NETWORK-BASED MIXTURE MODELS FOR GENOMIC DATA

A THESIS

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

PENG WEI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

June, 2009

©Peng Wei 2009

Acknowledgments

Over the past five years, I have received assistance from many people. Without their help, completion of this thesis would not have been possible. Here I would like to acknowledge those who supported and encouraged me during my graduate study at the University of Minnesota.

First of all, I would like to express my deepest gratitude to my thesis advisor Professor Wei Pan for his support and guidance since the very beginning of my graduate study five years ago. I would like to thank him for his excellent supervision, motivating me to strive for and achieve high standards in research, and supporting this work with his ideas. It was my great pleasure to work with him.

Heartful thanks also go to Professors Sudipto Banerjee, Jim Hodges, Cavan Reilly and Xiaotong Shen for agreeing to be in my thesis committee, taking the time to read and comment on my Ph.D. dissertation, and providing me with invaluable advice and guidance at various stages.

I would also like to take this opportunity to thank the faculty at the Division of Biostatistics and School of Statistics for offering me many great courses and making the learning enjoyable. In addition, I want to thank the University of Minnesota Graduate School for the Doctoral Dissertation Fellowship, which allowed me to devote full-time effort to my thesis research in the last year of my Ph.D. study.

Special thanks are extended to Professor Robert Hebbel, Drs. Judy Enenstein and Liming Milbauer at the Division of Hematology, Oncology and Transplantation, Univer-

sity of Minnesota Medical School. During my three year research assistantship at Dr. Hebbel's lab, I had the opportunities to participate in multiple cutting-edge biomedical research projects and learned an enormous amount of biological knowledge, which largely enhanced my thesis research.

Finally, I wish to express my sincerest thanks to my parents and parents-in-law who gave me their unconditional support and encouragement throughout. I would like to give special thanks to my wife Wei He for her deepest love, continuous support and enormous patience, without which I would not have been able to complete the doctoral degree smoothly.

Abstract

A common task in genomic studies is to identify genes satisfying certain conditions, such as differentially expressed genes between normal and tumor tissues or regulatory target genes of a transcription factor (TF). Standard approaches treat all the genes identically and independently *a priori* and ignore the fact that genes work coordinately in biological processes as dictated by gene networks, leading to inefficient analysis and reduced power. We propose incorporating gene network information as prior biological knowledge into statistical modeling of genomic data to maximize the power for biological discoveries.

We propose a spatially correlated mixture model based on the use of latent Gaussian Markov random fields (GMRF) to smooth gene specific prior probabilities in a mixture model over a network, assuming that neighboring genes in a network are functionally more similar to each other. In addition, we propose a Bayesian implementation of a discrete Markov random field (DMRF)-based mixture model for incorporating gene network information, and compare its performance with that based on Gaussian Markov random fields. We also extend the network-based mixture models to ones that are able to integrate multiple gene networks and diverse types of genomic data, such as protein-DNA binding, gene expression and DNA sequence data, to accurately identify regulatory target genes of a TF. Applications to high-throughput microarray data, along with simulations, demonstrate the utility of the new methods and the statistical efficiency gains over other methods.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the advent of high-throughput microarray technologies, biomedical researchers have been able to monitor changes in the expression levels of thousands of genes. Gene expression is the process of genetic information flow from DNA sequence to messager RNA (mRNA), called "transcription", and from mRNA to protein, called "translation". Proteins are the workhorse molecules of the cell and participate in every process within cells. Although every cell in an organism contains all the necessary DNA information for gene expression, different genes are expressed at different times and under various conditions, leading to distinct properties of cells, such as the differences between cancerous cells and normal cells. A fundamental question in biology is how gene expression is regulated. A general mechanism is through some regulatory proteins called transcription factors (TFs): a TF binds to one or more specific DNA subsequences (called motifs) in the regulatory region of a gene, then works with other TFs to activate or repress the binding target gene's expression. A biologically important question is to identify the binding target genes of a given TF. A new application of microarray technology, chromatin immunoprecipitation (ChIP) coupled with microarray (chip) analysis, hence named ChIP-chip (Ren *el al.* 2000), has enabled researchers to identify genome-wide binding locations of a TF in living cells.

A brief description of the ChIP-chip experiment is as follows: first, TF of interest binds to certain genome sequences in living cells; second, DNA sequences are chopped into small fragments, some of which are bound by the TF while the rest are not; third, those DNA fragments bound by the TF are isolated by chromatin immunoprecipitation

(ChIP) followed by separating the TF and its binding DNA fragments using reverse cross-linking; fourth, the separated DNA fragments are amplified and labeled with fluorescent dye Cy5 (red color), while some control DNA fragments, which are not enriched by the above immunoprecipitation (IP) process, are labeled with fluorescent dye Cy3 (green color); fifth, both pools of labeled DNA are hybridized to a microarray (chip). After hybridization, scanning, and image processing, intensity levels are obtained for both colors for all the spots on the microarray, with each spot corresponding to a gene. If a gene is the TF's target, the red intensity of the spot for the gene should be higher than the corresponding green one. Therefore, the ratio between the red and the green intensities measures how likely the gene is a binding target of the TF.

Because the resulting ChIP-chip binding data are in the usual format of DNA microarray expression data, it is technically possible to apply any of many existing statistical methods of detecting differentially expressed genes to binding data, including SAM (Tusher *et al.* 2001), Empirical Bayes (EB) methods (Efron *et al.* 2001; Newton *et al.* 2001), and mixture models (Pan *et al.* 2002; McLachlan *et al.* 2006); see Pan (2002), Cui and Churchill (2003) for reviews on statistical methods for gene expression data. Most methods, including all the above cited, treat the genes equally and independently *a priori* and ignore the fact that genes work coordinately in biological processes, leading to inefficient analysis and reduced power. In particular, due to high noise level inherent with high-throughput microarray technologies and the so called "large p, small n" problem, i.e., the large number of genes surveyed in contrast to the small number of

replicates, it is desirable to take advantage of existing biological knowledge to maximize the statistical power for genomic discovery, such as detecting differentially expressed genes or regulatory target genes of a TF.

There have been increasing efforts recently to incorporate biological knowledge into statistical analysis of microarray data to gain statistical efficiency. For example, Pan (2005, 2006a) proposed a stratified mixture model to incorporate gene functions as annotated in the Gene Ontology (GO) database (Ashburner *et al.* 2000); Xiao *et al.* (2005) developed a Hidden Markov Model to incorporate genomic location information into analysis, accounting for spatial patterns of gene co-expression. Another class of emerging methods is to analyze gene sets, rather than individual genes; the gene sets are formed based on biological pathways or gene functional groups (Subramanian *et al.* 2005; Tian *et al.* 2005; Efron *et al.* 2007; Newton *et al.* 2007; Milbauer *et al.* 2008). Nevertheless, each of the gene-set methods treats the genes equally and independently *a priori*, and the results depend on the specification of gene sets: a too large or too small gene set may lead to reduced statistical power; in fact, each gene set can be regarded as a gene subnetwork.

Gene networks represented by undirected graphs with genes as nodes and gene-gene interactions as edges provide a powerful means to concisely summarize biological knowledge accumulated over thousands of experiments. For example, Lee *et al.* (2004) employed a probabilistic approach to constructing a functional linkage (coupling) network for the yeast genome by integrating a variety of genomic data: a pair of genes that have

evidence of being co-functional in biological processes are connected on the resulting network. There are other types of gene networks besides functional linkage networks, such as TF-gene regulatory networks, protein-protein interaction networks, co-expression networks, biological pathways, e.g., Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa and Goto 2000), induced networks and so forth; see Futschik *et al.* (2007), Ideker and Sharan (2008) for comprehensive reviews on gene networks for human and other organisms. Since linked genes on a functional linkage network tend to be co-functional in biological activities, they are more likely to be co-regulated (or not regulated) by a TF. For example, genes that are connected to a TF's binding target gene on the network are also likely to be the TF's targets. In spite of large amounts of biological knowledge embedded in gene networks, there has been very limited work attempting to incorporate gene network information into statistical analysis of microarray data. An exception is the recent work by Wei and Li (2007) (see also references therein), who proposed integrating a KEGG pathway induced network or other gene networks into analysis of differential gene expression via a discrete Markov Random Field (MRF) model.

My thesis is devoted to developing, evaluating, and comparing statistical methods for incorporating gene networks into statistical modeling of high-throughput data, such as gene expression data, DNA-protein binding (ChIP-chip) data, and DNA sequence data. This is intended to meet the biological challenge to boost the statistical power for genomic discovery by maximizing the use of existing biological knowledge and data. In Chapter 2, we develop a spatially correlated mixture model based on the use of latent

Gaussian Markov random fields (GMRF) to smooth gene-specific prior probabilities in a mixture model over a network, assuming that neighboring genes in a network are functionally more similar to each other; Chapter 2 has been published as Wei and Pan (2008a). Chapter 3 proposes a Bayesian implementation of the discrete Markov random field (DMRF)-based mixture model of Wei and Li (2007) and compares its performance to the model based on GMRFs. This chapter also proposes two novel constraints on the prior specifications for the GMRF model and addresses direct posterior approach to estimating the False Discovery Rate (FDR) in the context of MRF models. Chapter 3 has been published as Wei and Pan (2009). Chapter 4 presents two mixture models, based on the use of GMRFs and DMRFs respectively, for integrating multiple gene networks and diverse types of genomic data to accurately identify regulatory target genes of a TF. Chapter 5 concludes a short discussion on future work. In Chapters 2-4, the Markov chain Monte Carlo (MCMC) algorithms for implementing the proposed methods are also detailed, together with simulation results and applications to high-throughput microarray data to illustrate the utility of the new methods and their advantages over other methods.

Chapter 2

Incorporating Gene Networks into Statistical Tests for Genomic Data via a Gaussian MRF Mixture Model

2.1 Introduction

In this chapter, we propose a spatially correlated mixture model based on the use of latent Gaussian Markov random fields (GMRF) to incorporate gene network information into statistical modeling of genomic data. The spatial mixture model was first proposed by Fernandez and Green (2002) in spatial statistics, and applied to analyze comparative genomic hybridization (CGH) data by Broet *et al.* (2006). This is in contrast to standard mixture models. For example, McLachlan et al (2006) proposed using a standard two-component normal mixture model to identify differentially expressed genes; although promising results have been obtained, one key assumption of the standard normal mixture model is that all genes share the same prior probability of coming from a component of the normal mixture without regard to their biological functions. In the context of CGH data analysis, Broet *et al.* (2006) proposed using a spatially correlated normal mixture model to introduce gene-specific prior probabilities and allow those prior probabilities to be correlated among neighboring genes on a chromosome, and gained more power to identify gene copy number changes. Extending the work of Broet *et al.* from one-dimensional chromosome locations to two-dimensional gene networks, we propose using the spatially correlated normal mixture model to improve power in identifying differentially expressed genes (for cDNA, Affymetrix or any other expression arrays) or target genes of a TF (for ChIP-chip data). A key difference from Broet *et al.* is that we use existing biological knowledge databases, such as KEGG pathways (Kanehisa and Goto 2000), or computationally predicted gene networks from integrated analysis (Lee *et al.*

2004), to construct gene functional neighborhoods and incorporate them in a spatially correlated normal mixture model. The basic rationale underlying the proposed model is that functionally linked genes tend to be co-regulated and co-expressed, which is thus incorporated into the analysis.

The rest of this chapter is organized as follows. We first review the standard mixture model (SMM), then propose a spatially correlated mixture model (Gaussian Markov random field-based mixture model or "GMRF-MM" for short). For illustration, we apply and compare the two methods using a ChIP-chip dataset to identify the target genes of TF GCN4. A simulated data set is also used to demonstrate the advantage of the proposed method over the standard mixture model. Finally, we summarize our results and outline some future work.

2.2 Methods

2.2.1 Problem

The goal of analysis is to identify which genes satisfy a certain condition, such as being differentially expressed (DE) or being a TF's transcriptional targets. This can be formulated as a formal or informal hypothesis testing problem: for each gene i, we test for a null hypothesis H_{i0} against an alternative hypothesis H_{i1}, usually the opposite of H_{i0}. For example, H_{i0} is that "gene i is equally expressed (EE)" for expression data, or that "gene i is not a target of the TF" for ChIP-chip data, while H_{i1} is the opposite of H_{i0}.

We assume that the data have been summarized as measurement Z_i for each gene

$i, i = 1, \ldots, G$; for example, Z_i can be a test statistic measuring the relative abundance of mRNA (or TF), or the statistical significance level (p-value) for rejecting H_{i0}. Define gene i's state $T_i = I(H_{i0}$ is false); that is, $T_i = 1$ or $T_i = 0$ corresponding to H_{i1} or H_{i0} holding respectively.

For our real data, we extracted a p-value for each gene, then transformed it to a z-score and subsequently modeled the z-scores (McLachlan et al 2006). The transformation is given by $z_i = \Phi^{-1}(1 - P_i)$, where Φ is the cumulative distribution function of the standard Normal distribution $N(0, 1)$, and P_i is the p-value for gene i. If P_i is properly calculated as a genuine p-value, then the null distribution of z_i is exactly the standard normal. In addition, the non-null distribution may model the right tail of the z-score distribution. The resulting two-component normal mixture model is

$$f(z_i) = \pi_0 \phi(z_i; 0, 1) + \pi_1 \phi(z_i; \mu_1, \sigma_1^2),$$

where $\phi(\cdot; \mu, \sigma^2)$ is the probability density function of $N(\mu, \sigma^2)$. However, sometimes the null distribution of the z-scores is not standard normal due to approximate p-values (e.g., resulting from possible correlations among the genes, in violation of the adopted independence assumption). In this situation, we need to estimate the null distribution $N(\mu_0, \sigma_0^2)$. We call $N(0, 1)$ the *theoretical null* and $\phi(\hat{\mu_0}, \hat{\sigma_0}^2)$ the *empirical null*. Furthermore, more than two components may be needed for f.

2.2.2 Gene networks

The types of gene networks that can be used here are not restricted; they can be any network as long as the basic assumption holds: based on a gene network, two neighboring genes (i.e., two genes with an edge between them) are more likely to satisfy H_{i1} or H_{i0} at the same time than two non-neighboring genes. As stated before, a gene network can be extracted from existing biological databases, such as KEGG pathways, or any computationally predicted gene network, possibly resulting from integrated analysis of multiple sources of genomic data. In this paper, we use the functional linkage network of yeast genes constructed by Lee *et al.* (2004) as an example. Lee *et al.* applied a naive Bayes method to assign a score to each possible gene pair by integrating a variety of genomic data, including mRNA co-expressions, gene co-citations, protein-protein interactions, gene fusions and phylogenetic profiles. Two genes with a score high enough are linked, suggesting that it is highly likely that they share some biological function. They obtained a gene network called "ConfidentNet" with high credibility. Represented by an undirected graph, the "ConfidentNet" consists of 4,681 nodes (genes) and 34,000 edges (gene-gene functional linkages). A summary of the distribution of the number of direct neighbors is: minimum=1, 25th-percentile=2, median=6, 75th-percentile=13 and maximum=188. We will use this yeast gene network in our real data example.

2.2.3 Standard mixture model

Suppose that the distribution functions of the data (e.g. z-scores) for the genes with $T_i = 1$ and $T_i = 0$ are f_1 and f_0, respectively. Assuming that *a priori* all the genes have an identical and independent distribution (i.i.d.), the marginal distribution of Z_i is a standard mixture model (SMM):

$$f(z_i) = \pi_0 f_0(z_i) + (1 - \pi_0) f_1(z_i), \tag{2.1}$$

where π_0 is the prior probability that H_{i0} holds. Note that the prior probabilities are the same for all genes. The standard mixture model has been widely used in microarray data analysis (e.g. Efron *et al.* 2001; Newton *et al.* 2001; Pan *et al.* 2002; McLachlan *et al.* 2006).

The null and non-null distributions f_0 and f_1 can be approximated by finite normal mixtures: $f_0 = \sum_{k_0=1}^{K_0} \pi_{0k_0} \phi(\mu_{k_0}, \sigma_{k_0}^2)$ and $f_1 = \sum_{k_1=1}^{K_1} \pi_{1k_1} \phi(\mu_{k_1}, \sigma_{k_1}^2)$, where $\phi(\mu, \sigma^2)$ is the density function for a Normal distribution with mean μ and variance σ^2. For z-scores, using $K_j = 1$ often suffices (McLachlan *et al.* 2006). In our real data example, we found that $K_0 = 2$ and $K_1 = 1$ worked well.

The standard mixture model can be fitted via maximum likelihood with the EM algorithm (McLachlan and Peel 2000). Once the parameter estimates are obtained, statistical inference is based on the posterior probability that H_{1i} holds: $Pr(T_i = 1|z_i) = \pi_1 f_1(z_i)/f(z_i)$. Because the spatially correlated mixture model is fitted in a Bayesian framework while it is unclear how to fit it in a frequentist approach, to facilitate comparison, we fit the standard mixture model in a similar Bayesian framework; see below

for prior specifications.

2.2.4 Gaussian Markov random field-based Mixture Model

In a Gaussian Markov random field-based Mixture Model (GMRF-MM), we introduce gene-specific prior probabilities $\pi_{ij} = Pr(T_i = j)$ for $i = 1, \ldots, G$ and $j = 0, 1$. Hence, the marginal distribution of z_i is

$$f(z_i) = \pi_{i0} f_0(z_i) + \pi_{i1} f_1(z_i). \tag{2.2}$$

Note that the prior probability specification in a stratified mixture model (Pan 2005, 2006a, 2006b) is a special case of (2.2): a group of the genes with the same function share a common prior probability π_{i0} while different groups have possibly varying prior probabilities; in fact, a partition of the genes by their functions can be regarded as a special case of a gene network.

Based on a gene network, we relate the prior probabilities π_{ij} to two latent Markov random fields $\mathbf{x}_j = \{x_{ij}; i = 1, ..., G\}$ by a logistic transformation:

$$\pi_{ij} = \exp(x_{ij})/[\exp(x_{i0}) + \exp(x_{i1})].$$

Each of the G-dimensional latent vectors $\mathbf{x}_j$ is distributed according to an intrinsic Gaussian conditional autoregression model (ICAR) (Besag and Kooperberg 1995). One key feature of ICAR is the Markovian interpretation of the latent variables' conditional distributions: the distribution of each spatial latent variable x_{ij}, conditional on $x_{(-i)j} =$

$\{x_{kj}; k \neq i\}$, depends only on its direct neighbors. More specifically, we have

$$x_{ij}|x_{(-i)j} \sim N\left(\frac{1}{m_i}\sum_{l\in\delta_i} x_{lj}, \frac{\sigma^2_{Cj}}{m_i}\right),$$

where δ_i is the set of indices for the neighbors of gene i, and m_i is the corresponding number of neighbors. To allow identifiability, we impose $\sum_i x_{ij} = 0$ for $j = 0, 1$. In this model, the parameter σ^2_{Cj} acts as a smoothing constant for the spatial field that controls the degree of dependency among the prior probabilities of the genes across the genome: a smaller σ^2_{Cj} induces more similar π_{ij}'s for those genes that are neighbors in the network.

In summary, it is biologically reasonable to assume that neighboring genes in a network are more likely to share biological functions and thus to participate in the same biological processes. Hence they should have similar prior probabilities of being DE or being targets of a TF at the same time.

2.2.5 Prior distributions

We have largely followed Broet *et al.*'s prior specifications. For either mixture model, we use vague or noninformative prior distributions: $\mu_0 = 0$, $\mu_1 \sim N(0, 10^6)I(n, 0)$, a truncated normal distribution between $n = \min_i z_i$ and 0; $\mu_2 \sim N(0, 10^6)I(0, m)$ and $m = \max_i z_i$. The two truncated normals are constructed to ensure unique labeling of the normal mixture components. In addition, we have $\sigma^2_j \sim$ Inverse Gamma$(0.1, 0.1)$ for $j = 0, 1, 2$. See Section 2.3.2.3 for discussions on selection of hyperparameters in the Inverse Gamma distrubution.

For the SMM, $(\pi_0, \pi_1, \pi_2) \sim$ Dirichlet$(1, 1, 1)$. For the GMRF-MM, $\sigma^2_{Cj} \sim$ Inverse Gamma$(0.01, 0.01)$ for $j = 0, 1, 2$. Notice that the precision parameter, $\tau_{Cj} = 1/\sigma^2_{Cj}$, has Gamma$(0.01, 0.01)$ with mean 1 and variance 100.

For completeness, the details of the model specifications for both spatial and standard normal mixture models are given in the Appendix.

2.2.6 Inference

Each of the above mixture models can be readily implemented in WinBUGS (Spiegelhalter *et al.* 2003). The posterior mean of any parameter based on Markov Chain Monte Carlo (MCMC) samples is used as its point estimate. In particular, if the point estimate $\widehat{Pr}(T_i = 0|Data)$ is smaller than a threshold t, we reject H_0. There is a correspondence between t and False Discovery Rate (FDR), which can be estimated based on $\widehat{Pr}(T_i = 0|Data)$. In this paper, we consider varying t, leading to various sensitivities and specificities and thus yielding an ROC plot.

2.3 Results

2.3.1 Real data

2.3.1.1 Data

To evaluate the performance of our proposed method, we applied the two methods - standard mixture model (SMM), and GMRF-based mixture model (GMRF-MM) - to a ChIP-chip data set for transcription factor GCN4. A TF is a protein that binds to

the promoter regions of its regulatory target genes and participates in the recruitment of RNA polymerase, thus regulating the transcription of its target genes into messenger RNA. ChIP-chip is a hybrid of chromatin-immunoprecipitation (ChIP) and microarray technology that is used to quantify the occupancy of genome-wide promoter regions by a TF. A typical ChIP-chip data set contains log binding ratios measuring the relative abundances of the TF bound to the genes, and possibly inferred statistical significance levels (p-values) for rejecting the null hypothesis that each of the genes is not bound by the TF. As a TF, GCN4 is involved in response to amino acid starvation in yeast. Lee *et al.* (2002) did ChIP-chip experiments for GCN4 with three replicates. Log binding ratios and p-values for 6,181 yeast genes were provided. Pokholok *et al.* (2005) constructed a set of 80 genes that are very likely to be the binding targets of GCN4 from multiple sources of data (including another set of more accurate ChIP-chip experiments based on a new generation of microarrays, a gene expression data set, and DNA motif analyses), as well as a set of 900 genes that are unlikely to be regulated by GCN4. Treating the positive and negative control sets as true positives and true negatives, we used them to evaluate the performance of the two methods; that is, sensitivity and specificity were estimated based on the two control sets. In addition, for the spatial normal mixture model, we used the yeast gene network constructed by Lee *et al.* (2004) to specify gene neighborhoods. After merging the ChIP-chip data set and the gene network, we ended up with a 4616-node network with 33,432 edges. We extracted those 4,616 genes' binding p-values and obtained their z-scores for final analysis; correspondingly, there were 66 and

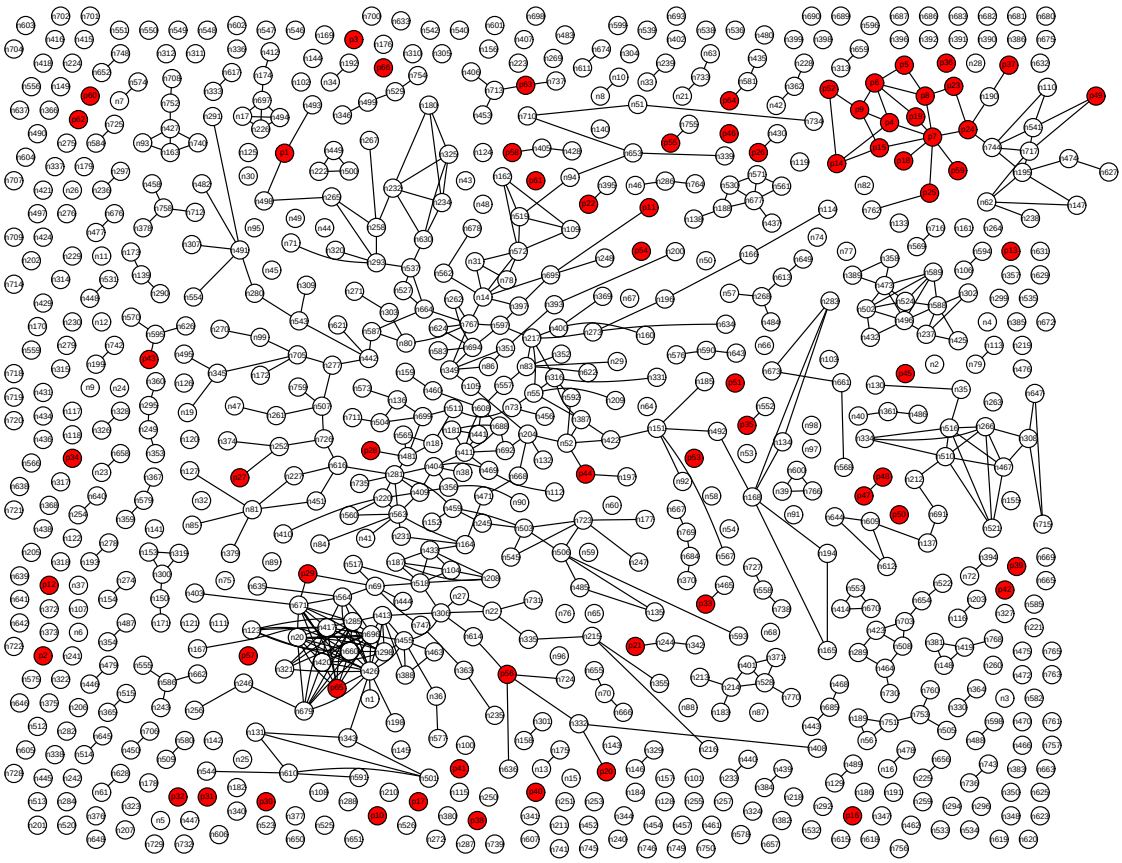

Figure 2.1: Subnetwork consisting of positive control genes (dark ones) and negative control genes (blank ones).

770 genes in the positive and negative control sets respectively.

For illustration, a subnetwork consisting of only the positive and negative control genes is shown in Figure 2.1, where dark nodes represent the target genes (positive controls) while blank ones are non-targets (negative controls). Several features are noticeable. First, there is a cluster of positive control genes in the upper-right corner, and there are quite a few clusters of negative control genes. Second, positive control

genes can be connected with negative control genes. Third, although there are many singletons (i.e. isolated genes without edges to other genes) in the subnetwork, they are not necessarily singletons in the whole network because they may be connected to other genes outside the control sets. We used the full network.

2.3.1.2 Model fitting

WinBUGS (Spiegelhalter *et al.* 2003) was used to implement the Bayesian models. Posterior means of the parameters were computed based on 4,000 MCMC samples after 6,000 burn-ins. First, a standard two-component normal mixture model with an empirical null distribution (i.e. its mean and variance parameters were unknown and needed to be estimated) was fitted; lack-of-fit of the mixture model against the data was observed (results not shown). To achieve better model fit, a third normal component with mean imposed to be negative was added into f_0; the fitted mixture model was

$$\begin{aligned}\hat{f}(z_i) &= 0.91\phi(z_i; 0, .80^2) + 0.037\phi(z_i; -1.98, .40^2) + \\ &\quad 0.058\phi(z_i; 1.67, 1.94^2),\end{aligned}$$

where the first two components with zero and negative means were treated as $\hat{f}_0$ and the third one with postive mean as $\hat{f}_1$. A visual examination revealed improved goodness-of-fit except at the peak of the data distribution (Figure 2.2). Similarly, a three-component spatial normal mixture model with an empirical null was fitted, yielding

$$\begin{aligned}\hat{f}(z_i) &= \hat{\pi}_{i,0,1}\phi(z_i; 0, .63^2) + \hat{\pi}_{i,0,2}\phi(z_i; -0.38, 1.02^2) + \\ &\quad \hat{\pi}_{i,1,1}\phi(z_i; 0.75, 1.53^2)\end{aligned}$$

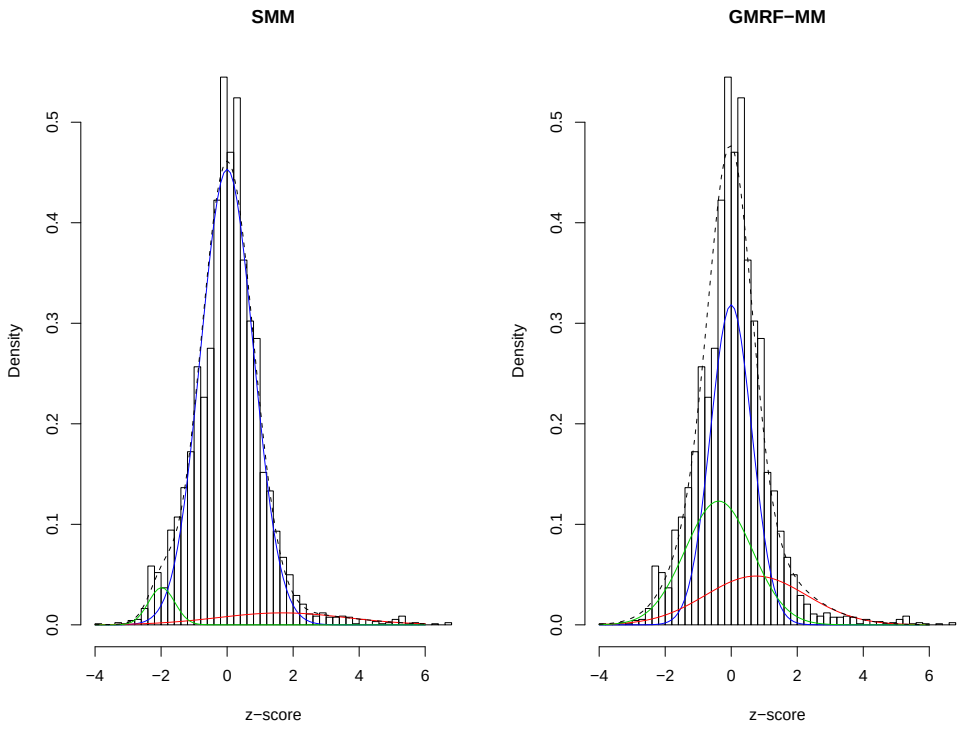

Figure 2.2: Fitted mixture models: on each panel, the dashed line is for the marginal/mixture distribution while the three solid lines are for the three components.

with the averages of $\hat{\pi}_{i,0,1}$, $\hat{\pi}_{i,0,2}$ and $\hat{\pi}_{i,1,1}$ as 0.500, 0.314 and 0.186. The fitted marginal and component-wise distributions are displayed in Figure 2.2. We used different initial values to look at the convergence of the MCMC simulations. Trace plots showed good convergence for all the models (Figure 2.3).

2.3.1.3 Statistical power

The ROC curves were constructed for the two methods based on the positive and negative control sets. As shown in Figure 2.4, when the specificity ranged from 0.9 to 0.4, as

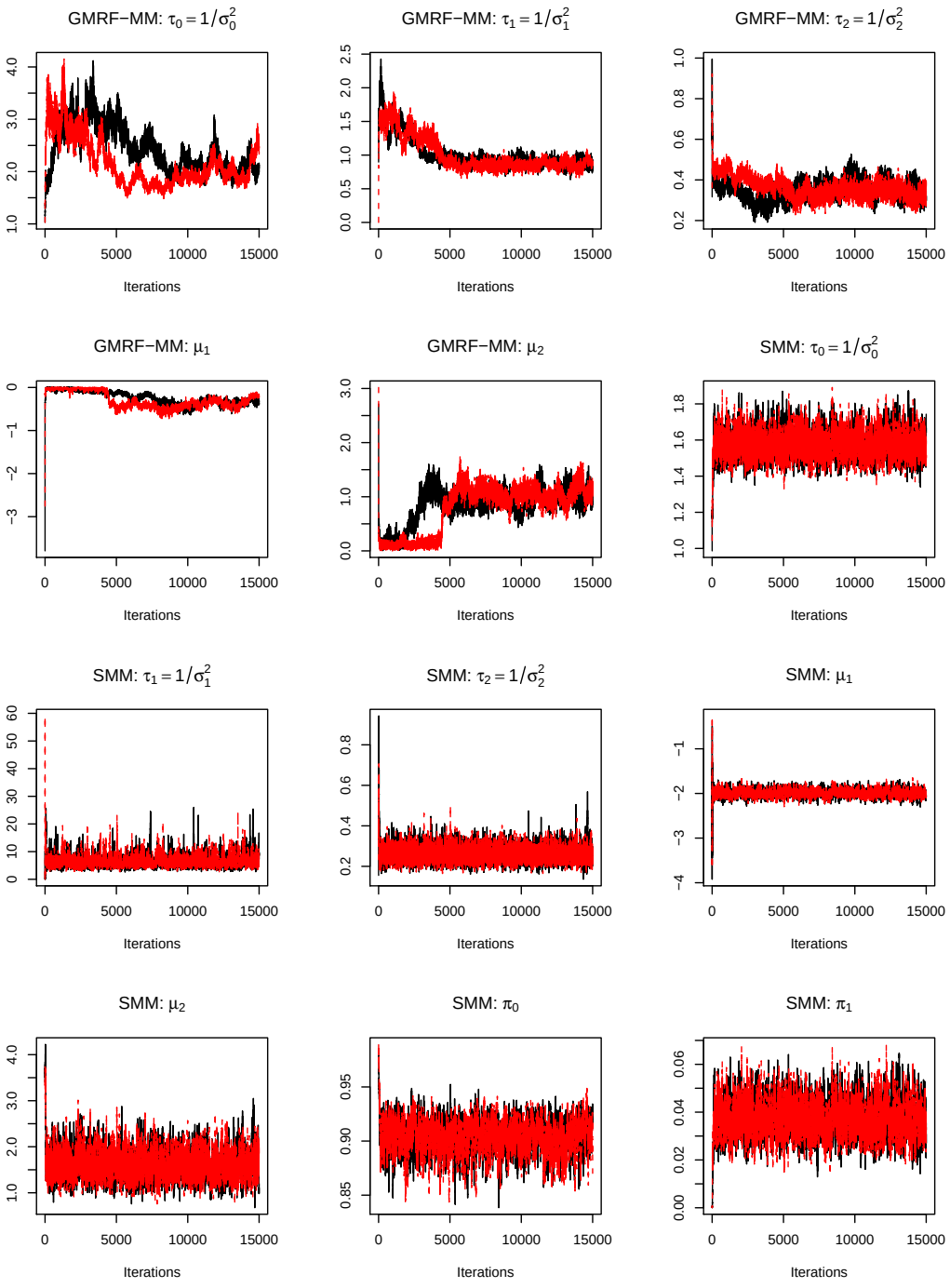

Figure 2.3: Convergence check.

desired in practice, the GMRF-MM gave a higher sensitivity than that of the SMM. Hence, at a high specificity (e.g. above 0.5 as usually desired), by using biological knowledge embedded in a gene network, the GMRF-MM had higher statistical power to detect the targets than did the SMM that ignored biological knowledge.

In addition, we compared the positive control genes' ranks by the posterior probabilies from the GMRF-MM and the original binding p-values. As shown in Figure 2.5(b), most of the positive control genes were ranked in the top 100 by both methods, while the GMRF-MM boosted a few more genes' ranks from moderate (ranked between 200 - 400) to relatively high (ranked between 100 - 200). There were about equally many positive control genes ranked low by either method, as illustrated by the upper-right part of Figure 2.5(a).

2.3.1.4 Representative gene evaluations

We examined several individual genes to gain more biological insights. First, for ARG8 (YOL140W), a gene in the positive control set, its posterior probability of being a target was estimated to be 0.728 by the GMRF-MM and 0.023 by the SMM. The binding ratio for this gene in Lee et al's rich medium ChIP-chip experiment was 1.02. However, Harbison *et al.*(2004) did more ChIP-chip experiments on GCN4 in amino acid starvation and nutrition deprivation conditions besides rich medium, and the binding ratio for ARG8 was 5.0 with p-value 10^{-11}. Because GCN4 is a transcriptional activator of amino acid biosynthetic genes in response to amino acid starvation, it is expected that genes involved in amino acid biosynthetic process are likely to be binding targets of

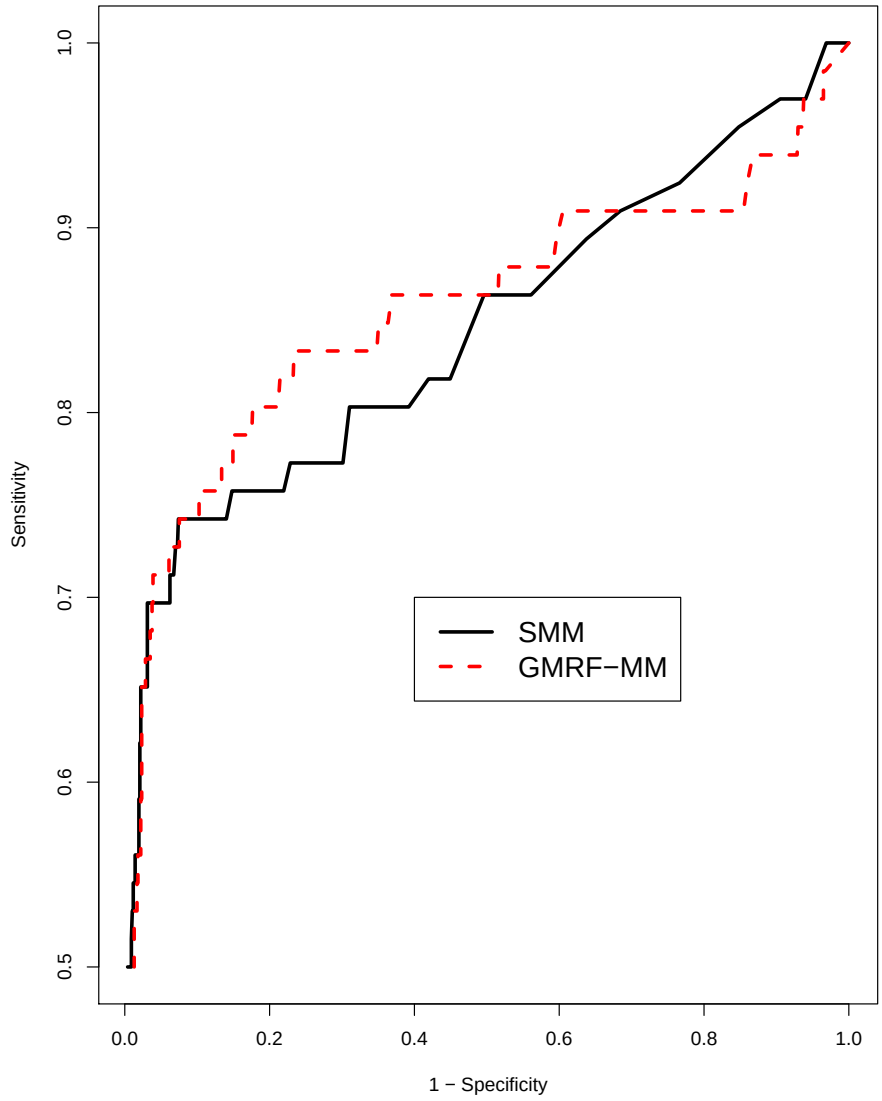

Figure 2.4: ROC curves for the two methods applied to the real data.

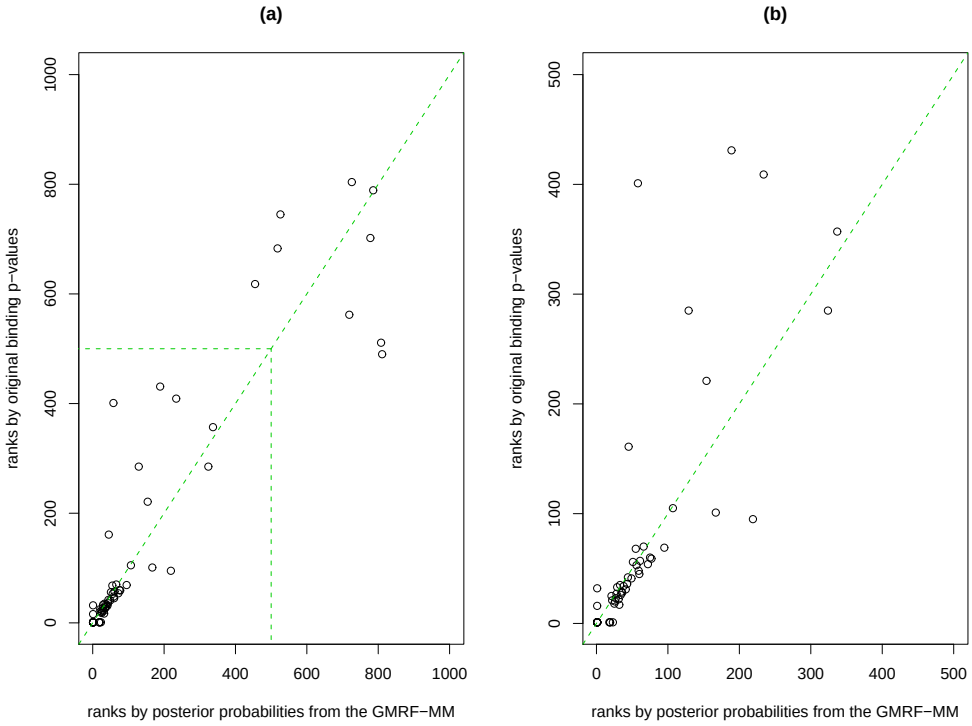

Figure 2.5: Comparison of positive control gene ranks by the posterior probabilities from the GMRF-MM and the original binding p-values. Dotted diagonal is the identity line. (a) all the positive control genes; (b) zoomed-in version of panel (a).

GCN4. In fact, ARG8 is known to be involved in arginine and ornithine biosynthetic processes (Pauwel *et al.* 2003; Jauniaux *et al.* 1978), and is annotated in GO Biological Process: amino acid biosynthetic process (GO ID:0008652). Note that ARG8, located in the upper-right cluster of positive control genes in Figure 2.1, is a direct neighbor of four other positive control genes but none of the negative control genes. We conjecture that its positive neighbors explain why ARG8's gene specific prior probability of being a target was estimated to be 0.733 by the GMRF-MM, in contrast to 0.058 by the SMM; the high prior probability boosted its posterior probability of being a target. Second, TRP5 (YGL026C) is a gene in neither control set, but it is a direct neighbor of ARG8. Its gene specific prior probability of being a target was estimated to be 0.716 by the GMRF-MM, as compared to 0.056 by the SMM, and the posterior probability was estimated to be 0.723 by the GMRF-MM and to be 0.032 by the SMM. The binding ratios for this gene were 1.15 and 1.21 in Lee *et al.*'s and Harbison *et al.*'s experiments respectively. However, Beyer *et al.* (2006) computationally predicted TRP5 to be a binding target of GCN4 with a high confidence level by integrating multiple data sources. In addition, TRP5 is known to participate in trytophan biosynthetic process (Elion *et al.* 1991; Toyn *et al.* 2000), and also annotated in GO Biological Process: amino acid biosynthetic process (GO ID:0008652). Hence, it is a likely target of GCN4. Finally, we looked at a positive control gene, ICY2 (YPL250C). It has six direct neighbors in the network: two of them are in the negative control set and none of them are in the postive control set. Its prior and posterior probabilities of being a target were estimated to be 0.668 and

0.836 respectively by the GMRF-MM; in contrast, the SMM gave the prior and posterior probabilities of 0.058 and 0.548 respectively. Two of its direct neighbors are negative control genes, ADY2 (YCR010C) and CRS5 (YOR031W), whose prior probabilities of being a target were estimated to be 0.08 and 0.12 respectively by the GMRF-MM, as compared to 0.058 for both by the SMM; their posterior probabilities of being a target were 0.06, 0.09 respectively by the GMRF-MM, and 0.02, 0.02 respectively by the SMM model. Therefore, although ICY2 is surrounded by non-target genes in the network, it was still identified as a binding target by the GMRF-MM, while its neighboring negative control genes were not identified as false positives. In summary, by taking use of biological knowledge embedded in a gene network, the GMRF-MM had higher statistical power for detecting the targets than did the SMM while maintaining a reasonable specificity.

2.3.2 Simulated data

2.3.2.1 Simulation set-up

To further investigate the property of our proposed method, we conducted a simulation study that mimicked real data: we simulated a gene network similar to the one used for the real data, and used data-generating distributions similar to the ones fitted to the real data. We used Wei and Li's discrete MRF (DMRF) model to generate the true binding (or differential expression) states for simulated data. Suppose for a network of G genes, we have the binding state vector $\mathbf{T} = (T_1, T_2, \ldots, T_G)$, which is modeled by a DMRF

with parameter $\Phi = (\gamma_0, \gamma_1, \beta)$. More specifically, we have

$$p(\mathbf{T}; \Phi) \propto exp(\gamma_0 n_0 + \gamma_1 n_1 - \beta n_{01}),$$

where $n_0 = \sum_i^G (1 - T_i)$ is the number of genes at state 0 with H_{i0} holding, $n_1 = \sum_i^G T_i$ is the number of genes at state 1 with H_{i1} holding, and n_{01} is the number of the network edges linking two genes at two different states. It follows that the conditional probability of gene i at state j given all the states of other genes is

$$p_i(j|.) \propto exp(\gamma_j - \beta u_i(1 - j)), \tag{2.3}$$

where $u_i(1 - j)$ is the number of the neighbors of gene i that have state $(1 - j), j = 0, 1$.

To generate simulated data, for simplicity we first removed 7 singletons from the yeast gene network and ended up with a 4609-node and 33,432-edge network. Second, to simulate $\mathbf{T}$, the latent binding states, we initialized the 66 genes in the positive control set to be binding targets and the rest of genes to be non-targets, giving an initial $\mathbf{T}$. Then we iterated the states 20 times based on Equation (2.3), with $\gamma_0 = 1, \gamma_1 = 1, \beta = 2$. It turned out that the number of binding targets became stable at about 170 after ten iterations, and we chose the states to be the ones right after the 10th iteration, giving 183 binding targets. Hence, the generated true state vector $\mathbf{T}$ was only an approximation to a MRF, lending the opportunity to investigate the robustness of the GMRF-MM. Note that our GMRF-MM assumes an exact MRF for the latent variables $\mathbf{x}$ related to prior probabilities, not a MRF for the latent binding states $\mathbf{T}$, giving our model another source of model mis-specification. Next, given $\mathbf{T}$, we simulated the z-scores according to

the fitted spatial mixture model from the real data; for simplicity, we only used the null and positive components, i.e., $\phi(0, 0.63^2)$ and $\phi(0.75, 1.53^2)$.

2.3.2.2 Simulation results

We simulated 5 data sets according to the above procedure, and fitted the two-component GMRF-MM and SMM as described in the previous section. To compare their performance, we plotted the ROC curves for these two methods as shown in Figure 2.6. Curves in the same color (or gray level) are for the same simulated dataset. Note that for each pair, the GMRF-MM won for any given specificity ranging from 10% to 95%. The average gain of sensitivity was about 10% at specificity 80%, while the average gain could reach 20% at specificity 50%. It was confirmed that again the GMRF-MM gave much higher sensitivity at a given specificity as compared to the SMM.

2.3.2.3 Sensitivity analysis

Because of imcomplete biological knowledge, it is likely that a gene network contains false positive edges while missing some true ones. To evaluate the impact of misspecified networks, we perturbed the network generated in the simulation. More specifically, we perturbed the network in three ways. In scenario 1, we randomly removed 5% (1672) edges from the original 33,432-edge network, which resulted in 46 singletons. We eliminated those singletons by randomly connecting each of them to another gene and ended up with a 31,806-edge network. In scenario 2, we randomly added 1672 edges to the original network, and thus had a 33,432-edge new network. Third, we removed the same

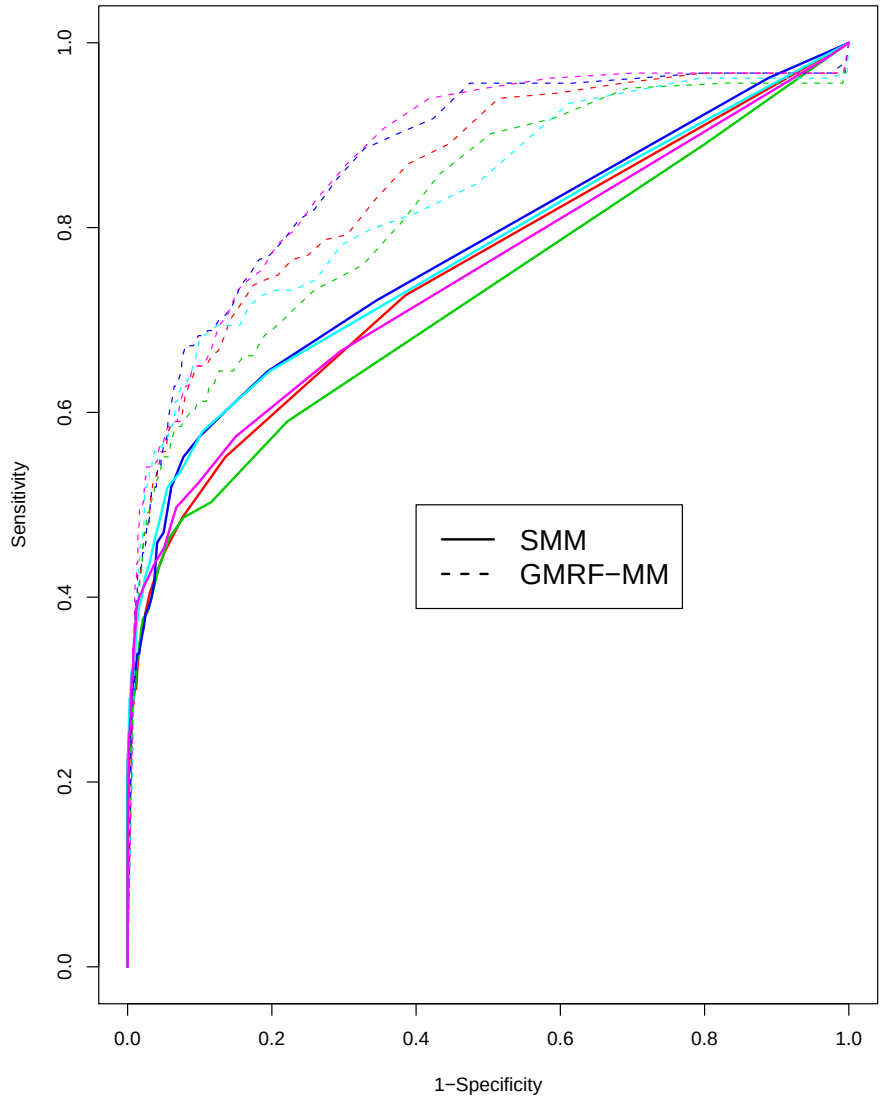

Figure 2.6: ROC curves for the two methods applied to five simulated data sets. Dashed lines are for the GMRF-MM; solid lines are for the SMM.

set of 1672 edges as in scenario 1 from, then added the same set of 1672 edges as in scenario 2 to the original network; furthermore, we eliminated 20 singletons by randomly connecting each of them of another gene, ending up with a 33,452-edge network. We applied the three perturbed networks as well as the true network to one of the aforementioned simulated data sets, and constructed the corresponding ROCs as shown in Figure 2.7. We see that removing a small percentage of edges did not seem to affect the results much (first scenario), while adding a small percentage of edges affected the results a bit more (second scenario). When the network contains both false negative and false positive edges as in the third scenario, the results seemed to be affected most substantially. Nevertheless, even in the third scenario the GMRF-MM performed no worse than the SMM. Consequently, based on our simulation study we conclude that our proposed GMRF-MM is reasonably robust to misspecified networks.

We also investigated how different hyperparameters of the prior distributions might influence the analysis results. In our current hyperparameter set-up, we imposed a moderately informative prior on the variance/precision parameters of the normal mixture components, i.e., the precision parameters had Gamma$(0.1, 0.1)$ prior distribution with mean 1 and variance 10, while other parameters had almost flat priors. We tried an almost noninformative prior distribution, Gamma$(0.0001, 0.0001)$, on the precision parameters as an alternative. This Gamma distribution has mean 1 and variance 10000. Congdon (2001) pointed out that if $p(\tau) \sim$ Gamma$(0.0001, 0.0001)$, the prior of τ will be approximately $p(\tau) \sim 1/\tau$, which is known as Jeffrey's prior and is a form of 'reference

prior' intended to reflect our ignorance about the parameter. We fitted the spatial model with these two hyperparameter set-ups, Gamma$(0.1, 0.1)$ and Gamma$(0.0001, 0.0001)$, to the same simulated dataset as used in the previous paragraph. The ROC curves are shown in Figure 2.8. We see that the two SMM curves are tied together all the way, while the two GMRF-MM curves are first coupled with each other and then separated a little. Either of the two curves from the GMRF-MM's is well above those from the SMM. In addition, the posterior distributions of the key parameters were also compared between the two GMRF-MM's and they were all very close (Tables 2.1 & 2.2).

2.4 Discussion

In this chapter, in contrast to the standard mixture model that treats the genes equally and independently *a priori*, we have proposed a spatially correlated mixture model (GMRF-MM) that allows incorporating gene network information into statistical modeling of complex inter-relationships among the genes. As expected, by borrowing information from a gene network to account for coordinated functioning of the genes, the new method has potential to improve statistical power for new discoveries with high-throughput genomic data. An application to a ChIP-chip data set and simulated data demonstrated the utility and advantage of the proposed method.

The proposed approach is in line with current efforts in integrating biological knowledge and multiple types of data (Dopazo 2006): the gene network being used can be extracted directly from existing biological databases, e.g. KEGG pathways, or can be

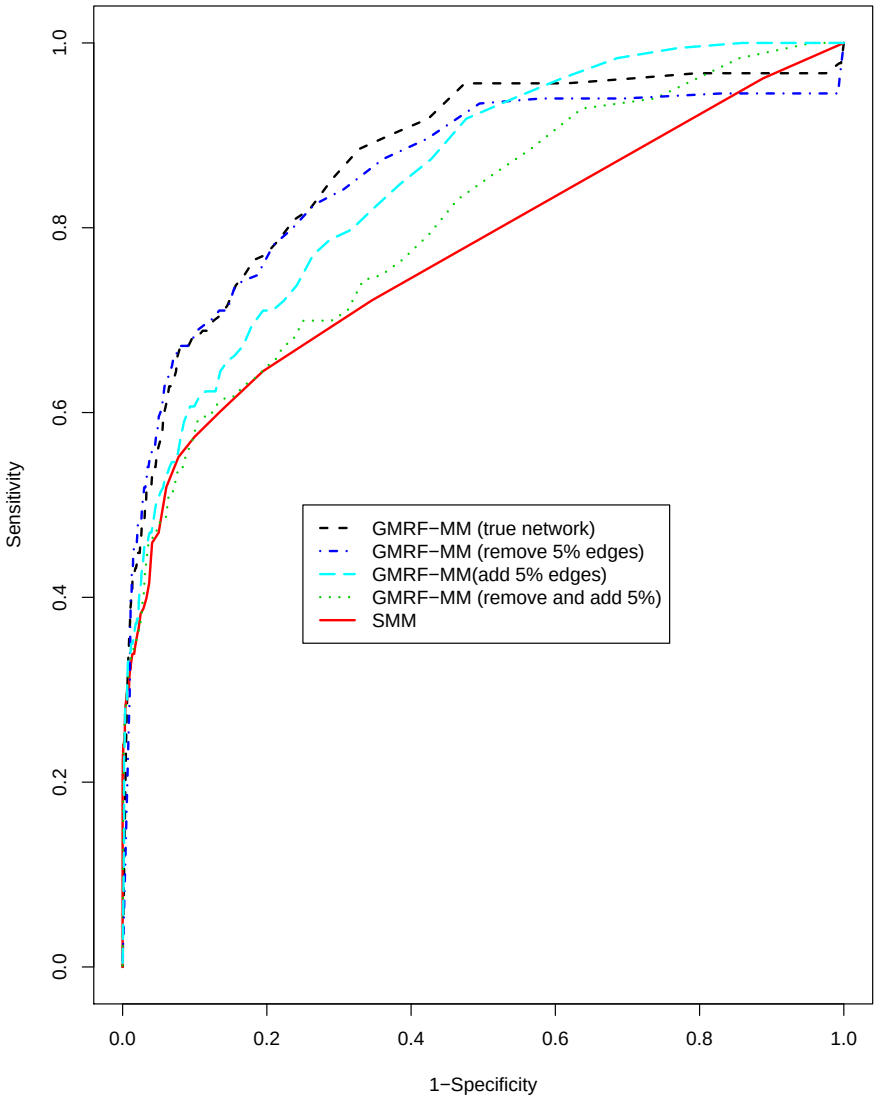

Figure 2.7: ROC curves for misspecified network structures.

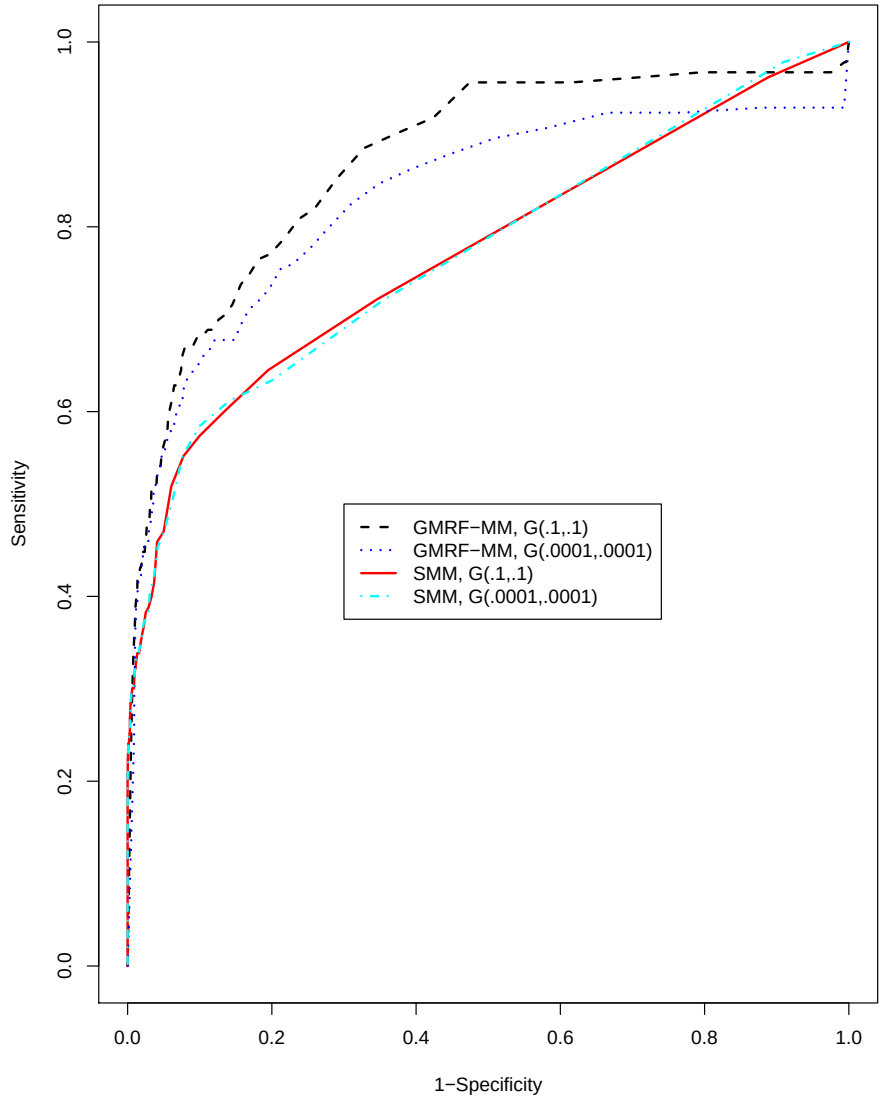

Figure 2.8: ROC curves for sensitivity analysis (two different priors for the precision parameters of the normal mixture components).

Table 2.1: Posterior distributions of key parameters in the GMRF-MM when using two different Gamma priors for the precision parameters.

Parameter	Prior	Mean	SD	2.5%	Median	97.5%
μ_2	G(.1,.1)	0.14	0.05	0.05	0.13	0.26
	G(.0001,.0001)	0.11	0.04	0.04	0.11	0.18
τ_1	G(.1,.1)	3.03	0.16	2.75	3.02	3.33
	G(.0001,.0001)	3.08	0.11	2.88	3.08	3.30
τ_2	G(.1,.1)	1.03	0.12	0.80	1.04	1.22
	G(.0001,.0001)	1.06	0.06	0.94	1.06	1.18
τ_{C1}	G(.1,.1)	0.52	0.28	0.18	0.42	1.11
	G(.0001,.0001)	0.80	0.34	0.31	0.72	1.49
τ_{C2}	G(.1,.1)	1.16	0.62	0.53	0.95	3.06
	G(.0001,.0001)	1.22	0.49	0.38	1.21	2.24

Table 2.2: Posterior distributions of key parameters in the SMM when using two different Gamma priors for the precision parameters.

Parameter	Prior	Mean	SD	2.5%	Median	97.5%
μ_2	G(.1,.1)	0.53	0.21	0.20	0.49	1.05
	G(.0001,.0001)	0.53	0.23	0.19	0.50	1.08
τ_1	G(.1,.1)	2.50	0.09	2.33	2.49	2.67
	G(.0001,.0001)	2.49	0.09	2.33	2.49	2.68
τ_2	G(.1,.1)	0.55	0.10	0.36	0.55	0.75
	G(.0001,.0001)	0.55	0.10	0.36	0.54	0.77
π_1	G(.1,.1)	0.95	0.02	0.92	0.96	0.98
	G(.0001,.0001)	0.95	0.02	0.91	0.96	0.98
π_2	G(.1,.1)	0.05	0.02	0.02	0.04	0.08
	G(.0001,.0001)	0.05	0.02	0.02	0.04	0.09

computationally predicted by integrative analysis of multiple types of data, as for the functional coupling gene network constructed by Lee *et al.* (2004) and used in our example. In addition, the proposed method covers the prior probability specification in a stratified mixture model used to incorporate gene functional annotations (Pan 2005, 2006a, 2006b) as a special case; in the latter, a corresponding gene network may be regarded as the follows: each functional group is a subnetwork consisting of the genes fully connected to each other, while there is no connection between any two functional groups, and the smoothing parameter σ_{Cj} in the ICAR model is small enough (e.g. $\sigma_{Cj} = 0$) to induce a constant prior probability for the genes within the same functional group. Of course, the special case may be too restrictive: for example, first, the genes within the same functional group may play different roles, not necessarily sharing the same prior probability; second, different gene functional groups may interact with each other, possibly through some genes with multiple functions. On the other hand, our idea differs from existing approaches to using gene networks (Wei and Li 2007, and references therein). For example, although there is a conceptual similarity between ours and that of Wei and Li (2007) in the use of MRF to account for correlations among the genes, we model the prior probabilities of genes' being in a certain state via a Gaussian MRF, in contrast to the underlying true states as modeled by Wei and Li, via a discrete(binary) MRF. As a consequence, by the general result of the consistency of the posterior probability relative to a specified prior distribution, we expect that our method is more robust to misspecified gene networks than that of Wei and Li, which may be more efficient if the gene network

is correctly specified; due to incomplete biological knowledge and prediction errors, it seems unlikely that a gene network can be specified completely correctly. In addition, we use a Normal distribution for each component of the mixture model while Wei and Li adopted the Gamma-Gamma model, although other parametric models in either can be adopted. Further studies to investigate the operating characteristics of our proposal, including comparing its performance with others, and to extend our proposal to more complex settings (e.g. Lewin *et al.* 2006) will be interesting.

2.5 Appendix

2.5.1 Model specifications

2.5.1.1 Model specification for a three-component standard mixture model

For $i = 1, \ldots, G; j = 0, 1, 2,$

$$(Z_i|T_i = j) \sim N(\mu_j, \sigma_j^2),$$

$$Pr(T_i = j) = \pi_j,$$

$$\mu_0 = 0; \;\; \mu_1 \sim N(0, 10^6)\mathrm{I}(n, 0), \;\; n = \min_i z_i,$$

$$\mu_2 \sim N(0, 10^6)\mathrm{I}(0, m), \;\; m = \max_i z_i,$$

$$\sigma_j^2 \sim \text{Inverse Gamma}(0.1, 0.1), \tau_j = 1/\sigma_j^2,$$

$$(\pi_0, \pi_1, \pi_2) \sim \text{Dirichlet}(1, 1, 1).$$

A two-component mixture model can be specified by dropping the negative component, as used in the simulation.

2.5.1.2 Model specification for a three-component GMRF-based mixture model

For $i = 1, \ldots, G; j = 0, 1, 2,$

$$
\begin{aligned}
&(Z_i|T_i = j) \sim N(\mu_j, \sigma_j^2),\\
&Pr(T_i = j) = \pi_{ij} = \exp(x_{ij})/[\exp(x_{i0}) + \exp(x_{i1}) + \exp(x_{i2})],\\
&x_{ij}|x_{(-i)j} \sim N\left(\frac{1}{m_i}\sum_{l\in\delta_i} x_{lj}, \frac{\sigma_{Cj}^2}{m_i}\right),\\
&\mu_0 = 0; \;\; \mu_1 \sim N(0, 10^6)\mathrm{I}(n, 0), \;\; n = \min_i z_i,\\
&\mu_2 \sim N(0, 10^6)\mathrm{I}(0, m), \;\; m = \max_i z_i,\\
&\sigma_j^2 \sim \text{Inverse Gamma}(0.1, 0.1), \tau_j = 1/\sigma_j^2,\\
&\sigma_{Cj}^2 \sim \text{Inverse Gamma}(0.01, 0.01), \tau_{Cj} = 1/\sigma_{Cj}^2,
\end{aligned}
$$

where $\sum_i x_{ij} = 0$, for $j = 0, 1, 2$; δ_i is the set of indices for the neighbors of gene i, and $m_i = |\delta_i|$. A two-component mixture model can be specified by dropping the negative component, as used in the simulation.

2.5.2 WinBUGS codes for implementing the two methods

2.5.2.1 For a three-component standard mixture model

```
model
{
for( i in 1:N ) {
Z[i] ~dnorm(muR[i], tauR[i]) # z-scores
```

```
muR[i] <- mu[T[i]]

tauR[i] <- tau[T[i]]

T[i] ~dcat(pi[ ]) # latent variable (zero/negative/postive components)

T1[i] <-equals(T[i],1) ;T2[i] <-equals(T[i],2); T3[i]<-equals(T[i],3);

}

# prior for mixing proportions

pi[1:3] ~ ddirch(alpha[])

# priors (means of normal mixture components)

mu[1] <- 0 # zero component

mu[2] ~dnorm(0, 1.0E-6)I(m,0.0) #negative component

mu[3] ~dnorm(0, 1.0E-6)I(0.0,n) #positive component

# priors (precision/variance of normal mixture components)

tau[1]~dgamma(0.1, 0.1)

tau[2]~dgamma(0.1, 0.1)

tau[3]~dgamma(0.1, 0.1)

sigma2[1] <-1/tau[1]

sigma2[2] <-1/tau[2]

sigma2[3] <-1/tau[3]

}
```

2.5.2.2 For a three-component GMRF-based mixture model

```
model

{

for( i in 1:N ) {

Z[i] ~dnorm(muR[i], tauR[i])  # z-scores

muR[i] <- mu[T[i]]

tauR[i] <- tau[T[i]]

# logistic transformation

pi[i,1] <-1/(1+exp(x2[i]-x1[i])+exp(x3[i]-x1[i]))

pi[i,2] <-1/(1+exp(x1[i]-x2[i])+exp(x3[i]-x2[i]))

pi[i,3] <-1/(1+exp(x1[i]-x3[i])+exp(x2[i]-x3[i]))

T[i] ~dcat(pi[i,1:3]) # latent variable (zero/negative/postive components)

T1[i] <-equals(T[i],1) ;T2[i] <-equals(T[i],2); T3[i] <-equals(T[i],3)

}

# Random Fields specification

x1[1:N] ~car.normal(adj[], weights[], num[], tauC[1])

x2[1:N] ~car.normal(adj[], weights[], num[], tauC[2])

x3[1:N] ~car.normal(adj[], weights[], num[], tauC[3])

# weights specification

 for(k in 1:sumNumNeigh) { weights[k] <- 1 }

# priors (precision/variance for MRF)
```

```
tauC[1] ~dgamma(0.01, 0.01)I(0.0001,)

tauC[2] ~dgamma(0.01, 0.01)I(0.0001,)

tauC[3] ~dgamma(0.01, 0.01)I(0.0001,)

sigma2C[1] <- 1/tauC[1]

sigma2C[2] <- 1/tauC[2]

sigma2C[3] <- 1/tauC[3]

# priors (means of normal mixture components)

mu[1] <- 0 # zero component

mu[2] ~dnorm(0, 1.0E-6)I(n,0.0) # negative component

mu[3] ~dnorm(0, 1.0E-6)I(0.0,n) # positive component

# priors (precision/variance of normal mixture components)

tau[1]~dgamma(0.1, 0.1)

tau[2]~dgamma(0.1, 0.1)

tau[3]~dgamma(0.1, 0.1)

sigma2[1] <- 1/tau[1]

sigma2[2] <- 1/tau[2]

sigma2[3] <- 1/tau[3]

}
```

Chapter 3

Network-based Genomic Discovery: Application and Comparison of Markov Random Field Models

3.1 Introduction

In Chapter 2, we proposed a Gaussian Markov random field-based mixture model (GMRF-MM) for incorporating gene network information into statistical analysis of genomic data. Specifically, we modeled the prior probabilities of the true states via GMRFs, in contrast to Wei and Li (2007), who modeled the latent true states of the genes using a discrete Markov random field (DMRF). Both methods were shown to be more powerful in detecting differentially expressed genes or regulatory targets based on real and simulated data than standard mixture models that do not capitalize on gene networks. However, the comparative performance of DMRF-based mixture model (DMRF-MM) and GMRF-based mixture model (GMRF-MM) is not yet clear, which motivated us to compare the two methods based on the GCN4 ChIP-chip data set, which was analyzed in Chapter 2, and simulated data in this chapter.

As our motivating example, the data were drawn from Lee *et al.* (2002), who did ChIP-chip experiments for a broad transcription regulator, General Control Nondepressible 4 (GCN4) in yeast *Saccharomyces cerevisiae*. It is known that GCN4 is a transcriptional activator of amino acid biosynthetic genes in response to amino acid starvation in yeast, and the purpose of the study was to identify the binding targets of GCN4 based on the ChIP-chip data. Specifically, Lee *et al.* did ChIP-chip experiments for GCN4 for 6,270 genes with three independent replicates and employed a parametric method called "single-array error model" (see Section 3.3.1 for more details) to obtain a p-value for each gene for testing the null hypothesis that the gene is not a binding target of GCN4.

Table 3.1: Some data from Lee *et al.*'s Chip-chip experiments.

Index	Binding ratio (IP-enriched/-unenriched)	Binding p-value	z-score
GENE1	2	0.00051	3.28
GENE2	1.5	0.019	2.08
GENE3	1.3	0.08	1.41
GENE4	0.9	0.67	-0.44
GENE5	0.77	0.89	-1.23

Table 3.1 shows a small portion (5 of 6,270 genes) of the GCN4 data, where the binding ratios and p-values were derived from the three replicates and the z-scores obtained from the p-values will be discussed in Section 3.2.1.

We point out some potential limitations with Wei and Li's approach to parameter estimation. First, they only obtained the maximum *a posteriori* (MAP) estimate via the iterated conditional modes (ICM) algorithm (Besag 1986), which only provides the most probable state of each gene, but not its posterior probability. As a result, user-specified cutoffs for claiming positive genes and estimating the False Discovery Rate (FDR) (Benjamini and Hochberg 1995; Newton *et al.* 2004) are not possible. Second, their approach does not take account of the uncertainty of the estimated spatial interaction parameter for the DMRF (Heikkinen and Hogmander 1994), which plays a central role in determining the smoothness of the DMRF. Finally, the ICM suffers from stopping at local maxima rather than the global one, and even starting from a set of "good" initial values does not guarantee that the ICM reaches the global maximum (Winkler 2003, p129).

Alternatively, we propose adopting a fully Bayesian approach to the DMRF to overcome the above drawbacks of Wei and Li's implementation. There is a body of literature on fully Bayesian approach to DMRF modeling in the context of image analysis and spatial statistics (Heikkinen and Hogmander 1994; Ryden and Titterington 1998; Green and Richardson 2002; Smith and Smith 2006; Smith and Fahrmeir 2007). In particular, Smith and Smith (2006) compared DMRF and GMRF using three image examples. Our proposed comparison is different from theirs. First, they related the GMRF to the latent states by thresholding, while here the latent states' prior probabilities are defined via a logistic transformation of some GMRFs. Second, unlike image analysis or traditional spatial statistics problems where the neighborhood structure is relatively simple, a gene network is essentially a very irregular lattice with a complicated structure, which may be mis-specified due to incomplete biological knowledge. Third, we evaluate the performance of a direct posterior probability approach to FDR estimation for these MRF-based models, which, to our knowledge, has not been studied elsewhere before. Therefore, it is informative to compare the performance of DMRF-MM and GMRF-MM in the context of microarray data and in particular, their robustness to mis-specified gene networks. In addition, we propose two novel constraints in the prior specifications for the GMRF-MM to improve its performance. Note that Wei and Li modeled the gene expression data (with replicates) directly by using Gamma mixtures; here, to facilitate comparison, we model a one-dimensional summary statistic vector using normal mixtures while modeling the dependency among latent states via DMRF or GMRF.

The rest of this chapter is organized as follows. We first briefly review the standard mixture model (SMM), GMRF-MM, and DMRF-MM, and then propose two modifications to the inference procedure of the GMRF-MM. We discuss statistical inference for the SMM and GMRF-MM in a fully Bayesian framework and the ICM approach to DMRF-MM parameter estimation. We also propose a fully Bayesian approach to DMRF-MM. We apply and compare the methods with the GCN4 ChIP-chip data as mentioned earlier. A simulation study was also conducted to compare the robustness of the two MRF-based methods to mis-specified gene networks. We end with a short discussion on some existing issues and future work. Note that Chapter 2 is referred as Wei and Pan (2008a) hereafter.

3.2 Methods

3.2.1 Notation

Our goal is to identify regulatory target genes of a TF. This can be formulated as a hypothesis testing problem: for each gene i, we test a null hypothesis H_{i0} against an alternative H_{i1}, usually the opposite of H_{i0}. For example, H_{i0} is that "gene i is not a target of the TF".

We assume that the data have been summarized by a scalar statistic Z_i for each gene $i, i = 1, \ldots, G$; for example, Z_i might be a test statistic measuring the relative abundance of the TF, the statistical significance level for rejecting H_{i0}, or z-scores as defined by $z_i = \Phi^{-1}(1 - P_i)$, where Φ is the cumulative distribution function of the

standard Normal distribution $N(0,1)$ and P_i is the p-value for gene i. Define the state of gene i by $T_i = \mathrm{I}(H_{i0}$ is false); that is, $T_i = 1$ or $T_i = 0$ corresponds to whether H_{i1} or H_{i0} holds respectively. Denote the distribution functions of Z_i for the genes when $T_i = 1$ and $T_i = 0$ as f_1 and f_0, respectively.

3.2.2 Standard mixture model

Assuming that *a priori* all the genes have an independent and identical distribution (iid), we have the marginal distribution of Z_i as a standard mixture model (SMM):

$$f(z_i) = \pi_0 f_0(z_i) + (1 - \pi_0) f_1(z_i), \tag{3.1}$$

where π_0 is the prior probability $Pr(T_i = 0)$. The prior probabilities are the same for all the genes.

The null and non-null distributions f_0 and f_1 may be approximated by finite normal mixtures: $f_0 = \sum_{k_0=1}^{K_0} \pi_{0k_0} \phi(\mu_{k_0}, \sigma_{k_0}^2)$ and $f_1 = \sum_{k_1=1}^{K_1} \pi_{1k_1} \phi(\mu_{k_1}, \sigma_{k_1}^2)$, where $\phi(\mu, \sigma^2)$ is the density function for a Normal distribution with mean μ and variance σ^2. For z-scores, if P_i is properly calculated as a genuine p-value, f_0 is exactly the standard normal, which, however, is usually not true in practice due to approximations (e.g., resulting from possible correlations among the genes, contrary to the adopted independence assumption). As a result, f_0 needs to be estimated in practice. In addition, f_1 may model the right-tail of the z-score distribution. McLachlan *et al.* (2006) demonstrated empirically that using $K_j = 1$ often suffices. In our real data example, we found that $K_0 = 2$ and $K_1 = 1$ worked well; also see Liang and Zhang (2008) for a more comprehensive discussion fn

practical issues in decomposing f into f_0 and f_1. For simplicity of exposition, we assume that $K_j = 1$ for $j = 0, 1$ in the following discussion; the Appendix gives an example of relaxing this restriction. The conditional distribution of z_i is thus:

$$p(z_i|T_i = j, \theta) = \phi(z_i; \mu_j, \sigma_j^2), \tag{3.2}$$

where $\theta = (\mu_0, \mu_1, \sigma_0, \sigma_1)$.

3.2.3 GMRF-based mixture model

In a GMRF-MM, gene-specific prior probabilities π_{ij} $(i = 1, \ldots, G$ and $j = 0, 1)$ are introduced, and are related to two latent GMRF's $\mathbf{x}_j = \{x_{ij}; i = 1, ..., G\}$ via a logistic transformation:

$$\pi_{ij} = Pr(T_i = j) = \exp(x_{ij})/[\exp(x_{i0}) + \exp(x_{i1})]. \tag{3.3}$$

Defined over a gene network, each of the G-dimensional latent vectors $\mathbf{x}_j$ is distributed according to an intrinsic Gaussian conditional autoregression model (ICAR) (Besag and Kooperberg 1995). A key feature of ICAR is the Markovian interpretation of the latent variables' conditional distributions: the distribution of each x_{ij}, conditional on $x_{(-i)j} = \{x_{kj}; k \neq i\}$, depends only on its direct neighbors. Specifically, we have

$$x_{ij}|x_{(-i)j} \sim N\left(\frac{1}{m_i}\sum_{l\in\partial i} x_{lj}, \frac{\sigma_{Cj}^2}{m_i}\right), \tag{3.4}$$

where ∂i is the set of indices for the neighbors of gene i, and m_i is the corresponding number of neighbors. Adding a constant to $\mathbf{x}_j$ does not change the full conditional distribution (3.4). Therefore, to allow identifiability, the sum-to-zero constraint $\sum_i x_{ij} =$

0 ($j = 0, 1$) is often imposed (Broet and Richardson 2006; Wei and Pan 2008a), which is also the default setting in WinBUGS (Spiegelhalter *et al.* 2003). In this model, the parameter σ^2_{Cj} acts as a smoothing prior for the spatial field and consequently controls the degree of dependency among the prior probabilities of the genes across the genome: smaller σ^2_{Cj} induces more similar π_{ij}'s for those genes that are neighbors in the network. Finally, the conditional distribution of z_i is

$$f(z_i|x_{i0}, x_{i1}) = \pi_{i0} f_0(z_i) + \pi_{i1} f_1(z_i). \tag{3.5}$$

3.2.3.1 Modifications

In the GMRF-MM, to allow identifiability, the following constraint is typically imposed

$$\sum_i x_{i0} = \sum_i x_{i1} = 0 \Longrightarrow \bar{x}_{.0} = \bar{x}_{.1} = 0,$$

where $\bar{x}_{.j} = (1/G) \sum_{i=1}^{G} x_{ij}$ for $j = 0, 1$. By (3.3), we have $\text{logit}(\pi_{i1}) = x_{i1} - x_{i0}$. It follows that

$$\frac{1}{G} \sum_{i=1}^{G} \text{logit}(\pi_{i1}) = \bar{x}_{.1} - \bar{x}_{.0} = 0.$$

Thus, the $\text{logit}(\pi_{i1})$ have mean 0. This implies that the posterior estimates of the $\text{logit}(\pi_{i1})$ will be shrunk towards 0, or roughly, the estimates of the π_{i1} are shrunk towards 0.5. This is consistent with our observation that the average of posterior means $\hat{\pi}_{i1}$ in the GMRF-MM, under this constraint, tends to be much larger than π_1 in the SMM. For example, for the GCN4 data, the former number was 0.186, while the latter was only 0.058 (Wei and Pan 2008a); correspondingly, the estimated mean for the right-tail (non-null) component in the GMRF-MM was much smaller than in the SMM. This

may lead to more false positive genes, which is undesirable. Because biologically only a small proportion of the genes (surely fewer than a half) could be targets of a TF (Lee *et al.* 2002), we propose shrinking $\text{logit}(\pi_{i1})$'s towards a negative constant c, e.g., an estimate of $\text{logit}(\pi_1)$. Specifically, we impose

$$\frac{1}{G}\sum_i \text{logit}(\pi_{i1}) = \bar{x}_{\cdot 1} - \bar{x}_{\cdot 0} = c,$$

which can be realized by imposing $\bar{x}_{\cdot 1} = c$ and $\bar{x}_{\cdot 0} = 0$. In practice, we found that the estimate $\hat{\pi}_1$ from the SMM performed reasonably well, and hence propose taking $c = \text{logit}(\hat{\pi}_1)$. Because of the choice of constant c, we call this modified method "GMRF-MM with the logit constraint", while we call the original model "GMRF-MM with the zero constraint".

The above proposed constraint targets the average of $\text{logit}(\pi_{ij})$, which may not have much direct effect on the average of π_{ij}. Alternatively, we propose shrinking π_{ij} towards $\hat{\pi}_j$ directly via the following weighted average constraint:

$$\pi_{ij} = \lambda \frac{\exp(x_{ij})}{\exp(x_{i0}) + \exp(x_{i1})} + (1-\lambda)\hat{\pi}_j, \tag{3.6}$$

where $0 \leq \lambda \leq 1$, and λ controls the extent to which the estimate of π_{ij} is shrunk towards $\hat{\pi}_j$. Note that when $\lambda = 0$, π_{ij} reduces to $\hat{\pi}_j$; in contrast, when $\lambda = 1$, π_{ij} is just as (3.3). For simplicity, we use $\lambda = 1/2$ in our data analysis, though other weights or even treating λ as a tuning parameter could be employed. For better performance, we still put the logit constraint on the GMRFs, i.e., $\bar{x}_{\cdot 1} = \text{logit}(\hat{\pi}_1)$ and $\bar{x}_{\cdot 0} = 0$. We call this modified model "GMRF-MM with the average constraint".

3.2.4 DMRF-based mixture model

In a DMRF-MM, the latent state vector $\mathbf{T} = (T_1, \ldots, T_G)'$ is directly modeled as a DMRF. Specifically, we assume the following auto-logistic model for the conditional distribution of T_i,

$$Pr(T_i = 1|T_{(-i)}, \Phi) = Pr(T_i = 1|T_{\partial i}, \Phi) = \frac{\exp\left(\gamma + \beta(n_i(1) - n_i(0))/m_i\right)}{1 + \exp\left(\gamma + \beta(n_i(1) - n_i(0))/m_i\right)}, \quad (3.7)$$

where $\Phi = (\gamma, \beta)$, γ and $\beta > 0$ are arbitrary real numbers, ∂i represents the (direct) neighbors of gene i and $n_i(j)$ is the number of gene i's neighbors having state j for $j = 0, 1$, and thus $(n_i(1) - n_i(0)) = \sum_{l \in \partial i}(2T_l - 1)$; m_i is the number of gene i's neighbors. The attraction parameter β corresponds to the spatial interaction strength in the DMRF, i.e., the tendency of sharing the same state as neighboring genes. Hence, the larger β is, the more probable it is that large clusters of common states appear. Due to the unknown normalizing constant $C(\Phi)$, the likelihood $l(\mathbf{T}; \Phi)$ does not have a closed-form. Instead, Besag (1986) proposed using the *pseudolikelihood*

$$pl(\mathbf{T}; \Phi) = \prod_{i=1}^{G} p(T_i|T_{\partial i}; \Phi) = \prod_{i=1}^{G} \frac{\exp\left(T_i(\gamma + \beta(n_i(1) - n_i(0))/m_i)\right)}{1 + \exp\left(\gamma + \beta(n_i(1) - n_i(0))/m_i\right)}, \quad (3.8)$$

The maximizer of the pseudolikelihood is often consistent (Winkler 2003, p272).

3.2.5 Comparison of the three mixture models

In this section, we compare the SMM, GMRF-MM and DMRF-MM by taking a close look at the full conditional distributions for T_i, i.e., the conditional distribution of T_i given the data and all other parameters in the model. For the SMM, GMRF-MM, and

DMRF-MM, we have

$$Pr(T_i = 1|\mathbf{z}, \theta, \pi_0, \pi_1) = \frac{1}{1 + \frac{\pi_0}{\pi_1} \frac{\phi(z_i;\mu_0,\sigma_0^2)}{\phi(z_i;\mu_1,\sigma_1^2)}}, \tag{3.9}$$

$$Pr(T_i = 1|\mathbf{z}, \theta, x_{i0}, x_{i,1}) = \frac{1}{1 + \frac{\pi_{i0}}{\pi_{i1}} \frac{\phi(z_i;\mu_0,\sigma_0^2)}{\phi(z_i;\mu_1,\sigma_1^2)}} = \frac{1}{1 + \exp\left(x_{i0} - x_{i1}\right) \frac{\phi(z_i;\mu_0,\sigma_0^2)}{\phi(z_i;\mu_1,\sigma_1^2)}}, \tag{3.10}$$

and

$$Pr(T_i = 1|\mathbf{z}, \theta, T_{\partial i}, \Phi) = \frac{1}{1 + \frac{1}{\exp(\gamma+\beta(n_i(1)-n_i(0))/m_i)} \frac{\phi(z_i;\mu_0,\sigma_0^2)}{\phi(z_i;\mu_1,\sigma_1^2)}} \tag{3.11}$$

respectively.

Although in practice inferences are based on the marginal posterior probability $Pr(T_i|\mathbf{z})$, the above conditional posterior probabilities provide a unique perspective to compare the three mixture models. First, the (conditional) posterior probability of being a target is jointly determined by the prior probability ratio and the data, i.e., the likelihood ratio $\frac{\phi(z_i;\mu_0,\sigma_0^2)}{\phi(z_i;\mu_1,\sigma_1^2)}$, in all three models. This sheds light on why mis-specified prior distributions, e.g., due to incomplete gene networks, may not have a large influence on the posterior probability if the likelihood ratio is large. Second, the GMRF-MM is more richly parameterized and thus more flexible because it has thousands of additional parameters (x_{ij}'s), as compared to DMRF-MM. However, these additional parameters are not treated as independent fixed effects but are linked by the adopted hierarchical structure of the GMRF's, which leads to borrowing information among the parameters via shrinkage (Carlin and Louis 2000). The extent of the shrinkage among x_{ij}'s is controlled by σ_{C0} and σ_{C1}. For example, when they are both 0, x_{ij} is a constant, and so is π_{ij}. Although the posterior distributions of σ_{C0} and σ_{C1}, which are jointly determined by

the data and the adopted priors, are obtained automatically in a Bayesian hierarchical modeling framework via Markov Chain Monte Carlo (MCMC), there may be potential overfitting problem. A possible solution to this is the "average constraint", which shrinks π_{ij} toward $\hat{\pi}_j$, leading to better model fitting and improved predictive performance, as illustrated in Sections 3 & 4. Finally, for the DMRF-MM, γ plays a role as log prior probability ratio when $n_i(1) = n_i(0)$.

3.2.6 Parameter estimation

Following Wei and Pan (2008a), we adopt a fully Bayesian approach to the SMM and GMRF-MM (with the zero, logit, or average constraint). Briefly, we use non-informative or moderately informative priors. MCMC is used to draw posterior samples for model parameters. See Appendix for complete Bayesian model specifications and MCMC algorithms.

3.2.6.1 ICM approach to DMRF

When inferring the true states $\mathbf{T}$ for the G genes, the parameter estimation must be carried out simultaneously. Wei and Li (2007) adopted the ICM algorithm of Besag (1986) to estimate the parameters in the DMRF-MM. ICM uses a "greedy" strategy in an iterative local maximization and its convergence is usually achieved after only a few iterations. See Wei and Li (2007) for details.

3.2.6.2 Bayesian approach to DMRF

Although the ICM approach is easy to implement and requires little computational effort, it has major drawbacks as mentioned before. Here we propose a Bayesian approach to DMRF-MM. Before we move on to a Bayesian model specification, we would argue that in the context of identifying binding target genes, it is more appropriate to make our inference based on the marginal posterior probability $p_i = Pr(T_i = 1|\mathbf{z})$ rather than the maximum *a posteriori* (MAP) $\mathbf{T}$, the mode of the joint posterior distribution $Pr(\mathbf{T}|\mathbf{z})$. Define the maximum marginal posterior (MMP) $\tilde{\mathbf{T}} = (\tilde{T}_1, \ldots, \tilde{T}_G)$, where $\tilde{T}_i = \mathrm{I}(p_i \geq 0.5) = \arg\max_{t_i \in \{0,1\}} Pr(T_i = t_i|\mathbf{z})$. In decision theory, the MMP corresponds to maximizing the expected number of correctly classified genes, or equivalently minimizing the expected mis-classification rate; in contrast, the MAP given by ICM corresponds to minimizing a zero-one loss function according to whether the classification is perfect or not, and is less appealing; see Appendix 3.6.3 for remarks on the above statements. Note that by adopting a Bayesian approach, we can obtain not only the MMP estimates $\tilde{T}_i$, but also p_i itself, which is more informative than the MMP estimate and allows user-specified cutoffs to infer T_i.

Our proposed Bayesian DMRF-MM can be specified as follows:

$$p(\mathbf{T}, \theta, \Phi|\mathbf{z}) \propto p(\mathbf{z}|\mathbf{T}, \theta)p(\mathbf{T}|\Phi)p(\theta)p(\Phi), \tag{3.12}$$

where $p(\mathbf{z}|\mathbf{T}, \theta)$ is as in (3.2); for $p(\mathbf{T}|\Phi)$, we adopt the pseudolikelihood (3.8); for θ, we have $p(\theta) = p(\mu_0)p(\mu_1)p(\sigma_0^2)p(\sigma_1^2)$, and we use the same prior distributions for θ as for GMRF-MM; for Φ, we have $p(\gamma) \propto 1$ and $p(\beta) \propto I(0 < \beta < \beta_{max})$, where β_{max}

is a prespecified maximum. Note that Ryden and Titterington (1998) showed that the pseudolikelihood $pl(\mathbf{T};\Phi)$ provides a reasonable approximation to $p(\mathbf{T}|\Phi)$. The complete model specifications and MCMC algorithm can be found in the Appendix. Because of our model specification, the full conditional distributions are straightforward to obtain for all parameters but Φ. As a result, we use *Metropolis within Gibbs* (Carlin and Louis 2000) to draw samples from the posterior distribution (3.12): we use a Metropolis algorithm for Φ and Gibbs samplers for the remaining parameters. The complete algorithm can be found in Appendix 3.6.2.

3.2.7 Inference

MCMC algorithms for the SMM and GMRF-MM with any of the three constraints can be implemented in WinBUGS V1.40 (Spiegelhalter *et al.* 2003), while Bayesian DMRF-MM, to our best knowledge, cannot be carried out in WinBUGS. As a result, we wrote an R program to implement the latter. Multiple starting values for MCMC were used to increase the chance that the chains had converged, which was monitored by trace plots. Depending on the model, the length of burn-in samples, i.e., MCMC samples before being used, varied. Generally, the SMM took the shortest burn-in time - less than 5000 iterations, while it usually took 10,000 iterations for GMRF-MM's chains to converge.

The posterior mean of any parameter based on 10,000 MCMC samples after burn-in was used as its point estimate. In particular, based on whether the point estimate $\hat{p_i} = \widehat{Pr}(T_i = 1|\mathbf{z})$ was larger than a threshold t, we determine whether to reject H_{i0}. There is a correspondence between t and FDR, which has become increasingly popular

for controlling multiple-test errors in microarray data analysis. A direct estimator of FDR can be constructed based on p_i (Newton *et al.* 2004):

$$\text{FDR}(t) = \sum_{i=1}^{G} q_i \text{I}(q_i \leq t) / \sum_{i=1}^{G} \text{I}(q_i \leq t), \tag{3.13}$$

where $q_i = Pr(T_i = 0|\mathbf{z}) = 1 - p_i$. Plugging in the estimates of the q_i's, we obtain an estimated FDR. Note that the denominator gives the estimated number of positive results. Also note that the above estimator is typically used with independent T_i's, which, however, are correlated here due to the imposed MRF structure. Therefore, it is not clear whether it works well in the current context. Although Wu (2008) proposed a Benjamini-Hochberg-like procedure to control FDR under dependence, it is not aimed at estimating FDR. As a result, we used the FDR estimator (3.13) in our real and simulated data examples, and assessed its accuracy under dependence.

3.3 Example

3.3.1 Data

We downloaded the GCN4 ChIP-chip data of Lee *et al.* (2002) from the authors' website (http://web.wi.mit.edu/young/index.html) in early March 2007. Binding ratios and p-values for 6,270 yeast genes were available. The p-values were derived based on three independent replicated experiments by a parametric method called the "single-array error model" (Hughes *et al.* 2000). A sample of the data is displayed in Table 3.1, where the z-scores were obtained as described in Section 3.2.1. Wei and Pan (2008a)

analyzed this data set by applying a SMM and a GMRF-MM with the zero constraint to the z-scores. Following Wei and Pan (2008a), we used the yeast functional linkage gene network "ConfidentNet" of Lee *et al.* (2004), which was shown to have high credibility. Specifically, Lee *et al.* applied a naive Bayes method to assign a score to each possible gene pair by integrating a variety of genomic data, including mRNA co-expressions, gene co-citations, protein-protein inter-actions, gene fusions and phylogenetic profiles. Two genes with a score high enough were linked, suggesting the high likelihood of their shared biological function. Represented by an undirected graph, the "ConfidentNet" consists of 4,681 nodes (genes) and 34,000 edges (gene-gene functional linkages). A summary of the distribution of the number of direct neighbors is: minimum=1, 25th-percentile=2, median=6, 75th-percentile=13 and maximum=188. An exceptional feature of the data is that Pokholok *et al.* (2005) constructed a set of 80 genes that were very likely to the regulatory targets of GCN4 from multiple sources of data (including another set of more accurate ChIP-chip experiments based on a new generation of microarrays, a gene expression data set, and DNA motif analyses), as well as a set of 900 genes that are unlikely to be regulated by GCN4. Treating the positive and negative control sets as the true positives and true negatives, we calculated the sensitivity and specificity for different statistical methods and subsequently constructed the Receiver Operating Characteristic (ROC) curves. After merging the ChIP-chip data set and the gene network, we ended up with a 4,609-node network with 33,432 edges. We extracted those 4,609 genes' binding p-values and obtained their z-scores for final analysis; correspondingly, there were 66 and

769 genes in the positive and negative control sets respectively.

3.3.2 Parameter estimates

We applied GMRF-MM with the logit and average constraints, and DMRF-MM (both Bayesian and ICM approaches) to the GCN4 ChIP-chip data, and compared the results with those by SMM and GMRF-MM with the zero constraint in Wei and Pan (2008a). Wei and Pan (2008a) reported that adding a mixture component with a negative mean improved the goodness-of-fit as well as statistical power as gauged by ROC curves; we fitted both two-component and three-component mixture models to the data and came to the same conclusion as theirs, except for the ICM-based DMRF-MM, for which a third component did not seem necessary. Nevertheless, we proceeded to compare different methods' performance based on three-component mixture models, and treated the normal component with a positive mean as the non-null one because of the use of z-scores, i.e., smaller p-values correspond to larger z-scores. Complete model specifications for three three-component SMM, GMRF-MM and DMRF-MM can be found in Appendix 3.6.1. Parameter estimates for all the models are shown in Tables 3.2 & 3.3. Note that the prior probabilities (π_j's) were the averages of gene-specific π_{ij}'s across the genes for any MRF-based model. Several features are noticeable. First, although all models seemed to give reasonable goodness-of-fit (by checking the fitted marginal and component-wise distributions against the data histograms, results not shown), Bayesian DMRF-MM was more similar to SMM in terms of model fitting: the negative components for SMM and DMRF-MM tended to capture the bump around -2, while the negative components for

GMRF-MM, with the zero, logit, or average constraint, tended to capture the peak area around zero and had much larger prior probabilities. Second, owing to the use of the modified prior constraint, the average prior probability for the positive component for GMRF-MM with the average constraint was 0.04, which was much closer to SMM's 0.06 compared to that with the zero constraint at 0.18. Additionally, the mean of the positive component for GMRF-MM with the average constraint was farther away from zero as compared with that for GMRF-MM with the zero constraint. The above differences resulted in the improved performance of GMRF-MM with the average constraint, which will be elaborated on later. In addition, the parameter estimates ($\hat{\mu}_1$ and $\hat{\bar{\pi}}_{\cdot 1}$) for GMRF-MM with the logit constraint lie between those for GMRF-MM with the zero and average constraints respectively. Finally, the ICM-based and Bayesian DMRF-MM parameter estimates were quite different, presumably because the former yielded joint MAP, while the latter gave marginal posterior means.

3.3.3 Predictive performance

The ROC curves were constructed for all the methods based on the positive and negative control sets except for ICM-based DMRF-MM, which only gave the most probable states, leading to one pair of sensitivity and specificity. As shown in Figure 3.1(a), at a very high specificity (e.g., above 0.95), all ROC curves were close to each other, resulting in similar performance. When the specificity ranged from 0.9 to 0.4, all MRF-based methods gave higher sensitivities than that of SMM, while Bayesian DMRF-MM and GMRF-MM with either the logit or average constraint had higher sensitivities than GMRF-MM with the

Table 3.2: Parameter estimates for the GCN4 ChIP-chip data (μ_0 is fixed at 0)

Models	π_0	σ_0	π_1	μ_1	σ_1	π_2	μ_2	σ_2
SMM	.91	.80	.06	1.67	1.94	.03	-1.98	.40
GMRF (zero)	.50	.63	.18	.75	1.53	.32	-.38	1.02
GMRF (logit)	.41	.58	.10	.98	1.03	.49	-.16	1.80
GMRF (average)	.66	.71	.04	2.26	1.84	.30	-.22	1.18
DMRF (Bayesian)	.91	.80	.05	1.83	1.90	.04	-2.00	.42
DMRF (ICM)	-	.89	-	4.26	1.02	-	-2.35	.05

Table 3.3: Parameter estimates for the Markov random fields

Models	Parameters in MRF
GMRF (zero)	$\sigma_{C0} = 76.70$,$\sigma_{C1} = 4.08$,$\sigma_{C2} = 10.54$
GMRF (logit)	$\sigma_{C0} = 0.89$,$\sigma_{C1} = 1.59$,$\sigma_{C2} = 84.52$
GMRF (average)	$\sigma_{C0} = 64.55$,$\sigma_{C1} = 0.15$,$\sigma_{C2} = 81.65$
DMRF (Bayesian)	$\beta = 1.54$, $\gamma_0 = 1.96$, $\gamma_1 = .30$
DMRF (ICM)	$\beta = 2.04$, $\gamma_0 = 2.39$, $\gamma_1 = -.02$

zero constraint. At a low specificity (e.g., below 0.2), both Bayesian DMRF-MM and GMRF-MM with the zero constraint deteriorated as compared to SMM, while GMRF-MM with either the logit or average constraint remained superior to SMM. In addition, ICM-based three-component DMRF-MM gave 25 true positives compared with 29 true positives given by a two-component ICM-based DMRF-MM, while the 0.001 cut-off for p-values as used in Lee *et al.* (2002) gave 23 true positives. All methods resulted in 2 false positives, leading to a sensitivity of 0.379, 0.439, and 0.348, respectively, and the same specificity of 0.997. From Figure 3.1(a), we can see that by applying a lower cut-off, the sensitivity for Bayesian DMRF-MM can be elevated from around 0.40 to as high as 0.50 at the same specificity, suggesting more flexibility of the Bayesian approach than the ICM.

In summary, at a high specificity (e.g., above 0.5 as usually desired), by taking use of biological knowledge embedded in a gene network, all MRF-based mixture models had higher statistical power to detect the targets than did SMM that ignored biological knowledge. In addition, GMRF-MM with either the logit or average constraint had significant improvement over GMRF-MM with the zero constraint. While the ROC curve of GMRF-MM with either the logit or average constraint did not dominate that of the Bayesian DMRF-MM, the former two had larger areas under the curve (AUC), suggesting potential superiority.

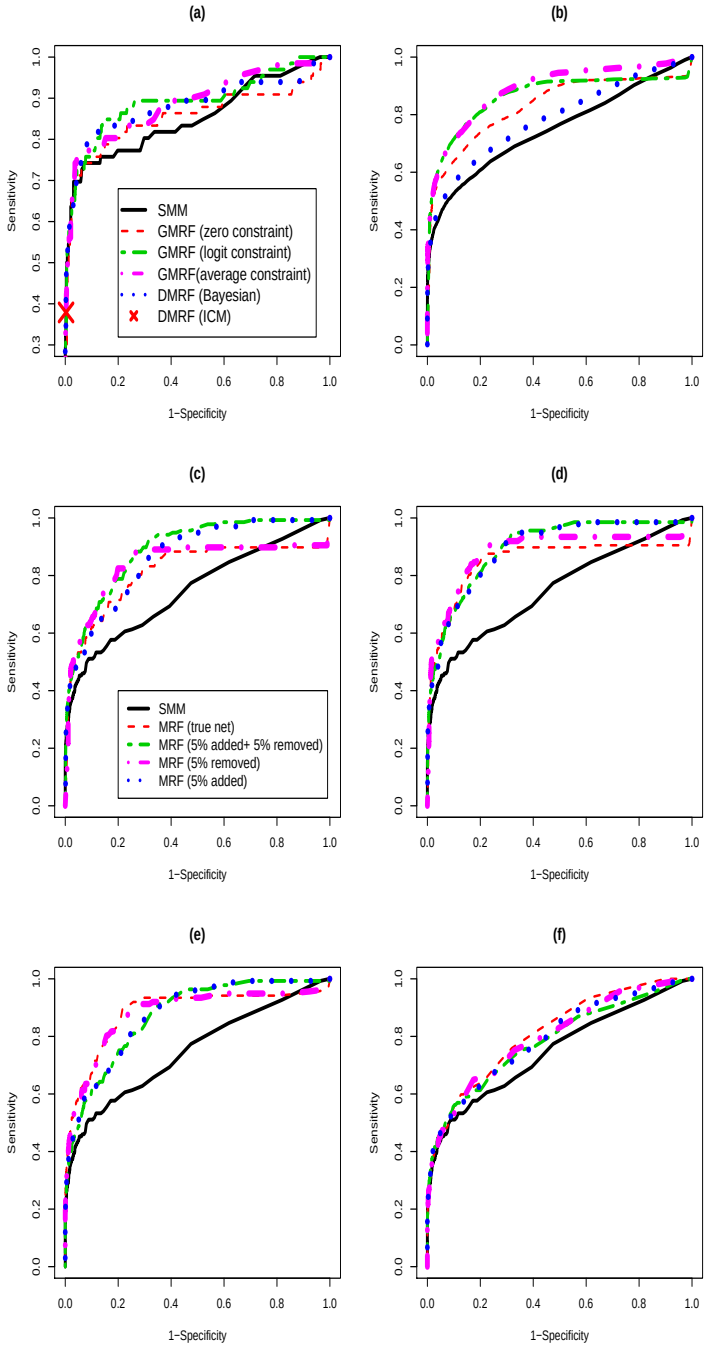

Figure 3.1: ROC curves for (a) GCN4 ChIP-chip data; (b) simulated data (averaged across 20 data sets); and perturbed networks (simulated data) for (c) GMRF-MM with the zero constraint, (d) GMRF-MM with the logit constraint, (e) GMRF-MM with the average constraint, and (f) Bayesian DMRF-MM.

3.3.4 Examples of genomic discoveries

Figure 3.2 shows the top 100 genes ranked by posterior probabilities by each of the four methods: SMM, Bayesian DMRF-MM, and GMRF-MM with either the logit or average constraint. Several features are noticeable. First, all methods achieved very high specificity (above 99%) and similar sensitivity (about 50%), corresponding to an indistinguishable part in the ROC plot (Figure 3.1(a)). Although decreasing the specificity to a slightly lower value, e.g., between 0.8 and 0.95, may help show the methods' differential performance as demonstated by the ROC curves, it becomes much harder to visualize hundreds of genes. Second, the genes selected by Bayesian DMRF-MM and GMRF-MM with the logit constraint were more connected with each other (42 and 49 edges, respectively) as compared to the SMM and GMRF-MM with the average constraint (34 and 31 edges respectively). This may suggest that the former two encouraged more spatial clustering, while GMRF-MM with the average constraint was more similar to SMM, possibly due to the shrinkage effect as induced by the adopted average prior constraint.

We examined a few individual genes in neither control set but predicted to be GCN4's targets (ranked among top 100) to gain more biological insights. First, ILV2 (YMR108W), ILV5 (YLR355C), and ILV6 (YCL009C) are connected with ILV2 as direct neighbor of the other two on the gene network. All of them are annotated in the Gene Ontology (GO) (Ashburner *et al.* 2000) Biological Process: branched chain family amino acid biosynthetic process (GO ID:0009082), which is a child term of amino acid biosynthetic process (GO ID:0008652). Because GCN4 is a transcriptional activator

Table 3.4: Ranks of selected un-annotated (in neither control set) genes.

Genes	SMM	DMRF	GMRF (logit)	GMRF(average)	Evidence of being GCN4's target
ILV2	114	79	65	75	Arndt *et al.* (1986)
ILV5	22	23	18	13	Beyer *et al.* (2006)
ILV6	70	57	47	54	Schuldiner *et al.* (1998)
TRP3	105	82	76	117	Martens *et al.* (1994)

of amino acid biosynthetic genes in response to amino acid starvation, it is expected that these three genes are likely to be binding targets of GCN4. In fact, they were confirmed by independent experiments (see Table 3.4). On the other hand, ILV5 and ILV6, with strong binding signals, were identified as GCN4's targets by all methods, but ILV2, with a relatively weak signal, was only identified by MRF-based methods but not SMM, suggesting the potential gains by incorporating gene network information. Second, TRP3 (YKL211C), surrounded by five positive control genes, is annotated in GO Biological Process: tryptophan biosynthetic process (GO ID: 0000162), also a child term of amino acid biosynthetic process (GO:0008652). It was confirmed as a binding target of GCN4 by Martens *et al.* (1994). Based on the ChIP-chip data, TRP3 was identified by Bayesian DMRF-MM and GMRF-MM with the logit constraint, but not by either SMM or GMRF-MM with the average constraint. It was ranked 105th and 117th by the latter two respectively, possibly due to the average constraint's shrinkage effect in the latter.

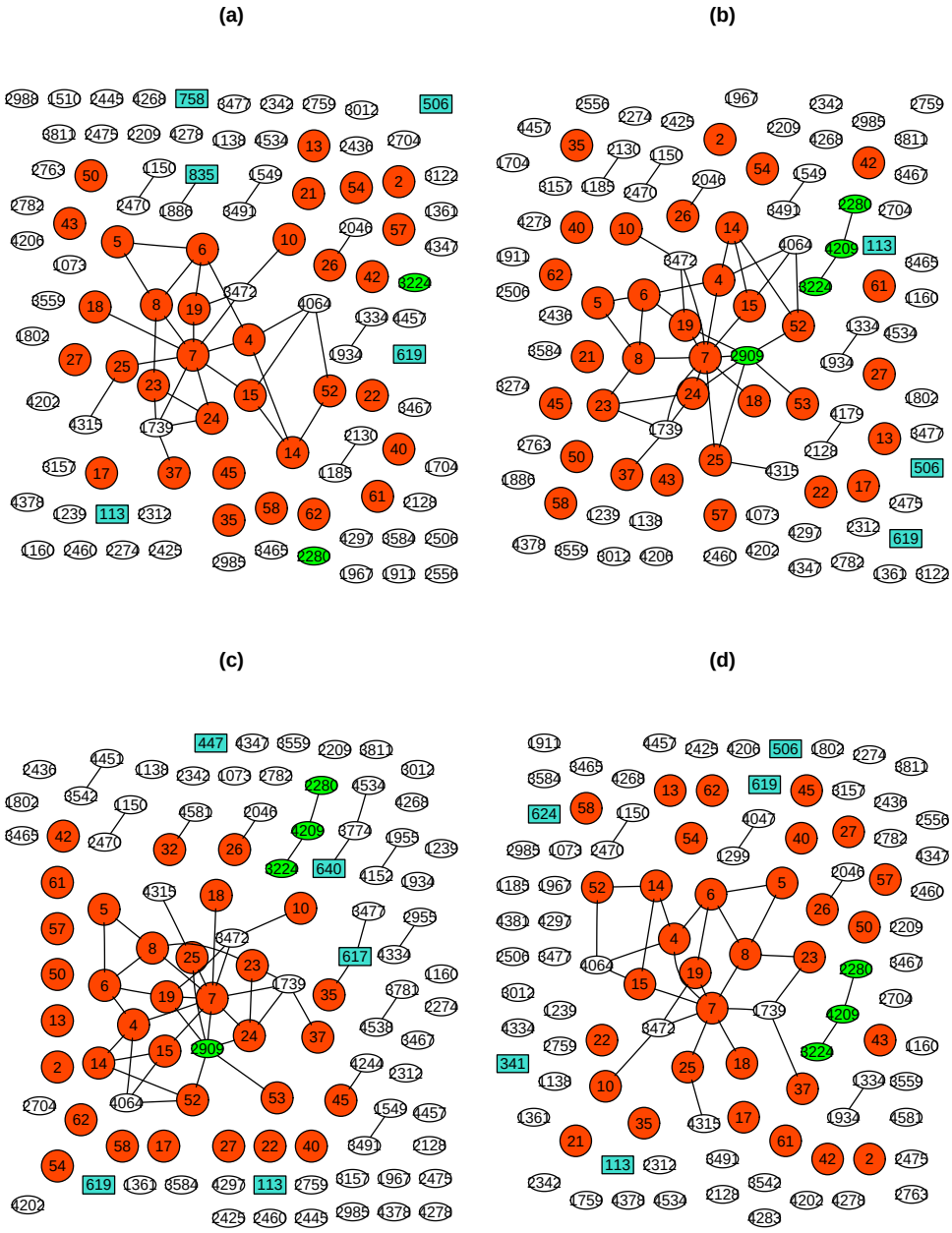

Figure 3.2: Sub-network of top 100 genes ranked by the posterior probabilities by each method (GCN4 ChIP-chip data): (a) SMM (33, 5); (b) Bayesian DMRF-MM (34, 3); (c) GMRF-MM with the logit constraint (33, 5); (d) GMRF-MM with the average constraint (32, 5). Numbers in the parentheses correspond to those of true positive and false positive genes, respectively. Positive control, negative control, and un-annotated (in neither control set) genes are represented by circle, rectangle, and ellipse, and numbered 1-66, 67-835, and 836-4609, respectively. Un-annotated genes discussed in Section 3.4 are represented by highlighted ellipse: 2280 (ILV6), 4209 (ILV2), 3224 (ILV5), and 2909 (TRP3).

3.3.5 FDR estimation

Figures 3.3(a)-(e) show realized FDR's (based on the control sets) versus estimated FDR's (based on (3.13)) for all the methods. Overall, the two curves matched well for SMM and Bayesian DMRF-MM except that when the number of claimed positives ranged from 25 to 50, the FDR was a bit under-estimated. In contrast, for GMRF-MM with the zero constraint, (3.13) systematically underestimated the FDR by 20% on average, while that with the logit constraint improved over it by about 10%, still under-estimating the FDR. Interestingly, GMRF-MM with the average constraint estimated the FDR quite well up to 50 claimed positives, and then over-estimated by around 5%. To sum up, GMRF-MM with either the logit or average constraint outperformed Bayesian DMRF-MM when we used ROC curves as criteria, while GMRF-MM with the logit constraint did not perform satisfactorily in terms of FDR estimation.

3.4 Simulation

3.4.1 Simulation set-up

To further compare the methods, particularly their robustness to mis-specified gene networks, we conducted a simulation study that mimicked real data: we used the same gene network as used for the real data, and used data-generating distributions similar to the ones fitted to the real data. We generated the true latent states based on a DMRF as specified by (3.7) and the yeast gene network. Specifically, to simulate T, the latent binding states, we initialized the 66 genes in the positive control set to be binding targets

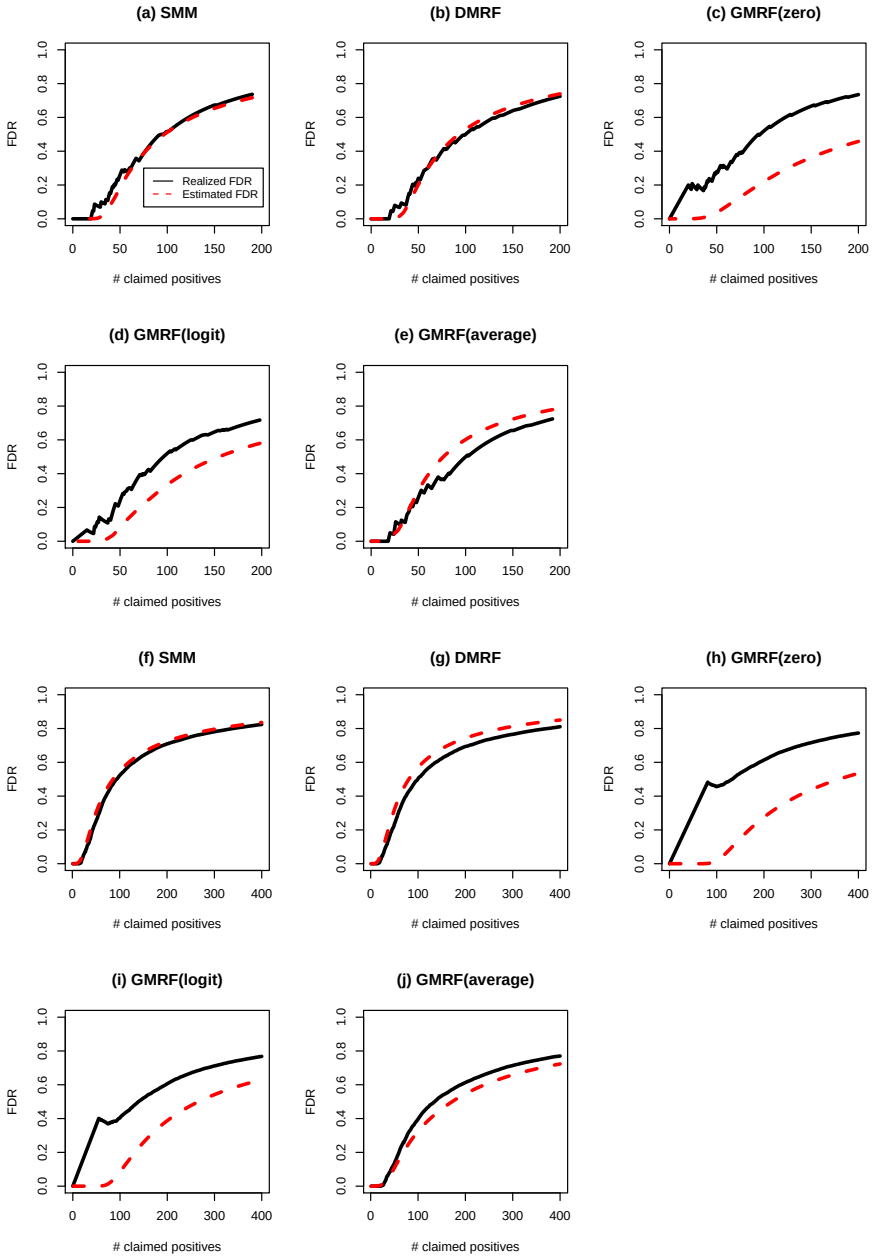

Figure 3.3: Estimated vs realized FDR's for (a)-(e) GCN4 ChIP-chip data, and (f)-(j) simulated data (averaged across 20 data sets).

and the rest of genes to be non-targets, giving an initial T. Then we iterated the states 20 times based on (3.7), with $\gamma = 0, \beta = 2$. It turned out that the number of binding targets became stable at about 130 after ten iterations, and we chose the states to be the ones right after the 10th iteration, giving 137 binding targets. Next, given T, we simulated 20 data sets with 4,609 z-scores according to the fitted GMRF-MM with the zero constraint from the real data; following Wei and Pan (2008a), for simplicity, we only used the null and positive components, i.e., $\phi(0, 0.63^2)$ and $\phi(0.75, 1.53^2)$.

In addition, to evaluate the impact of a mis-specified network, we perturbed the network used in the MRF-based methods for the simulated data. We perturbed the network in three ways. In scenario 1, we randomly removed 5% (1672) edges from the original 33,432-edge network, and it resulted in 46 singletons. We eliminated those singletons by randomly connecting each of them to another gene and ended up with a 31,806-edge network. In scenario 2, we randomly added 1672 edges to the original network, and thus had a 35,104-edge new network. Third, we removed the same set of 1672 edges as in scenario 1 from, then added the same set of 1672 edges as in scenario 2 to the original network; further more, we eliminated 20 singletons by randomly connecting each of them to another gene, ending up with a 33,452-edge network.

3.4.2 Simulation results

We applied the true network to each of the 20 simulated data sets, and constructed the ROC curves (averaged across the 20 simulated data sets) as shown in Figure 3.1(b). Based on the true network, GMRF-MM with any of the three constraints had higher

sensitivity than Bayesian DMRF-MM and SMM at a high specificity (e.g., above 0.5). Particularly the ROC curve for GMRF-MM with the average constraint dominated those for all other methods, suggesting its superiority.

In addition, we applied the three perturbed networks to a simulated data set. Figures 3.1(c)-(f) show each MRF-based mixture model's robustness to mis-specified networks. As we can see, the ROC curves, particularly at high specificities, were close to each other regardless of the gene networks used, indicating that all MRF-based methods considered here were reasonably robust to network mis-specifications.

Figures 3.3(f)-(j) show the comparison between the realized and the estimated FDR's (averaged across 20 simulated data sets). The estimated FDR was quite close to the realized one for SMM, Bayesian DMRF-MM, and GMRF-MM with the average constraint. The approximation performed much better for the GMRF-MM with the average constraint than GMRF-MM with either the zero or the logit constraint. The main reason for this is that the average constraint downweighted the high prior probabilities of being a target (some might be close to 1) for those true non-target genes that would otherwise be identified as targets by GMRF-MM with either the zero or logit constraint. It is consistent with Newton *et al.*'s comments that the performance of the FDR estimate based on (3.13) depends on the correctness of the fitted model. In other words, GMRF-MM with the average constraint fit better, leading to more accurate FDR estimation.

3.5 Discussion

TFs play a central role in the regulation of gene expression. Accurate identification of a given TF's regulatory target genes is thus a crucial step towards understanding gene regulation on a genome-wide scale and deciphering the principles of TF-gene regulatory networks. ChIP-chip technology provides a powerful tool for accomplishing such a task; however, challenges remain for statistical analysis due to high noise in high-throughput data and typically few replicates, resulting in a relatively high false positive or high false negative rate. In this paper, we have illustrated the extra power gained by incorporating gene network information into statistical analysis of a ChIP-chip data set for TF GCN4 in yeast *Saccharomyces cerevisiae*, a common model organism in molecular biology. As an important transcription regulator, GCN4 may directly and indirectly induce the expression of as many as 500 genes, more than 1/10 of the yeast genome (Hinnebusch and Natarajan 2002). Through integrating the functional gene network of Lee *et al.* and the ChIP-chip data, we were able to identify more biologically confirmed binding targets of GCN4 at no extra experimental cost, demonstrating the usefulness of the network-based methods.

In this article, we have formulated MRF-based mixture models and compared them to SMM. In particular, we have proposed two modifications to the identifiability constraint in a GMRF-MM to improve its parameter estimates and predictive performance, and a Bayesian approach to DMRF-MM. Application to the ChIP-chip real data together with a simulation study showed that in spite of different ways of incorporating gene networks,

all MRF-based mixture models had higher statistical power in detecting regulatory targets at a high specificity than did SMM treating all the genes i.i.d. *a priori*. In addition, the GMRF-MM with the average prior constraint was shown to be superior to both Bayesian DMRF-MM and GMRF-MM with either the zero or logit constraint. Estimating the FDR with (3.13) worked reasonably well for SMM, Bayesian DMRF-MM, and GMRF-MM with the average constraint, though the accuracy depended on parameter estimates and model fitting as expected; further study is needed to fully understand its performance under dependence. Finally, all network-based mixture models seemed to be reasonably robust to mis-specifications of gene networks, which is desirable in practice.

We note that for the two-component GMRF-MM described in Section 2.3, it seems that only one latent GMRF rather than two is needed because π_{i0} is determined by $(x_{i0} - x_{i1})$, and we may impose x_{i1}'s to be all 0. Similarly, for a K-component GMRF-MM, only $(K-1)$ latent GMRF's may be needed. In this way, we may reduce the number of parameters in the GMRF-MM by G. However, applying the above modification to the real data resulted in much worse performance than that of the original model (results not shown). Further study on this issue is needed.

In Bayesian modeling, improper posterior distributions may result from improper priors, while improper priors may still lead to proper posterior distributions (p.110, Gelman *et al.* 2004). In the DMRF-MM, we put improper priors on γ_0 and γ_1, whose marginal posterior distributions seemed to be proper based on the MCMC samples, suggesting that the joint posterior distribution probably exists. In the GMRF-MM, we used

improper GMRFs, which, however, become proper with the identifiability constraints imposed. In particular, Rodrigues and Assuncao (2008) showed that improper GMRF priors result in a proper posterior distribution even without the identifiability linear constraint in the context of Bayesian spatial modeling of normal response data, for which the joint posterior distribution is available in a closed form.

3.6 Appendix

3.6.1 Bayesian Model specifications for three-component standard and MRF-based mixture models

$$
\begin{aligned}
&(z_i|T_i = j, \theta_1) \sim N(\mu_j, 1/\tau_j),\\
&\mu_0 = 0, \;\; \mu_1 \sim N(0, 10^6)I(0 < \mu_1 < m), \;\; \mu_2 \sim N(0, 10^6)I(n < \mu_0 < 0),\\
&\tau_j \sim \text{Gamma}(0.1, 0.1),
\end{aligned}
$$

where $\theta_1 = (\mu_0, \mu_1, \mu_2, \tau_0, \tau_1, \tau_2)$, $\tau_j = 1/\sigma_j^2$, $m = \max_i z_i$ and $n = \min_i z_i$ for $i = 1, \ldots, G$ and $j = 0, 1, 2$. In addition, for SMM we have

$$(\pi_0, \pi_1, \pi_2) \sim \text{Dirichlet}(1, 1, 1)$$

with $\pi_j = Pr(T_i = j)$; for DMRF-MM we have

$$Pr(T_i = j|T_{(-i)}, \Phi) = \frac{\exp\{\gamma_j + \beta n_i(j)/m_i\}}{\sum_{k=0}^{2} \exp\{\gamma_k + \beta n_i(k)/m_i\}}$$

with $\Phi = (\gamma_0, \gamma_1, \gamma_2, \beta)$, $\gamma_2 = 0$, $\gamma_0 \propto 1$, $\gamma_1 \propto 1$, $\beta \propto I(0 \leq \beta < \beta_{max})$ and $\beta_{max} = 6$; for GMRF-MM we have

$$Pr(T_i = j) = \pi_{ij} = \frac{\exp(x_{ij})}{\exp(x_{i0}) + \exp(x_{i1}) + \exp(x_{i2})}, \;\; x_{ij}|x_{(-i)j} \sim N\left(\frac{1}{m_i}\sum_{l\in\partial i} x_{lj}, \frac{1}{\tau_{Cj} m_i}\right),$$

$$\tau_{Cj} \sim \text{Gamma}(0.01, 0.01),$$

where $\tau_{Cj} = 1/\sigma^2_{Cj}$, ∂i is the index set of the direct neighbors of gene i, $m_i = |\partial i|$, and $\frac{1}{G}\sum_{i=1}^{G} x_{ij} = 0$ for GMRF-MM with the zero constraint, while for GMRF-MM with the logit constraint, $\frac{1}{G}\sum_{i=1}^{G} x_{i0} = 0$, $\frac{1}{G}\sum_{i=1}^{G} x_{i2} = 0$, $\frac{1}{G}\sum_{i=1}^{G} x_{i1} = c$, and c is a negative number, e.g., $c = \text{logit}(0.05)$.

A two-component mixture model can be specified by dropping the normal component with the negative mean μ_2 from the above, as used in the simulation.

3.6.2 MCMC Algorithm

We denote by $(\alpha|\ldots)$ the full conditional of α, that is the distribution of α conditional on everything else in the model. For Bayesian DMRF-MM, the joint posterior distribution is

$$(\mathbf{T}, \theta_1, \Phi|\mathbf{z}) \propto p(\mathbf{z}|\mathbf{T}, \theta_1)p(\mathbf{T}|\Phi)p(\theta_1)p(\Phi)$$

- update μ_j $(j = 1, 2)$ by Gibbs sampling with proposal given by

  $$(\mu_j|\ldots) \sim N(\frac{\tau_j \sum_{\{i:T_i=j\}} z_i}{10^{-6} + n_j\tau_j}, \frac{1}{10^{-6} + n_j\tau_j})(\text{I}_{(0,m)}(\mu_1)\text{I}(j = 1) + \text{I}_{(n,0)}(\mu_2)\text{I}(j = 2)),$$

  where $n_j = |\{i : T_i = j\}|$.

- update τ_j $(j = 0, 1, 2)$ by Gibbs sampling with proposal given by

$$(\tau_j | \ldots) \sim \text{Gamma}(\tau_j | \frac{n_j}{2} + 0.1, \frac{\sum_{\{i:T_i=j\}} (z_i - \mu_j)^2}{2} + 0.1).$$

- update T_i by Gibbs sampling with proposal given by

$$(T_i | \ldots) \sim \text{Multinomial}(1; p_{i0}, p_{i1}, p_{i2}),$$

where

$$p_{ij} = \frac{\exp\{\gamma_j + \beta n_i(j)/m_i\} \phi(z_i; \mu_j, \sigma_j^2)}{\sum_{k=0}^{2} \exp\{\gamma_k + \beta n_i(k)/m_i\} \phi(z_i; \mu_k, \sigma_k^2)}.$$

- update $(\gamma_0, \gamma_1, \beta)$ using a random walk Metropolis algorithm with Gaussian proposal, which has diagonal covariance matrix. The acceptance ratio is calculated using the full conditional of $(\gamma_0, \gamma_1, \beta)$, which is proportional to

$$\frac{\exp\left\{n_0\gamma_0 + n_1\gamma_1 + \beta \sum_{j=0}^{2} \sum_{\{i:T_i=j\}} n_i(j)/m_i\right\}}{\prod_{i=1}^{G} \{\exp\{\gamma_0 + \beta n_i(0)/m_i\} + \exp\{\gamma_1 + \beta n_i(1)/m_i\} + \exp\{\beta n_i(2)/m_i\}\}}.$$

The Gaussian proposal was tuned such that the acceptance rate was around 0.23, the optimal one (Carlin and Louis 2000).

For GMRF-MM, the joint posterior distribution is

$$p(\mathbf{T}, \theta_1, \mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \tau_{C0}, \tau_{C1}\tau_{C2} | \mathbf{z}) \propto p(\mathbf{z}|\mathbf{T}, \theta_1) p(\mathbf{T}|\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2) p(\theta_1) \prod_{j=0}^{2} p(\mathbf{x}_j|\tau_{Cj}) p(\tau_{Cj})$$

- The full conditional distributions for $\mu_1, \mu_2, \tau_0, \tau_1$, and τ_2 are the same as those in the DMRF-MM.

- update T_i by Gibbs sampling with proposal given by

$$(T_i | \ldots) \sim \text{Multinomial}(1; p_{i0}, p_{i1}, p_{i2}),$$

where

$$p_{ij} = \frac{\exp(x_{ij})\phi(z_i; \mu_j, \sigma_j^2)}{\sum_{k=0}^{2} \exp(x_{ik})\phi(z_i; \mu_k, \sigma_k^2)}.$$

- update x_{ij} using Gibbs sampling with proposal given by

$$(x_{ij}|\ldots) \propto \frac{\exp\left\{\sum_{k=0}^{2} x_{ik}\mathrm{I}(T_i = k)\right\}}{\sum_{k=0}^{2} \exp(x_{ik})} \exp\left\{-\frac{m_i \tau_{Cj}}{2}(x_{ij} - \frac{1}{m_i}\sum_{l \in \partial i} x_{lj})^2\right\}$$

  The above full conditional is log-concave, hence slice sampling can be used to draw samples from it (Carlin and Louis 2000). The constraint for $\mathbf{x}_j$ is implemented by simply subtracting the current mean $\frac{1}{G}\sum_{i=1}^{G} x_{ij}^{(t)}$ from all of the $x_{ij}^{(t)}$ at the end of each iteration t.

- update τ_{Cj} using Gibbs sampling with proposal given by

$$(\tau_{Cj}|\ldots) \sim \mathrm{Gamma}(\tau_{Cj}|\frac{G-1}{2} + 0.01, \frac{\mathbf{x}_j^{'} Q \mathbf{x}_j}{2} + 0.01),$$

  where Q is $G \times G$ with non-diagonal entries $q_{kl} = -1$ if $k \sim l$ and 0 otherwise, and diagonal entries $q_{ii} = m_i$.

3.6.3 Bayes estimators: MAP and MMP

Given data $\mathbf{z} = (z_1, z_2, \ldots, z_G)$ and parameter $\mathbf{T} = (T_1, T_2, \ldots, T_G) \in \{0, 1\}^G$, we define a *loss function* $L(\mathbf{T}, \hat{\mathbf{T}}(\mathbf{z})) \geq 0$, where $\hat{\mathbf{T}}(\mathbf{z})$ is an estimator. The *Bayes risk* of the estimator $\hat{\mathbf{T}}$ under the loss function L is the mean loss

$$R(\hat{\mathbf{T}}) = E_{\mathbf{T},\mathbf{z}} L(\mathbf{T}, \hat{\mathbf{T}}(\mathbf{z})) = E_{\mathbf{z}} E_{\mathbf{T}|\mathbf{z}} L(\mathbf{T}, \hat{\mathbf{T}}(\mathbf{z})).$$

An estimator $\mathbf{T}^*$ is called a *Bayes estimator* if it minimizes the Bayes risk. Next, we introduce two loss functions: 0-1 loss and mis-classification rate. Define 0-1 loss

$L_1(\mathbf{T}, \hat{\mathbf{T}}(\mathbf{z})) = \mathrm{I}(\mathbf{T} \neq \hat{\mathbf{T}}(\mathbf{z}))$, where I(.) is an indicator function. Mis-classification rate loss is defined as $L_2(\mathbf{T}, \hat{\mathbf{T}}(\mathbf{z})) = \frac{1}{G}\sum_{i=1}^{G} \mathrm{I}(T_i \neq \hat{T}_i(\mathbf{z}))$.

Claim: Maximum *a posteriori* (MAP) and maximum marginal posterior (MMP) are the Bayes estimators corresponding to the 0-1 loss and the mis-classification rate loss, respectively.

Proof: The Bayes risk for L_1 is

$$\begin{aligned} R_1(\hat{\mathbf{T}}) &= E_{\mathbf{z}} E_{\mathbf{T}|\mathbf{z}} L_1(\mathbf{T}, \hat{\mathbf{T}}(\mathbf{z})) \\ &= E_{\mathbf{z}} E_{\mathbf{T}|\mathbf{z}} \mathrm{I}(\mathbf{T} \neq \hat{\mathbf{T}}(\mathbf{z})) \\ &= E_{\mathbf{z}}(1 - Pr(\mathbf{T} = \hat{\mathbf{T}}(\mathbf{z})|\mathbf{z})) \end{aligned}$$

For each $\mathbf{z}$, minimizing $R_1(\hat{\mathbf{T}})$ is equivalent to maximizing the posterior distribution $Pr(\mathbf{T}|\mathbf{z})$ in $\mathbf{T}$. Hence, MAP $\mathbf{T}^*(\mathbf{z}) = \arg\max_{t \in \{0,1\}^G} Pr(\mathbf{T} = t|\mathbf{z})$ is the Bayes estimator for the 0-1 loss function. Similarly, for the mis-classification rate loss, we have the Bayes risk

$$\begin{aligned} R_2(\hat{\mathbf{T}}) &= E_{\mathbf{z}} E_{\mathbf{T}|\mathbf{z}} L_2(\mathbf{T}, \hat{\mathbf{T}}(\mathbf{z})) \\ &= \frac{1}{G}\sum_{i=1}^{G} E_{\mathbf{z}}(1 - Pr(T_i = \hat{T}_i(\mathbf{z})|\mathbf{z})) \end{aligned}$$

For each $\mathbf{z}$, minimizing $R_2(\hat{\mathbf{T}})$ is equivalent to maximizing the marginal posterior distribution $Pr(T_i|\mathbf{z})$ in T_i for each i. It follows that MMP $\tilde{\mathbf{T}} = (\tilde{T}_1, \ldots, \tilde{T}_G)$ is the Bayes estimator for the mis-classification rate loss, where $\tilde{T}_i = \arg\max_{t_i \in \{0,1\}} Pr(T_i = t_i|\mathbf{z})$.

Chapter 4

Bayesian Joint Modeling of Multiple Gene Networks and Diverse Genomic Data to Identify Target Genes of a Transcription Factor

4.1 Introduction

In this chapter, we consider integrative modeling of multiple sources of genomic data and gene networks to accurately identify the regulatory target genes of a transcription factor (TF). TFs, a class of regulatory proteins, play a central role in controlling gene expression: a TF binds to one or more specific DNA subsequences in a gene's promotor region, called binding sites or motifs, and then works with other TFs to stimulate or inhibit the target gene's transcription. Thus, accurate identification of the genes that are regulated by a given TF is critical to elucidating gene regulation mechanisms and deciphering the principles of cell organization. In Chapters 2 and 3, we approached this task based on ChIP-chip data (also called DNA-protein binding data or genome-wide location analysis), which provide evidence about genome-wide physical binding sites of a specific TF in living cells. However, those DNA-TF interactions may not be functional in terms of regulating gene expression because other conditions such as binding of co-regulators and recruitment of RNA polymerase II complex are also needed to initiate the target gene's transcription. Two other types of genomic data provide complementary information about TF-gene regulation: microarray gene expression data comparing expression changes before and after knocking-out or mutating a TF-coding gene, and DNA sequence data which are aligned and scanned to find specific binding sites of a TF. Although extremely valuable, these two data sources provide only partial information: for expression data, genes that are directly or indirectly regulated by the TF will all show changes in expression levels, while DNA sequence data provide only potential

binding sites which may or may not eventually be bound by the TF. Because each data source measures different aspects of TF-gene regulation and high-throughput data are inherently associated with relatively high noise levels, using one type of data alone may result in high false positives or false negatives. In contrast, it is now widely recognized that an integrative analysis of multiple types of genomic data should be more efficient in identifying the target genes of a TF (see Wang *et al.* 2005; Xie 2006; Jasen *et al.* 2007; Pan, Wei and Khodursky 2008, and references therein). There are two main classes of joint modeling approaches in the literature: regression methods and mixture model methods. First, in a regression framework, one type of data (e.g. ChIP-chip binding data or DNA sequence data) is regressed on another type of data (e.g., gene expression data; Colon *et al.* 2003; Sun *et al.* 2006; Wei and Pan 2008b). In particular, Jasen *et al.* (2007) proposed a Bayesian regression model in a variable selection framework to combine all three sources of data to construct gene regulatory networks (i.e., a set of multiple TFs and their regulatory target genes). Note that regression-based methods require a large number of replicate expression arrays, which are not applicable to the *E. coli* data to be analyzed here. Second, in a mixture model framework, inference is based on the posterior probability of being a target given gene-specific measurements of different sources of data. Wang *et al.* (2005) proposed a parametric mixture model for both DNA sequence data and expression/binding data; Pan *et al.* (2008) extended the mixture model of Wang *et al.* to one that is able to integrate all three data sources to detect the targets of a TF. Conditional independence is commonly assumed in a mixture

joint model, i.e., different sources of data are independent given that a gene is or is not a target, which may or may not hold in practice. Here we propose to extend the parametric mixture model of Pan *et al.* to allow conditional dependence. In addition, by adopting a fully Bayesian approach, we are able to make inference about the conditional correlation structures for all three data sources based on Markov chain Monte Carlo (MCMC) samples.

In addition to relaxing the conditional independence assumption, another key contribution of our proposed method here is to allow incorporation of multiple gene networks into joint modeling of diverse types of genomic data to detect the targets of a TF. In Chapters 2 and 3, we proposed a Gaussian Markov random field (GMRF)-based mixture model to incorporate a gene network encoding gene-gene interactions into statistical analysis of ChIP-chip data to boost the power for detection of the target genes of a TF, and compared it with the Discrete Markov random field (DMRF)-based mixture model of Wei and Li (2007). The network-based methods are motivated by the biological fact that neighboring genes on a network, e.g., co-expression or functional coupling gene network, are more likely to be co-regulated by a TF than non-neighboring ones. As biological knowledge and experimental data accumulate rapidly, multiple gene networks become available. Interactions between two genes in different networks may have different biological implications. For example, for *E. coli* three gene networks will be used in our analysis of the motivating data example for TF LexA: (1) a co-expression network constructed based on a compendium of gene expression microarrays, where two

genes are direct neighbors if their expression levels are highly correlated across about 400 experimental conditions; (2) a functional coupling network induced by a Gene Ontology (GO; Ashburner *et al.* 2000) semantic similarity, where two genes are direct neighbors if their functional annotations are close enough in the GO, a database containing the most comprehensive existing knowledge about gene function; (3) a gene regulatory network, where two genes are connected if one gene's protein product is a TF and is known to regulate the other gene's expression according to RegulonDB (Salgado *et al.* 2006), a database containing all known TF-gene regulatory interactions in *E. coli*. Figure 4.1 shows subnetworks, one from each of the aforementioned networks, consisting of LexA's known and putative target genes as available from RegulonDB. As we can see, a gene may have different sets of direct neighbors according to different networks. This is in part because edges in different networks reflect different perspectives of gene-gene interactions, e.g., co-expression or co-function, and in part because of incomplete or simply wrong annotation shown by the network. Since the three gene networks contain partial yet complementary information about gene-gene interactions, integrating all of them with ChIP-chip binding, gene expression and DNA sequence data should boost the power for detecting the target genes of LexA. In this chapter, we propose two mixture models to address this problem based on the use of multiple GMRFs and DMRFs, respectively. Statistical inference is made in a fully Bayesian framework. The proposed methods can be easily extended to integrate more gene networks and more types of genomic data, and provide a general statistical framework for integrative analysis in genomic studies.

In this chapter, we apply our proposed methods to study the regulation by LexA in *E. coli*. LexA is an important transcriptional factor invloved in DNA repair and cell division: it is a repressor for genes involved in the "SOS" response whose transcription is induced in reponse to DNA damage due to ultraviolet (UV) or chemical exposures (Wade *et al.* 2005). Under normal growth conditions, LexA binds to the promoter regions of these "SOS" genes, repressing their transcription. When DNA becomes extensively damaged, the LexA repressor is cleaved and loses its function. As a result, the expression of "SOS" genes is induced, and DNA repair ability in the cells is enhanced. Recently, LexA was shown to be essential in the acquisition of bacterial mutations which lead to resistance to some antibiotic drugs (Cirz *et al.* 2005). Therefore, a thorough understanding of LexA regulation is not only crucial to the elucidation of DNA repair mechanism in *E. coli*, a common model microorganism, but also beneficial to antibiotic drug development.

The rest of this chapter is organized as follows. We first describe the LexA data including ChIP-chip binding, gene expression, DNA sequence data and three gene networks for *E. coli*. Next, we introduce two mixture models for integrating multiple sources of genomic data and gene networks based on the use of GMRFs and DMRFs, respectively. We discuss statistical inference for the proposed models in a fully Bayesian framework. Parameter estimates are based on MCMC samples. We apply the new methods to the LexA data to identify its regulatory target genes. We evaluate the proposed methods' predictive performance by comparing the results with the known and putative targets listed in RegulonDB (v5.8). We end with a discussion of some existing issues and possible

future work.

4.2 The Data

4.2.1 ChIP-chip binding, gene expression and DNA sequence data

The ChIP-chip binding data, gene expression data and DNA sequence data were obtained from Pan *et al.* (2008), who extracted and processed these three sources of data from Wade *et al.* (2005), Courcelle *et al.* (2001) and RegulonDB (v4.0), respectively.

The ChIP-chip data included two LexA samples (called LexA_1 and LexA_2 respectively) and two control samples (one Gal4 and one MelR (no Ab, no antibody) samples) hybridized on four Affymetrix Antisense Genome Arrays respectively. First, the arrays were background corrected with the MAS 5 algorithm, followed by quantile normalization. Second, four $\log_2$ intensity ratios (LIRs) were calculated, corresponding to the four combinations of any two arrays, for each probe: LexA_1/Gal4, LexA_1/no Ab, LexA_2/Gal4, LexA_2/no Ab; a large LIR indicated a locus containing enriched LexA, i.e, a binding site of LexA. Third, for each of the four array combinations, the LIRs were smoothed over all probes with a sliding window of 1250 bp along the chromosome. Finally, gene i's binding score B_i, a summary statistic measuring the relative abundance of the TF binding to the gene, was taken to be the average of its four LIR peaks from its coding region, or if there were probes from its intergenic region, B_i was the larger of i) the average of its four LIR peaks from its coding region and ii) that from its intergenic region.

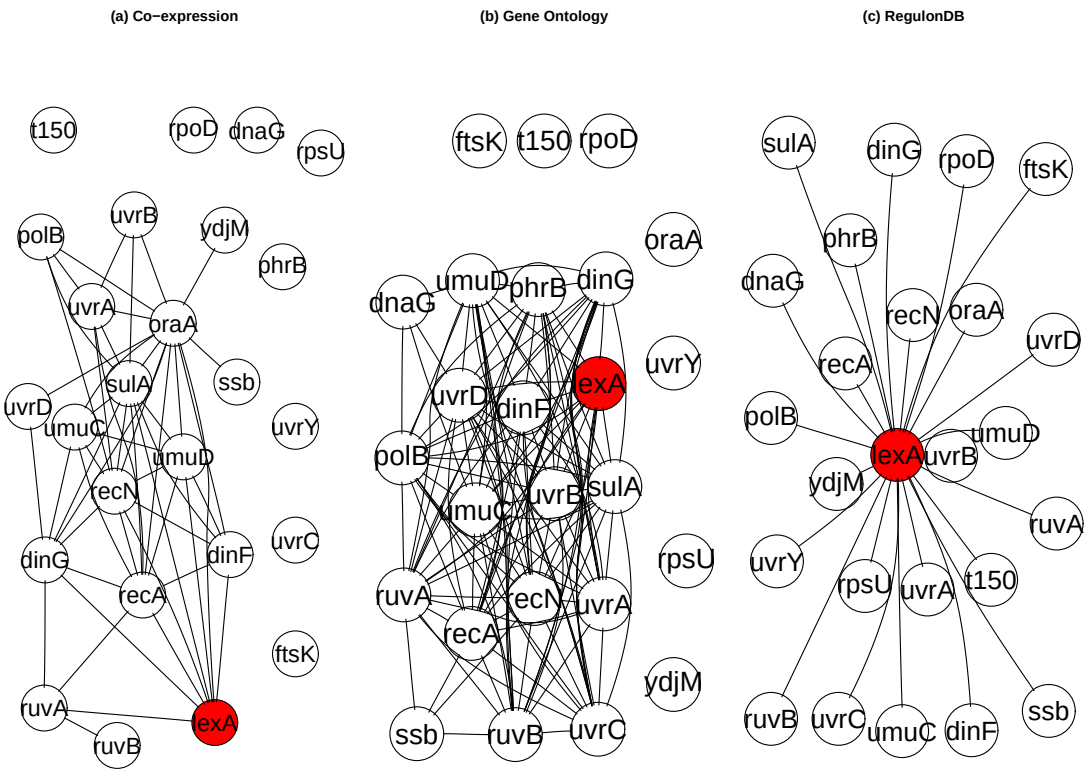

Figure 4.1: Subnetworks, one from each of the following three networks, consisting of LexA's known and putative target genes as available from RegulonDB. The three gene networks are: (a) co-expression network, (b) GO induced functional coupling network, and (c) RegulonDB gene regulatory network.

The expression data were drawn from four cDNA microarrays profiling gene expression levels for the wild type before and 20 minutes after UV treatment, and for the LexA mutant before and 20 minutes after UV treatment; a common control sample was used for each array. Two-channel intensities on each array were normalized using the loess local smoother to eliminate dye bias, as implemented in the R package `sma` (Yang *et al.* 2002). Suppose that normalized log-ratios of the two-channel intensities for gene i on the four arrays were M_{1i},...,M_{4i} respectively, then the summary statistic for gene expression data was taken as $E_i = (M_{2i} - M_{1i}) - (M_{4i} - M_{3i})$. Because LexA is known to be a repressor of some "SOS" response genes, it is expected that the regulatory targets of LexA should have larger values of E_i's (i.e., expression changes).

The DNA sequence data were obtained as following. Ten known binding sites of LexA were downloaded from RegulonDB (v4.0), involving nine genes each with one binding site and gene LexA with two binding sites. These ten binding sites were input into MEME (Bailey and Elkan 1995) to find a top consensus sequence (motif). scanACE (Roth *et al.* 1998) was then used to scan the whole genome with a very low threshold such that at least one subsequence matching the motif could be obtained for most genes; the maximum of all the matching scores for gene i was taken as S_i, the summary statistic for the sequence data.

After combining the three data sources and deleting genes with any missing values, we obtained $G = 3779$ genes in the combined data. Table 4.1 shows a small portion (5 of 3779 genes) of the resulting dataset.

Table 4.1: Some data from the LexA dataset.

Index	Binding (B_i)	Expression (E_i)	Sequence (S_i)
GENE1	-0.490	0.076	15.573
GENE2	2.275	2.777	23.968
GENE3	0.619	1.377	24.164
GENE4	0.210	-0.208	15.464
GENE5	0.120	-0.346	13.055

4.2.2 Gene networks for *E. coli*

Three gene networks were constructed for *E. coli* as mentioned before: co-expression network, functional coupling network, and gene regulatory network.

The co-expression gene network was derived from 380 microarray experiments across a variety of conditions, available at the Many Microbe Microarrays Database (M3D; Faith *et al.* 2008). Two genes are direct neighbors if the Pearson correlation coefficient of their expression profiles across the 380 experiments is greater than 0.65, resulting in a network with 3,208 nodes (genes) and 86,791 edges (interactions). The cutoff 0.65 is chosen so that the resulting network is neither too dense, including many false postive interactions, nor too sparse, failing to include many true positive interactions. As a comparison, a cutoff of 0.6 would lead to 147,563 interactions, while a cutoff of 0.7 would result in 46,666 interations.

The functional coupling gene network was induced from the Gene Ontology (GO), a

compendium of existing knowledge, derived from various sources, about gene function. GO is structured as a directed acyclic graph (DAG): each node corresponds to a GO category; a parent node represents a more general biological function whereas its child node is a subclass or a part of it; any gene in a child node is necessarily in its parent node. For example, GO category GO:0033554 with annotation "cellular response to stress" has a child node GO:0009432 with a more specific annotation "SOS response". The GO similarity between two genes in an ontology is defined as the maximum number of common nodes in all paths back to the root node of the ontology ("biological process") from all nodes to which those genes are assigned (see Wu *et al.* 2005 for more details). If the GO similarity between two genes is large, then at a very specific level the two genes are involved in at least one common biological process. Figure 4.2 illustrates a DAG induced from the GO. We computed the GO similarity for each pair of genes using MatrixMaker of Dvorkin (2007). Two genes are direct neighbors on the induced functional coupling network if their GO similarity is no less than 5, which means there are at least 5 common nodes in their shared longest path back to the root node "biological process" from all nodes in which they are annotated. Figure 4.2 shows an example of1 how to calculate the GO similarity between two genes. The induced network has 1,644 nodes and 116,422 edges.

In the gene regulatory network, two genes are direct neighbors if one gene's protein product is a TF and is known to regulate the other gene's expression according to RegulonDB (v5.8). The final network consists of 1,138 nodes and 2,399 edges.

Table 4.2: Summary statistics of the three gene networks used in the analysis.

Network	# of nodes	# of edges	percentiles of # of direct neighbors				
			0%	25%	50%	75%	100%
co-expression	3,208	86,791	1	5	20	64	424
functional coupling (GO)	1,644	116,422	1	48	102	249	708
gene regulatory (RegulonDB)	1,138	2,399	1	1	2	3	359

Some summary statistics and sample subnetworks of the three gene networks can be found in Table 4.2 and Figure 4.1, respectively. The networks differ substantially in the density of edges due to different definitions of gene-gene interactions.

4.3 Statistical Methods

4.3.1 Notation

Our goal is to identify regulatory target genes of a given TF based on ChIP-chip binding, gene expression and DNA sequence data. We assume that the three data sources have been summarized as (B_i, E_i, S_i) for each gene i, for $i = 1, \ldots, G$, as described in Section 4.2.1. Depending on the state of gene i, i.e., whether it is a target or not, we have $T_i = 1$ or $T_i = 0$ respectively. Denote the distribution functions of (B_i, E_i, S_i) when $T_i = 1$ and $T_i = 0$ as f_1 and f_0, respectively.

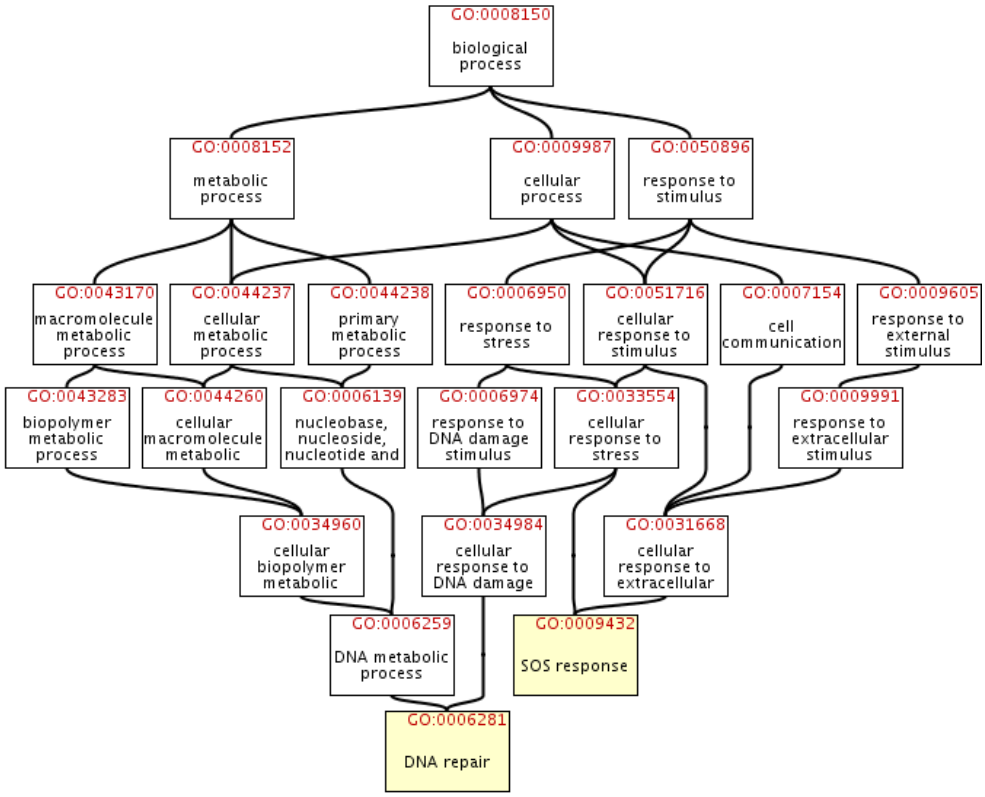

Figure 4.2: The combined directed acyclic graph (DAG) of DAGs induced from the GO terms "DNA repair" (GO:0006281) and "SOS response" (GO:0009432). lexA and dinG, two known target genes of TF LexA, are annotated in both terms. Because there are 6 and 5 nodes in the longest paths from "DNA repair" and "SOS response" to the root node "biological process", respectively (the root node itself is not counted), the GO similarity between lexA and dinG is 6. The graph was adapted from QuickGO GO Browser (http://www.ebi.ac.uk/QuickGO/).

4.3.2 Standard Mixture Joint Model

We first consider joint modeling of binding, expression and sequence data without incorporating gene networks. We have the following standard mixture joint model (SMJM):

$$f(B_i, E_i, S_i) = (1 - \pi_1) f_0(B_i, E_i, S_i) + \pi_1 f_1(B_i, E_i, S_i),$$

where $\pi_1 = Pr(T_i = 1)$ is the prior probability of gene i being a target, also called "mixing proportion". Note that it is the same for all the genes. We further specify the conditional distribution $f_j = \phi(.; \mu_j, \Sigma_j)$, a multivariate normal density function with mean vector μ_j and covariance matrix Σ_j for $j = 0, 1$. Here we allow the conditional covariance matrix Σ_j to have a general structure, i.e., the three data sources can be correlated given T_i. A special case is diagonal covariance matrix $\Sigma_j = Diag(\sigma_B^2, \sigma_E^2, \sigma_S^2)$, i.e., the three data sources are conditionally independent, as assumed in Pan *et al.*(2008).

4.3.3 GMRF-based Mixture Joint Model

Because neighboring genes on a network, e.g., co-expression or functional coupling network, tend to be co-regulated by a TF and there is more than one gene network available, each containing complementary yet partial information about gene-gene interactions, it is desired to incorporate multiple gene networks into joint modeling of genomic data. Here, we propose a GMRF-based mixture joint model (GMRF-MJM) to accomplish this goal. Specifically, we introduce gene-specific prior probabilities $\pi_{ij} = Pr(T_i = j)$, for $i = 1, \ldots, G$ and $j = 0, 1$. While allowing gene specific π_{i1}'s, we want π_{i1} for genes in the same state (target or non-target) to be more similar, i.e., the true targets have bigger

π_{i1}'s, while the true non-targets have smaller ones. To realize this, for each gene network $\mathcal{G}_k$ for $k = 1, \ldots, K$, we introduce two latent GMRF's, each corresponding to a state: $\mathbf{x}_j^{(k)} = \left\{ x_{ij}^{(k)}; i = 1, \ldots, G \right\}$ for $j = 0, 1$. We relate π_{ij}s and $\mathbf{x}_j^{(k)}$s via the following logit transformation:

$$\text{logit}Pr(T_i = 1 | \mathbf{x}_j^{(k)}\text{s}) = \text{logit}(\pi_{i1}) = c + \sum_{k=1}^{K} (x_{i1}^{(k)} - x_{i0}^{(k)}). \tag{4.1}$$

Then back on the π_{i1} scale, we have

$$\pi_{i1} = \frac{\exp(c + \sum_{k=1}^{K} x_{i1}^{(k)})}{\exp(\sum_{k=1}^{K} x_{i0}^{(k)}) + \exp(c + \sum_{k=1}^{K} x_{i1}^{(k)})}, \tag{4.2}$$

where $\mathbf{x}_j^{(k)}$ is distributed according to an intrinsic Gaussian conditional autoregression model (ICAR), which has the "local dependency" property (Besag and Kooperberg 1995). Specifically, we have

$$x_{ij}^{(k)} | x_{(-i)j}^{(k)} \sim N \left(\frac{1}{m_i^{(k)}} \sum_{l \in \partial i^{(k)}} x_{lj}, \frac{\sigma_{Cj(k)}^2}{m_i^{(k)}} \right), \tag{4.3}$$

where $\partial i^{(k)}$ is the set of indices for gene i's direct neighbors on network $\mathcal{G}_k$, and $m_i^{(k)}$ is the corresponding number of neighbors. The parameter $\sigma_{Cj(k)}^2$ controls spatial smoothness of the random field: smaller $\sigma_{Cj(k)}^2$ induces more similar $x_{ij}^{(k)}$s. Note that we assume the contribution of each network to $\text{logit}(\pi_{i1})$ is additive, as in (4.1). The advantage of our proposed model is to combine all available gene network information, and thus to boost the statistical power for detecting target genes as much as possible. For example, as shown is Figure 4.1, phrB is a true target that is not connected to any other target genes on the co-expression network, but is connected to other targets on the GO and

RegulonDB induced networks. As a result, in contrast to using the co-expression network alone, phrB's prior probability of being a target can still be boosted by using the proposed model here to combine all three networks.

To allow identifiability and for improved performance, we impose linear constraints $\frac{1}{G}\sum_{i=1}^{G} x_{ij}^{(k)} = 0$ for $j = 0, 1$, and let the intercept $c = \text{logit}(\hat{\pi}_1)$, where $\hat{\pi}_1$ is the estimated prior probability of being a target from the SMJM. The rationale is to shrink the estimate of π_{i1} towards $\hat{\pi}_1$, which we found is a good estimate of the overall proportion of target genes (see Chapter 3 for more details). For further improved performance, we let π_{i1} take the following shrinkage estimator form:

$$\pi_{i1} = \lambda \frac{\exp\left\{\text{logit}(\hat{\pi}_1) + \sum_{k=1}^{K} x_{i1}^{(k)}\right\}}{\exp\left\{\sum_{k=1}^{K} x_{i0}^{(k)}\right\} + \exp\left\{\text{logit}(\hat{\pi}_1) + \sum_{k=1}^{K} x_{i1}^{(k)}\right\}} + (1-\lambda)\hat{\pi}_1, \qquad (4.4)$$

where $0 \leq \lambda \leq 1$. For simplicity, we use $\lambda = 1/2$ in our data analysis (see Section 3.2.3.1 for more details).

Singleton genes, i.e., those without any neighbors in a network, are allowed in the proposed GMRF-MJM here. Denote $\mathcal{S}_k$ as the set of indices for singletons in gene network $\mathcal{G}_k$. We set $x_{ij}^{(k)} = 0$ for $i \in \mathcal{S}_k$ and $j = 0, 1$. If $i \in \cap_{k=1}^{K} \mathcal{S}_k$, then π_{i1} is simply $\hat{\pi}_1$.

Finally, the conditional distribution of the observed data (B_i, E_i, S_i) given T_i is the same as that in the SMJM.

4.3.4 DMRF-based Mixture Joint Model

Here we propose a DMRF-based Mixture Joint Model (DMRF-MJM) to integrate multiple gene networks. In the DMRF-MJM, we model the state vector $\mathbf{T} = (T_1, \ldots, T_G)'$ as a DMRF directly. Specifically, we propose the following auto-logistic model for the conditional distribution of T_i,

$$\begin{aligned} \text{logit}Pr(T_i = 1 | T_{(-i)}, \Phi) &= \text{logit}Pr(T_i = 1 | T_{(\cup_{k=1}^{K} \partial i^{(k)})}, \Phi) \\ &= \gamma + \sum_{k=1}^{K} \beta_k \left[n_i^{(k)}(1) - n_i^{(k)}(0)) \right] / m_i^{(k)} \end{aligned}$$

where $\Phi = (\gamma, \beta_1, \ldots, \beta_K)$, $n_i^{(k)}(j)$ is the number of gene i's neighbors having state j on network $\mathcal{G}_k$ for $j = 0, 1$, and $m_i^{(k)} = n_i^{(k)}(0) + n_i^{(k)}(1)$. The conditional probability of gene i being a target depends on the states of its neighbors, as defined on the K networks. In addition, we assume additive effects of the K gene networks, weighted by the non-negative parameters β_k's. As a result, β_k can be used to measure how informative network $\mathcal{G}_k$ is. For singleton gene $i \in \mathcal{S}_k$, we set $\left[n_i^{(k)}(1) - n_i^{(k)}(0)) \right] / m_i^{(k)} = 0$.

Due to the unknown normalizing constant $C(\Phi)$ in the joint distribution of $\mathbf{T} = (T_1, \ldots, T_G)'$, the likelihood $l(\mathbf{T}; \Phi)$ does not have a closed form. Instead, we propose to use the *pseudolikelihood* of Besag (1996):

$$pl(\mathbf{T}; \Phi) = \prod_{i=1}^{G} p(T_i | T_{(\cup_{k=1}^{K} \partial i^{(k)})}, \Phi) = \prod_{i=1}^{G} \frac{\exp\left\{ T_i (\gamma + \sum_{k=1}^{K} \beta_k \left[n_i^{(k)}(1) - n_i^{(k)}(0)) \right] / m_i^{(k)}) \right\}}{1 + \exp\left\{ \gamma + \sum_{k=1}^{K} \beta_k \left[n_i^{(k)}(1) - n_i^{(k)}(0)) \right] / m_i^{(k)} \right\}} \tag{4.5}$$

The conditional distribution of the observed data given T_i is the same as that in the SMJM. Note that our proposed DMRF defined on multiple neighborhoods is similar to

that used by Deng *et al.* (2004) in the context of protein function prediction rather than detection of the target genes of a TF here.

4.3.5 Prior distributions

We use vague or non-informative prior distributions. We denote by $MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and denote by $W((\rho R)^{-1}, \rho)$ the *Wishart* distribution with mean vector R^{-1}. Reparameterize the component-wise mean vector as: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \boldsymbol{\theta}$. We use the following priors for the parameters in the conditional distribution of the observed data: $\boldsymbol{\mu}_0 \sim MVN(\mathbf{0}, \boldsymbol{C})$, $\boldsymbol{\theta} \sim MVN(\mathbf{0}, \boldsymbol{C})\mathrm{I}(\boldsymbol{\theta} > \mathbf{0})$, where $\boldsymbol{C} = diag(10^{-6}, 10^{-6}, 10^{-6})$; $\Sigma_j^{-1} \sim W((3R)^{-1}, 3)$ for $j = 0, 1$, where R is taken as the estimated marginal covariance matrix of the three data sources whose off-diagonal elements are close to zero. This is considered as a very vague prior with respect to the correlation parameters (Carlin and Louis 2000). For the SMJM, we have $\pi_1 \sim Beta(1, 1)$.

For the GMRF-MJM, we have $\sigma^2_{Cj(k)} \sim$ Inverse Gamma$(0.01, 0.01)$, $j = 0, 1; k = 1, \ldots, K$. For the DMRF-MJM, we have $\gamma \propto 1$ and $\beta_k \propto \mathrm{I}(0 \leq \beta_k < 6)$, $k = 1, \ldots, K$.

4.3.6 Statistical inference

We do statistical inference in a fully Bayesian framework via MCMC sampling. MCMC algorithms for the SMJM and GMRF-MJM can be implemented in WinBUGS V1.40 (Spiegelhalter *et al.* 2003), while we wrote an R program to implement the MCMC algorithm for the DMRF-MJM. The WinBUGS code for the GMRF-MJM and the MCMC

algorithm for the DMRF-MJM can be found in the Appendix.

The posterior mean of any parameter based on 10,000 MCMC samples after 10,000 burn-ins is used as its point estimate. In particular, we rank genes based on the posterior probability of being a target $\hat{p_i} = \widehat{Pr}(T_i = 1|Data)$. False Discovery Rate (FDR) can be estimated based on $\hat{p_i}$ as discussed in Chapter 3, which is not pursued in this study.

4.4 Application to LexA data

4.4.1 Conditional independence assumption

We applied the SMJM to the ChIP-chip binding, gene expression and DNA sequence data. Table 4.3 shows the point and interval estimates for the parameters in the conditional correlation matrices of the three data sources. As we can see, for the non-target component, the three sources of data appear to be independent with each other. Interestingly, for the target component, binding and sequence data are highly correlated, in constrast to the other two pairs: binding and expression data, sequence and expression data, which turn out to be only slightly correlated and independent, respectively. This is consistent with the recent finding that LexA's binding affinity to its regulatory targets depends on the extent to which the binding site matches the canonical motif of LexA (Michel 2005). In addition, our results suggest that LexA is quite efficient in repressing its target genes' expression: weak binding only decreases its repression effect slightly.

4.4.2 Predictive performance

We evaluate the different methods' predictive performance by comparing the ranks given by each method for 25 LexA's known and putative targets annotated in RegulonDB (v5.8). Note that the gene regulatory network was also constructed from RegulonDB. As shown in Figure 4.1(c), LexA is connected with all other 24 known and putative targets, which is an ideal situation for the proposed network-based methods here. Therefore, analysis involving the RegulonDB gene regulatory network is somewhat like supervised-learning, and the corresponding predictive performance is like "training error" in machine learning. In spite of its potential bias or over-optimism, it is still informative as it gives us an idea how good the method could be under an ideal situation. On the other hand, to alleviate the problem of using training error, we also perform analysis in which the RegulonDB regulatory network is excluded and compare the results.

Table 4.4 shows the results. In general, incorporating gene networks into joint modeling of multiple sources of genomic data increased the chance of detecting the true targets as compared to using genomic data alone (binding, expression and sequence data only); this is evidenced by higher, in some cases substantially higher ranks based on network-based analyses than those based on using genomic data alone. For example, dinG was ranked 167th by the SMJM, but because of its connection to other highly ranked target genes on all three networks, its rank was boosted to one by all GMRF-MJM analyses that incorporated multiple gene networks. In addition, several features are noticeable. First, using a general conditional covariance structure in the SMJM did not lead to improved

rankings as compared to using diagonal conditional covariance structure. As a result, we used diagonal conditional covariance structure in all MRF-based analyses for better predictive performance. Second, when only incorporating one gene network, unsurprisingly, we see that the RegulonDB network resulted in much more ranking improvement over the SMJM than did the other two networks. Third, when integrating two or three gene networks, we observe that the predictive performance tends to be compromised. For example, the ranks based on both RegulonDB network and co-expression network were higher than those based on the co-expression network alone, but lower than those based on the RegulonDB network alone. Fourth, given the same networks, the ranks by the GMRF-MJM were higher than those by the DMRF-MJM. This is in agreement with the conclusion of our comparative study in Chapter 3. Finally, as shown in Table 4.5, the relative magnitude of the weights β's for the three gene networks in the DMRF-MJM are quite consistent: the co-expression network had the largest weight while the RegulonDB network second and the functional coupling network third. Whether β can be used to measure how "good" a gene network is needs further research.

4.5 Discussion

We have presented two mixture models, based on the use of GMRFs and DMRFs respectively, for integrating diverse types of genomic data and multiple gene networks to identify regulatory target genes of a TF. Rather than assume conditional independence of ChIP-chip binding, gene expression and DNA sequence data, we allow multiple sources

of data to be conditionally correlated. Due to a fully Bayesian approach, inference about model parameters can be easily carried out based on MCMC samples. Application to the LexA data demonstrates utility and statistical efficiency gains with the proposed joint models. An interesting biological finding is that the binding and sequence data are highly correlated for target genes only, which helps elucidate the regulation mechanism of LexA, this important TF involved in DNA repair in *E. coli*. Interestingly, ignoring the correlation even led to improved predictive performance. Further study on this problem is needed.

Although our application concerns identification of target genes of a TF in *E. coli*, it may be possible to adapt the proposed methods to address other problems for other organisms, for example, identifying complex disease genes by integrating multiple types of data such as SNP, gene expression, proteomic, metabolomic data and gene networks/pathways.

Based on the LexA data, we found that combining two or more types of gene networks may result in compromised predictive performance. This raises a question: shall we integrate as many gene networks as possible or choose to use the "best" gene network. If the latter, how to compare gene networks is still an open question. A possible perspective is to look at the structrual or topological differences between the networks. For example, as shown in Figure 4.1, the co-expression and the GO networks contain clusters of densely linked nodes, while the RegulonDB regulatory network contains spike-like clusters with TF-coding genes as the hubs. The latter may not be a good choice because of its lack

Table 4.3: Posterior estimates for component-wise (conditional) correlation matrices of binding (B), expression(E), and sequence(S) data. Numbers in the parentheses are 95% credible intervals.

non-target component				target component			
	B	E	S		B	E	S
B	1	0.012 (-0.028, 0.047)	-0.014 (-0.054,0.022)	B	1	**0.119 (0.033,0.184)**	**0.476 (0.428,0.514)**
E		1	0.010 (-0.029,0.045)	E		1	0.076 (-0.016, 0.147)
S			1	S			1

of robustness, i.e., misspecification of the hub genes may have a big impact on a large number of genes. On the other hand, the weight parameter β_k in the DMRF-MJM may be a candidate criterion for quantitative comparison. However, for the LexA data, the network rankings based on β_k's are not in agreement with those based on the predictive performance. Hence, model comparison criteria in the context of integrative and network modeling are needed. This could be a direction of future work.

Finally, a potential improvement over the current proposed GMRF-MJM is to assign weights to different gene networks as in the proposed DMRF-MJM. This would be an interesting topic for future investigation.

Table 4.4: Ranks given by various methods for known and putative target genes of LexA annotated in RegulonDB. "S":SMJM with diagonal covariance;"S.mul":SMJM with general covariance;"G":GMRF-MJM;"D":DMRF-MJM.

targets	Binding	Binding+Expression+Sequence															
				RegulonDB		GO		co-exp		GO.RegulonDB		co-exp.GO		co-exp.RegulonDB		co-exp.RegulonDB.GO	
	S	S	S.mul	G	D	G	D	G	D	G	D	G	D	G	D	G	D
uvrB	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
sulA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
umuD	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ydjM	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
recN	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
recA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
lexA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
uvrA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ssb	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ftsK	76	171	171	146	152	175	166	196	180	1	154	194	175	163	164	213	159
oraA	83	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
dinG	95	167	165	146	142	158	163	1	145	1	154	1	148	1	141	1	147
ruvA	127	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
polB	156	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
umuC	192	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
uvrD	260	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
rpsU	468	904	979	445	601	891	1091	1123	1056	472	738	1012	1221	603	683	903	782
phrB	1060	1455	1683	623	884	1318	1837	1405	1920	674	1089	1290	1955	842	1147	1068	1297
t150	1937	175	215	146	164	175	166	166	173	151	154	179	175	163	170	1	173
dinF	2217	1	148	1	1	1	1	1	1	1	1	1	1	1	1	1	1
dnaG	2553	2599	2782	862	1309	2111	2983	1990	3156	957	1936	2032	3359	1337	1783	1790	1896
rpoD	2553	2442	3360	841	1376	2610	2475	2759	2883	1028	1264	3058	2697	1315	1716	2601	1744
ruvB	3108	647	589	364	451	514	627	465	369	372	450	451	379	430	325	419	330
uvrC	3108	3069	3162	1027	1736	2111	2983	3682	1843	1165	1693	3058	2155	1706	1215	1830	1297
uvrY	3725	3322	2466	1039	1656	3551	2611	3352	3156	1206	1636	3546	2518	1642	1857	3514	1809

Table 4.5: Posterior means of parameters in the DMRF-MJM.

Networks	γ	$\beta_{co-expression}$	β_{GO}	$\beta_{RegulonDB}$
co-expression	-1.34	1.15	-	-
GO	-1.72	-	0.84	-
RegulonDB	-1.87	-	-	0.85
co-expression + GO	-1.2	1.06	0.61	-
co-expression + RegulonDB	-1.27	1.07	-	0.73
GO + RegulonDB	-1.67	-	0.69	0.73
co-expression + GO + RegulonDB	-1.16	1.02	0.49	0.64

4.6 Appendix

4.6.1 WinBUGS code for implementing the GMRF-MJM

```
model

{

for( i in 1:G) {

# Dat is a G by 3 Data matrix (binding, expression, sequence data)

Dat[i,1:3] ~ dmnorm(Mu[T[i],1:3],InvSigma[T[i],1:3,1:3])

pi[i,1] <-1 - pi[i,2]

pi[i,2] <-0.5/(1+exp(alpha1[i]-alpha2[i]+beta1[i]-beta2[i]+

gamma1[i]-gamma2[i]-logitpi2))+0.5*pi2

# latent variable (non-target(1)/target(2))
```

```
T[i] ~dcat(pi[i,1:2])

T1[i] <-equals(T[i],1) ;T2[i] <-equals(T[i],2)

}

# Random Fields specification

# for co-expression network

alpha1[1:N] ~car.normal(adjcoexp[], weightscoexp[], numcoexp[], tauCcoexp[1])

alpha2[1:N] ~car.normal(adjcoexp[], weightscoexp[], numcoexp[], tauCcoexp[2])

# for GO induced functional coupling network

beta1[1:N] ~car.normal(adjGO[], weightsGO[], numGO[], tauCGO[1])

beta2[1:N] ~car.normal(adjGO[], weightsGO[], numGO[], tauCGO[2])

# for RegulonDB gene regulatory network

gamma1[1:N] ~car.normal(adjReg[], weightsReg[], numReg[], tauCReg[1])

gamma2[1:N] ~car.normal(adjReg[], weightsReg[], numReg[], tauCReg[2])

# weights specification

for(k in 1:sumNumNeighcoexp) { weightscoexp[k] <- 1 }

for(l in 1:sumNumNeighGO) { weightsGO[l] <- 1}

for(l in 1:sumNumNeighReg) { weightsReg[l] <- 1}

# priors

# precision parameters for random fields)

tauCcoexp[1] ~dgamma(0.01, 0.01)

tauCcoexp[2] ~dgamma(0.01, 0.01)
```

```
tauCGO[1] ~dgamma(0.01, 0.01)

tauCGO[2] ~dgamma(0.01, 0.01)

tauCReg[1] ~dgamma(0.01, 0.01)

tauCReg[2] ~dgamma(0.01, 0.01)

# parameters in component distributions

Mu[1,1:3] ~ dmnorm(mu1[1:3],C[1:3,1:3]) # non-target

for(j in 1:3){

Mu[2,j] <- Mu[1,j] + theta[j] # target

}

theta[1:3] ~dmnorm(mutheta[1:3],prectheta[1:3,1:3])I(lo[1:3],up[1:3])

InvSigma[1,1:3,1:3]~dwish(R1[1:3,1:3],3)

InvSigma[2,1:3,1:3]~dwish(R2[1:3,1:3],3)

Sigma1[1:3,1:3]<-inverse(InvSigma[1,1:3,1:3])

Sigma2[1:3,1:3]<-inverse(InvSigma[2,1:3,1:3])

## estimated proportion of targets in the SMJM was 0.12

logitpi2<- logit(0.12)

pi2<-0.12

}
```

4.6.2 MCMC Algorithm for the DMRF-MJM

We denote by $(\alpha|\ldots)$ the full conditional of α, that is the distribution of α conditional on everything else in the model. In addition, we denote by $MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the multivariate

normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, by $\phi(.;\boldsymbol{\mu},\boldsymbol{\Sigma})$ the corresponding density function, and by $W((\rho R)^{-1},\rho)$ the Wishart distribution with mean R^{-1}. The observed data are denoted as $\mathbf{x}=\left\{x_i=(B_i,E_i,S_i)';i=1,\ldots,G\right\}$. Model specification and prior distributions for the DMRF-MJM can be found in Sections 4.3.4 & 4.3.5. The joint posterior distribution is

$$(\mathbf{T},\boldsymbol{\mu}_0,\boldsymbol{\theta},\boldsymbol{\Sigma}_0,\boldsymbol{\Sigma}_1,\Phi|\mathbf{x})\propto p(\mathbf{x}|\mathbf{T},\boldsymbol{\mu}_0,\boldsymbol{\theta},\boldsymbol{\Sigma}_0,\boldsymbol{\Sigma}_1)p(\mathbf{T}|\Phi)p(\boldsymbol{\mu}_0)p(\boldsymbol{\theta})p(\boldsymbol{\Sigma}_0)p(\boldsymbol{\Sigma}_1)p(\Phi)$$

- update $\boldsymbol{\mu}_0$ by Gibbs sampling with the proposal given by

$$(\boldsymbol{\mu}_0|\ldots)\sim MVN((n_0\boldsymbol{\Sigma}_0^{-1}+\boldsymbol{C}^{-1})^{-1}\boldsymbol{\Sigma}_0^{-1}\sum_{\{i:T_i=0\}}x_i,(n_0\boldsymbol{\Sigma}_0^{-1}+\boldsymbol{C}^{-1})^{-1}),$$

  where $n_0=|\{i:T_i=0\}|$.

- update $\boldsymbol{\theta}$ by Gibbs sampling with the proposal given by

$$(\boldsymbol{\theta}|\ldots)\sim MVN((n_1\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{C}^{-1})^{-1}\boldsymbol{\Sigma}_1^{-1}\sum_{\{i:T_i=1\}}(x_i-\boldsymbol{\mu}_0),(n_1\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{C}^{-1})^{-1}I(\boldsymbol{\theta}>0),$$

  where $n_1=|\{i:T_i=1\}|$.

- update $\boldsymbol{\Sigma}_j$, for $j=0,1$, by Gibbs sampling with the proposal given by

$$(\boldsymbol{\Sigma}_j|\ldots)\sim W((\sum_{\{i:T_i=j\}}(x_i-\boldsymbol{\mu}_j)(x_i-\boldsymbol{\mu}_j)'+3R)^{-1},n_j+3),$$

  where $\boldsymbol{\mu}_1=\boldsymbol{\mu}_0+\boldsymbol{\theta}$.

- update T_i by Gibbs sampling with proposal given by

$$(T_i|\ldots)\sim\text{Bernoulli}(\frac{d}{1+d}),$$

  where $d=\exp\left\{\gamma+\sum_{k=1}^{K}\beta_k\left[n_i^{(k)}(1)-n_i^{(k)}(0))\right]/m_i^{(k)}\right\}\frac{\phi(x_i;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)}{\phi(x_i;\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)}$.

- update $\Phi = (\gamma, \beta_1, \ldots, \beta_K)$ using a random walk Metropolis algorithm with Gaussian proposal, which has diagonal covariance matrix. The acceptance ratio is calculated using the full conditional of Φ, which is proportional to

$$\frac{\exp\left\{n_1\gamma + \sum_{j=0}^{1}\sum_{i:T_i=j}\sum_{k=1}^{K}\beta_k n_i^{(k)}(j)/m_i^{(k)}\right\}}{\prod_{i=1}^{G}\left\{\exp(\sum_{k=1}^{K}\beta_k n_i^{(k)}(0)/m_i^{(k)}) + \exp(\gamma + \sum_{k=1}^{K}\beta_k n_i^{(k)}(1)/m_i^{(k)})\right\}}.$$

  The Gaussian proposal was tuned such that the acceptance rate was around 0.23, the optimal one (Carlin and Louis 2000).

Chapter 5

Discussion and Future Work

5.1 Conclusion

We have proposed new Markov random field (MRF)-based mixture models for integrating gene networks and high-throughput genomic data. This research is intended to meet the biological challenge to boost statistical power for genomic discovery by maximizing the use of existing biological knowledge and diverse types of genomic and proteomic data. Chapter 2 develops a Gaussian Markov random field (GMRF)-based mixture model to smooth gene-specific prior probabilites over a gene network. Chapter 3 proposes a Bayesian implementation of the discrete Markov random field (DMRF)-based mixture model and compares its performance with that based on GMRFs. Furthermore, in Chapter 4, we extend the GMRF-based and DMRF-based mixture models to ones that allow integration of multiple gene networks and heterogeneous types of genomic data. Applications to real data, along with simulations, demonstrate the utility of the proposed methods and their superior performance over standard methods that do not capitalize on gene network information.

The basic assumption underlying the proposed network-based models is that any two neighboring genes in a network are more similar (i.e., more likely to be or not to be in the non-null state together) than non-neighboring ones. As illustrated by the numerical studies, the network-based methods improve the statistical efficiency dramatically when the above assumption roughly holds. In addition, the GMRF-based mixture model is more powerful than the DMRF-based mixture model. Due to incomplete biological knowledge, gene networks may include both false positive and false negative interactions.

Nevertheless, the proposed methods are relatively robust to misspecified networks thanks to the adopted mixture model framework: the posterior probability of a gene being in the non-null state is jointly determined by the prior probability of it, which is influenced by the gene network, and the likelihood, as determined by the observed experimental data. As a result, misspecified gene network information may not have a large influence on the posterior probability, based on which we claim significant genes, as long as the likelihood ratio is large.

5.2 Areas for Future Work

The proposed MRF-based mixture models may lack enough flexibility to deal with potential heterogeneity in the gene network. Gene-gene interactions are often condition-specific and dynamic. For example, two genes may interact with each other only in the process of meiosis. Therefore, it is expected that the spatial association strengths across a gene network are unlikely to be homogenous for a specific experimental condition. However, there are only two parameters, τ_{C0} and τ_{C1}, in the GMRF and one parameter, β, in the DMRF to control the spatial smoothness. Future work may include developing more flexible methods that are data adaptive, more robust to mis-specification of network structures, and allowing the incorporation of varying local structures of network topology.

We implement the network-based mixture models in a Bayesian framework and rely on MCMC samples to do inference. The MCMC algorithms are relatively easy to imple-

ment. However, because each MCMC iteration involves drawing random numbers from a large number of full conditional distributions and many iterations are needed to obtain reliable posterior estimates, it is slow to run the MCMC algorithms. It may be worth developing some more efficient computational approaches to the MRF-based mixture models via, for example, Variational Bayes (Smidl and Quinn 2005) or Empirical Bayes, as alternatives to MCMC.

Model selection criteria in the context of MRF-based mixture models are needed. Deviance Information Criteria (DIC) by Spiegelhalter *et al.* (2002) may be a candidate. However, for models with missing data, such as mixture models, DIC is not uniquely defined; see Celeux *et al.* (2006) for a comprehensive discussion of various different DIC definitions and their comparative performance with a standard mixture model. Extending DICs from SMM to MRF-based mixture models is not trivial and needs special treatments. This could be a direction of future work.

Chapter 6

Bibliography

Arndt K, Fink GR. (1986) GCN4 protein, a positive transcription factor in yeast, binds general control promoters at all 5' TGACTC 3' sequences. *Proc Natl Acad Sci U S A*, **83**(22):8516-20.

Ash Ashburner, M., et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, **25**, 25-29

Bailey T.L. and Elkan C. (1995). Unsupervised Learning of Multiple Motifs in Biopolymers using EM. *Machine Learning*, **21**, 51-80.

Benjamini, Y., Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing *Journal of the Royal Statistical Society: Series B*, **57**, 289-300.

Beyer, A., Workman, et al. (2006) Integrated assessment and prediction of transcription

factor binding. *PLoS Computational Biology*, **2**:e70.

Besag, J. (1986) On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society: Series B*, **48**, 259-302.

Besag, J. and Kooperberg, C. (1995) On conditional and intrinsic autoregressions. *Biometrika*, **82**, 733-746.

Broet, P., Richardson, S. (2006) Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model. *Bioinformatics*, **22**, 911-918.

Carlin, B.P., Louis, T.A. (2000) Bayes and Empirical Bayes methods for data analysis. Second edition. Chapman & Hall/CRC Press, New York.

Celeux,G., Forbes, F., Robert,C.P. and Titterington, D.M. (2006) Deviance information criteria for missing data models (with Discussion). *Bayesian Analysis*, **1**,651-706.

Cirz, R.T., Chin, J.K., Andes, D.R., et al. (2005) Inhibition of mutation and combating the evolution of antibiotic resistance. *PLoS Biol.* **3**(6): e176.

Conlon, E.M., Liu, X.S., Lieb, J.D. and Liu, J.S. (2003). Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl. Acad. Sci. USA* **100**, 3339-3344.

Congdon, P. (2001) *Bayesian Statistical Modelling*. Wiley, Chichester

Courcelle, J., Khodursky, A., Peter, B., Brown, P.O., Hanawalt, P.C. (2001) Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient

Escherichia coli. *Genetics*, **158**, 41-64.

Cui, X. and Churchill, G. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology*, **4**:210.

Deng, M.H., Chen, T., Sun, F.(2004) An Integrated Probabilistic Model for Functional Prediction of Proteins. *Journal of Computational Biology* **11(2/3)**, 463-475.

Dvorkin, D. (2007) Prediction of gene coexpression by integrating the Gene Ontology with microarray data. MS Thesis. Division of Biostatistics, University of Minnesota.

Dopazo, J. (2006) Functional Interpretation of Microarray Experiments. *OMICS: A Journal of Integrative Biology*, **10**, 398-410.

Efron, B., et al. (2001) Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, **96**, 1151-1160.

Efron, B. and Tibshirani, R. (2007) On testing the significance of sets of genes. *Annals of Applied Statistics*, **1**, 107-129.

Elion, E.A., Brill, J.A., Fink, G.R. (1991) FUS3 represses CLN1 and CLN2 and in concert with KSS1 promotes signal transduction. *Proceedings of National Academy of Science*, **88**, 9392-6.

Faith, J.J., Driscoll, M.E., Fusaro. V.A., Cosgrove, E.J., Hayete, B., Juhn, F.S., Schneider, S.J., and Gardner, T.S.(2008) Many Microbe Microarrays Database: uniformly

normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Research*, **36**(Database issue): D866-870.

Fernandez, C. and Green, P. (2002) Modelling spatially correlated data via mixtures: a Bayesian approach. *Journal of the Royal Statistical Society: Series B*, **64**, 805-826.

Futschik, M.E., Chaurasia, G., Herzel, H. (2007) Comparison of human protein-protein interaction maps. *Bioinformatics*, **23**(5):605-11.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004) Bayesian data analysis. Second edition. Chapman & Hall/CRC Press.

Green, P.J., Richardson, S. (2002) Hidden Markov models and disease mapping. *Journal of the American Statistical Association*, **97**, 1055-1070.

Harbison, C.T., Gordon, D.B. et al. (2004) Transcriptional Regulatory Code of a Eukaryotic Genome. *Nature*, **431**, 99-104.

Heikkinen, J., Hogmander, H. (1994) Fully Bayesian approach to image restoration with an application in biogeography. *Applied Statistics*, **43**, 569-582.

Hinnebusch, A.G., Natarajan, K. (2002) Gcn4p, a master regulator of gene expression, is controlled at multiple levels by diverse signals of starvation and stress. *Eukaryotic Cell*, **1**(1):22-32.

Hughes, T.R., et al. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**(1):109-126.

Ideker, T. and Sharan, R. (2008) Protein networks in disease. *Genome Research*, **18**(4):644-52.

Jauniaux, J.C., Urrestarazu, L.A., Wiame, J.M. (1978) Arginine metabolism in Saccharomyces cerevisiae: subcellular localization of the enzymes. *Journal of Bacteriology*, **133**, 1096-1107.

Jensen, S.T., Chen, G., Stoeckert, C. (2007) Bayesian Variable Selection and Data Integration for Biological Regulatory Networks.*Annals of Applied Statistics*, **1**, 612-633.

Kanehisa, M. and Goto, S. (2000) Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, **28**, 27-30.

Lee, I., Date, S.V., Adai, A.T., Marcotte, E.M. (2004) Probabilistic Functional Network of Yeast Genes. *Science*, **306**, 1555 - 1558.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I. et al. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799-804.

Lewin, A., Richardson, S., Marshall, C., Glazier, A., Aitman, T. (2006) Bayesian modeling of differential gene expression. *Biometrics*, **62**, 1-9.

Liang, F., Zhang, J. (2008) Estimating FDR under general dependence using stochastic approximation. *Biometrika*, Advance Access published on September 26, 2008.

Martens JA, Brandl CJ. (1994) GCN4p activation of the yeast TRP3 gene is enhanced by ABF1p and uses a suboptimal TATA element. *J Biol Chem.*, **269**(22):15661-7.

McLachlan, G.J., Bean, R.W., Ben-Tovim Jones, L. (2006) A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, **22**, 1608-1615.

McLachlan, G.J., Peel, D. (2000) *Finite Mixture Models.* Wiley, New York

Milbauer, L.C., Wei, P., Enenstein, J., Jiang, A., Hillery, C.A., Scott, J.P., Nelson, S.C., Bodempudi, V., Topper, J.N., Yang, R.B., Hirsch, B., Pan, W., Hebbel, R.P. (2008) Genetic endothelial systems biology of sickle stroke risk. *Blood*, **111**(7):3872-3879.

Michel, B. (2005) After 30 Years of Study, the Bacterial SOS Response Still Surprises Us. *PLoS Biology*, **3**(7): e255.

Newton, M.A., et al. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, **8**, 37-52.

Newton, M.A., Noueiry, A., Sarkar, D., Ahlquist, P. (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, **5**(2):155-76.

Newton, M.A., Quintana, F.A., den Boon, J.A., Sengupta, S. and Ahlquist, P. (2007) Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Annals of Applied Statistics*, **1**, 85-106.

Pan, W. (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **18**(4), 546-554.

Pan, W. (2005) Incorporating Biological Information as a Prior in an Empirical Bayes Approach to Analyzing Microarray Data. *Statistical Applications in Genetics and Molecular Biology*, **4**, Article 12.

Pan, W. (2006a) Incorporating gene functional annotations in detecting differential gene expression. *Applied Statistics*, **55**, 301-316.

Pan, W. (2006b) Incorporating gene functions as priors in model-based clustering of microarray gene expression data. *Bioinformatics*, 22, 795-801.

Pan, W., Lin, J., Le, C. (2002) Model-Based Cluster Analysis of Microarray Gene Expression Data. *Genome Biology*, **3**, research0009.1-8.

Pan, W., Wei, P., Khodursky, A. (2008) A parametric joint model of DNA-protein binding, gene expression and dna sequence data to detect target genes of a transcription factor. *Pacific Symposium on Biocomputing 2008* **13**, 465-476.

Pauwels, K., Abadjieva, A., Hilven, P., Stankiewicz, A., Crabeel, M. (2003) The N-acetylglutamate synthase/N-acetylglutamate kinase metabolon of Saccharomyces cerevisiae allows co-ordinated feedback regulation of the first two steps in arginine biosynthesis. *European Journal of Biochemistry*, **270**, 1014-24.

Pokholok, D.K., Harbison, C.T. et al. (2005) Genome-wide Map of Nucleosome Acetylation and Methylation in Yeast. *Cell.* **122**, 517-27.

Ren, B., Robert, F., *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306-2309.

Rodrigues, A. and Assuncao, R. (2008) Propriety of posterior in Bayesian space varying parameter models with normal data. *Statistics and Probability Letters*, **78**, 2408-2411.

Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotech.*, **16**, 939-945.

Ryden, T. and Titterington, D.M. (1998) Computational Bayesian analysis of hidden Markov models. *Journal of Computational and Graphical Statistics*, **7**, 194-211.

Salgado H, Gama-Castro S, Peralta-Gil M, Diaz-Peredo E, Sanchez-Solano F, Santos-Zavaleta A, Martinez-Flores I, Jimenez-Jacinto V, Bonavides-Martinez C, Segura-Salazar J, Martinez-Antonio A, Collado-Vides J. (2006). RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res.*, **34**, D394-D397.

Schuldiner,O., Yanover,C. and Benvenisty, N. (1998) Computer analysis of the entire budding yeast genome for putative targets of the GCN4 transcription factor. *Curr Genet*, **33**:16-20.

Smidl, V. and Quinn, A.P. (2005) The Variational Bayes Method in Signal Processing. Springer-Verlag, New York.

Smith, M., Fahrmeir, L. (2007) Spatial Bayesian Variable Selection With Application to Functional Magnetic Resonance Imaging. *Journal of the American Statistical Association*, **102**, 417-431.

Smith, D., Smith, M. (2006) Estimation of Binary Markov Random Fields Using Markov Chain Monte Carlo *Journal of Computational and Graphical Statistics*, **15**, 207-227.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and Van der Linde, A. (2002) Bayesian Measures of Model Complexity and Fit (with Discussion), *Journal of the Royal Statistical Society, Series B*, **64**(4):583-616.

Spiegelhalter, D., Thomas, A., Best, N., Lunn, D. (2003) WinBUGS User Manual, Version 1.4. Available at http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf

Subramanian, A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of National Academy of Science*, **102**, 15545-15550.

Sun, N., Carroll, R.J., Zhao, H. (2006) Bayesian error analysis model for reconstructing transcriptional regulatory networks. *Proc Natl Acad Sci USA*, **103**, 7988-7993.

Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. (2005) Discovering

statistically significant pathways in expression profiling studies. *Proceedings of National Academy of Science*, **102**, 13544-13549.

Toyn,J.H., Gunyuzlu, P.L., White, W.H., Thompson, L.A., Hollis, G.F.(2000) A counterselection for the tryptophan pathway in yeast: 5-fluoroanthranilic acid resistance. *Yeast*, **16**, 553-60.

Tusher, V.G., et al. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of National Academy of Science*, **98**, 5116-5121.

Wade, J.T., Reppas, N. B., Church, G. M. and Struhl, K. (2005). Genomic analysis of LexA binding reveals the permissive nature of the Escherichia coli genome and identifies unconventional target sites. *Genes Dev.*, **19**:2619-2630.

Wang, W., Cherry, J.M., Nochomovitz, Y., Jolly, E., Botstein, D. and Li, H. (2005). Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. Proc. Nat. Acad. Sci. USA, **102**, 1998-2003.

Wei, P., Pan, W. (2008a) Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. *Bioinformatics*, **24**(3):404-411.

Wei, P., Pan, W. (2008b) Incorporating gene functions into regression analysis of DNA-protein binding data and gene expression data to construct transcriptional networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **5**(3):401-415.

Wei, P., Pan, W. (2009) Network-based genomic discovery: application and comparison of Markov random field models. To appear in *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

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

Winkler, G. (2003) *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*. Springer-Verlag New York, Inc.

Xiao, G., Reilly, C., Martinez-Vaz, B., Pan, W., Khodursky, A.B. (2005) Improved detection of differentially expressed genes through incorporation of gene locations. Research Report 2005-028, Division of Biostatistics, University of Minnesota. Available at http://www.biostat.umn.edu/rrs.php

Xie, Y. (2006) Statistical analysis for microarray data: false discovery rate estimation, statistical testing and integrated analysis. PhD dissertation, University of Minnesota, Minneapolis, MN, USA.

Yang, Y.H., Dudoit, S., et al. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, **30**4:e15.

Wu, H., Su, Z., Mao, F., Olman, V. and Xu, Y. (2005) Prediction of functional modules based on comparative genome analysis and Gene Ontology application. *Nucleic Acids Res.*, **33**: 2822-2837.

Wu, W.B. (2008) On false discovery control under dependence. *Annals of Statistics*, **36**(1):364-380.