 3-Cytokine-cytokine receptor interaction, 4-Neuroactive ligand-receptor interaction, 5-Cell cycle, 6-Axon guidance, 7-Cell adhesion molecules (CAMs), 8-Complement and coagulation cascades, 9-Regulation of actin cytoskeleton, 10-Insulin signaling pathway, 11-Pathways in cancer. The threshold for pathway inclusion was set to a marginal posterior probability of 0.45.

ing simulated data, and a breast cancer gene expression data with survival outcomes was used to illustrate its application. The model can lead to the selection of significant genes that would have been missed otherwise, and, as shown in Section 2.4.2, it can achieve better prediction results compared to models that do not treat genes as connected elements that work in groups or pathways.

Several MRF priors for gene selection indicators have been proposed in the literature. It is interesting to compare the parametrization of the MRF used in this paper and in Wei & Li (2007) to the parametrization used in Li & Zhang (2009), where the prior on $\boldsymbol{\gamma}$ is defined as

$$P(\boldsymbol{\gamma}) \propto \exp(\boldsymbol{d}'\boldsymbol{\gamma} + \boldsymbol{\gamma}'\boldsymbol{G}\boldsymbol{\gamma}) \tag{2.24}$$

with $\boldsymbol{d} = d\mathbf{1}_p$, $\mathbf{1}_p$ the unit vector of dimension p and $\boldsymbol{G}$ a matrix with elements $\{g_{ij}\}$ usually set to some constants. While d plays the same role as μ in (2.17), the parametrization using $\boldsymbol{G}$ has a different effect from η on the probability of selection of a gene. This is evident from the conditional probability

$$P(\gamma_j|\gamma_i, i \in N_j) = \frac{\exp(\gamma_j(d + g\sum_{i\in N_j}\gamma_i))}{1 + \exp(d + g\sum_{i\in N_j}\gamma_i)}, \tag{2.25}$$

which can only increase as a function of the number of selected neighbor genes. In contrast, with the parametrization in (2.17), the prior probability of selection for a gene can decrease if none of the neighbors are selected. Although the parametrization is somewhat arbitrary, some care is needed in deciding whether to put a prior distribution on $\boldsymbol{G}$. Allowing $\boldsymbol{G}$ to vary can lead to a phase transition problem, that is, the dimension of the selected model can increase massively. To avoid such problem, the parametrization used in Li & Zhang (2009) requires fixing the hyperparameter $\boldsymbol{G}$, which could lead to the specification of a prior distribution that overwhelms the likelihood function. In the parametrization (2.17), however, η is sampled from its posterior distribution, and thus the influence of the gene network

prior will be mitigated.

Pathway databases are incomplete and the gene network information is often unavailable for many genes. Thus, there may be situations where the dependence structure and the MRF prior specification on the gene selection indicator, γ, cannot be used for all genes. When the only information available is the pathway membership of genes, the prior on γ could be elicited so as to capture other interesting characteristics. For example, we may want a gene to have *a priori* higher probability of being selected when several pathways that contain it are included in the model. We may also want to avoid favoring the selection of a large pathway just because of its size. In such cases, conditional on $\boldsymbol{\theta}$, independent Bernoulli priors can be specified for γ_j by relating the probability of selection to the proportion of included pathways that contain gene j and adjusting for the pathway sizes, p_k, that is,

$$\gamma_j|\boldsymbol{\theta} \sim \text{ Bernoulli}\left(c \cdot \frac{\sum_{k=1}^{K} \theta_k s_{kj}/p_k}{\sum_{k=1}^{K} s_{kj}/p_k}\right), \tag{2.26}$$

where the scalar c is a hyperparameter to be elicited. Notice that in this case it may be be difficult to specify an equivalent prior for the multinomial vectors λ_k in model (2.1) because of the overlapping genes among pathways.

Chapter 3

Bayesian Network for Genomic Data Integration

It has been estimated that about 30% of the genes in the human genome are regulated by microRNAs (miRNAs). These are small RNAs that can down-regulate the levels of mRNAs or proteins in animals and plants. Genes regulated by miRNAs are called targets. Typically, methods for target prediction are based solely on sequence data and on the structure information. In this paper, we propose a Bayesian graphical modeling approach that infers the miRNA regulatory network by integrating expression levels of miRNAs with their potential mRNA targets and, via the prior probability model, with their sequence/structure information. We use a directed graphical model with a particular structure adapted to our data based on biological considerations. We then achieve network inference using stochastic search methods for variable selection that allow us to explore the huge model space via MCMC. A time-dependent coefficients model is also implemented. We consider experimental data from a study on a very well known developmental toxicant causing neural tube defects, hyperthermia. Some of the pairs of target gene and miRNA we identify seem very plausible and warrant future investigation. Our proposed method is general and can be easily applied to other types of network inference by integrating multiple data sources.

3.1 Introduction

One of the major tasks and challenges in the post-genomics era is to decipher how genes and their products (proteins) are regulated. Regulation can happen at transcriptional, post-transcriptional, translational and post-translational level. Transcription is the process of synthesizing a stretch of ribonucleic acids (RNA) based on a specific DNA sequence. Transcriptional regulation can affect whether or not a specific RNA is transcribed as well as the amount of RNA produced. RNA can be regulated post-transcriptionally through degradation or modification of the RNA strand, which can affect its function. A segment of RNA can interact with other genes or proteins or can encode a protein. Translation, the process of forming a protein based on an RNA sequence, can also be positively or negatively regulated. Proteins often undergo post-translational modifications, which can affect their function. An abundant class of small ($\sim$22 nucleotide) RNAs, known as microRNAs (miRNAs), plays crucial regulatory roles in animals and plants, Farh et al. (2005). It has been estimated that at least 30% of the genes in human genomes are regulated by miRNAs, Rajewsky (2006). Genes regulated by miRNAs are generally called 'targets'. The actual mechanism of miRNA regulation is still an active area of research and the complete picture of the regulatory mechanism is still to be understood, Thermann & Hentze (2007). Currently, it is believed that miRNAs regulate their targets either by degrading mRNA post-transcriptionally, Bagga et al. (2005), or by suppressing initiation of protein synthesis, Pillai et al. (2005), and/or by inhibiting translation elongation after initiation of protein synthesis, Petersen et al. (2006).

Many algorithms have been developed to predict potential target sequences for miRNAs based on their specific sequence and structure characteristics. These algorithms mainly use sequence information, hybridization energy for structure prediction, and cross-species comparisons, Rajewsky (2006). Some of the more widely used algorithms include: TargetScanS of Lewis et al. (2005), PicTar of Krek et al. (2005), MIRANDA of John et al.

(2004) and DIANA-microT of Kiriakidou et al. (2004). A comprehensive review of these and other methods can be found in Yoon Yoon & Micheli (2006). Typically a large amount (e.g., hundreds to thousands) of potential targets are predicted by these algorithms, and it can be overwhelming for researchers to search through the candidate targets for those which play critical regulatory roles under particular experimental or clinical conditions.

Our goal is to develop a statistical approach to identify a small set of potential targets with high confidence, making future experimental validation feasible. Since miRNAs down-regulate the expression of their targets, expression profile of miRNAs and their potential targets can be used to infer their regulatory relationships. We propose a Bayesian graphical modeling approach that infers the miRNA regulatory network by integrating these two types of expression levels. We use a directed graphical model with a particular structure adapted to our data based on biological considerations. The model also integrates the sequence/structure information, as generated by the two widely used target prediction algorithms, via the prior probability model. We then achieve network inference using stochastic search methods for variable selection.

We consider experimental data from a study on a very well known developmental toxicant causing neural tube defects, hyperthermia. We have available 23 mouse miRNAs and a total of 1,573 potential targets. We infer their regulatory network under two different treatment conditions and also investigate time-dependent regulatory associations. Some of the pairs of target gene and miRNA we identify seem very plausible and warrant future investigation.

Huang et al. (2007, 2008) have proposed a Bayesian model for the regulatory process of targets and miRNAs which is similar to the one we propose here. However, in their model formulation the authors consider regression coefficients that are constant with respect to the mRNAs, while our formulation allows a more efficient way of selecting gene-miRNA pairs. Also, in

order to achieve posterior inference, we implement a full MCMC procedure while Huang et al. (2007) adopt a variational method that only approximates the posterior distribution. More important, Huang et al. (2007) restrict their search algorithm to a preselected subset of possible gene-miRNA relations, which they select based on the available sequence information, therefore excluding *a priori* a large number of associations that could instead occur in specific experimental conditions, such as hyperthermia.

The paper is organized as follows. Section 3.2 introduces the experimental study and describes the available data, i.e., the expression data of miRNAs and their potential mRNA targets, and the corresponding association scores. Section 3.3 illustrates the proposed modeling approach via a Bayesian graphical model and describes the prior model and the variable selection scheme. Section 3.4 describes how to perform posterior inference and Section 3.5 provides a detailed analysis of the miRNA regulatory network reconstruction based on the available data. Section 3.6 concludes the paper.

3.2 Neural Tube Defects

Neural Tube defects (NTDs) are some of the most common congenital defects with approximately 12 per day in the United States, Finnell et al. (2000). NTDs are generally related to failure of embryonic neural folds to fuse properly along the neuroaxis during development. Studies in both humans and animals suggest a complex genetic component to NTDs, likely involving multiple loci, together with environmental factors. MicroRNAs are believed to play important regulatory roles in mouse development and human disease, see for example Conrad et al. (2006), although detailed regulatory mechanisms are still unknown.

In this paper we consider experimental data from a study on a very well known developmental toxicant causing neural tube defects, hyperthermia. In the study mice are used as the animal model to study NTDs. Time-mated

female C57Bl/6 mice were exposed *in vivo* to a 10 minute hyperthermia or control treatment on gestational day 8.5, when the neural folds are fusing to form the neural tube. Four litters were collected for each treatment at 5, 10, and 24 hours after exposure. Each litter was treated as a single biological sample. MiRNAs and mRNAs were extracted from each sample for expression analysis.

3.2.1 miRNA Expression Levels

As the regulatory network can be very complex, we focus on a small sets of mRNA targets with high confidence. With limited budget available, a pilot study was performed to screen the expression profiles of most of the known (~ 240) mouse microRNAs based on one set of samples, for both heat shock and control at 4 different time points, and using TaqMan miRNA RTPCR assays available at the time (Applied Biosystems, Foster City, CA; provided in collaboration with Ambion, Austin, TX). Of the 240 miRNA evaluated, 50 had none or very low expression at all time points, while 86 had a 2-fold or greater change in expression in response to hyperthermia exposure at one or more time point. From this set of 86 miRNA, we chose a subset of 23 miRNA whose patterns of expression were interesting enough for further analysis and obtained replicate sample sets. The complete experiment was therefore carried out using only this set of 23 miRNAs.

MicroRNA was extracted from each sample at each time point under each experimental condition. Two technical replicates were prepared for RTPCR quantification to confirm the technical reproducibility. In RTPCR experiments, fluorescence techniques are used to detect the amplification of miRNAs to assess their abundance. A fluorescence threshold is determined for an experiment, and the cycle number, which reaches the predetermined threshold level of log2-based fluorescence, is defined as the Ct number. An inverse linear relationship exists between Ct number and the logarithm of input quantity of the gene when the amplification efficiency is perfect, Pfaffl

(2001). The Ct numbers of the miRNA technical replicates were averaged across the two technical replicates.

3.2.2 Target Prediction via Sequence and Structure Information

The sequence and structure information for target prediction was retrieved from the public domain by using the widely used algorithm PicTar of Krek et al. (2005)[1]. A zero or absent PicTar score indicates that the raw score did not exceed a pre-specified threshold, that is, the algorithm suggests no indication of a regulatory association. A matlab script was written to retrieve the RefSeq Ids of all potential targets for the 23 mouse miRNAs of interest. In addition, the current release (September 2009) of 1,209,841 predicted microRNA target sites in 26,697 mouse gene isoforms for 491 mouse miRNAs, generated by the Miranda algorithm of John et al. (2004) was downloaded from MICRORNA.ORG, see Betel et al. (2008). A matlab script was written to retrieve the scores between the 23 miRNAs in our study and putative target genes for the analysis.

3.2.3 Target mRNA Expression Levels

RNA was extracted from each sample at each time point and hybridized to GE Codelink Mouse Whole Genome Microarrays(GE Healthcare Life Sciences, Piscataway, NJ). The slides were scanned and mRNA expression levels were quantified. One biological sample was not prepared properly at hour 10 in the control group, and therefore discarded.

The RefSeq Ids of the probes spotted on the Codelink microarrays were linked to the retrieved potential targets of the 23 miRNAs previously identified. The mRNAs were included in the analysis only if they were among the potential targets predicted by the PicTar and Miranda algorithms. Genes

[1]available at http://pictar.bio.nyu.edu/cgi-bin/new_PicTar_mouse.cgi

with missing or negative values were excluded from the analysis. The expression levels of the remaining mRNAs were then log2 transformed so that both miRNA and mRNA expression were on the log2 scale. A total of 1,573 potential targets was included in the final analysis. The transformed expressions across the 3 time points were centered by subtracting their means.

3.3 Model

We have available expression levels on a set of miRNAs and their potential targets. For each target we are interested in identifying a small number of regulatory associations with high confidence. We have also available sequence information for target prediction in the form of scores of regulatory associations. We propose a Bayesian graphical modeling approach that infers the miRNA regulatory network by integrating the expression data and, via the prior probability model, the sequence/structure information. An important aspect of our methodology is the concept of sparsity, that is, we believe that most genes are regulated by a small number of miRNAs.

3.3.1 A Bayesian Network for Gene & miRNA Expression

We use a directed graphical model (Bayesian Network) with a particular structure adapted to our data that uses a predetermined ordering of the nodes based on biological considerations. This model is able to answer to the baseline question of '*which miRNAs regulate which targets*' and, in addition, allows us to build a fast computational procedure required in such a high-dimensional framework. A graphical representation of the full miRNA network is given in Figure 3.1. Our task is to find a significant subset of edges.

Graphical models are graphs in which nodes represent random variables and the lack of arcs represents conditional independence assumptions, see

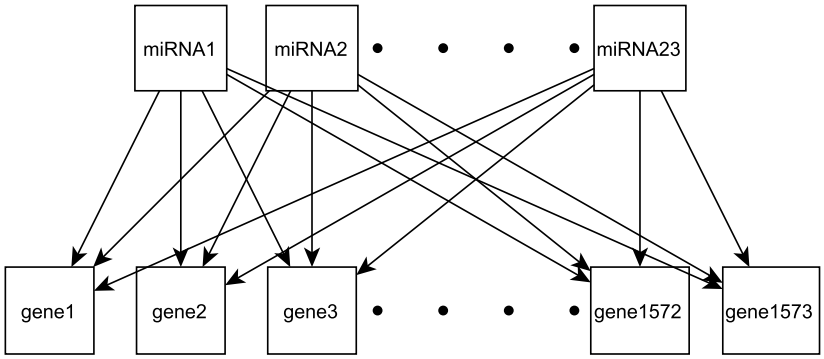

Figure 3.1: Graphical representation of the miRNA regulatory network.

for example Cowell et al. (1999). Graphical models provide a compact representation of joint probability distributions. Here we work with a multivariate normal distribution, and therefore with a Graphical Gaussian model (GGM). A graph $\mathcal{G}$ and the covariance matrix Ω entirely define a GGM $\mathcal{M}$, $\mathcal{M} \equiv (\mathcal{G}, \Omega)$. Arcs can be undirected, indicating symmetric dependencies, or directed, when there is a direction of the dependence. These dependencies can come from prior knowledge or from data analysis. Undirected graphical models have a simple definition of independence, e.g., two nodes A and B are conditionally independent given a third set, C, if all paths between the nodes in A and B are separated by a node in C. Directed graphical models need a specific ordering of the variables. Graphs that do not allow the presence of cycles are called directed acyclic graph (DAG). Conditional independencies in a DAG depend on the ordering of the variables.

We work with a DAG and impose an ordering of the variables such that each target can be affected only by the miRNAs and that the miRNAs can affect only the targets. Let $\mathbf{Z} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_G, \mathbf{X}_1, \ldots, \mathbf{X}_M)$ with $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_G)$ the matrix representing the targets and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_M)$ the miRNAs . Specifically, y_{ng} indicates the normalized averaged $\log_2$ gene expression of gene $g = 1, \ldots, G$ in sample $n = 1, \ldots, N$. These expression

values are biological replicates obtained by averaging two technical replicates. Similarly, x_{nm} indicates the expression of the m-th miRNA in sample n, with $m = 1, \ldots, M$. We have $G = 1,573$ and $M = 23$. In addition, we have $N = 11$ i.i.d. observations under the control status and $N = 12$ i.i.d. observations under hyperthermia. We infer the miRNA regulatory network separately under the two conditions.

Our assumptions are that $\mathbf{Z}$ is a matrix-variate normal variable with zero mean and a variance matrix Ω for its generic row, that is, following Dawid (1981) notation,

$$\mathbf{Z} - \mathbf{0} \sim \mathcal{N}(I_N, \Omega).$$

In addition, we assume that the target genes are independent conditionally upon the miRNAs, that is, $\mathbf{Y}_i \perp\!\!\!\perp \mathbf{Y}_j | \mathbf{X}_1, \ldots, \mathbf{X}_M$ and, without loss of generality, that the miRNAs are independent, that is, $\mathbf{X}_i \perp\!\!\!\perp \mathbf{X}_j$. Note that the marginal distribution of $(\mathbf{X}_1, \ldots, \mathbf{X}_M)$ does not affect the regulatory network. In a Bayesian Network framework these assumptions imply an ordering of the nodes and, consequently, a likelihood factorization of the type:

$$f(\mathbf{Z}) = \prod_{g=1}^{G} f(\mathbf{Y}_g | \mathbf{X}) \prod_{m=1}^{M} f(\mathbf{X}_m), \tag{3.1}$$

where $f(\mathbf{Y}_g|\mathbf{X}) \sim N(\mathbf{X}\beta_g, \sigma_g I_N)$ and $f(\mathbf{X}_m) \sim N(0, \sigma_m I_N)$, with $\beta_g = \Omega^{-1}_{\mathbf{XX}}\Omega_{\mathbf{XY}_g}$ and $\sigma_g = \omega_{gg} - \Omega^T_{\mathbf{XY}_g}\Omega^{-1}_{\mathbf{XX}}\Omega_{\mathbf{XY}_g}$. Here ω_{gg} indicates the g-th diagonal element of Ω and $\Omega_{\mathbf{XX}}$, $\Omega_{\mathbf{XY}}$ are the blocks of the covariance matrix according to the following partition

$$\Omega = \begin{pmatrix} \Omega_{\mathbf{YY}} & \Omega_{\mathbf{YX}} \\ \Omega_{\mathbf{XY}} & \Omega_{\mathbf{XX}} \end{pmatrix}.$$

For $m = 1, \ldots, M$ we have $\sigma_m = \omega_{mm}$.

According to current knowledge, miRNAs down-regulate gene exppression. It therefore seems appropriate to include this information into our

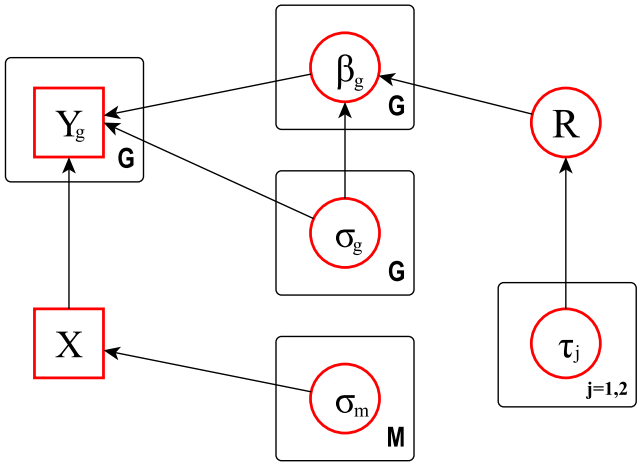

Figure 3.2: Structure of the graphical model

statistical model. This is achieved by specifying negative regression coefficients, i.e. by imposing the constraints $\beta_g > 0$, for $g = 1, \ldots, G$, via the prior model. First, we note that our model is equivalent to the following system of equations:

$$\begin{cases} \mathbf{Y}_1 &= -\mathbf{X}\beta_1 + \epsilon_{\sigma_1}, \\ \vdots & \\ \mathbf{Y}_G &= -\mathbf{X}\beta_G + \epsilon_{\sigma_G}, \end{cases} \tag{3.2}$$

where ϵ_{σ_g} is distributed as a multivariate normal with zero mean and covariance matrix $\sigma_g I_N$. Then, we complete the model specification by specifying prior distributions on the regressions coefficients and error variances. We impose our biological constraints by using Gamma distribution priors for the positive regressions coefficients, $(\beta_{gm}|\sigma_g) \sim Ga(1, c\,\sigma_g)$, and Inverse-Gamma distributions for error variances, $\sigma_g^{-1} \sim Ga((\delta+M)/2, c/2)$. Figure 3.2 shows a graphical representation of our model. Circles indicate parameters and squares observed random variables. The parameters $\boldsymbol{R}$ and τ are involved in the variable selection and are introduced in the Section below.

3.3.2 Prior Model for Variable Selection

The goal of the analysis is to find, for each target, a small subset of miRNAs that regulate that target with high probability. This can be framed into a variable selection problem. Specifically, we can introduce a $(G \times M)$ matrix $\boldsymbol{R}$ with elements $r_{gm} = 1$ if the mth miRNA is included in the regression of the gth target and $r_{gm} = 0$ otherwise. Conditioned upon $\boldsymbol{R}$ expression (3.2) is equivalent to a system of linear equations where the included regressors are only those miRNAs corresponding to $r_{gm} = 1$. To emphasize the variable selection nature of our model we write it as follows:

$$\mathbf{Y}_g = -\mathbf{X}_{(R)}\beta_{g(R)} + \epsilon_{\sigma_g}, \tag{3.3}$$

where $\beta_{g(R)}$ is the vector that is formed by taking only the non-zero elements of β_g and $\mathbf{X}_{(R)}$ is the matrix that is formed by taking only the corresponding columns of $\mathbf{X}$. The goal of our modeling is to infer which elements of the vectors β_g's are non-zero, indicating a relationship between the corresponding genes and miRNAs. This underlying regulatory network is encoded by the association matrix $\boldsymbol{R} = \{r_{gm}\}$. The elements of the vectors β_g's are then stochastically independent, given the regulatory network $\boldsymbol{R}$, and have the following mixture prior distribution:

$$\pi(\beta_{gm}|\sigma_g, r_{gm}) = r_{gm}N(0, c^{-1}\sigma_g) + (1 - r_{gm})I_{[\beta_{gm}=0]}. \tag{3.4}$$

In addition, taking into account the regulatory network, we obtain that $\sigma_g^{-1}|\boldsymbol{R} \sim Ga((\delta + k_g)/2, c/2)$, where k_g is the number of significative miRNAs in the regression of the g-th target.

Mixture priors have been used extensively for variable selection in linear regression settings, see George & McCulloch (1993) for univariate regression and Brown et al. (1998*b*) and Sha et al. (2004) for multivariate models. According to prior (3.4), when $r_{gm} = 0$ then β_{gm} is estimated by 0 and the corresponding column of $\mathbf{X}$ is excluded from the gth equation in model

(3.2). Notice that the dimensions of the matrix $\mathbf{X}$ are such that there are many more columns than rows. In the domain of classical regression, this results in insufficient degrees of freedom to fit the model unless constraints are placed on the regression coefficients β_g's. Conversely, this problem is readily addressed in the Bayesian paradigm and is known as the "small n, large p" framework. The variable selection formulation we adopt here overcomes the somehow rigid structure of the model in Brown et al. (1998*b*), which does not allow to select different predictors for different responses. See also Monni & Tadesse (2009) for an approach based on partition models.

3.3.3 Using Association Scores in the Prior Model

Scores of possible associations between gene-miRNA pairs obtained from sequence/structure information were used to estimate prior probabilities of miRNAs binding to their target genes. Let s_{gm} denote a generic score for gene g and miRNA m, obtained for example by the PicTar algorithm. As previously described, s_{gm} is either positive or, in the case of a regulatory association that is believed to be absent, equal to zero. Also, the PicTar algorithm shrinks small values to zero, setting $s_{gm} = 0$ if $s_{gm} < \varepsilon$ where ε is a pre-specified threshold used by the algorithm. In our model, the Bernoulli random variable r_{gm} indicates whether there is a relationship between gene g and miRNA m. We choose to model the success probability of r_{gm} as a function of the s_{gm} score as follows:

$$P(r_{gm} = 1|\tau) = \frac{\exp[\eta + \tau s_{gm}]}{1 + \exp[\eta + \tau s_{gm}]}, \tag{3.5}$$

where τ is an unknown parameter. We then assume that the elements of $\boldsymbol{R}$ are stochastically independent given τ. Notice that for $s_{gm} = 0$, we have that $P(r_{gm} = 1) = \exp[\eta]/(1 + \exp[\eta])$. Thus, the inverse logit transformation of η can be interpreted as the false negative rate associated with the PicTar thresholding scheme. For a score $s_{gm} > 0$ we have $P(r_{gm} = 1) > \eta$,

with higher scores yield higher prior probabilities of association. We further specify a hyperprior on τ as a gamma distribution $\tau \sim Ga(a_\tau, b_\tau)$, ensuring the positivity of the parameter.

Since we have available multiple prior sources of information, from different sequence/structure algorithms, it makes sense to combine them all by incorporating all scores into the prior distribution using additional τ parameters. For example, in the application of Section 5 we combine PicTar and Miranda scores as

$$P(r_{gm} = 1|\tau_1, \tau_2) = \frac{\exp[\eta + \tau_1 s_{gm} + \tau_2 q_{gm}]}{1 + \exp[\eta + \tau_1 s_{gm} + \tau_2 q_{gm}]}, \tag{3.6}$$

where the q_{gm}'s denote the Miranda scores.

3.3.4 Time-dependent Coefficients Model

The previous model implies that the relation between gene g and miRNA m is constant over time. In the experimental study for which we developed our model there is no dependence between the measurements at different time points, since these observations come from independent units. However, one may still wish to incorporate into the model the fact that relations may possibly change with time. This can be done by allowing different regression coefficients at different time points, as follows:

$$\begin{cases} \mathbf{Y}_1 &= -\mathbf{X}\beta_1 - \mathbf{X}^*_{\mathbf{2}}\beta'_1 - \mathbf{X}^*_{\mathbf{3}}\beta''_1 + \epsilon_{\sigma_1}, \\ \vdots & \\ \mathbf{Y}_G &= -\mathbf{X}\beta_G - \mathbf{X}^*_{\mathbf{2}}\beta'_G - \mathbf{X}^*_{\mathbf{3}}\beta''_G + \epsilon_{\sigma_G}, \end{cases} \tag{3.7}$$

where the $\mathbf{Y_g}$'s are $N \times 1$ vectors and

$$\mathbf{X} = \begin{pmatrix} \mathbf{X_1} \\ \mathbf{X_2} \\ \mathbf{X_3} \end{pmatrix}, \quad \mathbf{X_2^*} = \begin{pmatrix} \mathbf{0} \\ \mathbf{X_2} \\ \mathbf{0} \end{pmatrix}, \quad \mathbf{X_3^*} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{X_3} \end{pmatrix},$$

are the $N \times M$ matrices of the observed values, with $\mathbf{X_1}$, $\mathbf{X_2}$ and $\mathbf{X_3}$ the miRNA expressions collected at the first, the second and the third time point, respectively. The element $\beta_{gm} \in \beta_g$ represents the relation between gene g and miRNA m at the first time point, $\beta_{gm} + \beta'_{gm}$, with $\beta'_{gm} \in \beta'_g$, represents the relation at the second time point and $\beta_{gm} + \beta''_{gm}$, with $\beta''_{gm} \in \beta''_g$, at the third time point.

In order to do variable selection on the elements of β'_g and β''_g we introduce two additional binary matrices $\boldsymbol{R'}$ and $\boldsymbol{R''}$, with a similar role to $\boldsymbol{R}$ in the time-invariant model (3.3). We consider the elements of $\boldsymbol{R'}$ and $\boldsymbol{R''}$ independently distributed and following a Bernoulli distribution with parameter $P(r'_{gm} = 1) = \eta_b = P(r''_{gm} = 1)$. Because of the way we implement the MCMC, see Section 3.4, we do not need to impose the sequence information on the prior on $\boldsymbol{R'}$ and $\boldsymbol{R''}$.

As for the elements of the β_g's vectors, we assume that the elements of the β'_g's and β''_g's vectors are stochastically independent given the regulatory networks $\boldsymbol{R'}$ and $\boldsymbol{R''}$, respectively, and that they have the following prior distributions:

$$\pi(\beta'_{gm}|\sigma_g, r'_{gm}) = r'_{gm} N(0, c^{-1}\sigma_g\zeta) + (1 - r'_{gm}) I_{[\beta'_{gm}=0]},$$

$$\pi(\beta''_{gm}|\sigma_g, r''_{gm}) = r''_{gm} N(0, c^{-1}\sigma_g\zeta) + (1 - r''_{gm}) I_{[\beta''_{gm}=0]},$$

where the hyperparameter ζ, usually ≤ 1, reflects the prior information on the magnitude of the β'_g's and β''_g's.

We can reframe the time-dependent coefficients model in the same way

we have framed model (3.3), that is:

$$\mathbf{Y}_g = -\mathbf{X}_{(R)}\beta_{g(R)} - \mathbf{X}^*_{\mathbf{2}(R')}\beta'_{g(R')} - \mathbf{X}^*_{\mathbf{3}(R'')}\beta''_{g(R'')} + \epsilon_{\sigma_g},$$

where the columns of $\mathbf{X}^*_{\mathbf{2}}$ are selected if the corresponding elements of $\boldsymbol{R'}$ are equal to 1 and the columns of $\mathbf{X}^*_{\mathbf{3}}$ are selected if the corresponding elements of $\boldsymbol{R''}$ are equal to 1, for each equation.

3.4 Posterior Inference

For posterior inference the primary interest is in estimating the association matrix $\boldsymbol{R}$. Here we show that $\boldsymbol{R}$ can be estimated by designing a simple extension of the stochastic search procedures used for variable selection, see George & McCulloch (1993) and Sha et al. (2004), among many others.

We use a Metropolis-Hastings within Gibbs to explore the huge model space and find the most influential predictors. Our model has 23 regressors for each of 1,573 equations, that is a total of $36,179$ regression coefficients for the time invariant model (3.3) and $108,537$ for the time dependent model (3.7). Clearly, exploring such a huge posterior space is challenging. Here we exploit the sparsity of our model, i.e., the belief that most of the genes are well predicted by a small number of regressors, and resort to a Stochastic Search Variable Selection (SSVS) method. A stochastic search allows us to explore the posterior space in an effective way, quickly finding the most probable configurations, i.e., those corresponding to the coefficients that have high marginal probability of $r_{gm} = 1$, while spending less time in regions with low posterior probability.

In order to design this MCMC search we need to calculate the marginal posterior distribution of $\boldsymbol{R}$ by integrating out β_g from the posterior:

$$\begin{aligned} f(\mathbf{Y}_g|\mathbf{X}_{(R)}, \sigma_g, \boldsymbol{R}) \quad \propto \quad & \frac{1}{(2\pi)^{(N-k_g)/2}\sigma_g^{N/2}c^{k_g}}|U_g|^{1/2} \\ & \exp[\tfrac{1}{2\sigma_g}q_g]\Phi_{k_g}(0; -U_gC_g, \sigma_gU_g), \end{aligned}$$

where $U_g = (\mathbf{X^T}_{(R)}\mathbf{X}_{(R)})^{-1}$, $C_g = \mathbf{Y}_g^T\mathbf{X^T}_{(R)} - (\sigma_g^{1/2}/c)\mathbf{1}_{k_g}$ and $q_g = \mathbf{Y}_g^T\mathbf{Y}_g - C_gU_gC_g^T$ and with k_g the number of selected regressors. $\Phi_{k_g}(0; -U_gC_g, \sigma_gU_g)$ indicates the cdf of a multivariate normal, with mean $-U_gC_g$ and covariance matrix σ_gU_g, calculated at the zero vector.

Our algorithm consists of three steps. The first step is based on the marginal posterior distribution conditioned upon τ_1, τ_2, σ_g and consists of either the addition or the deletion of one arrows in our graphical model or the swapping of two arrows. The second step generates new values of τ_1 and τ_2 from their posterior distribution. In the last step values of all the error variances σ_g are updated. The un-normalized full conditionals needed for the Gibbs sampler can be derived from the conditional independencies of our model, as given in Figure 3.2. We now describe the three steps of the algorithm:

1. We use one of two types of moves to update $\boldsymbol{R}$:

   - with probability ϕ, we add or delete an element by choosing at random one component in the current $\boldsymbol{R}$ and changing its value;
   - with probability $1-\phi$, we swap two elements by choosing independently at random one 0 and one 1 in the current $\boldsymbol{R}$ and changing the value of both of them.

   The proposed $\boldsymbol{R}^{new}$ is then accepted with a probability that is the ratio of the relative posterior probabilities of the new versus the current model:

   $$\min\left[\frac{f(\mathbf{Y}|\mathbf{X}_{(R^{new})}, \boldsymbol{R}^{new}, \sigma_g)\pi(\boldsymbol{R}^{new}|\tau)}{f(\mathbf{Y}|\mathbf{X}_{(R^{old})}, \boldsymbol{R}^{old}, \sigma_g)\pi(\boldsymbol{R}^{old}|\tau)}, 1\right]. \tag{3.8}$$

   Because these moves are symmetric, the proposal distribution does not appear in the previous ratio.

2. In order to update τ_1 and τ_2 we employ another two Metropolis steps. The proposal is made via a truncated normal random walk kernel. The

proposed τ_1^{new} is then accepted with probability:

$$\min\left[\frac{\pi(\boldsymbol{R}|\tau_1^{new})\pi(\tau_1^{new})q(\tau_1^{old};\tau_1^{new})}{\pi(\boldsymbol{R}|\tau_1^{old})\pi(\tau_1^{old})q(\tau_1^{new};\tau_1^{old})},1\right], \tag{3.9}$$

where $q(\tau_1^{old};\tau_1^{new})$ is a truncated normal with mean τ_1^{new} and truncation at 0, given the constraint of positivity on τ_1. The variance of this distribution represents the tuning parameter and has to be set in such a way to explore the parameter space and have a good acceptance rate, see also Section 3.5. An analogous step is performed for $\tau 2$.

3. For $g = 1, \ldots, G$ we update the error variance σ_g using a Metropolis step where the proposal distribution $q(\sigma_g^{old};\sigma_g^{new})$ is a Gamma distribution with parameters a_σ and b_σ. The proposed new value is then accepted with probability:

$$\min\left[\frac{f(\mathbf{Y}|\mathbf{X}_{(R)},\boldsymbol{R},\sigma_g^{new})\pi(\sigma_g^{new})q(\sigma_g^{old};\sigma_g^{new})}{f(\mathbf{Y}|\mathbf{X}_{(R)},\boldsymbol{R},\sigma_g^{old})\pi(\sigma_g^{old})q(\sigma_g^{new};\sigma_g^{old})},1\right]. \tag{3.10}$$

To obtain an efficient exploration of the parameter space with set $a_\sigma = \sigma_g^{old}/b_\sigma$ and $b_\sigma = e/\sigma_g^{old}$, where e represents the variance of the proposal distribution and can be set to obtain wished acceptance ratio.

Posterior inference can then be performed based on the MCMC output using the marginal probabilities of the singles r_{gm}'s.

The MCMC algorithm for the time-dependent coefficient model (3.7) is pretty similar to the procedure described above, the main difference being that at the first step we update either $\boldsymbol{R}$, $\boldsymbol{R'}$ or $\boldsymbol{R''}$. We then derive the marginal posterior distribution $f(\mathbf{Y}_g|\mathbf{X}_{(R)},\boldsymbol{R})$ for the time dependent model obtaining:

$$
\begin{aligned}
f(\mathbf{Y}_g|\mathbf{X}_{(R)}, \mathbf{X}^*_{\mathbf{2}(R')}, \mathbf{X}^*_{\mathbf{3}(R'')}, \boldsymbol{R}, \boldsymbol{R'}, \boldsymbol{R''}, \sigma_g) &= (2\pi)^{-\frac{n-k_g}{2}} \sigma_g^{-\frac{n}{2}} c^{-k_g - \frac{k_{2g}+k_{3g}}{2}} \\
&\quad \zeta^{-\frac{k_{2g}+k_{3g}}{2}} |A_g|^{-\frac{1}{2}} |C_g|^{-\frac{1}{2}} \\
&\quad |E_g|^{-\frac{1}{2}} \exp[\tfrac{1}{2\sigma_g} q_g] \\
&\quad \Phi_{k_g}(0; -E_g^{-1}F_g, \sigma_g E_g^{-1}),
\end{aligned}
$$

with

$$
\begin{aligned}
q_g &= \mathbf{Y}_g^T\mathbf{Y}_g - \mathbf{Y}_{2g}^T\mathbf{X}_{2(R')}A_g^{-1}\mathbf{X}_{2(R')}^T\mathbf{Y}_{2g} - \mathbf{Y}_{3g}^T\mathbf{X}_{3(R'')}C_g^{-1}\mathbf{X}_{3(R'')}^T\mathbf{Y}_{3g} \\
&\quad -F_g^T E_g^{-1} F_g, \\
F_g &= -\mathbf{X}_{(R)}^T\mathbf{Y}_g + \mathbf{X}_{3(R)}^T\mathbf{X}_{3(R'')}C_g^{-1}\mathbf{X}_{3(R'')}^T\mathbf{Y}_{3g} + \mathbf{X}_{2(R)}^T\mathbf{X}_{2(R')}A_g^{-1}\mathbf{X}_{2(R')}^T\mathbf{Y}_{2g} \\
&\quad -\sigma_g^{1/2} c^{-1}\mathbf{1}_{k_g}, \\
E_g &= \mathbf{X}\mathbf{X}^T - \mathbf{X}_{2(R)}^T\mathbf{X}_{2(R')}A_g^{-1}\mathbf{X}_{2(R')}^T\mathbf{X}_{2(R)} - \mathbf{X}_{3(R)}^T\mathbf{X}_{3(R'')}C_g^{-1}\mathbf{X}_{3(R'')}^T\mathbf{X}_{3(R)}, \\
A_g &= (\mathbf{X}_{2(R')}^T\mathbf{X}_{2(R')} + (c\zeta)^{-1}I_{k_{2g}}) \\
C_g &= (\mathbf{X}_{3(R'')}^T\mathbf{X}_{3(R'')} + (c\zeta)^{-1}I_{k_{3g}})
\end{aligned}
$$

and $\mathbf{Y}_g^T = (\mathbf{Y}_{1g}^T, \mathbf{Y}_{2g}^T, \mathbf{Y}_{3g}^T)$; k_{2g} and k_{3g} are the number of selected β'_{gm} and β''_{gm}. We can now write the first step of the MCMC as:

1′. We first select which of the three matrices to update. We choose $\boldsymbol{R}$ with probability λ and $\boldsymbol{R'}$ (or $\boldsymbol{R''}$) with probability $(1-\lambda)/2$. We then use the same add/delete or swap scheme described above and we accept the proposed $\boldsymbol{R}^{new}$ (or $\boldsymbol{R'}^{new}$ or $\boldsymbol{R''}^{new}$). For $\boldsymbol{R}$ the acceptance probability is:

$$
\min\left[\frac{f(\mathbf{Y}|\mathbf{X}_{(R^{new})}, \mathbf{X}^*_{\mathbf{2}(R'^{old})}, \mathbf{X}^*_{\mathbf{3}(R''^{old})}, \boldsymbol{R}^{new}, \boldsymbol{R'}^{old}, \boldsymbol{R''}^{old})\pi(\boldsymbol{R}^{new}|\tau)}{f(\mathbf{Y}|\mathbf{X}_{(R^{old})}, \mathbf{X}^*_{\mathbf{2}(R'^{old})}, \mathbf{X}^*_{\mathbf{3}(R''^{old})}, \boldsymbol{R}^{old}, \boldsymbol{R'}^{old}, \boldsymbol{R''}^{old})\pi(\boldsymbol{R}^{old}|\tau)}, 1\right]
$$

and similarly if $\boldsymbol{R'}$ or $\boldsymbol{R''}$ is selected. Note that to perform this step we need to use only the prior distribution of the selected matrix.

This algorithm can be run either without any constraint on the moves relative to $\boldsymbol{R}$, $\boldsymbol{R'}$ and $\boldsymbol{R''}$ or with the constraint that the elements of $\boldsymbol{R'}$ (or $\boldsymbol{R''}$)

can be selected only when the corresponding element of $\boldsymbol{R}$ is already selected and that the elements of $\boldsymbol{R}$ can be eliminated only when the corresponding element of $\boldsymbol{R'}$ and $\boldsymbol{R''}$ are not selected. For our application we adopted the constrain strategy. To implement this we do not need to add the ratio of the proposal distributions into (3.8), since we use symmetric moves. This choice, jointly with some empirical results (not reported here), led us to not use the PicTar-Miranda information for the prior distribution of $\boldsymbol{R'}$ and $\boldsymbol{R''}$, because the selecting constraints imply that the prior probability of selecting the generic element r'_{gm} (or r''_{gm}) already depends on the PicTar-Miranda information through the prior probability on the corresponding element r_{gm}. This also implies a faster computational procedure in comparison with the option of including the PicTar-Miranda information into the prior of $\boldsymbol{R'}$ and $\boldsymbol{R''}$.

3.5 Neural Tube Defects Application

We now apply our model to analyze the data described in Section 3.2, combining miRNA and mRNA expression levels with sequence information. Our model allows us to identify significant miRNAs for each target, possibly along the time.

3.5.1 Parameter Settings

We first need to set the values of the hyperparameters of the model. A normal distribution with variance σ^2 corresponds to a normal truncated at zero with variance $\approx 0.7\sigma^2$. The parameter c of the prior distribution of the regression coefficients β_{gm} can be interpreted as a correction factor. We therefore decided to set $c = 0.7$. To complete the specification of the prior distribution on σ_g we set $\delta = 3$, obtaining a distribution centered on half the sample variance of $\mathbf{Y}_g$, expecting our model to be able to explain around 50% of the variability of the data.

In our variable selection framework, the parameter η of the Bernoulli distribution (3.6) reflects the prior belief about the percentage of significant coefficients in the model. In this application, having 23 regressors for each of the 1,573 equations, we set $\eta = -2.5$ to obtain a prior expected number of regressors between 1 and 2. For the more computational expensive time dependent model we set $\eta = -2.5$ and $\eta_b = 0.05$, to avoid memory problems. We also set the hyperparameters $a_\tau = 1.5$ and $b_\tau = 0.2$ to obtain a Gamma distribution that gives high probability to a broad set of values of τ_1 and τ_2, taking into account the scale of values that come from PicTar and Miranda algorithms. However the posterior distributions we obtained, in all the different chains we ran, showed that this parameter setting is not strongly informative. When running MCMC's we have set the variance of the truncated normal proposal distribution of τ_1 and τ_2 equal to 0.01 to obtain a acceptance rate close to 25%.

We ran two different chains for each of the four possible combinations, the time invariant model for the control and the hyperthermia group and the time dependent model for the control and hyperthermia group. We used either adding/deleting or swapping moves with equal probability at each step of the chain; we assigned a probability of $\lambda = 0.5$ to the move that updates $\boldsymbol{R}$ and then probability 0.25 to each of the moves that update $\boldsymbol{R'}$ and $\boldsymbol{R''}$. In all cases, after the initial burn-in, both chains mostly explored the same region of the parameter space corresponding to configuration of $\boldsymbol{R}$ with high posterior probability. In general, we found good agreement between the two chains, which were run from different starting points. To be more precise, correlations between the posterior probabilities of the two chains ranged from 0.84 to 0.88.

Figure 3.3 gives the summary trace plots for the number of selected coefficients β_{gm} and corresponding log-posterior probabilities for the time invariant model on the control group. In this case the chain was run for one million iterations, from a starting randomly chosen set of 1,000 arrows, and mostly

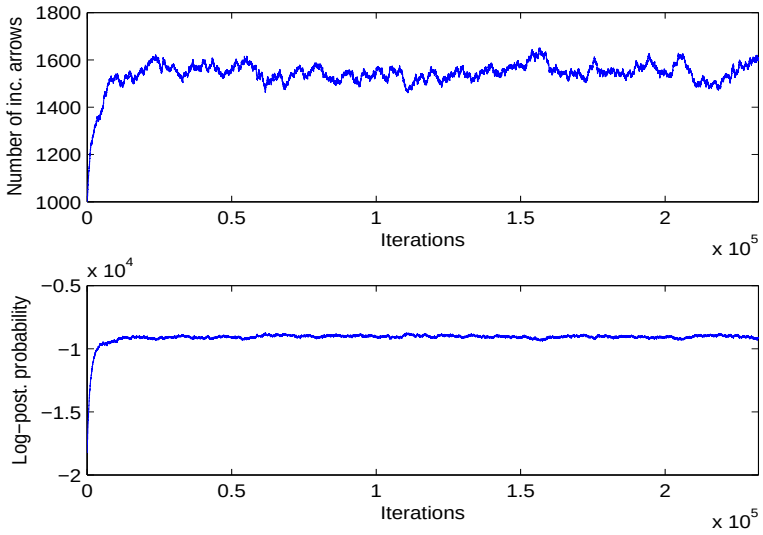

Figure 3.3: Trace plot for number of selected arrows and for the log-posterior probability for the time invariant model

visited models with roughly 1,500 edges, i.e., on average almost 1 edge per gene, a number not too far from the prior specification.

3.5.2 Results

The huge number of potential coefficients in the model implies that the weight of a single coefficient toward the posterior probability of the entire model can be potentially very small. Also, due to sparsity there may be many models with almost the same (small) posterior probability. Because of this, it is good practice to perform posterior inference based on the marginal posterior probability of the single coefficients, rather than on their joint distribution. These posterior probabilities of the presence of single interactions, that is, $P(r_{gm} = 1|\boldsymbol{Y}, \boldsymbol{X})$ can be estimated directly from the MCMC samples by taking the proportion of MCMC iterations for which $r_{gm} = 1$.

The small sample size of our experimental groups does not allow us to

create a validation set, and therefore all the samples are used to fit the model. Selected models are then evaluated based on the R^2 statistic, calculated using the posterior mean of regression coefficients.

As expected, when more covariates are included into the model, based on their posterior probabilities, the statistics R^2 increases, indicating that the ordering created by the posterior probabilities correctly indicates the significative variables. For the time invariant model a threshold of 0.15, corresponding to 1,720 included edges, gave an R^2 of 0.31, for the control group, and of 0.32 for the hyperthermia group, with 1,864 included edges. An identical behavior was observed for the additional coefficients of the time dependent model, i.e., when the number of included β''s and β'''s increases then the quality of the fitting improves; with a threshold of 0.15 for β's and a threshold of 0.5 for β''s and β'''s we obtain a $R^2 = 0.32$ for the control group, including 1,919 β's, 164 β''s and 165 β'''s, and a $R^2 = 0.37$ for the hyperthermia group, including 2,053 β's, 210 β''s and 210 β'''s.

In an effort to assess whether our model correctly selects miRNAs that under-regulate targets gene, we also calculated the ordinary least square estimates of the regression coefficients and checked how many of them were negative, see Appendix B for the calculation of the OLS estimates. Notice that this approach does not impose the negative constraint on β's. By including all coefficients with posterior probability greater than .2 we obtain that, 96.0% and 96.3% of the coefficients, for the control and hyperthermia group, respectively, were correctly negative.

By exploring the regulatory network as a function of the posterior probability of the arrows we found that, for the time invariant model on the control group, a posterior probability cut-off of 0.8 selected 43 arrows between 41 target genes and 11 miRNAs. These correspond to an expected rate of false detection (Bayesian FDR) of 10.5%, that we calculated, following Newton et al. (2004), as

$$FDR = \mathrm{C}(\kappa)/|J_\kappa|$$

where $\mathrm{C}(\kappa) = \sum_{g,m} \psi_{gm} I_{[\psi_{gm} \leq \kappa]}$ and $\psi_{gm} = 1 - P(r_{gm} = 1 | \boldsymbol{Y}, \boldsymbol{X})$, with $|J_\kappa|$ the size of the list ($|J_\kappa| = \sum_{g,m} I_{[\psi_{gm} \leq \kappa]}$). We set $\kappa = 1 - k$ with k the chosen threshold (i.e., 0.8). For the hyperthermia-treated group, the same cut-off led to 59 selected arrows, between 59 target genes and 6 miRNAs, corresponding to a a Bayesian FDR of 8.9%. Increasing the cut-off value to 0.9 identified 20 target genes and 7 miRNAs (corresponding to 22 arrows, with a Bayesian FDR of 5.3%) in the control group and 31 target genes and 4 miRNAs (corresponding to 31 arrows, with a Bayesian FDR of 3.8%) in the hyperthermia group.

Figure 3.4, produced using GraphExplore of Wang et al. (2004), displays the selected network for the control group using a threshold of 0.8 on the posterior probability under the time invariant model. A close look at the pairs of target genes and miRNAs with high posterior probabilities reveals that some of the regulatory relationships seem plausible and warrant future investigation. Using a posterior probability cut-off of 0.9 for both time-dependent and independent analysis of control-treatment data, 7 miRNAs (miR-32, 181d, 213, 223, 299-5p, 367, 375) had at least one target gene for a total of 70 gene targets. The same cut-off for both time-dependent and independent analyses of hyperthermia data produced four miRNAs (miR-142-3p, 299-5p, 367, 423) with at least one target gene for a total of 45 genes. Overall there were 9 miRNAs and 86 gene targets identified with posterior probability of at least 0.9. Of the 107 miRNA-gene target associations with a posterior probability of at least 0.9, 10 were predicted by Miranda only, 2 by PicTar only, and 7 by both Miranda and PicTar. 70 of the gene targets identified were associated with miR-367, a pluripotency-specific marker in human and mouse ES cells, Li et al. (2009). MiR-367 expression steadily decreased over time in control and hyperthermia-treated embryos, which is consistent for a marker of pluripotency in a differentiating embryo. Also, 20 of the gene targets were associated with miR-299-5p, which has been shown to regulate de novo expression of osteopontin, a protein that plays a role in

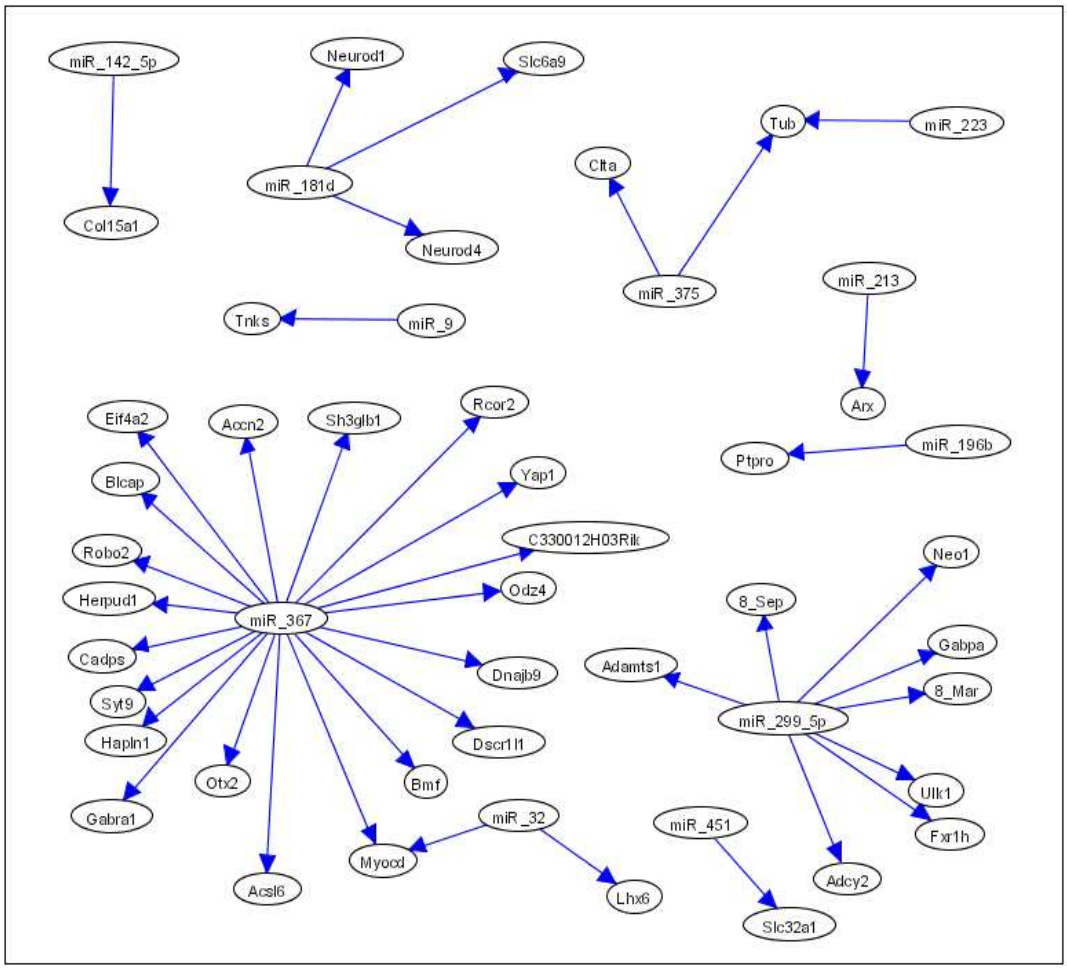

Figure 3.4: Selected network for the control group using a threshold of 0.8 on the posterior probability

enhancing proliferation and tumorigenicity, Shevde et al. (2009). The gene-miRNA associations identified by our methods are potentially interesting for follow-up NTD studies.

It is also interesting to look at the inference on the regression coefficients. Figure 3.5 shows the estimates of the significant β_{gm}'s for the time invariant model under hyperthermia condition. Each bar in the plot represents the 1,573 regression coefficients for one of the 23 miRNAs. Non-zero values correspond to the posterior mean estimates of the best β_{gm}'s with posterior inclusion probability above 0.15 (all other β's are estimated by zero). Notice, for example, how miRNAs miR-423, corresponding to the 22^{nd} bar, and miR-375, corresponding to the 16^{th} bar, play an important role into the down-regulatory mechanism.

Let us finally comment on the posterior inference on τ_1 and τ_2. These parameters measure the influence of the prior information on the posterior inference. Posterior inference on these parameters depends on the value given to η. When selecting edges the hyperparameter η represents the weight assigned to the data and, consequently, τ_1 and τ_2 play the role of the weight of the prior sequence information derived from the PicTar and Miranda algorithms, respectively. The bigger the value of η the more the posterior distribution of τ_j will be concentrated around small values. Besides this general rule, inference on the τ_j's generally depends on the concordance between data and prior information, the number of observations and the number of parameters in the model. According to our results, information extracted from PicTar plays a much bigger role than the one from Miranda. With $\eta = -2.5$, the posterior distribution of τ, for the control group, is concentrated around values that imply a 12% increase on the prior probability of $r_{gm} = 1$ for edges with corresponding high PicTar scores. For the hyperthermia group the corresponding percentage is 14%; when using the time dependent model the prior probability of $r_{gm} = 1$ increase by 13% and 15%, respectively for the control and hyperthermia group. The behaviour of the

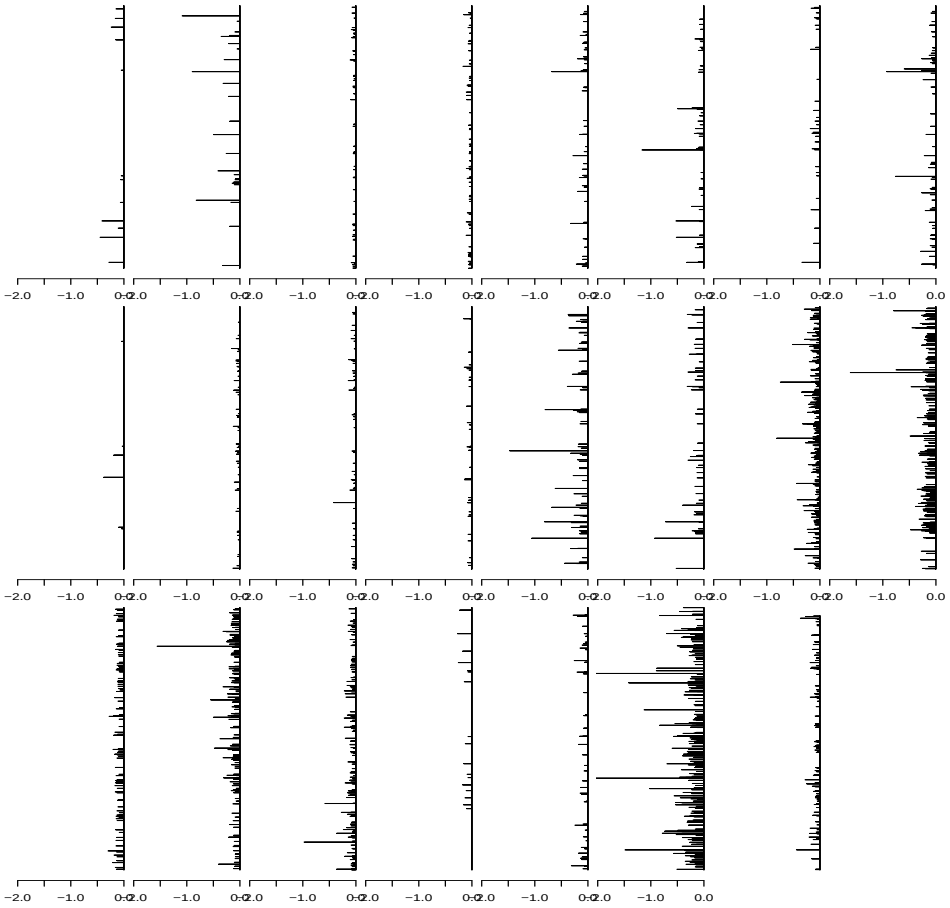

Figure 3.5: Estimation of the significant β_{gm} for the time invariant model under hyperthermia condition

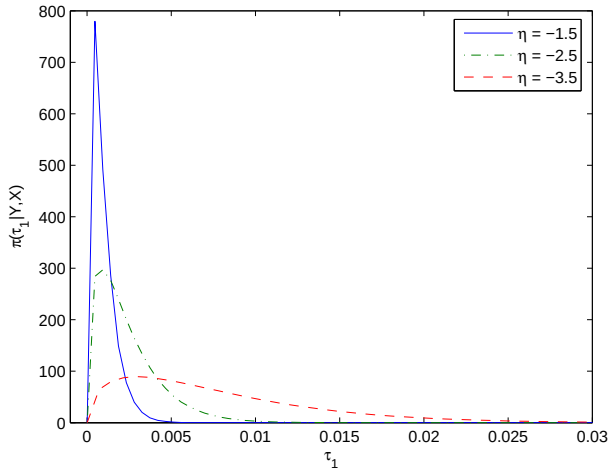

Figure 3.6: Density Kernel estimate using the not time dependent model for the control group

posterior distribution of τ_1, setting different values of η, is summarized in Figure 3.6. The scale of the estimates compensates the very large values we observe for some of the PicTar scores. We can clearly see how the posterior distribution concentrates to bigger values when η decreases. With $\eta = -3.5$ the prior probability of $r_{gm} = 1$ increase by 48% and, setting $\eta = -4.5$, the probability of selecting edges with corresponding high PicTar scores is more than 2 times bigger of the corresponding prior probability for edges with $s_{gm} = 0$.

3.6 Conclusions

We have proposed a Bayesian graphical modeling approach that infers the miRNA regulatory network by integrating expression levels of miRNAs with their potential mRNA targets and, via the prior probability model, with their sequence/structure information. Our model is able to incorporate multiple data sources directly into the prior distribution avoiding arbitrary prior data

synthesis. We have used stochastic search variable selection methods to infer the miRNA regulatory network. We have considered experimental data from a study on a very well known developmental toxicant causing neural tube defects, hyperthermia. The analysis has involved 23 mouse miRNAs and a total of 1,573 potential targets. Our goal was to identify a small set of potential targets with high confidence. Some of the pairs of target gene and miRNA selected by our model seem promising candidates for future investigation. In addition, the time-dependent model has achieved significant improvement in the percentage of explained variance, only slightly increasing the size of the selected model. Our proposed modeling strategy is general and can easily be applied to other types of network inference by integrating multiple data sources.

An interesting feature of our inference is the somehow poor concordance of the prior information, i.e., PicTar and Miranda scores, with the data. This phenomenon has been observed by other authors in models for data integration. Wei & Li (2008), for example, attribute this to the fact that our knowledge of biological processes is not complete and can potentially include errors and therefore induce misspecified edges on the networks. They also suggest to first check the consistency of the prior information with the available data. In our case, if the correlation between a miRNA and a target gene is very small we may want to remove the edge from the network. On the other hand, given the limited number of observations typical of experimental studies in genomics, it would seem important to retain as much, possibly accurate, prior information as possible. This important aspect of models for data integration certainly deserves future investigation.

Extensions and generalizations of our model are possible. One future avenue we intent to pursue is trying to relax the assumption on the conditional independence of the targets given the miRNAs. This assumption is necessary in order to integrate out the covariance matrix, as in Brown et al. (1998*b*), and still allow the selection of individual relations between a gene and a

miRNA. Looking at this as a computational issue, it may be possible to still sample the values of this huge covariance matrix in the MCMC, perhaps by reducing the number of non-zero elements via the prior information on the gene network.

Appendix A

MCMC scheme for sampling (θ, γ)

We now describe the MCMC steps for (θ, γ) in more detail. For interpretability, as previously described, no empty pathways or orphan genes are proposed during sampling and, for identifiability, selecting the same set of genes for different pathways is not allowed. At each iteration, only one pathway and/or a gene are proposed to be added or removed.

(1) Change inclusion status of both gene and pathway – randomly choose between addition (move 1.i in Figure 2.2) or removal(1.ii).

(1.i) Add a pathway and a gene:
First select a pathway that is not included in the model and has none of its member genes in the model ($\theta_k^{old} = 0$ and $p_{k\gamma}^{old} = 0$). Randomly choose one gene from the pathway ($\gamma_j^{old} = 0$) and propose including both the pathway and the gene, i.e., set $\theta_k^{new} = 1,\ \gamma_j^{new} = 1$. The move is accepted with probability

$$\min\left\{1, \frac{f(\boldsymbol{\theta}^{new}, \boldsymbol{\gamma}^{new}|\boldsymbol{T}, Y)}{f(\boldsymbol{\theta}^{old}, \boldsymbol{\gamma}^{old}|\boldsymbol{T}, Y)} \cdot \frac{p_k \cdot \sum_{r=1}^{K} I\{\theta_r^{old}=0, p_{r\gamma}^{old}=0\}}{\sum_{r=1}^{K} I\{\theta_r^{new}=1, p_{r\gamma}^{new}=1, cond1, condId1\}}\right\}, \quad (11)$$

where *cond1* and *condId1* are explained in move type (1.ii) below.

(1.ii) Remove a pathway and a gene:
This move is the reverse of (1.i) described above. We first select a pathway that is included in the model and has only one of its

member genes in the model ($\theta_k^{old} = 1$ and $p_{k\gamma}^{old} = 1$). In addition, this included gene ($\gamma_j^{old} = 1$) may not be the sole representative for other included pathways, to ensure that no empty pathway is created. Furthermore, identical sets of genes from different selected pathways cannot be created. These constraints corresponds, respectively, to *cond1* and *condId1* in the proposal ratios (11) and (12). We attempt to remove both the pathway and the gene, i.e., set $\theta_k^{new} = 0,\ \gamma_j^{new} = 0$ and accept the move with probability

$$\min\left\{1, \frac{f(\boldsymbol{\theta}^{new}, \boldsymbol{\gamma}^{new}|\boldsymbol{T}, Y)}{f(\boldsymbol{\theta}^{old}, \boldsymbol{\gamma}^{old}|\boldsymbol{T}, Y)} \cdot \frac{\sum_{r=1}^{K} I\{\theta_r^{old} = 1, p_{r\gamma}^{old} = 1, cond1, condId1\}}{p_k \cdot \sum_{r=1}^{K} I\{\theta_r^{new} = 0, p_{r\gamma}^{new} = 0\}}\right\}, \tag{12}$$

(2) Change the inclusion status of gene but not pathway – randomly choose between addition (2.i) or removal(2.ii).

(2.i) Add a gene in an already included pathway:
First select a pathway already included in the model and that has some member genes that could potentially be added ($\theta_k^{old} = 1$ and $p_k > p_{k\gamma}^{old}$). Let G be the set of pathways that satisfy these conditions. Choose one of the non-included genes from this pathway ($\gamma_j^{old} = 0$) and attempt to add it, i.e, set $\theta_k^{new} = \theta_k^{old} = 1,\ \gamma^{new} = 1$. The proposal is accepted with probability

$$\min\left\{1, \frac{f(\boldsymbol{\theta}^{new}, \boldsymbol{\gamma}^{new}|\boldsymbol{T}, Y)}{f(\boldsymbol{\theta}^{old}, \boldsymbol{\gamma}^{old}|\boldsymbol{T}, Y)} \cdot \frac{\sum_{r=1}^{K} I\{\theta_r^{old} = 1, p_r > p_{r\gamma}^{old}\} \cdot \sum_{r\in G} \frac{1}{p_{r\gamma}^{new(cond2\gamma, condId2\gamma)}}}{\sum_{r=1}^{K} I\{\theta_r^{new} = 1, p_{r\gamma}^{new} > 1, cond2_\theta, condId2_\theta\} \cdot \sum_{r\in G} \frac{1}{p_r - p_{r\gamma}^{old}}}\right\}, \tag{13}$$

where '$cond2_\theta$', '$cond2_\gamma$', '$condId2_\theta$' and '$condId2_\gamma$' are explained in move type (2.ii) below.

(2.ii) Remove a gene from an already included pathway:
This move is the reverse of (2.i) described above. We first select a pathway already included in the model and that has more than one of its member genes included in the model ($\theta_k^{old} = 1, p_{k\gamma}^{old} > 1$). In addition, at least one of the included genes from this pathway may not be the sole representative for other included pathways and its removal would not create an identifiability problem – this corresponds to constraints '$cond2_\theta$' and '$condId2_\theta$'in the proposal ratios of (13) and (14). Once the pathway is selected, choose a gene among the eligible candidates, that is, an included member gene ($\gamma_j^{old} = 1$) which is not the sole representative for other included pathways and whose removal does not create an identifiability problem – this corresponds to constraints '$cond2_\gamma$' and '$condId2_\gamma$'. Constraints '$cond2_\theta$' for pathways, and '$cond2_\gamma$' for genes, will ensure that no empty pathways are created after the proposed move. Leave the pathway status unchanged and attempt to remove the selected gene, i.e., set $\theta_k^{new} = \theta_k^{old} = 1,\ \gamma_j^{new} = 0$. The proposed move is accepted with probability

$$\min\left\{1, \frac{f(\boldsymbol{\theta}^{new}, \boldsymbol{\gamma}^{new}|\boldsymbol{T}, Y)}{f(\boldsymbol{\theta}^{old}, \boldsymbol{\gamma}^{old}|\boldsymbol{T}, Y)} \cdot \frac{\sum_{r=1}^{K} I\{\theta_r^{old} = 1, p_{r\gamma}^{old} > 1, cond2_\theta, condId2_\theta\} \cdot \sum_{r\in G} \frac{1}{p_r - p_{r\gamma}^{new}}}{\sum_{r=1}^{K} I\{\theta_r^{new} = 1, p_r > p_{r\gamma}^{new}\} \cdot \sum_{r\in G} \frac{1}{p_{r\gamma}^{old(cond2_\gamma, condId2_\gamma)}}}\right\}. \tag{14}$$

(3) Change inclusion status of pathway but not gene – randomly choose between addition (3.i) or removal(3.ii).

(3.i) Add a pathway but leave genes' status unchanged:
First select a pathway that is not included in the model but has some of its member genes included in the model through other pathways ($\theta_k^{old} = 0$ and $p_{k\gamma}^{old} \geq 1$). Attempt to add the pathway but leave the status of its member genes unchanged, i.e., set $\theta_k^{new} = 1$. The proposed move is accepted with probability

$$\min\left\{1, \frac{f(\boldsymbol{\theta}^{new}, \boldsymbol{\gamma}^{new}|\boldsymbol{T}, Y)}{f(\boldsymbol{\theta}^{old}, \boldsymbol{\gamma}^{old}|\boldsymbol{T}, Y)} \cdot \frac{\sum_{r=1}^{K} I\{\theta_r^{old} = 0, p_{r\gamma}^{old} \geq 1, condId3\}}{\sum_{r=1}^{K} I\{\theta_r^{new} = 1, p_{r\gamma}^{new} \geq 1, cond3\}}\right\},$$

where *condId3* means that it is not possible to select a pathway whose selected genes form the entire set of selected genes for another selected pathway, and *cond3* is explained in move type (3.ii) below.

(3.ii) Remove a pathway but leave genes' status unchanged:
This move is the reverse of (3.i) described above. First select a pathway included in the model that has all of its $p_{k\gamma}^{old}$ included member genes associated with other included pathways ($\theta_k^{old} = 1$ and 'cond3'). This will ensure that no orphan gene is created. Attempt to remove the pathway but leave the status of the genes unchanged, i.e., set $\theta_k^{new} = 0$ and accept the move with probability

$$\min\left\{1, \frac{f(\boldsymbol{\theta}^{new}, \boldsymbol{\gamma}^{new}|\boldsymbol{T}, Y)}{f(\boldsymbol{\theta}^{old}, \boldsymbol{\gamma}^{old}|\boldsymbol{T}, Y)} \cdot \frac{\sum_{r=1}^{K} I\{\theta_r^{old} = 1, p_{r\gamma}^{old} \geq 1, cond3\}}{\sum_{r=1}^{K} I\{\theta_r^{new} = 0, p_{r\gamma}^{new} \geq 1, condId3\}}\right\}.$$

Appendix B

Posterior Inference on Regression Coefficients

If inference on regression coefficients is desirable, these can estimate either via the posterior distributions or the least squares estimates. For model (3.1) we have the following posterior distribution:

$$\pi(\beta_g|\mathbf{Y}, \mathbf{X}_{(R)}, \omega^2) \sim HN^+(U_g C_g, \sigma_g U_g) \tag{15}$$

where HN^+ indicates a k_g-variate half-normal distribution that gives positive probability only to vectors formed by elements bigger than zero.

For the more general time-dependent model we have the following posterior distributions:

$$\begin{cases} \pi(\beta_g|\mathbf{Y}, \mathbf{X}_{(R)}, \omega^2) \sim HN^+(E_g^{-1}F_g, \sigma_g E_g^{-1}) \\ \pi(\beta_g''|\mathbf{Y}, \mathbf{X}_{(R)}, \omega^2) \sim N(J_g^{-1}H_g, \sigma_g J_g^{-1}), \end{cases} \tag{16}$$

with

$$\begin{aligned} J_g &= \mathbf{X}_{3(R'')}^T \mathbf{X}_{3(R'')} - \mathbf{X}_{3(R'')}^T \mathbf{X}_{3(R)} L_g^{-1} \mathbf{X}_{3(R)}^T \mathbf{X}_{3(R'')} + (\zeta c)^{-1} I_{k_{3g}}, \\ H_g &= \mathbf{Y}_{3g}^T \mathbf{X}_{3(R'')} + (\mathbf{Y}_g^T \mathbf{X}_{(R)} - \mathbf{Y}_{2g}^T \mathbf{X}_{2(R')} D_g^{-1} \mathbf{X}_{2(R')}^T \mathbf{X}_{2(R)} + \sigma_g^{1/2} c^{-1} \mathbf{1}_{k_g}) \\ &\quad L_g^{-1} \mathbf{X}_{3(R)}^T \mathbf{X}_{3(R'')}, \\ D_g &= \mathbf{X}_{2(R')}^T \mathbf{X}_{2(R')} + (\zeta c)^{-1} I_{k_{2g}}, \\ L_g &= \mathbf{X}_{(R)}^T \mathbf{X}_{(R)} - \mathbf{X}_{2(R)}^T \mathbf{X}_{2(R')} D_g^{-1} \mathbf{X}_{2(R')}^T \mathbf{X}_{2(R)}. \end{aligned}$$

The posterior distribution of β' has the same form as the posterior distribution of β''. Using the least squares approach, instead, we obtain the following equations for β, β' and β'':

$$\left\{ \begin{array}{lcl} \hat{\beta}_{gLS} & = & (\mathbf{X}_{(R)}^T \mathbf{X}_{(R)})^{-1} \mathbf{X}_{(R)}^T (Y_g - \mathbf{X}_{2(R')}^* \beta'_g + \mathbf{X}_{3(R'')}^* \beta''_g), \\ \hat{\beta'}_{gLS} & = & (\mathbf{X}_{2(R')}^T \mathbf{X}_{2(R')})^{-1} \mathbf{X}_{2(R')}^T (Y_{2g} - \mathbf{X}_{2(R)} \beta_g), \\ \hat{\beta''}_{gLS} & = & (\mathbf{X}_{3(R'')}^T \mathbf{X}_{3(R'')})^{-1} \mathbf{X}_{3(R'')}^T (Y_{3g} - \mathbf{X}_{3(R)} \beta_g), \end{array} \right.$$

and then

$$\begin{array}{rcl} \hat{\beta}_{gLS} & = & K_g^{-1} [\hat{\beta}_{gOLS} - (\mathbf{X}_{(R)}^T \mathbf{X}_{(R)})^{-1} \mathbf{X}_{(R)}^T (\mathbf{X}_{2(R')}^* (\mathbf{X}_{2(R')}^T \mathbf{X}_{2(R')})^{-1} \mathbf{X}_{2(R')}^T Y_{2g} \\ & & + \mathbf{X}_{3(R'')}^* (\mathbf{X}_{3(R'')}^T \mathbf{X}_{3(R'')})^{-1} \mathbf{X}_{3(R'')}^T Y_{3g})], \end{array}$$

with

$$\begin{array}{rcl} K_g & = & I_{k_g} - (\mathbf{X}_{(R)}^T \mathbf{X}_{(R)})^{-1} \mathbf{X}_{(R)}^T (\mathbf{X}_{2(R')}^* (\mathbf{X}_{2(R')}^T \mathbf{X}_{2(R')})^{-1} \mathbf{X}_{2(R')}^T \mathbf{X}_{2(R)} \\ & & + \mathbf{X}_{3(R'')}^* (\mathbf{X}_{3(R'')}^T \mathbf{X}_{3(R'')})^{-1} \mathbf{X}_{3(R'')}^T \mathbf{X}_{3(R)}). \end{array}$$

Bibliography

Ahmed, O., Adel, A., Diab, D. & Gobran, N. (2006), 'Prognostic value of serum level of interleukin-6 and interleukin-8 in metastatic breast cancer patients', *Egypt J Immunol* **13**, 61–68.

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

Bagga, S., Bracht, J., Hunter, S., Massirer, K., Holtz, J., Eachus, R. & Pasquinelli, A. (2005), 'Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation', *Cell* **122**, 553–563.

Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006), 'Prediction by supervised principal components', *Journal of the American Statistical Association* **101**, 119–137.

Besag, J. (1974), 'Spatial interaction and the statistical analysis of lattice systems', *Journal of the Royal Statistical Society, Ser. B* **36**, 192–225.

Besag, J. (1986), 'On the statistical analysis of dirty pictures', *Journal of the Royal Statistical Society, Ser. B* **48**, 259–302.

Betel, D., Wilson, M., Gabow, A., Marks, D. & Sander, C. (2008), 'The microRNA.org resource: targets and expression', *Nucleic Acids Research* **36**.

Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., Joshi, M., Harpole, D., Lancaster, J. M., Berchuck, A., Jr, J. A. O., Marks, J. R., Dressman, H. K., West, M. & Nevins, J. R. (2006), 'Oncogenic pathway signatures in human cancers as a guide to targeted therapies', *Nature* **439**, 353–357.

Bottolo, L. & Richardson, S. (2009), 'Evolutionary stochastic search', *Journal of Computational and Graphical Statistics* (under revision) .

Brown, P., Vannucci, M. & Fearn, T. (1998*a*), 'Bayesian wavelength selection in multicomponent analysis', *J. of Chemiometrics* **12**, 173–182.

Brown, P., Vannucci, M. & Fearn, T. (1998*b*), 'Multivariate bayesian variable selection and prediction', *J. of the Royal Statistical Society, Series B* **60**, 627–641.

Carvalho, C., Chang, J., Lucas, J., Nevin, J., Wang, Q. & West, M. (2008), 'High-dimensional sparse factor modeling: Applications in gene expression genomics', *Journal of the American Statistical Association* **103**, 1438–1456.

Chipman, H., George, E. & R.E. McCulloch, R. (2001), *The Practical Implementation of Bayesian Model Selection*, IMS Lecture Notes - Monograph Series Volume 38.

Conrad, R., Barrier, M. & Ford, L. (2006), 'Role of miRNA and miRNA processing factors in development and disease', *Birth Defects Research (Part C)* **78**, 107–117.

Cowell, R., Dawid, A., Lauritzen, S. & Spiegelhalter, D. (1999), *Probabilistic Networks and Expert Systems*, Springer Verlag.

Dahlquist, K., Salomonis, N., Vranizan, K., Lawlor, S. & Conklin, B. (2002), 'GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways', *Nature Genetics* **31**, 19–20.

Dawid, A. P. (1981), 'Some matrix-variate distribution theory: Notational considerations and a Bayesian application', *Biometrika* **68**, 265–274.

Doniger, S., Salomonis, N., Dahlquist, K., Vranizan, K., Lawlor, S. & Conklin, B. (2003), 'MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile for microarray data', *Genome Biology* **41**, R7.

Downward, J. (2006), 'Signatures guide drug choice', *Nature* **439**, 274–275.

Farh, K.-H., Grimson, A., Jan, C., Lewis, B., Johnston, W., Lim, L., Burge, C. & Bartel, D. (2005), 'The widespread impact of mammalian microRNAs on mRNA repression and evolutions', *Science* **310**, 1817–1821.

Finnell, R., Waes, J.-V., Bennett, G., Barber, R., Wlodarczyk, B., Shaw, G., Lammer, E., Piedrahita, J. & Eberwine, J. (2000), 'Genetic basis of susceptibility to environmentally induced neural tube defects', *Annals of the New York Academy of Sciences* **919**, 261–277.

Frankel, L., Lykkesfeldt, A., Hansen, J. & Stenvang, J. (2007), 'Protein Kinase C alpha is a marker for antiestrogen resistance and is involved in the growth of tamoxifen resistant human breast cancer cells', *Breast Cancer Res. Treat.* **104**, 165–179.

George, E. & McCulloch, R. (1993), 'Variable selection via Gibbs sampling', *J. Am. Statist. Assoc.* **88**, 881–9.

George, E. & McCulloch, R. (1997), 'Approaches for bayesian variable selection', *Statistica Sinica* **7**, 339–373.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. & Lander, E. (1999), 'Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring', *Science* **286**, 531–537.

Huang, J., Frey, B. & Morris, Q. (2008), 'Comparing sequence and expression for predicting microrna targets using genmir3', *Pacific Symposium for Biocomputing (PSB)* **13**.

Huang, J., Morris, Q. & Frey, B. (2007), 'Bayesian inference of microrna targets from sequence and expression data', *Journal of Computational Biology* **14(5)**.

John, B., Enright, A., Aravin, A., Tuschl, T., Sander, C. & Marks, D. (2004), 'Human microrna targets', *PLoS Biology* **2(11)**.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G., Matthews, L., Lewis, S., Birney, E. & Stein, L. (2005), 'Reactome: a knowledgebase of biological pathways', *Nucleic Acids Research* **33**, D428–32.

Kamb, A., Futreal, P., Rosenthal, J., Cochran, C., Harshman, K., Liu, Q., Phelps, R., Tavtigian, S., Tran, T., Hussey, C., Bell, R., Miki, Y., Swensen, J., Hobbs, M., Marks, J., Bennett, L., Barrett, J., Wiseman, R. & Shattuck-Eidens, D. (1994), 'Localization of the vhr phosphatase gene and its analysis as a candidate for brca1', *Genomics* **23**, 163–167.

Keyse, S. (2008), 'Dual-specificity MAP kinase phosphatases (MKPs) and cancer', *Cancer and Metastasis Reviews* **27**, 253–261.

Kim, D., Akcakanat, A., Singh, G., Sharma, C. & Meric-Bernstam, F. (2009), 'Regulation and localization of ribosomal protein s6 kinase 1 isoforms', *Growth Factors* pp. 12–21.

Kiriakidou, M., Nelson, P., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z. & Hatzigeorgiou, A. (2004), 'A combined computational-experimental approach predicts human microRNA targets', *Genes and Development* **18**, 1165–78.

Krek, A., Grn, D., Poy, M., Wolf, R., Rosenberg, L., Epstein, E., MacMenamin, P., Piedade, I., Gunsalus, K., Stoffel, M. & Rajewsky, N. (2005), 'Combinatorial microRNA target predictions', *Nature Genetics* **37**, 495–500.

Krieger, C., Zhang, P., Mueller, L., Wang, A., Paley, S., Arnaud, M., Pick, J., Rhee, S. & Karp, P. (2004), 'MetaCyc: a multiorganism database of metabolic pathways and enzymes', *Nucleic Acids Research* **32**, D438–442.

Kwon, D., Tadesse, M., Sha, N., Pfeiffer, R. & Vannucci, M. (2007), 'Identifying biomarkers from mass spectrometry data with ordinal outcome', *Cancer Informatics* **3**, 19–28.

Lee, S., Jeong, Y., Im., H. G., Kim, C., Chang, Y. & Lee, I. (2007), 'Silibinin suppresses PMA-induced MMP-9 expression by blocking the AP-1 activation via MAPK signaling pathways in MCF-7 human breast carcinoma cells', *Biochemical and Biophysical Research Communications* **354(1)**, 65–171.

Lewis, B., Burge, B. & Bartel, D. (2005), 'Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets', *Cell* **120**, 15–20.

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

Li, F. & Zhang, N. (2009), 'Bayesian variable selection in structured high-dimensional covariate space with application in genomics', *Journal of the American Statistical Association* p. to appear.

Li, S., Yu, S., Kao, L., Tsai, Z., Singh, S., Chen, B., Ho, B., Liu, Y. & Yang, P. (2009), 'Target identification of microRNAs expressed highly in human embryonic stem cells', *Journal of Cellular Biochemistry* **106(6)**, 1020–1030.

Madigan, D. & Raftery, A. (1994), 'Model selection and accounting for model uncertainty in graphical models using Occam's window', *Journal of the American Statistical Association* **89**, 1535–1546.

Madigan, D. & York, J. (1995), 'Bayesian graphical models for discrete data', *International Statistical Review* **63(2)**, 215–232.

Monni, S. & Tadesse, M. (2009), 'A stochastic partitioning method to associate high-dimensional responses and covariates', *Bayesian Analysis* **4(3)**.

Nakao, M., Bono, H., Kawashima, S., Kamiya, T., Sato, K., Goto, S. & Kanehisa, M. (1999), 'Genome-scale gene expression analysis and pathway reconstruction in KEGG', *Genome Informatics Series: Workshop on Genome Informatics* **10**, 94–103.

Newton, M., Noueiry, A., Sarkar, D. & Ahlquist, P. (2004), 'Detecting differential gene expression with a semiparametric hierarchical mixture method', *Biostatistics* **5**, 155–176.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. & Kanehisa, M. (1999), 'KEGG: Kyoto Encyclopedia of Genes and Genomes', *Nucleic Acids Research* **27**, 29–34.

Pan, W., Xie, B. & Shen, X. (2009), 'Incorporating predictor network in penalized regression with application to microarray data', *Biometrics* p. to appear.

Park, M. Y., Hastie, T. & Tibshirani, R. (2007), 'Averaged gene expressions for regression', *Biostatistics* **8**, 212–27.

Petersen, C., Bordeleau, M.-E., Pelletier, J. & Sharp, P. (2006), 'Short RNAs repress translation after initiation in mammalian cells', *Molecular Cell* **4**, 533–542.

Pfaffl, M. (2001), 'A new mathematical model for relative quantification real-time RT-PCR', *Nucleic Acid Research* **29**, 2002–7.

Pillai, R., Bhattacharya, S., Artus, C., Zoller, T., Cougot, N., Basyuk, E., Bertrand, E. & Filipowicz, W. (2005), 'Inhibition of translational initiation by let-7 microRNA in human cells', *Science* **209**, 1573–6.

Pittman, J., Huang, E., Dressman, H., Horng, C., Cheng, S., Tsou, M., Chen, C., Bild, A., Iversen, E., Huang, A., Nevins, J. & West, M. (2004), 'Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes', *Proceedings of the National Academy of Sciences* **101**, 8431–6.

Rajewsky, N. (2006), 'Microrna target predictions in animals', *Nature Genetics* **38**, S8–13.

Sha, N., Tadesse, M. & Vannucci, M. (2006), 'Bayesian variable selection for the analysis of microarray data with censored outcomes', *Bioinformatics* **22**, 2262–2268.

Sha, N., Vannucci, M., Tadesse, M., Brown, P., Dragoni, I., Davies, N., Roberts, T., Contestabile, A., Salmon, N., Buckley, C. & Falciani, F. (2004), 'Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage', *Biometrics* **60(3)**, 812–819.

Shevde, L. A., Metge, B., Mitra, A., Xi, Y., Ju, J., King, J. & Samant, R. (2009), 'Spheroid-forming sub-population of breast cancer cells demonstrates vasculogenic mimicry via hsa-mir-299-5p regulated de novo expression of osteopontin(p)', *Journal of Cellular and Molecular Medicine* p. to appear.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S. & *et*

al. (2002), 'Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning', *Nature Medicine* **8**, 68–74.

Smith, M. & Kohn, R. (1996), 'Nonparametric regression using bayesian variable selection', *Journal of Econometrics* **75**, 317–344.

Solit, D. B., Garraway, L. A., Pratilas, C. A., Sawai, A., Getz, G., Basso, A., Ye, Q., Lobo, J. M., She, Y., Osman, I., Golub, T. R., Sebolt-Leopold, J., Sellers, W. R. & Rosen, N. (2006), 'BRAF mutation predicts sensitivity to MEK inhibition', *Nature* **439**, 358–362.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. & Mesirov, J. P. (2005), 'Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles', *Proceedings of the National Academy of Sciences* **102**, 15545–50.

Telesca, D., Muller, P., Parmigiani, G. & Freedman, R. (2008), Modeling dependent gene expression, Technical report, University of Texas M.D. Anderson Cancer Center, Department of Biostatistics.

The Gene Ontology Consortium (2000), 'Gene Ontology: tool for the unification of biology', *Nature Genetics* **25**, 25–29.

Thermann, R. & Hentze, M. (2007), 'Drosophila miR2 induces pseudo-polysomes and inhibits translation initiation', *Nature* **447**, 875–879.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. & Altman, R. (2001), 'Missing value estimation methods for DNA microarrays', *Bioinformatics* **17**, 520–525.

van't Veer, L., Dai, H., van de Vijver, M., He, Y., Hart, A., Mao, M., Peterse, H., van der Kooy, K., Marton, M., Witteveen, A., Schreiber, G.,

Kerkhoven, R., Roberts, C., Linsley, P., Bernards, R. & Friend, S. (2002), 'Gene expression profiling predicts clinical outcome of breast cancer', *Nature* **415**, 530–536.

Wang, Q., Dobra, A. & West, M. (2004), 'Graphexplore: A software tool for graph visualization', *ISDS Discussion Paper* **04-22**.

Wei, L. (1992), 'The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis', *Statistics in Medicine* **11**, 1871–1879.

Wei, Z. & Li, H. (2007), 'A Markov random field model for network-based analysis of genomic data', *Bioinformatics* **23**, 1537–44.

Wei, Z. & 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* **02(1)**, 408–429.

Whittaker, J. (1990), *Graphical models in applied multivariate statistic*, Wiley.

Yoon, S. & Micheli, G. (2006), 'Computational identification of microRNAs and their targets', *Birth Defects Research (Part C)* **78**, 118–128.

Zhang, J. & Wiemann, S. (2009), 'KEGGgraph: a graph approach to KEGG PATHWAY in R and Bioconductor', *Bioinformatics* **25**, 1470–1471.

Zhou, Q., Effati, R., Talvinen, K., Pospiech, H., Syvoja, J. & Collan, Y. (2008), 'Genomic changes of the 55 kda subunit of DNA polymerase epsilon in human breast cancer', *Cancer Genomics Proteomics* **5**, 287–292.

Zou, H., Hastie, T. & Tibshirani, R. (2006), 'Sparse principal component analysis', *Journal of Computational and Graphical Statistics* **15**, 265–286.