Published in final edited form as:

J R Stat Soc Ser C Appl Stat. 2010 January 1; 59(1): 105–125. doi:10.1111/j.1467-9876.2009.00686.x.

Network-based genomic discovery: application and comparison of Markov random field models

Peng Wei* and Wei Pan

University of Minnesota, Minneapolis, USA

Abstract

As biological knowledge accumulates rapidly, gene networks encoding genome-wide gene-gene interactions have been constructed. As an improvement over the standard mixture model that tests all the genes *iid a priori*, Wei and Li (2007) and Wei and Pan (2008) proposed modeling a gene network as a Discrete- or Gaussian-Markov random field (DMRF or GMRF) respectively in a mixture model to analyze genomic data. However, how these methods compare in practical applications is not well understood and this is the aim here. We also propose two novel constraints in prior specifications for the GMRF model and a fully Bayesian approach to the DMRF model. We assess the accuracy of estimating the False Discovery Rate (FDR) by posterior probabilities in the context of MRF models. Applications to a ChIP-chip data set and simulated data show that the modified GMRF models have superior performance as compared with other models, while both MRF-based mixture models, with reasonable robustness to misspecified gene networks, outperform the standard mixture model.

Keywords

Bayesian hierarchical model; ChIP-chip; Conditional autoregression (CAR); Discrete Markov random field; Gaussian Markov random field; Gene networks; Mixture models

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 messenger 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 *et al.* 2000), has enabled researchers to identify genome-wide binding locations of a TF in living cells. In

*Address for correspondence: Peng Wei, Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building (MMC 303), Minneapolis, MN 55455-0378, USA. pengw@biostat.umn.edu.

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 is to identify the binding targets of GCN4 based on ChIP-chip data.

A brief description of the ChIP-chip experiment is as follows: first, GCN4 binds to certain genome sequences in living cells; second, DNA sequences are chopped into small fragments, some of which are bound by GCN4 while the rest are not; third, those DNA fragments bound by GCN4 are isolated by chromatin immunoprecipitation (ChIP) followed by separating GCN4 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 GCN4'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 GCN4. 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 3.1 for more details) to obtain a p-value for each gene for testing the null hypothesis of not a binding target of GCN4. Table 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 2.1.

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; see Pan (2002) for a review on statistical methods for gene expression data. For example, a standard mixture model (McLachlan *et al.* 2006) could be applied to the binding z-scores, which, however, treats all the genes identically and independently *a priori* and ignores the fact that genes work coordinately in biological processes as dictated by gene networks, 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 (6,270 versus 3 for the GCN4 data), it is desirable to take advantage of existing biological knowledge to maximize statistical power for genomic discovery. 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 network for the yeast genome by integrating a variety of genomic data: a pair of genes that are evidenced to be co-functional in biological processes are connected on the resulting network. There are other types of gene networks besides functional linkage network, such as TF-gene regulatory network, protein-protein interaction network, co-expression network 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 GCN4's binding target gene on the network are also likely to be GCN4's targets. An emerging theme in genomic studies is thus to incorporate gene networks as prior biological knowledge into statistical modeling of microarray data to maximize the power for

biological discoveries (Wei and Li 2007 & 2008, Wei and Pan 2008, and Sanguinetti *et al.* 2008). Wei and Li (2007) modeled the latent true states of the genes using a discrete Markov random field (DMRF), while Wei and Pan (2008) modeled the prior probabilities of the true states via Gaussian Markov random fields (GMRF). 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 and simulated data in this paper. It is worth noting that the GMRF-MM is similar to the models used by Fernandez and Green (2002) and Broet and Richardson (2006), but the concept of spatial neighborhood is extended from physical neighboring to gene-gene interactions as embedded in a gene network.

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 allowed. 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 Titterton 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 GMRF's; 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 compare 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.

This paper 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.

2 Methods

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 , $i = 1, \dots, G$; 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 $T_i = I(H_{i0} \text{ is false})$; 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.

2.2 Standard mixture model

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

$$f(z_i) = \pi_0 f_0(z_i) + (1 - \pi_0) f_1(z_i), \quad (1)$$

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:

$f_0 = \sum_{k_0=1}^{K_0} \pi_{0k_0} \varphi(\mu_{k_0}, \sigma_{k_0}^2)$ and $f_1 = \sum_{k_1=1}^{K_1} \pi_{1k_1} \varphi(\mu_{k_1}, \sigma_{k_1}^2)$, where $\varphi(\mu, \sigma^2)$ 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, in 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 on 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 discussions; the Appendix gives an example of relaxing this restriction. The conditional distribution of z_i is thus:

$$p(z_i | T_i = j, \theta) = \varphi(z_i; \mu_j, \sigma_j^2), \quad (2)$$

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

2.3 GMRF-based mixture model

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

$$\pi_{ij} = Pr(T_i = j) = \exp(x_{ij}) / [\exp(x_{i0}) + \exp(x_{i1})]. \quad (3)$$

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

$$x_{ij} | x_{(-i)j} \sim N \left(\frac{1}{m_i} \sum_{l \in \partial i} x_{lj}, \frac{\sigma_{Cj}^2}{m_i} \right), \quad (4)$$

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 (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 2008), which is also the default setting in WinBUGS (Spiegelhalter *et al.* 2003). In this model, the parameter σ_{Cj}^2 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: the smaller σ_{Cj}^2 induces more similar π_{ij} 's for those genes that are neighbors in the network. In addition, the conditional distribution of z_i is

$$f(z_i | x_{i0}, x_{i1}) = \pi_{i0} f_0(z_i) + \pi_{i1} f_1(z_i). \quad (5)$$

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

$$\sum_i x_{i0} = \sum_i x_{i1} = 0 \Rightarrow \bar{x}_{\cdot 0} = \bar{x}_{\cdot 1} = 0,$$

where $\bar{x}_{\cdot j} = (1/G) \sum_{i=1}^G x_{ij}$ for $j = 0, 1$. By (3), we have $\text{logit}(\pi_{i1}) = x_{i1} - x_{i0}$. It follows that

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

Thus, $\text{logit}(\pi_{i1})$'s have mean 0. This implies that the posterior estimates of $\text{logit}(\pi_{i1})$'s will be shrunk towards 0, or roughly, the estimates of π_{i1} 's are shrunk towards 0.5. It is consistent with our observation that the average of posterior means π_{i1} 's 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 2008); correspondingly, the estimated mean for the right-tail (non-null) component in the GMRF-MM was much smaller than the one 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

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

which can be realized by imposing $x_{.1} = c$ and $x_{.0} = 0$. 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 at 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:

$$\pi_{ij} = \lambda \frac{\exp(x_{ij})}{\exp(x_{i0}) + \exp(x_{i1})} + (1 - \lambda) \hat{\pi}_j, \quad (6)$$

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). 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., $x_{.1} = \text{logit}(\pi_1)$ and $x_{.0} = 0$. We call this modified model “GMRF-MM with the average constraint”.

2.4 DMRF-based mixture model

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

$$Pr(T_i=1|T_{(-i)}, \Phi) = Pr(T_i=1|T_{\partial i}, \Phi) = \frac{\exp(\gamma + \beta(n_i(1) - n_i(0))/m_i)}{1 + \exp(\gamma + \beta(n_i(1) - n_i(0))/m_i)}, \quad (7)$$

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 states j for $j = 0, 1$, and thus $(n_i(1) - n_i(0)) = \sum_{l \in i} (2T_l - 1)$; 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 for neighboring genes. Hence, the larger β , the more probable do 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*

$$pl(\mathbf{T}; \Phi) = \prod_{i=1}^G p(T_i|T_{\partial i}; \Phi) = \prod_{i=1}^G \frac{\exp(T_i(\gamma + \beta(n_i(1) - n_i(0))/m_i))}{1 + \exp(\gamma + \beta(n_i(1) - n_i(0))/m_i)}, \quad (8)$$

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

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

$$Pr(T_i=1|\mathbf{z}, \theta, \pi_0, \pi_1) = \frac{1}{1 + \frac{\pi_0}{\pi_1} \frac{\varphi(z_i; \mu_0, \sigma_0^2)}{\varphi(z_i; \mu_1, \sigma_1^2)}}, \quad (9)$$

$$Pr(T_i=1|\mathbf{z}, \theta, x_{i0}, x_{i1}) = \frac{1}{1 + \frac{\pi_{i0}}{\pi_{i1}} \frac{\varphi(z_i; \mu_0, \sigma_0^2)}{\varphi(z_i; \mu_1, \sigma_1^2)}} \frac{1}{1 + \exp(x_{i0} - x_{i1}) \frac{\varphi(z_i; \mu_0, \sigma_0^2)}{\varphi(z_i; \mu_1, \sigma_1^2)}}, \quad (10)$$

and

$$Pr(T_i=1|\mathbf{z}, \theta, T_{\partial i}, \Phi) = \frac{1}{1 + \frac{1}{\exp(\gamma + \beta(n_i(1) - n_i(0))/m_i)} \frac{\varphi(z_i; \mu_0, \sigma_0^2)}{\varphi(z_i; \mu_1, \sigma_1^2)}} \quad (11)$$

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, in all three models. This sheds light on that mis-specified prior distributions, e.g., due to incomplete gene networks, may not have a large influence on the posterior probability if the data likelihood ratio is large. Second, GMRF-MM is more richly parameterized and thus more flexible by introducing 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 π_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 $n_i(1) = n_i(0)$.

2.6 Parameter estimation

Following Wei and Pan (2008), we adopt 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.

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.

2.6.2 Bayesian approach to DMRF—Although the ICM approach is easy to implement and requires little computational effort, it has major drawbacks 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) of $\mathbf{T}$, the mode of the joint posterior distribution $Pr(\mathbf{T}|\mathbf{z})$. Define the maximum marginal posterior (MMP) $\hat{\mathbf{T}} = (T_1, \dots, T_G)$, where $T_i = I(p_i > 0.5) = \arg \max_{t_i \in \{0,1\}} Pr(T_i = t_i|\mathbf{z})$. 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 A.3 for remarks on the above statements. Note that by adopting a Bayesian approach, we can obtain not only the MMP estimates 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:

$$p(\mathbf{T}, \theta, \Phi|\mathbf{z}) \propto p(\mathbf{z}|\mathbf{T}, \theta)p(\mathbf{T}|\Phi)p(\theta)p(\Phi), \quad (12)$$

where $p(\mathbf{z}|\mathbf{T}, \theta)$ is as (2); for $p(\mathbf{T}|\Phi)$, we adopt the pseudolikelihood (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 those in 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 Titterton (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 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 (12): we use a Metropolis algorithm for Φ and Gibbs samplers for the remaining parameters. The complete algorithm can be found in Appendix A.2.

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):

$$\text{FDR}(t) = \sum_{i=1}^G q_i I(q_i \leq t) / \sum_{i=1}^G I(q_i \leq t), \quad (13)$$

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 (13) in our real and simulated data examples, and assessed its accuracy under dependence.

3 Example

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 "single-array error model" (Hughes *et al.* 2000). A sample of the data is displayed in Table 1, where the z-scores were obtained as described in Section 2.1. Wei and Pan (2008) analyzed this data set by applying a SMM and a GMRF-MM with the zero constraint to the z-scores. Following Wei and Pan, 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 interactions, 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.

3.2 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 (2008). Wei and Pan 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, with an exception to 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. A complete model specifications for a three-component SMM, GMRF-MM and DMRF-MM can be found in Appendix A.1. Parameter estimates for all the models are shown in Tables 2 & 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 $\hat{\pi}_{\cdot,1}$) 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 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 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 even higher sensitivities compared with that of GMRF-MM with the 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 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.

3.4 Examples of genomic discoveries

Figure 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 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 demonstrated 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 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 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.

3.5 FDR estimation

Figures 3(a)–(e) show realized FDR's (based on the control sets) versus estimated FDR's (based on (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, GMRF-MM with the zero constraint 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% slightly. 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.

4 Simulation

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 (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 and the rest of genes to be non-targets, giving an initial T . Then we iterated the states 20 times based on (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 (2008), 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.

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 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 1(c)–(f) show each MRF-based mixture model's robustness to mis-specified network. 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(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 average constraint performed much better 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 (13) depends on the correctness of the fitted model. In other words,

GMRF-MM with the average constraint had better model fitting, leading to more accurate FDR estimation.

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 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 (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 result from improper priors, while improper priors may still lead to proper posterior distributions (p110, 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.

In this paper, we have used ROC curves to compare different statistical methods. However, in most real problems, control gene sets like the ones we used are not available. As a result, other model comparison criteria are needed in such situations. 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.

Acknowledgments

The authors are grateful to the two reviewers and the Joint Editor for helpful comments. PW and WP were supported by NIH grant HL65462. PW was also partially supported by a Doctoral Dissertation Fellowship from the University of Minnesota Graduate School.

References

- Arndt K, Fink GR. 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.* 1986; 83(22):8516–20. [PubMed: 3464968]
- Ashburner M, et al. Gene ontology: tool for the unification of biology. *Nature Genetics.* 2000; 25:25–29. [PubMed: 10802651]
- Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B.* 1995; 57:289–300.
- Beyer A, Workman C, et al. Integrated Assessment and Prediction of Transcription Factor Binding. *PLoS Computational Biology.* 2006; 2(6):e70. [PubMed: 16789814]
- Besag J. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society: Series B.* 1986; 48:259–302.
- Besag J, Kooperberg C. On conditional and intrinsic autoregressions. *Biometrika.* 1995; 82:733–746.
- Broet P, Richardson S. Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model. *Bioinformatics.* 2006; 22:911–918. [PubMed: 16455750]
- Carlin, BP.; Louis, TA. Bayes and Empirical Bayes methods for data analysis. 2. Chapman & Hall/CRC Press; New York: 2000.
- Celeux G, Forbes F, Robert CP, Titterton DM. Deviance information criteria for missing data models (with Discussion). *Bayesian Analysis.* 2006; 1:651–706.
- Fernandez C, Green PJ. Modelling spatially correlated data via mixtures: a Bayesian approach. *Journal of the Royal Statistical Society: Series B.* 2002; 64:805–826.
- Futschik ME, Chaurasia G, Herzel H. Comparison of human protein-protein interaction maps. *Bioinformatics.* 2007; 23(5):605–11. [PubMed: 17237052]
- Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. Bayesian data analysis. 2. Chapman & Hall/CRC Press; 2004.
- Green PJ, Richardson S. Hidden Markov models and disease mapping. *Journal of the American Statistical Association.* 2002; 97:1055–1070.
- Hinnebusch AG, Natarajan K. Gcn4p, a master regulator of gene expression, is controlled at multiple levels by diverse signals of starvation and stress. *Eukaryotic Cell.* 2002; 1(1):22–32. [PubMed: 12455968]
- Heikkinen J, Hogmander H. Fully Bayesian approach to image restoration with an application in biogeography. *Applied Statistics.* 1994; 43:569–582.
- Ideker T, Sharan R. Protein networks in disease. *Genome Research.* 2008; 18(4):644–52. [PubMed: 18381899]
- Lee I, Date SV, Adai AT, Marcotte EM. Probabilistic Functional Network of Yeast Genes. *Science.* 2004; 306:1555–1558. [PubMed: 15567862]

- Lee TI, Rinaldi NJ, et al. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*. 2002; 298:799–804. [PubMed: 12399584]
- Liang F, Zhang J. Estimating FDR under general dependence using stochastic approximation. *Biometrika*. 2008 Advance Access published on September 26, 2008.
- Martens JA, Brandl CJ. GCN4p activation of the yeast TRP3 gene is enhanced by ABF1p and uses a suboptimal TATA element. *J Biol Chem*. 1994; 269(22):15661–7. [PubMed: 8195216]
- McLachlan GJ, Bean RW, Ben-Tovim Jones L. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*. 2006; 22:1608–1615. [PubMed: 16632494]
- Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*. 2004; 5(2):155–76. [PubMed: 15054023]
- Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*. 2002; 18(4):546–554. [PubMed: 12016052]
- Pokholok DK, Harbison CT, et al. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*. 2005; 122:517–27. [PubMed: 16122420]
- Ren B, Robert F, et al. Genome-wide location and function of DNA binding proteins. *Science*. 2000; 290:2306–2309. [PubMed: 11125145]
- Rodrigues A, Assuncao R. Propriety of posterior in Bayesian space varying parameter models with normal data. *Statistics and Probability Letters*. 2008; 78:2408–2411.
- Ryden T, Titterton DM. Computational Bayesian analysis of hidden Markov models. *Journal of Computational and Graphical Statistics*. 1998; 7:194–211.
- Sanguinetti G, Noirel J, Wright PC. MMG: a probabilistic tool to identify sub-modules of metabolic pathways. *Bioinformatics*. 2008; 24(8):1078–1084. [PubMed: 18292114]
- Schuldiner O, Yanover C, Benvenisty N. Computer analysis of the entire budding yeast genome for putative targets of the GCN4 transcription factor. *Curr Genet*. 1998; 33:16–20. [PubMed: 9472075]
- Smith M, Fahrmeir L. Spatial Bayesian Variable Selection With Application to Functional Magnetic Resonance Imaging. *Journal of the American Statistical Association*. 2007; 102:417–431.
- Smith D, Smith M. Estimation of Binary Markov Random Fields Using Markov Chain Monte Carlo. *Journal of Computational and Graphical Statistics*. 2006; 15:207–227.
- Spiegelhalter DJ, Best NG, Carlin BP, Van der Linde A. Bayesian Measures of Model Complexity and Fit (with Discussion). *Journal of the Royal Statistical Society, Series B*. 2002; 64(4):583–616.
- Spiegelhalter D, Thomas A, Best N, Lunn D. 2003WinBUGS User Manual, Version 1.4. Available at <http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf>
- Wei P, Pan W. Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. *Bioinformatics*. 2008; 24(3):404–441. [PubMed: 18083717]
- Wei Z, Li H. A Markov Random Field Model for Network-based Analysis of Genomic Data. *Bioinformatics*. 2007; 23:1537–1544. [PubMed: 17483504]
- Wei Z, Li H. A Hidden Spatial-temporal Markov Random Field Model for Network-based Analysis of Time Course Gene Expression Data. *Annals of Applied Statistics*. 2008; 2(1):408–429.
- Winkler, G. Image Analysis, Random Fields and Markov Chain Monte Carlo Methods. Springer-Verlag New York, Inc; 2006.
- Wu WB. On false discovery control under dependence. *Annals of Statistics*. 2008; 36(1):364–380.

A APPENDIX

A.1 Bayesian Model specifications for three-component standard and MRF-based mixture models

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

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

$$\Pr(T_i = j) = \pi_{ij} = \frac{\exp(x_{ij})}{\exp(x_{i0}) + \exp(x_{i1}) + \exp(x_{i2})}, \quad x_{ij} | x_{(-i)j} \sim N\left(\frac{1}{m_i} \sum_{l \in \partial_i} x_{lj}, \frac{1}{\tau_{C_j} m_i}\right), \quad \tau_{C_j} \sim \text{Gamma}(0.01, 0.01),$$

where $\tau_{C_j} = 1/\sigma_{C_j}^2$, i is the index set of the direct neighbors of gene i , $m_i = |\partial_i|$, and $\sum_{i=1}^G x_{ij} = 0$ for GMRF-MM with the zero constraint, while for GMRF-MM with the logit constraint, $\sum_{i=1}^G x_{i0} = 0$, $\sum_{i=1}^G x_{i2} = 0$, $\sum_{i=1}^G x_{i1} = c$, 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.

A.2 MCMC Algorithm

We denote by $(\alpha | \dots)$ 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

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

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

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

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

- update T_i by Gibbs sampling with proposal given by

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

where

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

- 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

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

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

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

where

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

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

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

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

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

A.3 Bayes estimators: MAP and MMP

Given data $\mathbf{z} = (z_1, z_2, \dots, z_G)$ and parameter $\mathbf{T} = (T_1, T_2, \dots, T_G) \in \{0, 1\}^G$, 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 $L_1(\mathbf{T}, \hat{\mathbf{T}}(\mathbf{z})) = I(\mathbf{T} \neq \hat{\mathbf{T}}(\mathbf{z}))$, where $I(\cdot)$ is an indicator function. And mis-classification rate loss is

defined as $L_2(\mathbf{T}, \hat{\mathbf{T}}(\mathbf{z})) = \frac{1}{G} \sum_{i=1}^G I(T_i \neq \hat{T}_i(\mathbf{z}))$.

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

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

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 $\hat{\mathbf{T}} = (\hat{T}_1, \dots, \hat{T}_G)$ is the Bayes estimator for the mis-classification rate loss, where $\hat{T}_i = \arg \max_{t_i \in \{0,1\}} \Pr(T_i = t_i | \mathbf{z})$.

Figure 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.

Figure 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).

Figure 3. Estimated vs realized FDR's for (a)–(e) GCN4 ChIP-chip data, and (f)–(j) simulated data (averaged across 20 data sets).

Table 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 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

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)	$\beta = 2.04, \gamma_0 = 2.39, \gamma_1 = -.02$

Table 4

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

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