STATISTICAL METHODS IN GENOME SEQUENCE ANALYSIS

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

XIAOXIAO KONG

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Advised by

Cavan S. Reilly, Ph.D.

September, 2011

Acknowledgments

So many encouraged and helped me through the most joyful and challenging period of my life in the past three years to create this dissertation. There are so many to thank.

First and foremost I want to thank and show my deepest respect to Dr. Cavan Reilly, my doctoral advisor, who always inspires himself and his students keeping the dream of making big contributions to the Biostatistics field. Without his insight, persistent and unrelenting support and guidance, this dissertation would not have been possible. I shall never forgot his encouragement and patience that helped me through when I struggled to create something new to conquer the problems. I shall never forgot these joyful discussions with him. He is one of the rare advisors that students dream that they will find.

I want to thank my doctoral committee–Dr. James Hodges, Dr. Wei Pan, Dr. Baolin Wu, Dr. Hui Zou–who I appreciate for their time to read my thesis and provide me with invaluable advice at various stages. Without their support, I could not have done what I was able to do.

I would also like to give special thanks to Dr. James Hodges, who made (maybe not on purpose) me believe that I do have some potential which provides me the confidence of never giving up in conquering all these difficulties during my research. He very carefully read my dissertation and provide so many critical and valuable comments and suggestions. Not only does he build a role model, but also shows me how to be a good Biostatistician by honestly pointing out my weakness.

I would also like to give special thanks to Dr. Bradley Carlin, who helped me through the life difficulty during the past 4 months. Without his understanding, effort and support, I could not done what I was able to do so far.

I would also like to give special thanks to Dr. Wei Pan, Dr. Hui Zou, Dr. Saonli Basu, Dr. Na Li, Dr. Baolin Wu, Dr. Sudipto Banerjee and Dr. Weihong Tang, who opened the door to Biostatistics research to me!

Special thanks are also extended to the Division of Biostatistics, University of Minnesota for the financial support and the Graduate School, University of Minnesota for the Doctoral Dissertation Fellowship, which allowed me to devote full-time effort to the research and writing of the dissertation in the last year of my Ph.D. study.

I would also like to extend my thanks to all other faculty members, staff and fellow students in the Biostatistics Division. The past five years studying in University of Minnesota has been the one of most enjoyable experience in my life.

Most importantly, thank you mom, dad, Min for love, trust and patience!

Abstract

Mass spectral data alignment study

The first part of this thesis deals with the need to align spectra to correct for mass-to-charge experimental variation in clinical applications of mass spectrometry (MS). Proteomics is the large-scale study of proteins. The term "proteomics" was first coined in 1997 to make an analogy with genomics, the study of genes. Most MS-based proteomic data analysis methods involve a two-step approach, identify peaks first and then do the alignment and statistical inference on these identified peaks only. However, the peak identification step relies on prior information on the proteins of interest or a peak detection model, both of which are subject to error. Also numerous additional features such as peak shape and peak width are lost in simple peak detection, and these are informative for correcting mass variation in the alignment step. Here we present a novel Bayesian approach to align the complete spectra. The approach is based on a parametric model which assumes the spectrum and alignment function are Gaussian processes, but the alignment function is monotone. We show how to use the expectation-maximization algorithm to find the posterior mode of the set of alignment functions and the mean spectrum for a patient population. After alignment, we conduct tests while controlling for error attributable to multiple comparisons on the level of the peaks identified from the absolute mean spectra difference of two patient populations.

Motif discovery study

In the second part of this thesis we show how to reformulate the usual model-based approach to motif detection as a conditional log-linear model and how this reformulation of the problem allows one to use the lasso to build complex dependency structures into the motif probability model in a fashion that is not overparameterized. We illustrate the performance of the approach with a set of simulations and show that it can dramatically outperform existing methods when there is dependence in the motif and is comparable in cases where there is no dependence. By not marginalizing out the parameters that govern the probability distribution of the motif (as is usually done), we can characterize the motif in a more rigorous fashion.

In the final part of the thesis we describe how to incorporate the Bayesian group lasso, the Bayesian adaptive lasso, and the Bayesian group adaptive lasso into conditional log-linear modeling for motif discovery. If an explanatory factor is represented by a group of derived input variables, the lasso tends to select individual derived input variables from the grouped variables, while the group lasso could overcome this difficulty and still do variable selection at the group level. Also the lasso shrinkage produces biased estimates for the large coefficients, while the adaptive group lasso can overcome this difficulty and maintain the oracle property. Finally the group adaptive lasso enjoys both the advantage of the group lasso and the adaptive lasso.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Statistical methods in mass spectrometry analysis

Proteomics is the large-scale study of proteins, particularly their structure and function. The term "proteomics" was first coined in 1997 to make an analogy with genomics, the study of genes. Proteomics has the potential to have a profoundly positive impact on biology and medicine, especially clinical diagnostics and drug discovery.

Mass spectrometric based proteome profiling is one current active area in proteomics. Mass spectrometry (MS) is an analytical technique for the determination of the protein composition of a biological sample. It is being used intensively and increasingly to identify proteins for early diagnosis, prognosis, monitoring disease progression and response to treatment, or to identify which patients are most likely to benefit from particular treatments.

Matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) MS is the

most commonly used MS instrument. In MALDI-TOF MS, a sample is mixed with a crystal forming matrix, placed on an inert metal target, and struck by a pulsed laser to produce gas phase ions. This process takes place in an electric field, which accelerates the ions into a field-free flight tube where they are separated by their mass-over-charge ratios (m/z) and drift until they strike a detector that records the time of flight. Systematic sources of variability are often present in the MALDI-TOF MS. Within a set of measured mass spectra, m/z axes can be variously shifted, compressed, and expanded, in complex, non-linear ways by approximately 0.1% to 0.2% of the m/z range. Therefore alignment of spectra within datasets is often required such that signals corresponding to the same proteins share the same m/z values. This matching is done either on the original intensity levels in so called profile-based methods or on identified peaks in feature-based methods. In feature-based alignment methods, the peak extraction step relies on prior information on the proteins of interest or a peak detection model, which is error-prone. In profile-based alignment methods, the spectra comparison is based on the original intensity values and numerous additional features such as peak shape and peak width, which are lost in simple peak detection, are available for correcting technical distortions in the alignment phase. Thus profile-based alignment methods have the potential to improve the sensitivity for detecting proteins that can be used to distinguish between patient populations. Currently available profile-based alignments either are not flexible enough to capture complex m/z distortions or require manual specification of important model parameters, which is subjective and may induce bias.

Here we focus on proposing a profile-based method which could quickly and effectively increase the correlation between each patient spectrum and the corresponding mean spectrum after alignment, which circumvents the limitations of current alignment methods, and supplies researchers with a tool to exploit the full potential of MS data.

1.2 Statistical methods in motif discovery

A number of problems in computational biology can be cast as attempting to identify examples of some known sequence feature. One widely studied problem in this context is the identification of transcription factor binding sites (TFBSs). A gene is a heredity unit in DNA that influences a particular characteristic in an organism. Gene expression is the process whereby a gene is transcribed into RNA and then with RNA as a transient template translated to make proteins, the basic building blocks of cellular life. Gene expression begins when certain proteins, called transcription factors (TFs), bind to specific DNA subsequences, known as transcription factor binding sites. TFs control gene expression by promoting or blocking the recruitment of RNA polymerase II (the enzyme which activates the transcription of genetic information from DNA to RNA) to genes. Characterizing and locating the TFBSs are crucial tasks in molecular biology for understanding how the cell regulates its genes to accomplish its tasks, such as response to developmental and environmental changes. TFBSs are short DNA segments, typically about 5 to 20 base pairs long. There is substantial variability in TFBSs that a given TF can bind to. For example, the ROX1 TF is able to bind to the TFBSs with sequences

CCCATTGTTCTC, CCAATTGTTTTG, and CTCATTGTTGTC (here the letters indicate different bases in a DNA sequences). The nature of the variability itself is not well understood. More than one hundred computational tools have been developed for TFBS discovery. For all of them, the key step is to detect one or more groups of oligonucleotides similar enough to each other (i.e., differing in some nucleotide substitutions, insertions or deletions (indels)) in the sequences. Two main approaches have been proposed for this so far: non-model based (pattern-driven) and model based (alignment-driven) methods. Most model-based algorithms ignore the dependency structure of positions within a TFBS. The pattern-driven methods search for rigid TFBSs whose instances vary only by substitutions, and not by indels. Real TFBSs are flexible, and do vary by indels. This partially explains why so many associated computational tools have already been developed while their success in detecting TFBSs is still limited. Besides improving the algorithms themselves, incorporating additional biological information into the current methods might be useful in guiding algorithms to achieve better sensitivity.

Here we focus on modeling the dependency structure with a motif, in order to more accurately capture the nature of the variability itself of motif patterns.

1.3 Summary

The rest of the dissertation addresses these statistical problems in detail and is organized as follows. In Chapter 2 we developd a novel profile-based alignment method, using an explicit probability model in conjunction with a Bayesian inferential approach. In

Chapter 3 we reformulated the product-multinomial model used for motif detection in most model-based methods as a conditional log-linear model, which allows us to use tools for linear models for modeling the dependency structure within a motif, in order to more accurately capture the nature of the variability of motif patterns. We use the lasso to build complex dependency structures into the motif probability model in a fashion that is not overparameterized. In Chapter 4, we described how to incorporate Bayesian generalizations of the lasso, the Bayesian group lasso, the Bayesian adaptive lasso and the Bayesian group adaptive lasso, into a conditional log-linear model for motif discovery.

Chapter 2

A Bayesian approach to the alignment of mass spectra

2.1 Introduction

One approach to the characterization of the proteome of a tissue sample from an organism is to use the mass spectrum of the tissue sample. In mass spectrometry (MS), a biological sample is fragmented into ions, and the paired mass-to-charge ratio (m/z) versus the intensity of the resulting set of ions is measured. Several technical approaches to obtaining a representation of the spectrum are in common use, such as matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) MS, and surface-assisted laser desorption-ionization time-of-flight (SALDI-TOF) MS, but all these approaches involve the alignment of the m/z axis to correct for experimental variations, if one wants to

compare the spectra from different tissues. What these different approaches have in common is that they produce a set of intensity measurements for m/z values, or since we will present examples using the MALDI-TOF technology which usually produces singly charged ions, we can just think of this as a set of intensity measurements as they depend on mass. One common goal is to compare mean spectra across different patient populations to identify biomarkers.

All alignment approaches fall into one of two very broad categories (Vandenbogaert *et al.* (2008)): they are based either on peak data, where the data have been processed to identify important signals (peaks) and distinguish these signals from noise, or on profile data, where the MS spectra are taken as recorded. We will now briefly describe each of these general strategies and discuss several implementations of each strategy.

2.1.1 The peak-based alignment approach

In the peak-based approach, one tries to distinguish between relevant signals from peptides and irrelevant noise in the data in an initial peak detection step, and then relies only on these peaks for the subsequent analysis. In the peak detection step, one applies a number of data preprocessing steps. Data preprocessing is considered a necessary step before characterization of the proteome of a set of samples for all applications of MS due to the noise in the measured intensity and also the slight variation in the measured locations of the peaks. This data preprocessing typically consists of the following steps: baseline correction, smoothing, intensity normalization, peak identification, and peak alignment. Following this data preprocessing, analysis of different patient populations

can be carried out by using the detected and aligned peaks. Peak detection typically excludes a large amount of noise and thereby reduces the data considerably. However, the performance of the overall alignment process strongly depends on the performance of the peak detection step. Peak-based alignment methods typically only take m/z variation of aligned peaks into account; only a few of them also use intensity. Several peak-based alignment methods have been proposed, as we now discuss.

Tibshirani *et al.* (2004) aligned peaks extracted from smoothed raw spectra via a dendrogram constructed using agglomerative hierarchical clustering with complete linkage. The idea is that tight clusters should represent the same biological peak that has been horizontally shifted in different spectra, thus one can extract the centroid of each cluster to represent the consensus position for that peak across all spectra. A peak in an individual spectrum is deemed to be one of the common peaks (cluster) if its center is close enough to the estimated central position of the common peak. The assumption that motivates such a procedure is true only if global or nonlinear local shifts in the m/z scale are uniformly small enough so that no peak deviates far from the mean spectrum of the population. Finally, selection of the number of clusters (i.e., the dendrogram cut off value) is in general difficult to determine.

Yu *et al.* (2006) proposed to address the multiple peak alignment problem via a sequential approach based on Gaussian scale-space theory. They assume that multiple sets of detected peaks are the observed samples of a set of common peaks. They convert the problem of estimating locations of the unknown number of common peaks from

multiple sets of detected peaks into the much simpler problem of searching for local maxima in the scale-space representation. The optimization of the scale parameter σ is achieved through minimizing the energy function $E(\sigma) = D(\sigma)+\lambda R(\sigma)$. The data-fitting term $D(\sigma)$ is the sum-of-square-differences between the sample peaks and the scale-space-based representation. The regularization term $R(\sigma) = 1 - \exp\{-(\sigma - \mu_s)^2/(2\sigma_s^2)\}$ prevents overfitting (too small σ) and oversmoothing (too large σ). The parameter λ allows the user to adjust the relative importance between these two terms. From the Bayesian point of view, the data-fitting term is the log likelihood, the regularization term is the log prior. They use the m/z distances between neighboring peaks in all samples to estimate σ_s and μ_s in the regularization term. In their simulations, the energy function achieved the same minimum, namely around $\hat{\sigma} = \mu_s$, for all the λ values they tried: 0.5, 1, 1.5, and 2.0. They claimed that the value of the relative importance coefficient λ does not have much influence on the minimum energy and simply set $\lambda = 1$. The above claim is true if and only if $R(\hat{\sigma}) = 0$, which corresponds to $\hat{\sigma} = \mu_s$. The μ_s happened to be a good estimate for σ in their simulations, but this will not generally be the case. Thus how to set the value of λ is still an unsolved problem. In the next step, they obtain the number and locations of the common peaks at the estimated scale level by searching for a local maximum, and align peaks with respect to the common peaks using a closest point matching method (i.e., for every common peak, its counterpart in a sample has the smallest distance among all peaks in the same sample), which is valid only if the majority of peaks locate close to the true locations with only a few outliers.

The fundamental assumption of the above Gaussian scale-space approach is that the locations of the observed peaks follow symmetric unimodal distributions (e.g., normal distributions) with their means equal to the corresponding locations of the common peaks and variances reflecting the width of the peak, but that strongly depends on whether or not the preprocessing steps deal well with the frequently-observed global and local nonlinear shifts of peaks.

2.1.2 The profile-based alignment approach

Alternatively, one can attempt to compare the complete MS spectra directly. Methods that use this approach usually attempt to find an alignment such that the overall difference between intensities of all MS spectra and the reference spectrum (e.g., the mean spectrum or any spectrum among the studied ones) after alignment is minimized.

Some try to find a polynomial warping function that applies to all m/z values using least squares, as proposed by Eilers (2004), which we call the parametric time warping (PTW) method. Typically, quadratic functions are used in practice, which are not flexible enough to capture non-quadratic m/z distortion, thereby introducing bias.

Listgarten *et al.* (2005) proposed what they call the continuous profile model (CPM). This model employs a hidden Markov model based approach, in which each observed mass spectrum is a non-uniformly subsampled version of a latent spectrum, to which local rescaling and additive noise are applied. They estimated the latent spectrum using the expectation-maximization (EM) algorithm, then align each observed spectrum to the latent spectrum using the Viterbi algorithm, a dynamic programming algorithm

for finding the most likely sequence of hidden states. One potential advantage of this approach is the consideration of the intensity information in the alignment, and another is to combine all the preprocessing, peak detection and peak alignment steps into one computation. On the other hand, one must manually set many parameters, and this approach only works for extracting the latent spectrum for one sample by aligning multiple technical replicates. The implicit assumption behind this approach is that there is one-to-one correspondence among peaks across sequences to be aligned (Yu *et al.* (2006)). This is usually not true for samples from different persons. Thus this method can not be directly applied to clinical datasets which usually have no technical replicates for each sample. In addition, it allows very large cumulative shifts, which can potentially lead to alignment of unrelated peaks thereby creating artificial features. This method is successful in its original context, speech recognition, to align speech energy signal spectra with different speaking speeds from one person.

Here we develop a profile-based alignment approach, using an explicit probability model in conjunction with a Bayesian inferential approach. The approach to alignment described here builds on a previously developed method for curve registration by Reilly *et al.* (2004). Not only does this approach avoid the peak selection step of peak-based methods, it strikes a balance between the profile-based methods described above in that, while being far more flexible than quadratic time warping method, it avoids the pitfalls of methods that require manual specification of many parameters.

2.2 Methods

2.2.1 Baseline subtraction and normalization

The quantitative intensity measurements output by most MALDI-TOF mass spectrometers already have the baseline subtracted from the intensity by the machine's internal algorithm (Yasui *et al.* (2006)). Thus further baseline subtraction will not be considered here.

Due to variation in sample preparation and deposition on the target, matrix crystallization, and ion detection, overall abundance measurements often vary by at least 4-fold for different samples (Wu *et al.* (2004)). The proposed alignment method depends on the spectra's abundance, therefore abundance variations need to be minimized. Many normalization methods have been proposed to solve this problem. Here we simply apply one scaling factor to each MS run so that area under the spectrum is the same for all the samples (Wu *et al.* (2004)). After this we take the logarithms of the abundance to reduce the variance's dependence on the mean level.

2.2.2 Alignment in the raw TOF measurements instead of the calibrated m/z values

MALDI-TOF MS is a technique in which a co-precipitate of a UV-light absorbing matrix and a biomolecule is irradiated by a nanosecond laser pulse. Biomolecule i is turned into gas phase ions with initial velocity v_{0i} (the initial velocity approximately follows a normal distribution), which then are accelerated in an electric field with voltage U and enter a

field-free flight tube. During the flight in this tube, different ionized biomolecules are separated according to their mass to charge ratio and reach the detector at different times. The m_i/z_i of an ion can be approximated by a quadratic function of its drift time t_i. Suppose L is the drift length, while β_0, β_1 and β_2 are calibration constants. Then,

$$m_i/z_i = \frac{2U}{(\frac{L}{t_i})^2 - v_{0i}^2} \approx \beta_0 + \beta_1 t_i + \beta_2 t_i^2.$$

The values of m/z in a TOF-MS are determined by estimating β_0, β_1 and β_2 using an internal or external calibration to known m/z values. Typically, the calibration involves measuring a standard sample that contains several species with known m/z values and fitting quadratic curves to the observed versus known m/z values of the several species to estimate the coefficients β_0, β_1 and β_2. The estimates are then used to determine the corresponding m/z value of each TOF point. The errors in this approximation can become fairly large (more than 2%) when the calibration equation is extrapolated beyond the range of masses of the calibrants (Coombes *et al.* (2005)).

Figure 2.1 shows a segment of MS spectra of bronchio-alveolar lavage samples of 7 patients receiving lung transplants who experienced transplant rejections (solid lines) and 5 patients who did not reject for at least 5 years (dashed lines). The x axis of Figure 2.1 (a) is the TOF scale, while the x axis of Figure 2.1 (b) is the calibrated m/z scale. Comparing Figure 2.1 (a) and (b), the proteins are not indexed properly by their calibrated m/z values across multiple spectra. Meanwhile, the calibration potentially introduces bias via misspecification of the quadratic model and introduces variability through the use of the parameter estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$. It may be advantageous

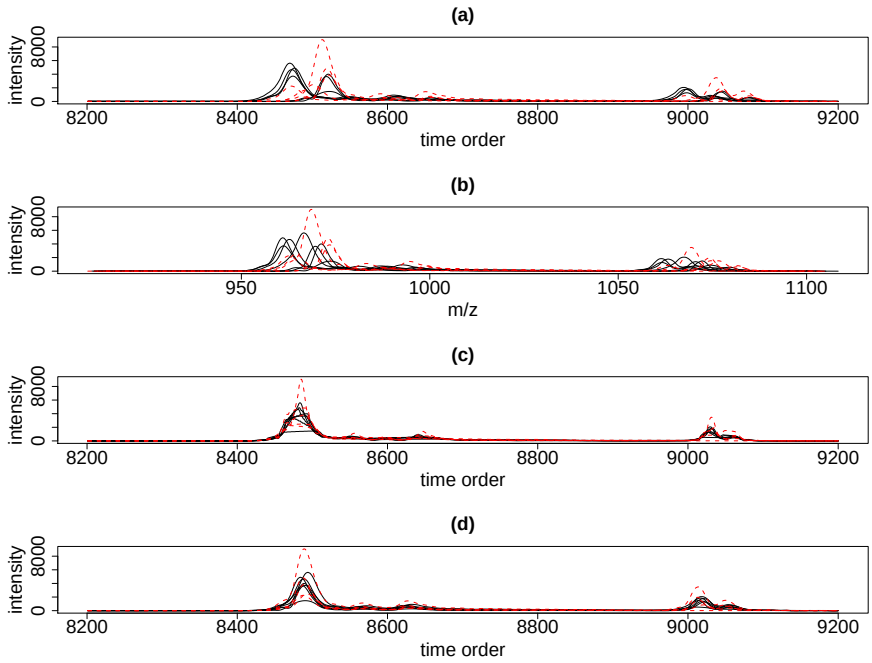

Figure 2.1: A segment of MS spectra of bronchio-alveolar lavage samples of 7 patients receiving lung transplants who experienced transplant rejections (solid lines) and 5 patients who did not reject for at least 5 years (dashed lines). (a) Before alignment, TOF scale. (b) Before alignment, m/z scale. (c) After alignment without the second constraint, TOF scale. (d) After alignment with the second constraint, TOF scale.

to forego calibration and align the spectra using the TOF scale directly. Also the TOF measurements are equal-spaced measurements, which is slightly faster for alignment with respect to computation, as we will note in section 2.2.5.2.

2.2.3 Alignment Model

Let $x_i(t)$ represent the height (intensity on the log scale) of the spectrum for subject i at time t (in the TOF measurement scale). If $\theta(t)$ is the average spectrum for this patient population at t, $\xi_i(t)$ is the deforming function for this subject at t in TOF measurement, and $\epsilon_i(t)$ is random error, then we assume

$$x_i(t) = \theta\left(\xi_i(t)\right) + \epsilon_i(t),$$

for all t. Here we assume these deforming functions are bijective, continuous and strictly monotone increasing, otherwise $\xi_i(t)$ is able to completely erase observed peaks in the spectra. These assumptions are natural since we do not believe the observed spectra are cut up and reassembled versions of the underlying signal; rather, the need for alignment arises because there are slight distortions in the TOF scale. Note that we only observe data on some finite set $\mathcal{T}(t_1, t_2, \ldots, t_T)$ and so $\xi_i(t)$ need only be defined for $t \in \mathcal{T}$, but given that ξ_i is assumed monotone, it has an inverse and so $E[x_i(\xi_i^{-1}(t))] = \theta(t)$. Thus we need to define $x_i(t)$ for all t in the range of $\xi_i^{-1}(t)$. Since $\xi_i(t)$ is also assumed continuous, we need to define $x_i(t)$ for all t. Let E_j for $j = 1, \ldots, T-1$ be the partition of $[t_1 = \min_{t\in\mathcal{T}}, t_T = \max_{t\in\mathcal{T}}]$ defined by the locations of the knots of $\xi_i(t)$ (use the same partition for all i), $t_1, t_2, \ldots, t_T$. So here we define $x_i(t) = \sum_{j=1}^{T-1} x_i(t_j) I_{\{t\in E_j\}}$. In

addition $\theta(t)$ will be assumed constant within each E_j segment.

We parameterize $\xi_i(t)$ as piecewise linear with knots positioned at the locations (or a subset of the locations) where we have observed data. We also restrict the range of $\xi_i(t)$ to be in a finite set with cardinality strictly greater than the number of knots. Hence, there are only finitely many possible $\xi_i(t)$, but since we can increase the cardinality of the range of $\xi_i(t)$, we can obtain an approximation to any desired level of accuracy.

2.2.4 Estimation

We assume that the L_2 norms of $x_i(t) - \theta(\xi_i(t))$ over E_j, $\int_{E_j} [x_i(t) - \theta(\xi_i(t))]^2 \, \mathrm{d}t$, are independently distributed as $\sigma^2 |E_j| \chi_1^2$ (where $|I|$ denotes the length of the interval I). We also assume that the L_2 norms of $\xi_i(t) - t$ over E_j, $\int_{E_j} [\xi_i(t) - t]^2 \, \mathrm{d}t$, are independently distributed as $\tau^2 |E_j| \chi_1^2$. Then the -2 log posterior is given up to a constant by

$$
\begin{aligned}
= \; & -\sum_{i=1}^{n} \sum_{j=1}^{T-1} \Bigg\{ \log(\sqrt{(|E_j|)}\sigma) \\
& + \frac{1}{|E_j|\sigma^2} \Big\{ \int_{E_j} [x_i(t) - \theta(\xi_i(t))]^2 \, \mathrm{d}t + \frac{\sigma^2}{\tau^2} \int_{E_j} [\xi_i(t) - t]^2 \, \mathrm{d}t \Big\} \Bigg\}.
\end{aligned}
$$

Assuming $\frac{\sigma^2}{\tau^2}$ is known, we then find the posterior modes of $\xi_1(t), \xi_2(t), \ldots, \xi_n(t)$ by minimizing

$$
\sum_{i=1}^{n} \sum_{j=1}^{T-1} \frac{1}{|E_j|} \Bigg\{ \int_{E_j} [x_i(t) - \theta(\xi_i(t))]^2 \, \mathrm{d}t + \frac{\sigma^2}{\tau^2} \int_{E_j} [\xi_i(t) - t]^2 \, \mathrm{d}t \Bigg\} \tag{2.1}
$$

subject to

$$
\begin{aligned}
&1) \quad \xi_i(t_j) < \xi_i(t_{j+1}) \\
&2) \quad \text{if} \min\left(x_i(t_{j+1}), x_i(t_j)\right) > r_i(t_j), \\
&\qquad \text{then } |\xi_i(t_{j+1}) - \xi_i(t_j)| = t_{j+1} - t_j \\
&\qquad i = 1, 2, \ldots, n \quad , j = 1, 2, \ldots, T-1.
\end{aligned}
$$

The first constraint will guarantee that $\xi_i(t)$ is strictly monotone increasing. The motivation for the second constraint is that we want to maintain the shape of the peaks during the alignment process. Because our method aligns the spectra point by point, the peaks can be arbitrarily stretched or compressed along TOF axis as long as the objective function is minimized. By this constraint, the distance (along TOF axis) could be conserved between consecutive times at which with both intensities are no less than r_i. Thus the shape of those peaks are conserved. For example, if we believe that most of the biologically meaningful peaks' intensities are greater than the 80^{th} percentile of the local spectrum, then we should set r_i to the 80^{th} percentile of the local spectrum here. The larger r_i is, the less the shape of the spectrum will be conserved. Figure 2.1 (c) shows the spectra after alignment without the second constraint; note that the bell shape of the peaks is lost.

For MALDI-TOF MS, setting r_i as a local percentile instead of a global percentile is necessary as shown in Figure 2.2 where the dashed line represents the measured spectrum of a lung transplant patient, the solid curve displays the 80^{th} percentile of the nearest

400 points in a symmetric window, and the dotted horizontal line is the global 80^{th} percentile. The amplitude of the mass spectrum intensity varies greatly along the TOF scale, so only part of the spectrum (between roughly 10000 to 20000, in time order) has intensities higher than the global 80^{th} percentile. If we set r_i as the global 80^{th} percentile, the proposed algorithm can only guarantee to keep the shapes of these highest peaks during the alignment procedure, while if use the local 80^{th} percentile as r_i, we could keep the shapes of peaks across the whole spectrum.

If we treat θ as missing data and $\xi_i(t)$ as the parameter, we can use the EM algorithm to search for posterior modes. At the $(k+1)^{th}$ iteration, in the E-step, we calculate the expectation of the log likelihood with respect to deforming functions from the k^{th} iteration, $\{\xi_1^{(k)}, \ldots, \xi_n^{(k)}\}$,

$$\begin{aligned} & Q(\xi_1, \ldots, \xi_n; \xi_1^{(k)}, \ldots, \xi_n^{(k)}) \\ = & \sum_{i=1}^{n} \sum_{j=1}^{T-1} \frac{1}{|E_j|} \mathrm{E}_{\{\xi_1^{(k)}, \ldots, \xi_n^{(k)}\}} \left[\int_{E_j} \left[x_i(t) - \theta\left(\xi_i\left(t\right)\right) \right]^2 \mathrm{d}t \right]. \end{aligned}$$

In the M-step, maximize

$$-Q(\xi_1, \ldots, \xi_n; \xi_1^{(k)}, \ldots, \xi_n^{(k)}) - \frac{\sigma^2}{\tau^2} K(\xi_1, \ldots, \xi_n)$$

with respect to $\xi_1, \ldots, \xi_n$, where

$$K(\xi_1, \ldots, \xi_n) = \sum_{i=1}^{n} \sum_{j=1}^{T-1} \int_{E_j} \left[\xi_i\left(t\right) - t \right]^2 \mathrm{d}t.$$

To do so we need only compute the expected value of $\theta(t)$ and $\theta(t)^2$, but given all the $\xi_i(t)^{(k)}$ we can estimate these quantities using $\frac{1}{n}\sum_i x_i(\xi_i^{-1}(t)^{(k)})$. Once these expectations are computed we maximize (minimize) the expected log posterior (expected -2

log posterior) with respect to $\xi_i(t)$. Given the form of the expected log posterior we can maximize it with respect to each $\xi_i(t)$ independently of the rest. These maximization steps are carried out using the dynamic programming (DP) algorithm developed in Reilly *et al.* (2004), but here there is an expectation in the function to optimize that was not present there. The part of the expected log-likelihood relevant for the the M-step is

$$\mathrm{E}_{\{\xi_1^{(k)},\ldots,\xi_n^{(k)}\}}\left[\int_{E_j}\left[x_i(t)-\theta\left(\xi_i\left(t\right)\right)\right]^2\mathrm{d}t\right],$$

whereas in Reilly et al. (2004) there was no expectation. This makes only a slight difference since

$$\begin{aligned}
&\mathrm{E}_{\{\xi_1^{(k)},\ldots,\xi_n^{(k)}\}}\left[\int_{E_j}\left[x_i(t)-\theta\left(\xi_i\left(t\right)\right)\right]^2\mathrm{d}t\right]\\
=\;&\int_{E_j}\Bigg\{x_i(t)^2-2x_i(t)\mathrm{E}_{\{\xi_1^{(k)},\ldots,\xi_n^{(k)}\}}\theta(\xi_i(t))\\
&+\mathrm{E}_{\{\xi_1^{(k)},\ldots,\xi_n^{(k)}\}}\theta(\xi_i(t))^2\Bigg\}\mathrm{d}t
\end{aligned}$$

and given $\xi_i^{(k)}(t)$ we can estimate $\mathrm{E}[\theta(\xi_i(t))]$ and $\mathrm{E}[\theta(\xi_i(t))^2]$ using the corresponding moments of the $x_i(t)$ averaging over i using the current values of the deforming function $\xi_i^{(k)}(t)$ to first deform all curves.

If we treat each of the n observed spectra as the spectrum to which all other spectra are aligned, then we can obtain n initial values for the EM algorithm, all of which are likely to be quite good. We could randomly pick any one as the initial value and this is in theory fine because the initial value should not impact the final solution one obtains. Typically, the EM algorithm obtains convergence within 20 iterations here.

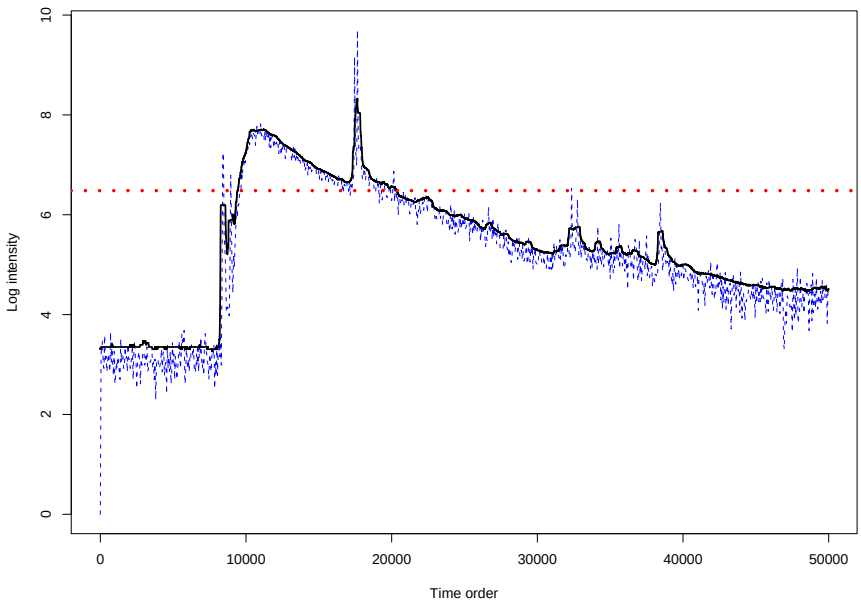

Figure 2.2: The mass spectrum of a lung transplant patient. Dashed line: spectrum on the log scale. Solid curve: 80^{th} percentile of the nearest 400 points in a symmetric window. Dotted line: global 80^{th} percentile.

2.2.5 Computational considerations

2.2.5.1 Rough scale

One aspect of the spectra investigated here that makes the straightforward implementation of the previously outlined computational strategy difficult is the number of observations for each spectrum. On the other hand, we expect $\xi(t)$ to be a slowly varying function, so we conduct our alignment procedure on a much rougher scale than we actually have data. Hence we first obtain a kernel smoothed estimate of the spectrum at every K points (for example $K = 8$), then apply the algorithm to estimate $\xi(t)$ on this rough scale. Then we calibrate the data by defining $\xi(t)$ on the fine scale via linear interpolation.

2.2.5.2 Dynamic programming

We use DP to obtain an approximation to the solution to the objective function (2.1). Our DP algorithm implementation only allows determination of an approximate solution because the values of the range of $\xi(t)$ must be a discrete set, but since this discrete set can approach the set of real numbers the solution can reach any desired level of accuracy.

Here we follow the approach of Reilly *et al.* (2004). The basic idea of the application of the DP algorithm to our problem is to relate the minimized value of the approximation to objective function (2.1) when there are k observations, $\mathcal{F}_k$, to the value of this discrete approximation when there are $k-1$ observations, $\mathcal{F}_{k-1}$.

The TOF observations are equally spaced, and if we label them from 1 to T and

also assume that $x_i(t)$ and $\theta(t)$ are constant within each E_i interval, then the objective function (2.1) could be simplified as

$$\sum_{i=1}^{n}\sum_{j=1}^{T-1}\int_{j}^{j+1}\left\{\left[x_i(t)-\theta\left(\xi_i\left(t\right)\right)\right]^2+\frac{\sigma^2}{\tau^2}\left[\xi_i\left(t\right)-t\right]^2\right\}\mathrm{d}t. \tag{2.2}$$

As we mentioned in section 2.2.2, using the TOF scale and thus using objective function (2.2) instead of (2.1), we avoid calculating the length and related division for each E_j, and also the difference of $t_{j+1}-t_j$, which slightly speeds up the computation. Let $\mathcal{C}_m$ be the set of piecewise linear, strictly increasing functions with nodes at $1,2,\ldots,T$ such that the values at the nodes are of the form $\frac{m+v}{m}$ with v taking value from set $\{0,1,\ldots,m(T-1)\}$ and $m>1$. Suppose deformation function ξ_i maps the kth point of the ith spectrum to the value $\frac{m+l_i}{m}$. To simplify the notation, we let $l=(l_1,l_2,\ldots,l_n)$ be the vector of the deformed values of the kth point. Such a collection of ξ_i is denoted as $\mathcal{C}_{k,l}=\{\xi_i\in\mathcal{C}_m:\xi_i(k)=\frac{m+l_i}{m},i=1,2,\ldots,n\}$, and the approximation to objective function (2.2) with k nodes in each spectrum is

$$\begin{aligned}\mathcal{F}_k(l) \;=\; & \min_{\xi_i\in\mathcal{C}_{k,l}}\sum_{i=1}^{n}\sum_{j=1}^{k}\int_{j}^{j+1}\Big\{\left[x_i(t)-\theta\big(\xi_i\left(t\right)\big)\right]^2\\ & +\frac{\sigma^2}{\tau^2}\left[\xi_i\left(t\right)-t\right]^2\Big\}\mathrm{d}t.\end{aligned}$$

The cardinality of the domain of ξ_i is T, the cardinality of the range of it is $mT-m+1$. Because the only bijective map from one set to another with the same cardinality is the identity, we need to set $m>1$ to guarantee $mT-m+1$ greater than T. As m increases, the gap between possible values of $\xi_i(t)$ decreases, and this gap can be made arbitrarily small by choosing a sufficiently large value of m. For ξ_i to be bijective we

require $\xi_i(1) = 1$ and $\xi_i(T) = T$.

Let $\hat{\xi}_i$ represent the optimal deformation for $\xi_i \in \mathcal{C}_m$. Let $\hat{\xi}_i(k) = \frac{m+l_i}{m}$, which means the optimal deformation function $\hat{\xi}_i$ maps the kth point in the ith spectrum to the value $\frac{m+l_i}{m}$. Let $\hat{\xi}_i(k-1) = \frac{m+s_i}{m}$ and the requirement that the deformation does not tear the predictions implies that $s_i \in [1, l_i)$. The DP principle demands that from any point on an optimal deformation, the remaining deformation is optimal over the remaining number of time interval initiated at that point. Hence

$$\begin{aligned}\mathcal{F}_k(l) &= \min_{s_i \in [1,l_i)} \left\{ \mathcal{F}_{k-1}(s_1, s_2, \ldots, s_n) + \sum_{i=1}^{n} \int_k^{k+1} \Big\{ \left[x_i(j) - \theta\left(\xi_i(j)\right)\right]^2 \right. \\ &\quad \left. + \frac{\sigma^2}{\tau^2} \left[\xi_i(j) - j\right]^2 \Big\} \mathrm{d}t \right\}\end{aligned}$$

for $k = 2, \ldots, T-1$ and $l_i = \lceil 1 + \frac{k-1}{m} \rceil, \ldots, \lfloor T - \frac{T-k}{m} \rfloor$. Therefore the optimal deformation for $\mathcal{F}_k(l)$ could be achieved through this recurrence relation, which takes far less time than considering every possible deformation. ξ_i in this expression is just a function of k, l_i, s_i since the expression entails $\xi_i(k-1) = \frac{m+s_i}{m}$ and $\xi_i(k) = \frac{m+l_i}{m}$.

2.2.5.3 Sakoe-Chiba Band

In reported quality-control experiments (Yasui *et al.* (2006)), observed m/z values fluctuate from experiment to experiment by approximately $\pm 0.1\%$ to $\pm 0.2\%$ of the true m/z value. According to $m/z \approx \beta_2 t^2$, the relative mass error and the relative TOF error satisfy $\frac{\triangle m/z}{m/z} \approx 2\frac{\triangle t}{t}$. The largest expected shift in the TOF scale is about $\pm\frac{1}{2}\max(\frac{\triangle m/z}{m/z})\max(t)$. Thus we expect to see that the deformation function $\xi_i(t)$ looks

like (a) in Figure 2.3 instead of (b). That means the optimal solution path should be near the diagonal area of the plot. Thus the DP algorithm could restrict its searching region within that area, instead of the whole rectangular array. A global constraint, the Sakoe-Chiba band (Sakoe *et al.* (1978)), could be used here. In our context, if we let $b = \frac{1}{2}\max(\frac{\triangle m/z}{m/z})\max(t)$, then the Saloe-Chiba band implies

$$|\xi_i(j) - j| \leq b$$

Suppose $\xi_i(k) = \frac{m+l_i}{m}$. Without using this constraint, the DP algorithm will search l_i from k to $T(m-1)-(T-k)$, $k = 2, \ldots, T-1$ and $i = 1, 2, \ldots, n$. The DP algorithm's complexity is thus $O(mnT^2)$. After using this constraint, the DP algorithm will search l_i from $m[\max(2, k-b)-1]$ to $m[\min(b+k, T-1)-1]$, thus the DP algorithm's complexity in each EM iteration is reduced to $O(bmnT)$. In the MS context, T is length of the measurement sequence, around 50,000 measurement points each spectrum. Thus using the Sakoe-Chiba band here speeds up the computation substantially.

The parameter b, the largest shift allowed, could be estimated based on the mass accuracy ($\frac{\triangle m/z}{m/z}$) reported in the MALDI-TOF instrument manual or quality control experiments. For example, Bruker Biflex III MALDI-TOF Mass Spectrometer measures up to $50,000$ spectrum points, and the dwell time (DW) per data point can be set as 1, 2, 4 and 10 nanosecond (ns). Suppose we measure $50,000$ spectrum points with DW 1 ns. If the mass accuracy, $\frac{\triangle m/z}{m/z}$, is 0.15%, then the maximal allowed shift, b, is $\frac{1}{2} \times 0.15\% \times 50000 = 37.5$ ns. In the algorithm, $\frac{\sigma^2}{\tau^2}$ is the only parameter that needs to be set manually. Here $\frac{\sigma^2}{\tau^2}$ is the ratio of the intensity variance (on the log scale) to

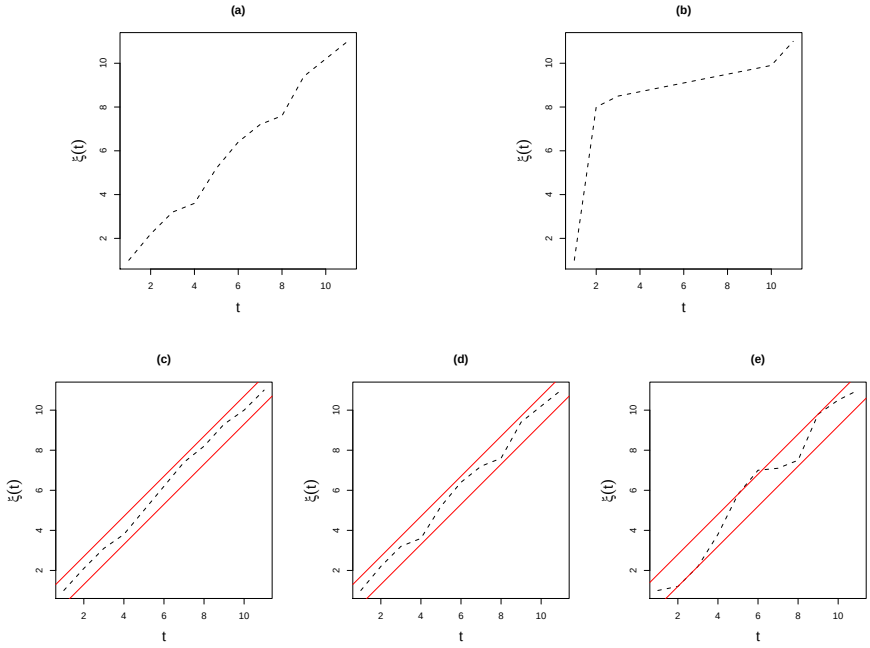

Figure 2.3: Sakoe-Chiba Band. (a) a possible deformation path, (b) a non-allowed deformation path, (c) $\frac{\sigma^2}{\tau^2}$ is too large, (d) $\frac{\sigma^2}{\tau^2}$ is suitable, (e) $\frac{\sigma^2}{\tau^2}$ is too small. In (c), (d) and (e), solid lines define the Sakoe-Chiba band.

the deformation variance. If $\frac{\sigma^2}{\tau^2}$ is relatively large, then only very small deformations (shifts) are allowed. In this case, the optimal solution will look like Figure 2.3 (c), and the deformation will not be large enough to find a suitable alignment. If $\frac{\sigma^2}{\tau^2}$ is relatively small, then very large shifts are allowed. The optimal solution will look like Figure 2.3 (e), and the alignment will entail shifts that are inconsistent with the magnitude of shifts reported in MALDI-TOF quality control experiments. Figure 2.3 (d) shows the ideal situation, where enough deforming is allowed within the predetermined band b. This motivates us to set $\frac{\sigma^2}{\tau^2}$ as the smallest value so that the solution lies within the band defined by the parameter b. In practice, we start from a relatively small $\frac{\sigma^2}{\tau^2}$ value. If one of the solution curves hits the band boundary, the algorithm will stop the alignment procedure, increase $\frac{\sigma^2}{\tau^2}$ value by 10 percent, and then redo the alignment. The algorithm repeats the previous step until the solution lies within the band defined by the parameter b.

2.2.6 Identification of peak features corresponding to differentially expressed proteins while controlling the FDR

After aligning a set of MS spectra, researchers first attempt to compare the aligned mass spectra to identify features (peaks) corresponding to underlying differentially expressed proteins or protein fragments between different patient populations. Suppose the mean spectrum of the first population is $\theta_1(t)$, the mean spectrum of the second population is $\theta_2(t)$, and their absolute difference is $\theta_{\text{diff}}(t) = |\theta_1(t) - \theta_2(t)|$. Clearly our scientific interest, the underlying differentially expressed proteins, corresponds to peaks in $\theta_{\text{diff}}(t)$.

Two natural values to quantify the size of a peak are its height and its area. The area of a peak is a more accurate measure of the corresponding protein's abundance than the height. If there were no variability in the initial velocities, then all ions with the same mass and charge would strike the detector at the same instant, the TOF spectrum would show a spike at the TOF corresponding to each protein, with a height equal to the protein's concentration. In actual MALDI-TOF spectra with the presence of a spread of initial velocities, we observe bell-shaped peaks rather than one-dimensional spikes (Coombes *et al.* (2005), House *et al.* (2006)). Thus our focus is on the area under the peak.

We identify peaks by use of the continuous wavelet transform (CWT) method proposed by Du *et al.* (2006), which allows wavelet transforms at every scale with continuous translation. CWT uses the Mexican hat wavelet as the mother wavelet. The Mexican hat wavelet with scale w best matches peaks with a width of $2w$. Thus for a peak in an MS spectrum, the corresponding CWT coefficient reaches a maximum around the peak center when the scale best matches the peak width. The peak width can be estimated based on the CWT scale corresponding to the maximum point of the CWT coefficient.

Suppose there are n_1 patients from population 1 and n_2 from population 2. The corresponding spectrum of each patient after alignment is $X_{ij}^{\text{aligned}}(t)$, $i = 1, 2$, $j = 1, \ldots, n_i$. The procedure of identifying peak features corresponding to differentially expressed proteins or protein fragments between two patient populations is as follows:

1. After alignment (align all the spectra from different patient populations together),

calculate the mean spectrum of each population, $\hat{\theta}_i(t) = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}^{\text{aligned}}(t)$. And then calculate the absolute mean spectrum difference, $\hat{\theta}_{\text{diff}}(t) = |\hat{\theta}_1(t) - \hat{\theta}_2(t)|$.

2. Identify peaks in the absolute mean spectrum difference $\hat{\theta}_{\text{diff}}(t)$ by CWT method. Record the center location c_k and the scale w_k of each peak from the CWT output, the corresponding peak area is approximately within $[c_k - w_k, c_k + w_k]$, $k = 1, \ldots, m$, where m is the number of identified peaks.

3. For each identified peak, calculate its abundance in each spectrum $X_{ij}^{\text{aligned}}(t)$, $a_{ij}^k = \sum_{l=c_k-w_k}^{c_k+w_k} X_{ij}^{\text{aligned}}(l)$, $i = 1, 2$, $j = 1, \ldots, n_i$, $k = 1, \ldots, m$.

4. For each identified peak, conduct a two sample t-test (or some non-parametric test) to compare a_{1j}^k, $j = 1, \ldots, n_1$ and a_{2j}^k, $j = 1, \ldots, n_2$, and record the corresponding p-value, p_k, $k = 1, \ldots, m$.

5. Apply the Benjamini Hochberg procedure (Benjamini and Hochberg (1995)) to $p_1, \ldots, p_m$. The corresponding ordered p-values are $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$. For controlling the FDR at level α, identify k such that,

$$k = \max\left\{j : p_{(j)} \leq \frac{j}{m}\alpha\right\}.$$

Then report all peaks corresponding to $p_{(j)}$, $j = 1, \ldots, k$ as differentially expressed, which can subsequently be mapped to differentially expressed proteins.

2.3 Examples and results

2.3.1 Example 1: Application to a set of bronchio-alveolar lavage samples

We aligned the bronchio-alveolar lavage sample mass spectra of 12 patients receiving lung transplants. A segment of the spectra after alignment is shown in Figure 2.1 (d). We denote spectrum j in group i as S_j^i. Group 1 includes 7 patients receiving lung transplants who experienced transplant rejections (i.e., bronchiolitis obliterans syndrome (BOS)) and Group 2 includes 5 controls (i.e., lung transplant recipients who did not reject for at least 5 years; most rejections occur by this point). We calculate the square of Pearson's correlation between each sample and the corresponding group mean spectrum before and after the alignment, denoted r^2_{raw} and r^2_{aligned} respectively. The ratios of them, $r^2_{\text{aligned}}/r^2_{\text{raw}}$, range from 1.043 to 1.330 as shown in Table 2.1. The correlation between each patient spectrum and the corresponding mean spectrum after alignment has increased, thereby indicating that a better alignment has been achieved. We also used the PTW method to align the spectra. The PTW Matlab routine, by P. Eilers (Medical Centre, University Leiden, The Netherlands), is available in the warpGUI Matlab program. The ratios of the square of Pearson's correlation between each sample with the corresponding group mean spectrum before and after the alignment using PTW, $r^2_{\text{aligned-PTW}}/r^2_{\text{raw}}$, range from 1.003 to 1.008 as shown in Table 2.1. The superior performance of the proposed method demonstrates by its greater flexibility to capture nonlinear distortion of the TOF (or m/z) axis.

Sample	S_1^1	S_2^1	S_3^1	S_4^1	S_5^1	S_6^1	S_7^1
r^2_{raw}	0.604	0.604	0.673	0.574	0.620	0.657	0.565
r^2_{aligned}	0.773	0.756	0.702	0.733	0.759	0.791	0.654
$r^2_{\text{aligned-PTW}}$	0.609	0.607	0.676	0.578	0.625	0.661	0.567
$r^2_{\text{aligned}}/r^2_{\text{raw}}$	1.280	1.251	1.043	1.278	1.224	1.203	1.158
$r^2_{\text{aligned-PTW}}/r^2_{\text{raw}}$	1.007	1.006	1.004	1.008	1.008	1.006	1.003
Sample	S_1^2	S_2^2	S_3^2	S_4^2	S_5^2		
r^2_{raw}	0.783	0.649	0.695	0.691	0.533		
r^2_{aligned}	0.826	0.792	0.800	0.810	0.709		
$r^2_{\text{aligned-PTW}}$	0.786	0.651	0.699	0.696	0.537		
$r^2_{\text{aligned}}/r^2_{\text{raw}}$	1.055	1.220	1.151	1.172	1.330		
$r^2_{\text{aligned-PTW}}/r^2_{\text{raw}}$	1.003	1.004	1.005	1.006	1.007		

Table 2.1: Comparing the correlation with the mean spectra before and after alignment.

2.3.2 Example 2: Application to a second set of bronchio-alveolar lavage samples

We also have data from another study that used a slightly different protocol for obtaining the spectra using a MALDI-TOF instrument. The protein profile was obtained in the Bruker Biflex III mass spectrometer operating in linear mode with external calibration to the +1 and +2 charge states of Cytochrome C. The mass accuracy in linear mode with external calibration is 1500 ppm ($\frac{\triangle m/z}{m/z} = 0.15\%$). Fifty-thousand spectrum points were measured for each patient. We have 32 chronic lung transplant recipients who ex-

perienced rejections and 47 who did not have a rejection. After alignment, we use the procedure described in section 2.2.6 to identify peak features corresponding to differentially expressed proteins. 93 peaks are identified, and 80 peaks are significant while controlling the FDR at level 0.05. Figure 2.4 shows a segment of the spectra. There are 8 identified peak areas within this segment, of which 7 are significant while controlling the FDR at level 0.05 (Figure 2.4 (c)). The three peaks with the highest intensities are Human neutrophil defensin peptides (HNP) peaks (m/z=3371, 3441 and 3485), and have been shown to be highly expressed in lung transplant recipients with BOS, while infrequently detected in the control population that did not develop BOS (Zhang *et al.* (2005)).

We performed the 1000 permutation tests, each time randomly assign 32 spectra to the disease group and 47 spectra to the control group. 958 out of 1000 permutations do not identify any significant peaks, while we expect to get false discoveries about 5% of the time because we are controlling the FDR at 5%.

We also did sensitivity analysis to investigate how the value of parameter b influenced the peak identification result. Since $b = \frac{1}{2}\max(\frac{\triangle m/z}{m/z})\max(t)$ and the number of measured spectrum points is fixed, it only varies with mass accuracy, $\frac{\triangle m/z}{m/z}$. In our experiment, the mass accuracy varied from 0% (no alignment) to 0.6%, and the results are shown in Table 2.2. When the value of $\frac{\triangle m/z}{m/z}$ used to determine the parameter b departs slightly (0.1% and 0.2%) from the mass accuracy value given by the equipment's manual (here 0.15%), the results show hardly any difference. When the value of $\frac{\triangle m/z}{m/z}$

	0%	0.05%	0.10%	0.15%	0.20%	0.25%	0.30%	0.40%	0.50%	0.60%
0%	78	78	77	77	76	74	70	59	57	52
0.05%	78	78	77	77	76	74	70	59	57	52
0.10%	77	77	78	77	77	74	70	59	58	53
0.15%	77	77	77	80	79	76	72	62	60	55
0.20%	76	76	77	79	80	76	74	65	61	55
0.25%	74	74	74	76	76	78	72	63	60	57
0.30%	70	70	70	72	74	72	81	72	72	69
0.40%	59	59	59	62	65	63	72	80	72	72
0.50%	57	57	58	60	61	60	72	72	82	73
0.60%	52	52	53	55	55	57	69	72	73	92

Table 2.2: Number of commonly identified significant peaks using different $\frac{\triangle m/z}{m/z}$ in the alignment. The diagonal values are the numbers of identified significant peaks when using the corresponding $\frac{\triangle m/z}{m/z}$ value in the alignment. The off-diagonal ones are the numbers of commonly identified significant peaks when using different $\frac{\triangle m/z}{m/z}$ values in the alignment.

is far away (like 0.5% and 0.6%) from 0.15%, some peaks are lost while some artificial peaks are created.

2.4 Discussion

Most MS-based proteomic data analyses use a two-step approach: identify peaks first and then do the alignment and statistical inference on these identified peaks only. However numerous additional features such as peak shape or width are lost in the peak detection

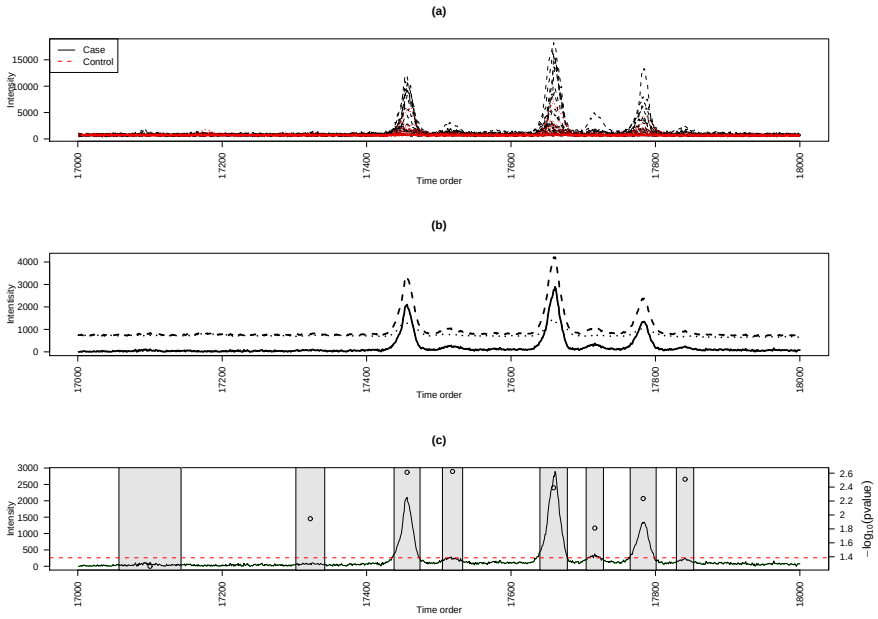

Figure 2.4: A segment of spectra of 32 chronic lung transplant recipients who experienced rejection versus 47 recipients without rejection. (a) Spectra after alignment. Recipients with rejection: dashed curves. Recipients without rejection: dotted curves. (b) Mean spectrum of each group and their absolute difference. Recipients with rejection: dashed curve. Recipients without rejection: dotted curve. The absolute mean spectra difference: solid curve. (c) Identified peaks from the absolute mean spectra difference. The absolute mean spectra difference: solid curve. Identified peaks: shadowed area. p-value (in the minus logarithm to the base 10 scale) of two sample t-test at each identified peak: circle. p-value threshold of controlling FDR at level 0.05: dashed horizontal line. Note that one peak identified in this region was not significant when the FDR is controlled at the 0.05 level.

step, which are informative for correcting TOF axis distortions in the alignment step. Also the performance of the overall alignment process and statistical analysis strongly depends on the performance of the peak detection step. It has been shown that the use of inadequate or ineffective methods in the peak detection step may make it difficult to extract meaningful biological information in the subsequent analysis (Morris *et al.* (2005)). In fact, important differences in low-intensity peaks or on shoulders of peaks can be missed by peak detection algorithms (Morris *et al.* (2007)). An alternative way is to compare the data-rich MALDI-TOF MS spectra directly. A major technical difficulty in the direct comparison is the variation in the TOF axis; only after this alignment can individual features in different experiments be compared directly. Here we present a novel Bayesian alignment approach to align the complete mass spectra. The proposed method is capable of accurately capturing various distortions of the m/z (or TOF) axes and does not require manual specification of important parameters (although we must provide initial values to the EM algorithm). It circumvents the limitations of current alignment methods and supplies researchers a tool to exploit the full potential of MS data.

After the alignment, researchers can directly compare the data-rich MALDI-TOF MS spectra. Pointwise comparison has been proposed (Morris *et al.* (2007), Datta and DePadilla (2006)). However due to the large number (a typical acquisition involves tens of thousands of points) of simultaneous tests, controlling error for multiple comparisons must be addressed in such a situation. Morris *et al.* (2007) computed the pointwise

posterior probabilities of at least a δ-fold intensity change at each spectral location. The threshold on the posterior probabilities for flagging a location as significant is based on setting the expected Bayesian FDR at a prespecified level α. Datta and DePadilla (2006) used marginal p-values obtained from t-tests for testing the intensity differences at each m/z ratio in cancer versus non-cancer samples. They studied the effect of selecting a cutoff FDR on the performance of the clustering and classification algorithms using the significant features.

There is a problem with the above pointwise comparison and controlling error for multiple comparisons procedure (a similar problem arises in neuroimaging, Chumbley and Friston (2009) and Heller *et al.* (2006)). First, the key issue here is that we are not making inferences about points but about the regions corresponding to an underlying differentially expressed protein. This means controlling the false discovery rate on the level of points is not relevant for controlling the FDR when looking for difference in protein expression. The crucial thing to control is the false discovery rate of the features we are making inferences about, which is the peaks identified from the mean spectra difference. Second, though the expected number of falsely discovered points can be controlled, the expected number of falsely discovered peaks can not be controlled at all. Third, dealing with tens of thousands of simultaneous statistical tests requires adjusting the p-values for multiple comparisons, imposing high statistical thresholds that may uncover only the points with the very highest intensity difference but mask others that do differentiate patients populations. For example, in our lung transplant data, if we

control the pointwise false discovery rate at level 0.05, none of the adjusted p-values are significant at all. Here we propose to do comparison and multiple-comparison adjustment on the level of identified peaks in the absolute mean spectra difference. Also detecting peaks from mean spectra is superior to detecting peaks from individual spectra, which has been thoroughly discussed in Morris *et al.* (2005). First, the noise in the mean spectrum decreases by $\sqrt{n}$, where n is the sample size, thus detecting peaks from the mean spectrum is intrinsically more sensitive. Second, small consistent peaks are more easily shown in the mean spectrum. Morris *et al.* (2005) demonstrated that using the mean spectrum will increase the sensitivity of peak detection both in real data examples and simulation studies.

The proposed alignment method is very effective for low mass accuracy data with respect to reaching a better alignment. For data with 5-50 ppm accuracy levels of recent instruments, a better alignment of the spectra is still possible using our approach. This method is still relevant for old datasets and may have value for those using older machines or pushing the limits of mass spectrometry with novel extensions. Additionally, sometimes we need to compare spectra from different laboratories (with machines that have different accuracy) or obtained at different times, which will frequently be severely misaligned. These misalignments will be handled with our method.

Chapter 3

Motif discovery with conditional log-linear models

3.1 Introduction

Nucleotide and amino acid sequences utilized by cells for the same function typically have short subsequences that are nearly common across distinct sequences (we will then use the term *letter* to refer to either a nucleotide or amino acid when we discuss these methods generally). This commonality typically arises due to random divergence of ancestral sequences which have been subjected to selection, although other mechanisms are known that lead to nearly common subsequences residing in different genes or in the same gene in a different organism. This divergence can introduce gaps in one subsequence compared with another and can lead to the substitution of one letter by another. Knowledge of

what these subsequences are and their location can help us understand the function of genes and the regulation of the genome. As experimental identification and verification of such weakly conserved subsequences are very challenging, computational discovery of these weakly conserved subsequences has attracted considerable interest. In particular, the discovery of transcription factor binding sites (TFBS) in DNA has received much attention. A TFBS is a subsequence of a DNA molecule to which proteins bind and thereby regulate the operation of the DNA molecule.

More than a hundred computational tools have been developed for TFBS discovery. Most of these algorithms involve a two-step approach. First, one or more groups of oligos similar enough to each other (i.e. differing in some nucleotide substitutions) are detected in the sequences. Second, an algorithm estimates how likely each group would be to appear in a set of sequences by comparing the number and degree of conservation of the occurrences of each group with "background" expected values that would be obtained by picking at random regulatory regions from the same organism. The most likely groups of oligos found are in turn likely to be motif instances.

For the first step, it is helpful to delineate two main approaches that have been developed for identification of these weakly conserved subsequences: model based (or alignment-driven) and non-model based (or pattern-driven) methods.

In the model-based literature a *motif* is the probability distribution for an ungapped sequence of letters which is believed to be weakly conserved across functionally or structurally related sequences, such as TFBS or functional sites in proteins. In most model-

based algorithms (such as the EM algorithm by Lawrence and Reilly (1990) and the Gibbs sampler by Liu *et al.* (1994)), instances of motifs are assumed to be independent observations from a product multinomial model with width w parameterized by a matrix ϕ where $\phi_{a,b}$ is the probability of finding letter b at location a within an instance. The usual product multinomial model is that the distribution of letters is independent across different positions within the motif, and a Dirichlet prior is typically used for reasons of conjugacy. The matrix ϕ is called the position weight matrix (PWM), or the position-specific weight matrix (PSWM) or even the position-specific scoring matrix (PSSM). The block motif model is based on the idea behind PSSMs and was first developed by Liu (1995). The regulatory element is described by a PSSM and sites outside of the regulatory element are described by a multinomial distribution, but we don't know the start location of the PSSM in each sequence. The goal of the block motif model is to determine the start location of the PSSM in each sequence and determine the probabilities of observing the different letters at the various locations in the PSSM.

A drawback of the simple product multinomial model is that it ignores the dependency structure of positions within an instance, which has been shown to exist for TFBS in several cases where sufficient binding sites data are available for detailed analysis (Bulyk *et al.* (2002), Man and Stormo (2001), Zhou and Liu (2004)). Such dependence is also widely recognized by methods used to detect exon boundaries in the gene detection literature (e.g., Burge and Karlin (1997)) and is typically modeled in that context. As further evidence, Li *et al.* (2006) did an experiment using the program MEME that im-

plements the EM algorithm for the simple product multinomial model described above. In this experiment the authors planted real binding sites into DNA sequences and compared the scores (i.e., the negative log of the p-values of the log likelihood ratio test) of MEME's predictions and the planted binding sites for each dataset from Tompa *et al.* (2005). For most datasets, the predictions of MEME have higher scores than the planted motifs, which provides further evidence that the product multinomial model does not accurately capture the nature of the binding sites and thus can miss the true binding sites in these datasets. Nonetheless, the literature on the product multinomial model has seen many extensions of the basic model that include complex distributions for the non-motif regions of the sequences, inclusion of gaps in the motifs, multiple instances of the motif in a sequence, multiple motifs within each or some sequences and more. The approach proposed here is model based, however we see our contribution as quite distinct from these many refinements of the product-multinomial-model based basic algorithm, and indeed one could incorporate a great many of these extensions into the approach developed here. However we will not do so in our comparisons to existing methods, to highlight the novel aspects of our proposed approach.

Weeder (Pavesi *et al.* (2004)) is currently the most popular non-model based method for finding weakly conserved subsequences. First, the Weeder algorithm enumerates all possible subsequences of (or up to) a given length with an allowed maximum number of substitutions, like one substitution for motifs of length 6, two for length 8, three for length 10 and four for length 12. The choice of the maximum number of substitutions

is made on an *ad hoc* basis, but the previously described scheme is typical of what users specify. Second, it calculates the sequence specificity score for each subsequence according to the number of sequences in which it appears and how well conserved it is in each sequence, with respect to expected values derived from the frequency analysis of all the available upstream sequences of the same organism. Though Weeder can capture the dependency structure of positions within motifs due to the nature of the algorithm, it uses a greedy and heuristic search method, the user must specify a number of parameters with little basis, and the sequence specificity score does not have any rigorous rationale. Moreover, since it is not based on any probability model one can not use the usual tools of statistical inference to assess the subsequences discovered in terms of the confidence in any subsequence the algorithm finds.

The strengths and limitations of the current model-based and non-model-based methods motivate us to first reformulate the usual product multinomial model used for motif detection as a conditional log-linear model. This allows us to use tools for linear models that allow for modeling the dependency structure within a motif. In particular we can use the Bayesian lasso to shrink parameter estimates in a way that potentially prevents overparameterization while allowing for the detection of important interactions. We note that by considering interactions among positions in a motif we are proposing a paradigm shift in the field of motif discovery. For example, the most widely used database for documentation of TFBSs (the TRANSFAC database) is constructed in a manner that assumes there is no interaction among the positions since the database is not equipped

to maintain this information.

Others have also considered dependence in motifs. Barash *et al.* (2003) proposed the use of tree Bayesian networks, mixtures of PSSMs and mixtures of trees as dependency models and showed that these models perform better than PSSM on real sequence data. Most of these models can be cast as special cases of the model proposed here except those involving mixtures. However the mixture-based approaches face fundamental difficulties regarding the specification of the number of components that our approach avoids via the use of the Bayesian lasso. Zhou and Liu (2004) also considered the problem and noted that the approaches of Barash *et al.* (2003) are all overparameterized and involve difficult prior specifications. Due to the overparameterization problem, they modeled a motif using a generalized PWM that could incorporate certain pairwise dependencies (they have constraints on these sets of correlations). The approach proposed here can also model such pairwise correlation, however by using the lasso our method of penalizing the model for including extra parameters is similar to soft thresholding, and the use of soft thresholding is likely superior as in certain contexts it has been shown to be nearly optimal in a minimax sense for a wide variety of loss functions (Donoho *et al.* (1995)). The advantage of the use of a conditional log-linear model is that these models are well understood and allow us to introduce interactions among positions as interactions among variables in a log-linear model.

Moreover we incorporate information about the regulatory regions from the same organism into the Bayesian prior specification of the conditional log-linear model, therefore

our program can simultaneously detect groups of oligos similar enough to each other (i.e. differing in some nucleotide substitutions) in the sequences and estimate how likely each group would be to appear in a set of sequences by comparing the number and degree of conservation of the occurrences of each group with "background" expected values that would be obtained by picking at random regulatory regions from the same organism.

3.2 Model specification

3.2.1 Data structure and notation

First we will lay out the notation we will use for the data and the model parameters. Let R_k denote a single observed biopolymer sequence, where $R_k = (r_{k,1}, \ldots, r_{k,n_k})$ with $r_{k,j}$s as letters. $\mathbf{R}$ denotes a collection of multiple sequences, $R_1, \ldots, R_M$, each written as a row vector, so we can write $\mathbf{R} = (R_1^T, \ldots, R_M^T)^T$. Next, let $h()$ represent the counting function, whose domain is a set of letters and whose range is a vector. The length of this vector is equal to the number of letters and the ith element of this vector is the number of times the ith letter appears in the set of letters on which it is acting. For example, if R is a DNA sequence, $h(R)$ returns a vector of length 4 with counts of each type of nucleotide base (A, C, G, T) in R.

Let $\mathbf{R} = (R_1^T, \ldots, R_M^T)^T$ be a set of M biopolymers believed to share some weakly conserved subsequence. Suppose the length of the conserved motif is w. Here we assume that each biopolymer contains one realization of this weakly conserved motif, and we use a vector $\alpha = \{\alpha_1, \ldots, \alpha_M\}$ to denote the starting location of the motif realization

in each sequence: $\alpha_i = j$ if there is a motif realization starting at location j in R_i, in which $1 \leq \alpha_k \leq n_k - w + 1$. We call α the alignment variable for $\mathbf{R}$. We use $\{\alpha\} = \{\alpha_k + j - 1 : k = 1, \ldots, M, j = 1, \ldots, w\}$ to represent the set of letter indices occupied by the motif elements with alignment variable α. We use $\{\alpha\}^c$ to represent the set of letter indices occupied by elements outside of motif instances with alignment variable α. For any set C of indices, $\mathbf{R}_C$ represents the collection of the letters indexed by elements of C. For example, given any alignment variable α, we have $\mathbf{R}_{\{\alpha\}} = \{r_{k,\alpha_k+j-1} : j = 1, \ldots, w; k = 1, \ldots, M\}$

Letters not included in the conserved motif element are treated as i.i.d. observations from a common multinomial distribution called the non-motif sites model with p categories (equals 4 for DNA or RNA), which can be represented by the probability vector $\theta_0 = (\theta_{1,0}, \ldots, \theta_{p,0})^T$, where $\theta_{1,0} + \ldots + \theta_{p,0} = 1$ and $\theta_{i,0} \geq 0$ for all i. For vectors $v = (v_1, \ldots, v_p)^T$ and $\theta = (\theta_1, \ldots, \theta_p)^T$, we use the following notation: $\theta^v = {\theta_1}^{v_1} \cdots {\theta_p}^{v_p}$ and $\Gamma(v) = \Gamma(v_1) \cdots \Gamma(v_p)$

3.2.2 The conditional log-linear model for motif detection

The basic idea that motivates the current model is that given start locations of the realization of the motif in each of the sequences, we can construct a w-way contingency table using the sequence data in the motif region, whose ith factor is the letter in the ith position and each factor has as many levels as the sequence data has letters. Each sequence contributes a count of 1 to the table depending on what letters are in each of its motif positions. Log-linear models for contingency tables are generalized

linear models using the log link function with a Poisson response, which treat the n cell counts as independent observations (conditional on the parameters of the model) of a Poisson distribution. Log-linear models identify the data as the n cell counts rather than the individual classifications of the M sequences. Log-linear models specify how the expected count depends on the levels of the categorical variables for that cell as well as associations and interactions among those variables. The purpose of log-linear modeling is the analysis of association and interaction patterns. Note that by using a log-linear model we can include other variables as predictors if these variables are observed for each of the sequences. We are not aware of any other approach that allows one to include this information. In applications this would likely involve interactions between the factors representing motif position and the other covariates. For example, if we had a dichotomous variable representing, say species, one could test if the motif differs between species in a rigorous fashion.

Let $y = (y_1, y_2, \ldots, y_n)$ denote the observed counts in the $n = p^w$ cells of this contingency table. Since y is determined by $\mathbf{R}_{\{\alpha\}}$, we will express y as $y(\mathbf{R}_{\{\alpha\}})$ later. We assume the counts $\{y_i\}$ conditional on $\{\alpha\}$ are independent Poisson random variables with parameters μ_i, thus we have

$$y_i|\mu_i \sim \text{Poisson}(\mu_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}, \tag{3.1}$$

where $i = 1, \ldots, n$.

We express the log mean parameters of a log-linear model as

$$\log(\mu_i) = \eta_i = x_i^T \boldsymbol{\beta}, \tag{3.2}$$

where x_i^T is the ith row of the $n \times q$ design matrix X and $\boldsymbol{\beta}$ is a $q \times 1$ vector. The variables investigated by log-linear models are all treated as response variables. In other words, no distinction is made between independent and dependent variables. Each letter in the jth position in a motif element is treated as a level of a categorical variable and can be reparameterized as $p-1$ indicator variables $X^j = (x_2^j, \ldots, x_p^j)$, where X^j is a $n \times (p-1)$ matrix and x_m^j is a length-n vector. There is a computational advantage of using indicator variables to construct the design matrix compared to using other contrasts, like poly-contrasts. This advantage is due to the fact that about 10% of the entries in the resulting design matrix take the value 1 while the rest are 0. Hence when conducting the most time-consuming part of the algorithm, namely computing the expected logarithm of the cell means, η_i, from $x_i^T\boldsymbol{\beta}$, we need only sum up the β_j's corresponding to the entries taking the value 1 in x_i^T, which saves at least 90% of the computational time. Interactions are denoted by the colon operator (:), with up to w-way interactions, the design matrix X is composed as

$$X = [1, \underbrace{X^1, \ldots, X^w}_{\text{main effects}}, \underbrace{X^1 : X^2, \ldots, X^{(w-1)} : X^w}_{\text{first order interactions}}, \ldots, \quad \underbrace{X^1 : \cdots : X^w}_{(w-1) \text{ order interactions}} \quad].$$

Here, $q+1$, the number of columns of X, equals $p^w = n$, which means that X is a square matrix. In practice one would not likely consider every possible interaction. We return to this issue in the context of our application.

3.2.3 The Bayesian lasso

Consider a linear model with q covariates

$$y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where y is the $n \times 1$ vector of responses, $\mathbf{X}$ is the $n \times (q+1)$ matrix of regressors with the first column corresponding to the intercept, $\boldsymbol{\beta}$ is a $(q+1) \times 1$ column vector of $(\beta_0, \beta_1, \ldots, \beta_q)^T$, and $\boldsymbol{\epsilon}$ is the $n \times 1$ column vector of independent and identically distributed normal errors with mean 0 and unknown variance σ^2.

The least absolute shrinkage and selection operator (lasso) method of Tibshirani (1996) estimates linear regression coefficients $(\beta_1, \ldots, \beta_q)^T$ through L_1-penalized least squares estimates. They compute

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (y - \mathbf{X}\boldsymbol{\beta})^T (y - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{k=1}^{q} |\beta_j|,$$

for some $\lambda \geq 0$.

We use the lasso idea to specify priors for the regression coefficients in the conditional log-linear model that we use to specify the probability distribution of the motif. The lasso estimate for linear regression parameters can be interpreted as a Bayesian posterior mode when the regression parameters have independent Laplace (i.e. double-exponential) priors, which is given by the expression

$$p(\boldsymbol{\beta}) = \prod_{j=1}^{q} \frac{\lambda}{2} e^{-\lambda|\beta_j|}. \tag{3.3}$$

The lasso can shrink some coefficients exactly to 0 and hence is often thought of as a method that simultaneously performs variable selection and shrinkage. Though the shape

of the solution path of the Bayesian lasso is very similar to that of lasso (Park and Casella (2008)), these authors don't suggest the use of posterior modes as coefficient estimates. The posterior median is taken to be the point estimate of a coefficient and thus the Bayesian lasso will not shrink coefficient estimates exactly to 0. Therefore the Bayesian lasso cannot perform variable selection in the sense of producing a parameter estimate of 0. This is not of concern here since our interest lies in the posterior distribution of the motif and its start location in each sequence, not the actual estimates of the β_js.

3.2.4 Hierarchical models and full conditionals

Given the alignment variable α, the observed counts y are determined by $\mathbf{R}_{\{\alpha\}}$. We model non-motif sites as draws from the multinomial distribution with parameter $\boldsymbol{\theta_0} = (\theta_{01}, \ldots, \theta_{0p})^T$. Thus the model for non-motif sites is given as follows:

$$h(\mathbf{R}_{\{\alpha\}^c}) \sim \text{Multinomial}(\boldsymbol{\theta_0})$$

$$\boldsymbol{\theta_0} \sim \text{Dirichlet}(\boldsymbol{\gamma_0})$$

where $\boldsymbol{\gamma_0} = (\gamma_{01}, \ldots, \gamma_{0p})$.

The Bayesian lasso analysis of a w-way contingency table is given by the following hierarchial model,

$$y(\mathbf{R}_{\{\alpha\}})|\eta, \alpha \propto \prod_{i=1}^{n} \text{Poisson}(e^{\eta_i}),$$

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta},$$

$$\boldsymbol{\beta} \sim \prod_{j=1}^{q} \frac{\lambda}{2} e^{-\lambda|\beta_j|},$$

$$\lambda \sim \text{Gamma}(a, b)$$

where we have selected a conjugate prior for λ for computational convenience.

We use the uniform prior for the alignment variable α, therefore

$$p(\alpha_k = i) = \frac{1}{n_k - w + 1}$$

where $k = 1, \ldots, N$ and $i = 1, \ldots, n_k - w + 1$, independently of $\boldsymbol{\beta}$, λ.

With the model fully specified we can use the Gibbs sampler to obtain samples from $p(\boldsymbol{\beta}, \alpha, \lambda | \mathbf{R}, a, b)$. First the full conditional distributions of $\boldsymbol{\beta}$ and λ are straightforward to obtain from the joint posterior. For $\boldsymbol{\beta}$ we find that

$$p(\boldsymbol{\beta}|\alpha, y, \lambda, \mathbf{R}) \propto \prod_{i=1}^{n} p(y_i(\mathbf{R}_{\{\alpha\}})|\boldsymbol{\beta}) \prod_{j=1}^{q} p(\boldsymbol{\beta}_j) \tag{3.4}$$

$$\propto \prod_{i=1}^{n} \frac{e^{x_i^T \boldsymbol{\beta} y_i} e^{-e^{x_i^T \boldsymbol{\beta}_i}}}{y_i!} \prod_{j=1}^{q} \frac{\lambda}{2} e^{-\lambda|\beta_j|}, \tag{3.5}$$

where x_i^T is the ith row of $\mathbf{X}$. We use the Metropolis algorithm to sample $\boldsymbol{\beta}$ with a normal proposal distribution whose covariance matrix is a scaled identity matrix, with the scalar selected to achieve an acceptance rate around 0.4.

For λ we find that

$$\begin{aligned} p(\lambda|\alpha, y, \boldsymbol{\beta}, a, b) &\propto \prod_{j=1}^{q} p(\boldsymbol{\beta}_j|\lambda) p(\lambda|a, b) \\ &\propto \prod_{j=1}^{q} \frac{\lambda}{2} e^{-\lambda|\beta_j|} (\lambda)^{a-1} \exp(-\frac{\lambda}{b}) \\ &\propto (\lambda)^{q+a-1} \exp(-\lambda(\frac{1}{b} + \sum_{j=1}^{q} |\beta_j|)) \\ &\propto \text{Gamma}(q + a, (\frac{1}{b} + \sum_{j=1}^{q} |\beta_j|)^{-1}), \end{aligned}$$

hence the full conditional of λ can be sampled using a standard routine. We set the parameters (a, b) with values such as $(1, 1)$. Because of the parameterization, a Gamma(1,1) has large enough support to accommodate non-negative λ.

To sample α, we will sample each of the α_i from their full conditionals. To this end we now describe the strategy we use to sample the α_i. We can write the full conditional of α as

$$p(\alpha|\mathbf{R}, \boldsymbol{\theta}_0, \boldsymbol{\eta}) \propto p(\mathbf{R}, \alpha|\boldsymbol{\theta}_0, \boldsymbol{\eta}) = \boldsymbol{\theta}_0^{h(\mathbf{R}_{\{\alpha\}^c})} \prod_{j=1}^{n} \frac{(e^{\eta_j})^{y_j(\mathbf{R}_{\{\alpha\}})} \exp(-e^{\eta_j})}{y_j(\mathbf{R}_{\{\alpha\}})!}.$$

Let $\alpha_{[-k]}$ denote the set of start locations in all sequences but sequence k. We can integrate out the parameter vector $\boldsymbol{\theta}_0$ to obtain the posterior distribution $p(\alpha_k = i|\alpha_{[-k]}, \mathbf{R}, \boldsymbol{\eta})$ of the start location in sequence k conditional on $\alpha_{[-k]}$ and the current log-linear model parameter $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$. To this end

$$\begin{aligned}
p(\alpha_k = i|\alpha_{[-k]}, \mathbf{R}, \boldsymbol{\eta}) &\propto p(\mathbf{R}, \alpha|\boldsymbol{\eta}) \\
&\propto \int p(\mathbf{R}, \alpha|\boldsymbol{\theta}_0, \boldsymbol{\eta}) f(\boldsymbol{\theta}_0) d\boldsymbol{\theta}_0 \\
&\propto \int \theta_0^{h(\mathbf{R}_{\{\alpha\}^c})+\boldsymbol{\gamma_0}} \prod_{j=1}^{n} \frac{(e^{\eta_j})^{y_j(\alpha_{[-k]}, \alpha_k=i)} \exp(-e^{\eta_j})}{y_j(\alpha_{[-k]}, \alpha_k = i)!} d\boldsymbol{\theta}_0 \\
&\propto \Gamma(h(\mathbf{R}_{\{\alpha\}^c}) + \boldsymbol{\gamma_0}) \prod_{j=1}^{n} \frac{(e^{\eta_j})^{y_j(\alpha_{[-k]}, \alpha_k=i)} \exp(-e^{\eta_j})}{y_j(\alpha_{[-k]}, \alpha_k = i)!}.
\end{aligned}$$

Here we set γ_{0i} as 1 for $i = 1, \ldots, p$. To speed up the program, we calculate the posterior for each location relative to the posterior for the current location $\frac{p(\alpha_k=i|\alpha_{[-k]}, \mathbf{R}, \boldsymbol{\eta})}{p(\alpha_k=i_c|\alpha_{[-k]}, \mathbf{R}, \boldsymbol{\eta})}$, where i_c is the current start location in sequence k. This allows us to compute a quantity

that is proportional to $p(\alpha_k = i|\alpha_{[-k]}, \mathbf{R}, \boldsymbol{\eta})$ for all i. Since there are only finitely many possible values for α_k and we have something proportional to these probabilities we can use the inverse cdf method for simulating draws from the full conditional of α_k.

For a w-way contingency table with p levels for each factor, there are p^w cells. Each cell corresponds to a combination of these w factors $(t_1, t_2, \ldots, t_w)$. Here t_i (for $i = 1, \ldots, w$) takes value from 1 to p. We index these cells as from 1 to p^w according to the formula

$$I(t_1 \ldots t_w) = \sum_{i=1}^{w} (t_i - 1)p^{(i-1)} + 1.$$

When updating the start location in sequence k, suppose that the current and the proposed motif element within this sequence are expressed as the current: $(r_{k,i_c}, \ldots, r_{k,i_c+w-1})$ and the proposed: $(r_{k,i}, \ldots, r_{k,i+w-1})$. There are only two cell counts which will change compared with the current cell counts if we adopt the proposed start location. To be specific, the index of the cell corresponding to $(r_{k,i}, \ldots, r_{k,i+w-1})$ is $I_p = \sum_{j=1}^{w}(r_{k,i+j-1} - 1)4^{(j-1)} + 1$ and the count of that cell will increase one while the index of the cell corresponding to $(r_{k,i_c}, \ldots, r_{k,i_c+w-1})$ is $I_c = \sum_{j=1}^{w}(r_{k,i_c+j-1} - 1)4^{(j-1)} + 1$ and the count of that cell will decrease one. All the other cell counts remain the same.

Suppose y_i^c records the count in the ith cell for the current start location, and y_i^p records the count in the ith cell for the proposed start location. Then $y_{I_p}^p = y_{I_p}^c + 1$, $y_{I_c}^p = y_{I_c}^c - 1$ and $y_i^p = y_i^c$ $(i \neq I_p, I_c)$. Thus

$$\frac{p(a_k = i|\alpha_{[-k]}, \mathbf{R}, \boldsymbol{\eta})}{p(\alpha_k = i_c|\alpha_{[-k]}, \mathbf{R}, \boldsymbol{\eta})}$$

can be simplified to

$$\frac{\Gamma(h(\mathbf{R}_{\{\alpha_{[-k]},\alpha_k=i\}^c})+\boldsymbol{\gamma}_0)}{\Gamma(h(\mathbf{R}_{\{\alpha_{[-k]},\alpha_k=i_c\}^c})+\boldsymbol{\gamma}_0)}\frac{e^{\eta_{I_p}}\cdot y_{I_c}^p}{e^{\eta_{I_c}}\cdot y_{I_p}^c}.$$

We use this formula to iteratively sample through $\alpha_1,\ldots,\alpha_M$.

3.2.5 Selection of initial values and existence of local posterior modes

The Gibbs sampler can become trapped in a local maximum by converging to a location shifted several bases from the true motif start location (Liu *et al.* (1994)). For example, in each of 40 simulated sequences, there exists one and only one motif instance TATAGCTT. We ran the log-linear motif discovery program 9 times with different randomly generated MCMC initial values (motif start locations and regression coefficients). Then 2 chains converged to the true locations; 2 chains converged to locations 1-base-pair shifted from the true locations; 4 chains converged to locations 2-base-pairs shifted from the true locations; 1 chain converged to locations 3-base-pairs shifted from the true locations. This example shows that different MCMC initial values can become stuck in distinct local maxima. Thus we are seeking a way to generate MCMC initial values which lead the chain to converge to the global maximum instead of being trapped at the local maxima.

The idea is that we try to get a guess of the motif start location in each sequence, then based on this guess get the estimates of the regression coefficients in the log-linear model, then put them together as the MCMC initial values. There is a simple and intuitive way to get a rough guess of the motif start locations. We scan the sequence

dataset and find the most conserved w-mers with the simplest scoring function (score is 1 if match and 0 if not), then use the locations of the most conserved w-mers as the initial start locations.

For the example above, we use locations of the most conserved 8-mers as the initial start locations. In the next step, we get the initial value of the regression coefficients. We run the Bayesian log-linear model (the Bayesian log-linear motif discovery model with fixed motif start locations) with randomly generated regression coefficient initial values, and use the posterior mean of the regression coefficients as the initial value for the motif discovery program. In one experiment where we used 9 sets of initial values we found that the use of the above 2 steps lead all 9 chains to converge to the true start locations.

3.3 Simulation studies

For all the simulated sequence datasets, sequences outside of motif instances are generated from the multinomial distribution [A, C, G, T]=[0.25, 0.25, 0.25, 0.25]. For each sequence dataset generated, we ran the Bayesian log-linear model with double exponential prior (log-linear) algorithm allowing all possible two-way interactions and the block motif algorithm for 20000 iterations (burn in the first 15000 iterations) and picked the posterior mode motif configurations to compare.

We compare the algorithms in terms of seven statistics defined in Tompa *et al.* (2005), which are defined at Table (3.1). They are nucleotide-level sensitivity (nSn), nucleotide-level positive predictive value ($nPPV$), nucleotide-level performance coefficient (nPC),

nucleotide-level correlation coefficient (nCC), site-level sensitivity (sSn), site-level positive predictive value ($sPPV$) and site-level average site performance ($sASP$). The sensitivity gives the fraction of known sites (or site nucleotides) that are predicted, and the positive predictive value gives the fraction of predicted sites (or site nucleotides) that are known. Let Q be the set of known binding positions in a sample and let P be the set of predicted positions. Then the performance coefficient is defined as the ratio of $|Q \cap P|$ and $|Q \cup P|$ ($|S|$ measures the size of S). The correlation coefficient nCC is the Pearson product-moment coefficient of correlation between the predicted and the known binding sites nucleotide positions. The value of nCC ranges from -1 (indicating perfect anticorrelation) to $+1$. Thus, if the predicted motifs exactly coincide with the known binding sites, nCC will be $+1$. If each nucleotide position were predicted to be in the motif randomly and independently, then the expected value of nCC would be 0, indicating no correlation.

In our simulation settings, nSn and $nPPV$ are equal, because when the motif width parameter in the algorithm equals the true motif width and there exist one and only one motif instance in each sequence, the number of false negatives equals the number of false positives. Similarly, sSn, $sPPV$ and $sASP$ are equal here. Thus we only present the results of nSn, nPC, nCC and sSn in the following tables.

We are using our own C code for the block motif algorithm in the simulation comparisons. The existing block-motif-based algorithms all have used certain procedures for preprocessing sequences or postprocessing predicted motifs, and some implementa-

nTP	the number of nucleotide positions in both known sites and predicted sites
nFN	the number of nucleotide positions in known sites but not in predicted sites
nFP	the number of nucleotide positions not in known sites but in predicted sites
nTN	the number of nucleotide positions in neither known sites nor predicted sites
site overlap	a predicted site overlaps a known site if they overlap by at least one-quarter the length of the known site
sTP	the number of known sites that overlap predicted sites
sFN	the number of known sites that do not overlap predicted sites
sFP	the number of predicted sites that do not overlap known sites
nSn	$\frac{nTP}{nTP+nFN}$
$nPPV$	$\frac{nTP}{nTP+nFP}$
nPC	$\frac{nTP}{nTP+nFN+nFP}$
nCC	$\frac{nTP \cdot nTN - nFN \cdot nFP}{\sqrt{(nTP+nFN)(nTN+nFP)(nTP+nFP)(nTN+nFN)}}$
sSn	$\frac{sTP}{sTP+sFN}$
$sPPV$	$\frac{sTP}{sTP+sFP}$
$sASP$	$\frac{sSn+sPPV}{2}$

Table 3.1: Definition of seven statistics suggested by Tompa *et al.* (2005) for evaluation of motif discovery algorithms.

tions also have used complicated non-motif site models in place of our i.i.d. multinomial model (such as a third-order Markov model with various prior choices (Liu*et al.* (2001))). Here we would like to focus on comparing models with and without interactions, hence we use an implementation that only differs from our model in terms of the presence of interactions.

3.3.1 Simulation 1: No interactions, no weakly conserved positions

The sequence datasets were generated using motif model (M1) given as follows. The motif width is 8. Each position has probability of 1 having one of the letters (A, C, G, T). Ten sequence datasets were generated from this model, and each dataset contained 40 100-base-pairs long sequences with one motif instance per sequence.

If we use randomly generated initial values, the average performance of the two algorithms for Model M1 is shown in Table 3.2. The block motif algorithm outperforms the log-linear algorithm in both the nucleotide level and the site level, due to the phase-shift issue discussed in Section 3.2.5. If instead we use the initial values generated in the way suggested in Section 3.2.5, then both the algorithms identified the true start locations ($nSn = nPC = nCC = sSn = 1$).

3.3.2 Simulation 2: No interactions, two weakly conserved positions

The sequence datasets were generated using motif model (M2) given as follows. The motif width is 8. Model M2 has 2 weakly conserved independent positions (i.e. the most conserved letter has a frequency less than 0.7) at the first two positions, $(1, 2)$. Other

Models	Algorithms	nSn	nPC	nCC	sSn
M1	log-linear	0.923	0.867	0.916	1.000
	block motif	1.000	1.000	1.000	1.000

Table 3.2: Comparing the performance of the log-linear model with a double exponential prior to the block motif algorithm when there are no interactions between positions and are no weakly conserved positions in the data. This table presents the results averaged over 10 simulated sequence datasets.

positions have probability of 1 having one of the letters (A, C, G, T). Ten sequence datasets were generated for this model, and each dataset contained 40 100-base-pairs long sequences with one motif instance per sequence.

The average performance of the two algorithms for each model is shown in Table 3.3. The log-linear algorithm outperforms the block motif algorithm at both the nucleotide level and the site level.

3.3.3 Simulation 3: One fixed pair of interactions with two weakly conserved positions

The sequence datasets were generated using three motif models (M3-L, M3-M and M3-R) given as follows. The motif width is 8. These models have a fixed pair of correlated positions, $(3, 6)$, which have joint distributions [AA, CC, GG, TT]=$[0.25, 0.25, 0.25, 0.25]$, but 2 weakly conserved independent positions (i.e. the most conserved letter has a frequency less than 0.7). Without considering interaction, either 3 or 6 has marginal

Models	Algorithms	nSn	nPC	nCC	sSn
M2	log-linear	0.927	0.872	0.920	0.988
	block motif	0.809	0.717	0.792	0.910

Table 3.3: Comparing the performance of the log-linear model with a double exponential prior to the block motif algorithm when there are no interactions between positions and two weakly conserved positions in the data. This table presents the results averaged over 10 simulated sequence datasets.

distribution [A, C, G, T]=$[0.25, 0.25, 0.25, 0.25]$, which is the same as non-motif sites distributions. Model M3-L has the 2 weakly conserved independent positions at the first two positions, $(1, 2)$. Model M3-M has the 2 weakly conserved independent positions at the middle two positions, $(4, 5)$. Model M3-R has the 2 weakly conserved independent positions at the last two positions, $(7, 8)$. Ten sequence datasets were generated for each model, and each dataset contained 40 100-base-pairs long sequences with one motif instance per sequence.

The average performance of the two algorithms for each model is shown in Table 3.4. The log-linear model based algorithm outperforms the block motif algorithm for the M3-L, M3-M and M3-R models at both the nucleotide level and the site level.

3.3.4 Simulation summary

The log-linear model can dramatically outperform the block motif model at both the nucleotide level and the site level when there is dependence in the motif and is comparable

Models	Algorithms	nSn	nPC	nCC	sSn
M3-L	log-linear	0.946	0.900	0.942	0.963
	block motif	0.449	0.307	0.401	0.580
M3-M	log-linear	0.980	0.962	0.978	0.980
	block motif	0.533	0.433	0.493	0.595
M3-R	log-linear	0.892	0.810	0.883	0.968
	block motif	0.504	0.405	0.461	0.585

Table 3.4: Comparing the performance of the log-linear model with a double exponential prior to the block motif algorithm. Positions $(3, 6)$ have interactions, which have joint distributions [AA, CC, GG, TT]=$[0.25, 0.25, 0.25, 0.25]$. There are two weakly conserved positions. Model M3-M has the 2 weakly conserved independent positions at the middle two positions, $(4, 5)$. Model M3-L has the 2 weakly conserved independent positions at the first two positions, $(1, 2)$. Model M3-R has the 2 weakly conserved independent positions at the last two positions, $(7, 8)$. This table presents the results averaged over 10 simulated sequence datasets.

in cases where there is no dependence.

3.4 Real TFBSs data analysis

Transcription factor binding sites are located in promoters, which are regulatory regions of DNA that facilitate the transcription of particular genes. A promoter sequence is typically 1000 bp upstream of the gene it regulates (Tompa *et al.* (2005)). The TFBS discovery problem is to identify motifs that are over-represented in promoter sequences of a group of co-regulated genes compared to all promoter sequences in the genome of the same organism (Pavesi *et al.* (2004) and Tompa *et al.* (2005)).

Many TFBS discovery algorithms involve a two-step approach (Pavesi *et al.* (2004)). First, one or more groups of oligos similar enough to each other (i.e. differing in some nucleotide substitutions) are detected in the promoter sequences of a group of co-regulated genes using some algorithm. We call these identified groups of oligos TFBS candidates and refer to this step as the "Identifying TFBS candidates on the gene level" step. In this step, the most conserved motifs within the promoter sequences are identified. Motifs occurring ubiquitously in a genome (e.g. A-rich or T-rich motifs in Saccharomyces cerevisiae) will be identified as top motifs in this step, but are not likely to be relevant to the specific set of genes we considered. Therefore a second step is used to estimate how well a given motif targets the promoter regions of the genes used to find it relative to the promoter regions of all genes in the genome of the same organism, which can be gauged by how much a motif is over-represented in the promoter sequences we consider

compared to the all promoter regions in the genome. We refer to this step as the "Scoring or ranking the degree of over-representation of motifs on the genome level" step. The most over-represented group of oligos found is in turn likely to be a TFBS for the set of co-regulated genes of interest (Pavesi *et al.* (2004), Hughes *et al.* (2000)).

We developed a log-linear model based method for "Identifying TFBS candidates on the gene level" in previous sections if we treat the position posterior mode as the identified TFBS candidate. Though it identifies only one group (posterior mode) of oligos similar enough to each other at each run, we can identify multiple groups of TFBS candidates through an iterative-masking approach: the sites of discovered TFBS candidates are masked out of the sequence dataset and then the log-linear model based algorithm is re-applied to this masked dataset to find additional groups of TFBSs candidates. Masking a certain position in a sequence can be achieved by forcing its probability of being TFBS instance start position to be zero. By masking the positions, it will be impossible to find the same TFBSs candidate twice.

The above iterative-masking approach usually will first identify certain repetitive simple genomic features, such as poly-A subsequences, poly-T subsequences, and their variants (with some nucleotide substitutions from poly-A or poly-T), since they are much more conserved than the TFBS pattern we are looking for. These simple repetitive patterns are common in the genome and reflect some general property of promoter regions of the organism considered. After masking out these simple repeats and their variants, the algorithm can start identifying the target TFBSs. Because the target TFBS is usually

much less conserved than repetitive simple genomic features, the log-linear model based algorithm with all 2-way interactions included sometimes is poorly identified and fails to converge. For example, our MCMC algorithm failed to converge in 1000000 iterations after masking out simple repeats and their variants when we analyzed the yest09r dataset (details about this dataset are available in the later parts of this Section).

Here we propose an approach to incorporate the "Scoring or ranking the degree of over-representation of motifs on the genome level" step into the "Identifying TFBS candidates on the gene level" step. There are two advantages of doing so. First we could identify the target TFBSs in one run instead of through an iterative-masking approach (Roth *et al.* (1998)). Second, the MCMC algorithm appears to converge much faster. The log-linear model including 2-way interactions can be a poorly identified model in the TFBSs discovery context, which slows convergence in MCMC and increases the importance of the prior. The key element of the proposed approach is to add an informative inverse frequency weighted prior for the regression coefficients, which is now described.

3.4.1 The inverse frequency weighted prior

We call a w-mer over-represented in promoter sequences ($S = (s_1, \ldots, s_N)$) of N co-regulated genes if the observed frequency (o^w) of this w-mer in S is larger than its expected frequency (f^w) in promoter sequences (A) of all genes in the genome of the same organism. Suppose we identify one potential w-mer motif (w_i) instance from each s_i and use $f^w(w_i)$ to denote its expected frequency in A. We call this motif over-represented

in S if the average ratio (over-representation score) of the observed against the expected frequency ($\frac{1}{N}\sum_{i=1}^{N}\frac{1/n_w}{f^w(w_i)}$) is larger than 1, where n_w is the total number of w-mers in S.

In our log-linear model based motif discovery algorithm, these N w-mers can be represented as the observed cell counts of a w-way contingency table, which are denoted as $y = (y_1, y_2, \ldots, y_n)$ ($n = p^w$, the number of cells in this contingency table). If we use f_j^w to denote the expected frequency of the jth cell in A, then $\frac{1}{N}\sum_{i=1}^{N}\frac{1/n_w}{f^w(w_i)}$ can be re-expressed as $\frac{1}{n}\sum_{j=1}^{n}\frac{y_j/n_w}{f_j^w}$. We know that the expectation of y_j, $\mathrm{E}[y_j]$, is equal to $e^{(x_j^T\boldsymbol{\beta})}$, therefore the expectation of the over-representation score, $\mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\frac{y_j/n_w}{f_j^w}\right]$, is equal to $\frac{1}{n}\sum_{j=1}^{n}\frac{e^{(x_j^T\boldsymbol{\beta})}/n_w}{f_j^w}$.

A direct way to incorporate the "Scoring or ranking the degree of over-representation of motifs on the genome level" step into the "Identifying TFBS candidates on the gene level" step is through adding the expectation of this over-representation score to the prior for the regression coefficients,

$$\boldsymbol{\beta} \sim \prod_{j=1}^{q}\frac{\lambda}{2}e^{-\lambda|\beta_j|}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{e^{(x_j^T\boldsymbol{\beta})}/n_w}{f_j^w}\right)^{\rho} \tag{3.6}$$

where ρ is a positive-value "tuning" parameter to link the likelihood (which gauges how conserved a motif is on the gene level) and the over-representation score (which measures how specific a motif is on the genome level) onto a comparable scale.

However as this over-representation score prior does not lead to a proper posterior, we propose another prior, which could be interpreted as an approximation to it in the sense

of incorporating the "Scoring or ranking the degree of over-representation of motifs on the genome level" step into the "Identifying TFBS candidates on the gene level" step. In our log-linear model based motif discovery algorithm, let $\{e^{(X\boldsymbol{\beta})_{(i)}}, i = 1, \ldots, n\}$ denote the decreasingly ordered $\{e^{(x_i^T\boldsymbol{\beta})}, i = 1, \ldots, n\}$. Let $f^w_{(i)}$ denote the expected frequency corresponding to $e^{(X\boldsymbol{\beta})_{(i)}}$. We propose to add an inverse frequency weighted component, $\prod_{i=1}^K \frac{e^{(X\boldsymbol{\beta})_{(i)}}}{f^w_{(i)}}$, to the prior for the regression coefficients,

$$\boldsymbol{\beta} \sim \left[\prod_{j=1}^{q} \frac{\lambda}{2} e^{-\lambda|\beta_j|}\right] \left[\prod_{i=1}^{K} \frac{e^{(X\boldsymbol{\beta})_{(i)}}}{f^w_{(i)}}\right]. \tag{3.7}$$

The above prior is an improper informative prior in the sense that it is not always integrable and contains frequency information of w-mers in the promoter regions of all genes in the genome of the organism we considered. For example, when $K = 1$,

$$\begin{aligned} & \prod_{j=1}^{q} \frac{\lambda}{2} e^{-\lambda|\beta_j|} \frac{e^{(X\boldsymbol{\beta})_{(1)}}}{f^w_{(1)}} \\ \geq & \prod_{j=1}^{q} \frac{\lambda}{2} e^{-\lambda|\beta_j|} \frac{e^{(x_i^T\boldsymbol{\beta})}}{f^w_{(1)}} \end{aligned}$$

for all x_i. Therefore

$$\begin{aligned} & \prod_{j=1}^{q} \frac{\lambda}{2} e^{-\lambda|\beta_j|} \frac{e^{(X\boldsymbol{\beta})_{(1)}}}{f^w_{(1)}} \\ \geq & \prod_{j=1}^{q} \frac{\lambda}{2} e^{-\lambda|\beta_j|} \frac{e^{(\beta_0+\beta_1)}}{f^w_{(1)}} \text{for } x_i^T = (1, 1, 0, \ldots, 0) \\ = & e^{\beta_0} \frac{\lambda}{2 f^w_{(1)}} \prod_{j=2}^{q} \frac{\lambda}{2} e^{-\lambda|\beta_j|} e^{(-\lambda|\beta_1|+\beta_1)}. \end{aligned}$$

When $\lambda < 1$, we have $\int_{-\infty}^{+\infty} e^{(-\lambda|\beta_1|+\beta_1)}\mathrm{d}\beta_1 > \int_{0}^{+\infty} e^{(-\lambda|\beta_1|+\beta_1)}\mathrm{d}\beta_1 > +\infty$. Therefore $\prod_{j=1}^{q} \frac{\lambda}{2} e^{-\lambda|\beta_j|} \frac{e^{(X\boldsymbol{\beta})_{(1)}}}{f_{(1)}^{w}}$ is not integrable. But this prior can lead to a proper posterior, of which the proof is provided in the appendix to this chapter.

The frequencies of 6-mers and 8-mers in promoter regions (taken from the regulatory sequence tools RSAT Database, http://rsat.ulb.ac.be/rsat/, Van Helden (2003)) of different organisms have been calculated by the authors of Weeder and made publicly available as frequency tables, which are used as the estimate for f_i^w in our analysis. For example, about 6000 regulatory region sequences were used to calculate tables for yeast, and about 26000 for human and mouse.

3.4.2 An ad hoc approach to selecting K in the inverse frequency weighted prior

The parameter K in Equation (3.7) serves the same purpose as the parameter ρ in Equation (3.6), which links the likelihood (which gauges how conserved a motif is on the gene level) and the over-representation score (which measures low specific a motif is on the genome level) onto a comparable scale. Similar to the shrinkage parameter λ, our model assumes that the value of K is data dependent and cannot be determined in advance before the analysis. The correct value of K should lead to identify the most over-represented motif on the genome level. Therefore, K is selected as the integer which can lead to the maximal posterior over-representation score ($\int \frac{1}{N} \sum_{i=1}^{N} \frac{1/n_w}{f_i(w_i)} p(\alpha|R)\mathrm{d}\alpha$). As shown in Figure 3.1, the analysis with $K = 4$ has the highest posterior over-representation score for both the yst09r and the hm24r dataset (details about these datasets are given in

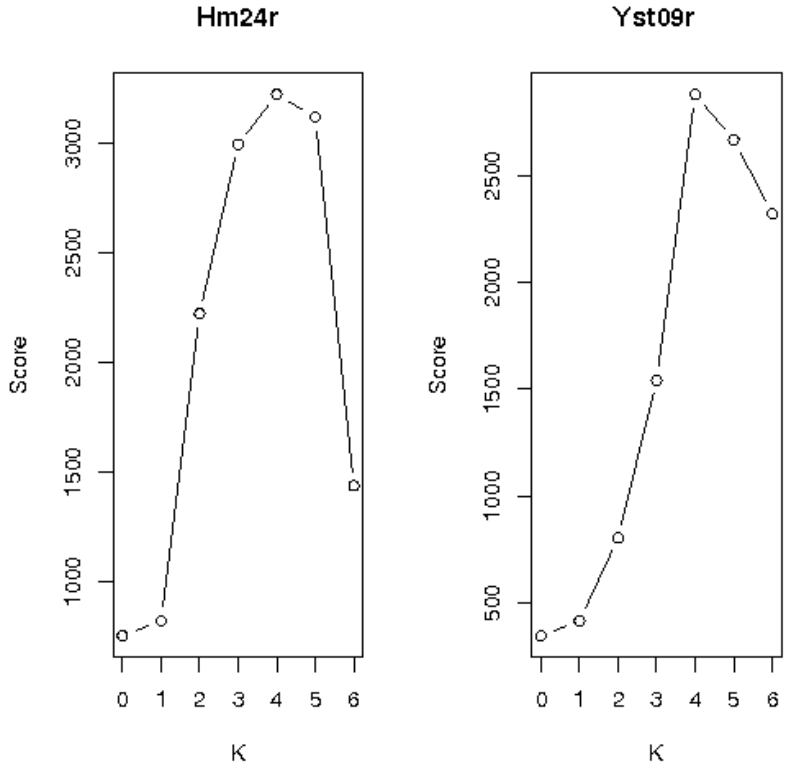

Figure 3.1: Posterior over-representation score as a function of k.

Section 3.4.3).

3.4.3 The Benchmark datasets

To compare the accuracy of motif discovery programs for the specific task of discovering novel transcription factor binding sites in DNA sequences, Tompa *et al.* (2005) created a benchmark of datasets from the TRANSFAC database for assessing current

and future motif discovery tools. TRANSFAC is a database on eukaryotic cis-acting regulatory DNA elements and trans-acting factors. It covers the whole range from yeast to human and contains data on transcription factors, their experimentally-proven binding sites, and regulated genes. Tompa's datasets come from four species: Homo sapiens (human), Mus musculus (mouse), Drosophila melanogaster (fruitfly), and Saccharomyces cerevisiae (yeast). There are three benchmarks which contain the same binding sites, but differ from each other in how the sequences outside of the binding sites were constructed. The "real" benchmark has the binding sites in their real genomic promoter sequences. The "generic" benchmark has the binding sites planted in randomly chosen genomic promoter sequences from the same organism. The "Markov" benchmark has the binding sites planted in sequences randomly generated according to a Markov chain of order 3 that was constructed from the promoter sequences of the same organism. The drawback of using the "real" and "generic" benchmark is that no one knows the complete 'correct' answer: there could be unannotated binding sites, and programs that correctly predict these would necessarily be penalized. The drawback of using the "Markov" benchmark is that no one knows the 'correct' stochastic process that nature uses, and so we would be introducing biases that favor certain tools over others. By providing the above benchmark datasets, Tompa et al assessed the current motif discovery tools. They conclude that Weeder outperformed the other tools in most circumstances and by most measures in that assessment. Many tools perform much better on the yeast dataset than on datasets from other species.

In this section, we analyze two datasets from Tompa's benchmark datasets, yst09r and hm24r to compare the accuracy of the log-linear based motif discovery program with Weeder and MEME. There are three types of possible motif models which could be chosen in Weeder and MEME. The OOPS model is for One Occurrence Per Sequence, which means that each sequence in the dataset contains exactly one occurrence of each motif. The ZOOPS model is for Zero or One Occurrence Per Sequence, which means that each sequence contains at most one occurrence of each motif. The ANR model is for Any Number of Repetitions, which means that each sequence contains any number of non-overlapping occurrences of each motif. Only the OOPS model is available in our current log-linear based motif discovery program. The Weeder algorithm can look for motifs of length from 6 to 12. The MEME algorithm can look for motifs of length from 2 to 300. Our current log-linear based motif discovery program can look for motifs of up to length 10 if all two-way interactions are included (using a computer having at least 3.87 Gb memory). Weeder and MEME can search for motifs on the given DNA strand only or on both the given DNA strand and the reverse complement strand. Our current log-linear based motif discovery program can search for motifs on the given DNA strand only, though we can construct the reverse strand and use our program on that.

3.4.4 TFBS detected in human

3.4.4.1 Dataset background

First we test the performance of the proposed method on the hm24r dataset. The hm24r dataset is part of Tompa's "real" human benchmark dataset. The hm24r dataset contains 8 sequences, where each sequence is 500-base-pairs long. Of these binding sites, 6 out of 8 sequences have been experimentally confirmed to contain one or two known motif instances. The start locations, instances and instance length of the known motifs are listed in Table 3.5.

3.4.4.2 Motif discovery

We modeled the motif as 8-mers on the given strand using the OOPS motif model and include both main effects and all two-way interactions in our log-linear model based algorithm. The parameter K in Equation (3.7) was set to 3, which leads to the maximum posterior over-representation score. The initial start locations were generated from uniform distributions and the initial regression coefficients were set to 0. Here we didn't use the method provided in Section 3.2.5 to set up initial values, since that method leads to initial values corresponding to repetitive simple genomic features or their variants, which we want to discount eventually through the inverse frequency weighted prior. Here we ran 2 chains of 500,000 iterations (the first 100,000 iterations are used as a burn in period). The Gelman-Rubin diagnostic method and the associated R statistics (Gelman and Rubin (1992)) were used to assess if the chains had failed to converge. After 100000

Sequence	Start position	Instance	Instance length
1	439	TTTGGCGC	8
2	413	CCCCGCCCCGCGCTCCCC	18
2	452	CTCGTGGCGCCCCAGGG	17
3	NA	NA	NA
4	405	TTTGGCGC	8
5	NA	NA	NA
6	466	TTTCGCGGCAAA	12
6	483	TTTGGCGCGTAA	12
7	451	TTTCGCGCC	9
8	452	TTTCCCGC	8

Table 3.5: The locations of the binding sites of the hm24r dataset, as given by the TRANSFAC database.

iterations, all the R statistics are less than 1.2, which indicates that the chains have not failed to converge. We used Weeder and MEME to analyze the same dataset with the same assumptions (modeling the motif as 8-mers on the given strand using the OOPS model). The results of the three methods are shown in Table 3.6, Figure 3.2, and Figure 3.3. For the hm24r dataset, the log-linear based model with two-way interactions identified 5 known motif instances, Weeder identified 4 known motif instances, while MEME failed to identify any motif instances.

We also ran the Weeder and MEME algorithms with the default parameters. We ran the Weeder algorithm to search for motifs of length 6, 8, 10 and 12 on the given strand using the ZOOPS model and ran the MEME algorithm to search for motifs of length up to 50 on both the given DNA strand and the reverse complement strand. The results, as shown in Table 3.6 and Figure 3.4, are the almost same as with the previous parameter settings.

For the hm24r dataset, 2 out of 8 sequences (Sequence 3 and Sequence 5) do not contain any binding site instances. Both Sequence 2 and Sequence 6 contain 2 instances. Thus the default parameters for Weeder and MEME seem to outperform the previous settings. Hence we see that for this dataset the log-linear model based algorithm with two-way interactions outperforms Weeder and MEME for both parameter settings at both the nucleotide level and the site level, as shown in Table 3.6.

We can examine the dependency structure of this motif from the posterior samples of $\boldsymbol{\beta}$. First we look at the regression coefficients, β_is, with more than 60% of their

Algorithm	Motif model	Strand	Motif length	nSn	$nPPV$	nSp	nPC	nCC	sSn	$sPPV$	$sASP$
Log-linear (Interaction)	OOPS	Given	8	0.435	0.625	0.994	0.345	0.512	0.625	0.625	0.625
Log-linear (Main effect)	OOPS	Given	8	0.348	0.500	0.992	0.258	0.406	0.500	0.500	0.500
Weeder	ZOOPS	Given	6, 8, 10 and 12	0.348	0.400	0.988	0.229	0.359	0.500	0.500	0.500
Weeder	OOPS	Given	6, 8	0.356	0.484	0.992	0.258	0.404	0.500	0.500	0.500
MEME	ZOOPS	Both	up to 50	0.000	0.000	0.973	0.000	-0.025	0.000	0.000	0.000
MEME	OOPS	Given	8	0.000	0.000	0.973	0.000	-0.025	0.000	0.000	0.000

Table 3.6: Assessment scores for the hm24r dataset.

posterior mass being positive. Seven main effects and twenty eight two-way interactions fall in this category. The seven main effects are $\beta_{1(T)}(70\%)$ (this notation means that the letter in Position 1 is T, and 70% of the posterior mass of $\beta_{1(T)}$ is positive), $\beta_{2(T)}(66\%)$, $\beta_{3(T)}(60\%)$, $\beta_{4(G)}(61\%)$, $\beta_{6(C)}(60\%)$, $\beta_{7(G)}(67\%)$, $\beta_{8(C)}(60\%)$, which is exactly the form of the 8-mer from Position 1 to Position 8 except Position 5 of the posterior mode pattern TTTGGCGC, while only 54% of the posterior mass of $\beta_{5(G)}$ is positive. We also examined the set of β_is having more than 75% of their posterior mass being positive. While no main effect falls in this category there are 7 two-way interaction effects falling in this category as listed in Table 3.7.

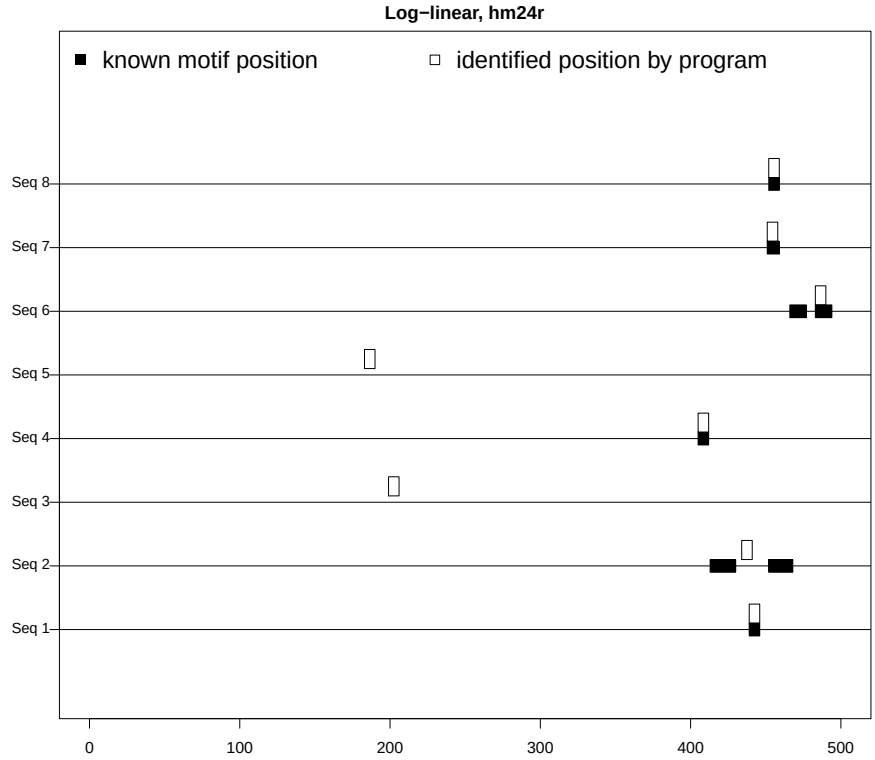

Figure 3.2: The result of the log-linear model with two-way interactions for the hm24r dataset. We assume the motif width is 8. We assume the OOPS motif model on the given strand.

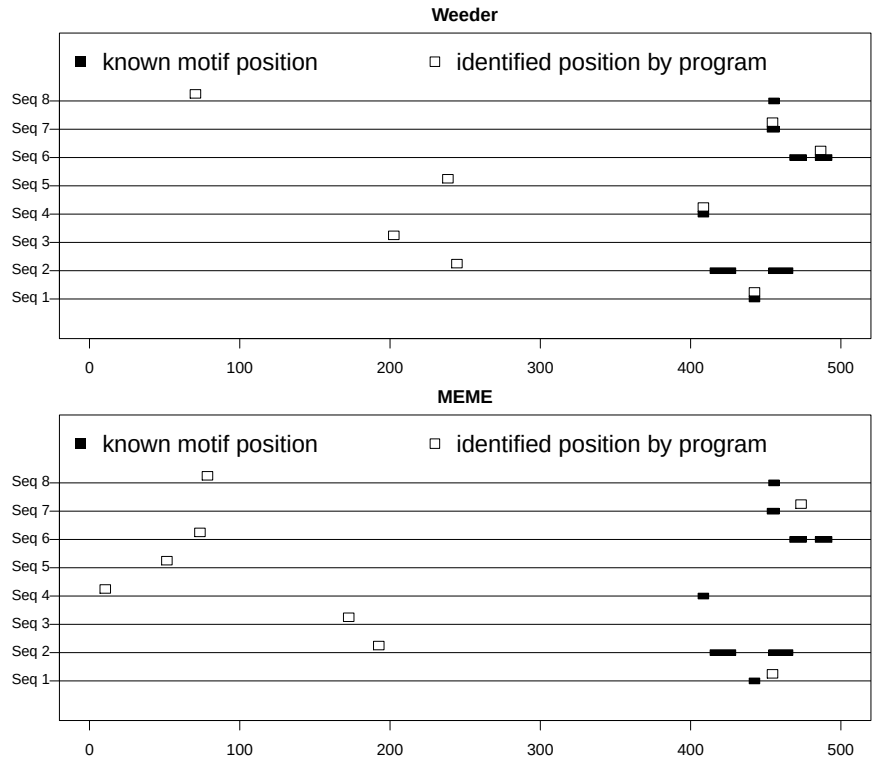

Figure 3.3: The results of Weeder and MEME for the hm24r dataset. We assume the motif width is 8. We assume the OOPS motif model on the given strand.

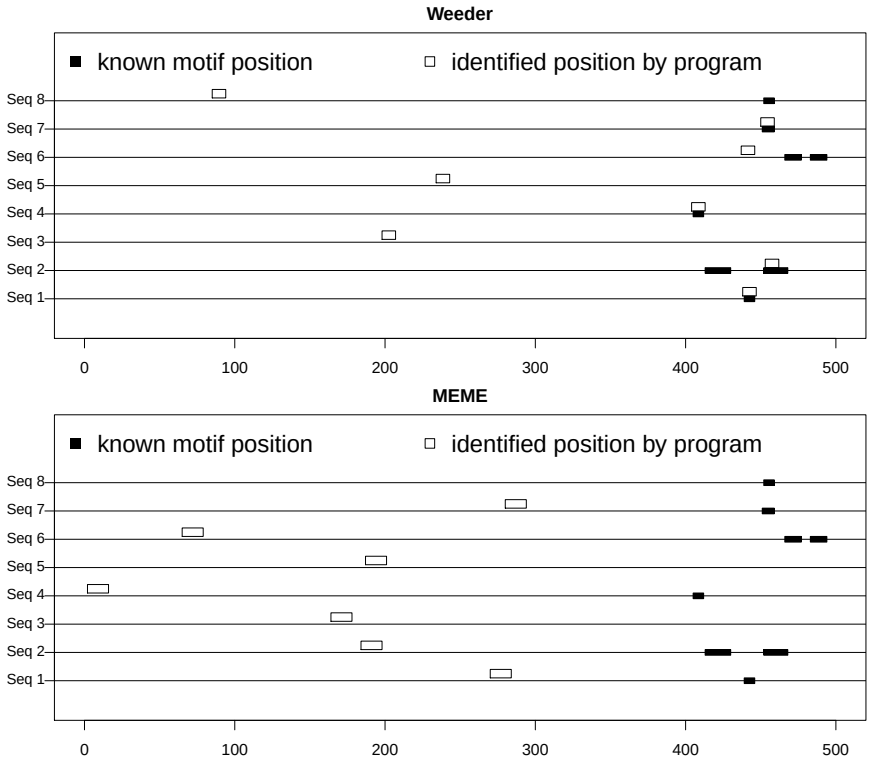

Figure 3.4: The results of Weeder and MEME for the hm24r dataset. We ran the Weeder program assuming the ZOOPS motif model and looked for motifs of length 6, 8, 10 and 12. We ran the MEME program assuming the ZOOPS motif model, allowed instances on both the given DNA strand and the reverse complement strand, and allowed the motif width to be up to 50.

	1	2	3	4	5	6	7	8	% posterior samples > 0
$\beta_{1(T)2(T)}$	T	T							80
$\beta_{1(T)3(T)}$	T		T						76
$\beta_{1(T)6(C)}$	T					C			82
$\beta_{1(T)7(G)}$	T						G		76
$\beta_{2(T)3(T)}$		T	T						85
$\beta_{3(T)7(G)}$		T					G		76
$\beta_{6(C)7(G)}$						C	G		80

Table 3.7: Coefficients with at least 75 % of posterior samples being positive.

3.4.4.3 The log-linear based model with 2-way interactions vs. without interactions

It's interesting to look at the result of the log-linear based model without including 2-way interactions as compared to our results with interactions. The result is shown in Figure 3.5. The program failed to identify the motif instance, TTTCCCGC, in Sequence 8 which was successfully found by the model with 2-way interactions.

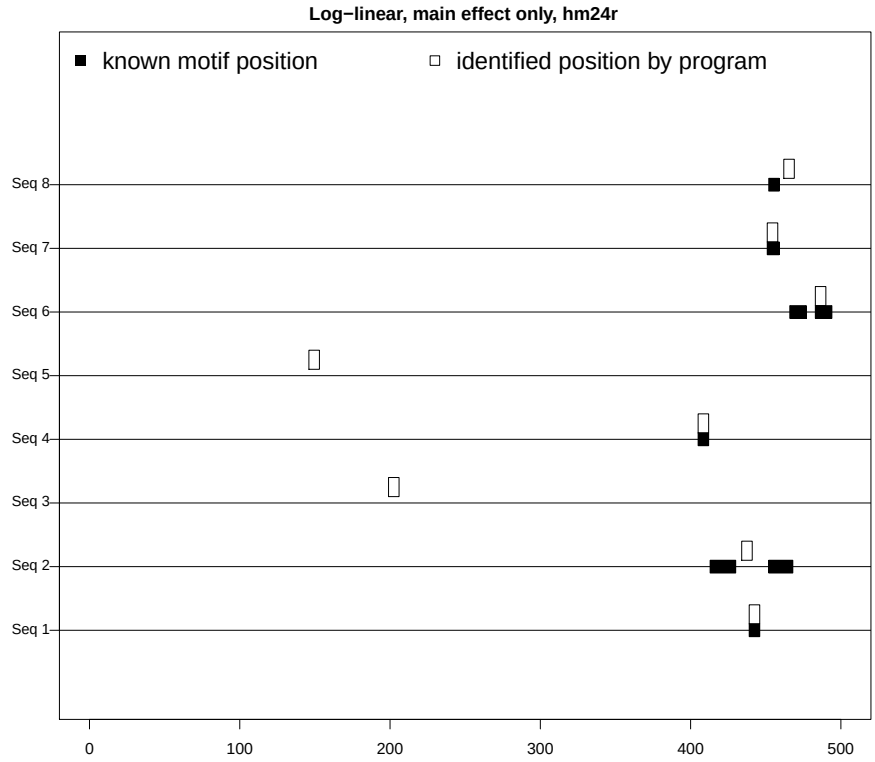

Figure 3.5: The result of the log-linear based model without interactions for the hm24r dataset.

3.4.5 TFBS detected in yeast

3.4.5.1 Dataset background

Next we used the yst09r dataset as an example to test the performance of the proposed method. The yst09r dataset is part of the "real" benchmark dataset and comes from Saccharomyces cerevisiae. The yst09r dataset contains 16 binding sites' real promoter sequences, where each sequence is 1000-base-pairs long. Of these binding sites, 13 out of 16 have been experimentally confirmed to contain one known motif instance. The start locations, instance and instance length of the known motifs are listed in Table 3.8.

3.4.5.2 Motif discovery

We modeled the motif as 8-mers on the given strand using the OOPS model and included both main effects and all two-way interactions in our log-linear model based algorithm. The parameter K in Equation (3.7) was set to 4, which leads to the maximum posterior over-representation score. The initial start locations were generated from uniform distributions and the initial regression coefficients were set to 0. Here we ran 2 chains of 1000000 iterations (the first 100000 iterations were used as a burn in period) each chain. The Gelman-Rubin diagnostic method and the associated R statistics were used to check the chain convergence. After 100000 iterations, all the R statistics were less than 1.2, which indicates that the chains have not failed to converge. We used Weeder and MEME to analyze the yst09r dataset with the same assumptions (modeling the motif as 8-mers on the given strand only using the OOPS model). The results of the three methods are

Sequence	Start position	Instance	Instance length
1	768	TCTTGTGGTGGTTACTC	17
2	NA	NA	NA
3	845	AGCCGCCGA	9
4	NA	NA	NA
5	748	AATTAGCCGCGCAAGTT	17
6	798	CTCTGGCTGCAGGCTAG	17
7	830	CAAGAACCGCCAAGAAC	17
8	778	TCCTAGCCACCTCAAGG	17
9	NA	NA	NA
10	399	TATCCCTGCGCGGCTAAAG	19
11	386	TTGGAGCCGCCAAAAAA	17
12	514	GCCTAGCCGCCGGAGCC	17
13	753	TGTTAGCCGCCGAAACG	17
14	141	AAATAGCCGCCATGACC	17
15	751	TCCATCGGCGGCAAAAG	17
16	550	CTCTAGCCGCCGACGAC	17

Table 3.8: The locations of the binding sites of the yst09r dataset, as given by the TRANSFAC database.

shown in Figure 3.6, and Figure 3.7. For the yst09r dataset, the log-linear based model with two-way interactions identified 8 known motif instances, Weeder and MEME failed to identify any motif instances.

We also ran the the Weeder and MEME algorithms with the default parameters. We ran the Weeder algorithm to search for motifs of length 6, 8, 10 and 12 on the given strand using the ZOOPS model and ran the MEME algorithm to search for motifs of length up to 50 on both the given DNA strand and the reverse complement strand. The results are as shown in Figure 3.8. Weeder identified 4 known motif instances. MEME identified the poly-T pattern and its variants as the top patterns. The next most likely pattern is the target TFBSs, of which MEME identified 10 known TFBSs. For the yst09r dataset, 3 out of 16 sequences (Sequence 2, Sequence 4, and Sequence 9) do not contain any binding site instance. Thus the default parameters for Weeder and MEME outperform the previous settings for this dataset. The success of MEME is largely due to its allowance of instances on both the given DNA strand and the reverse complement strand, in fact 8 out of 10 identified known motif instances are on the reverse complement DNA strand. If we ran MEME assuming the ZOOPS model, allowing instances only on the given DNA strand, and allowing the maximum motif width to be 50, then only 4 known instances are identified by MEME as shown in Figure 3.9.

We also can examine the dependency structure within this motif from the posterior samples of $\boldsymbol{\beta}$. First we look at the regression coefficients, β_is, with more than 60% of their posterior mass being positive. Five main effects and twenty eight two-way

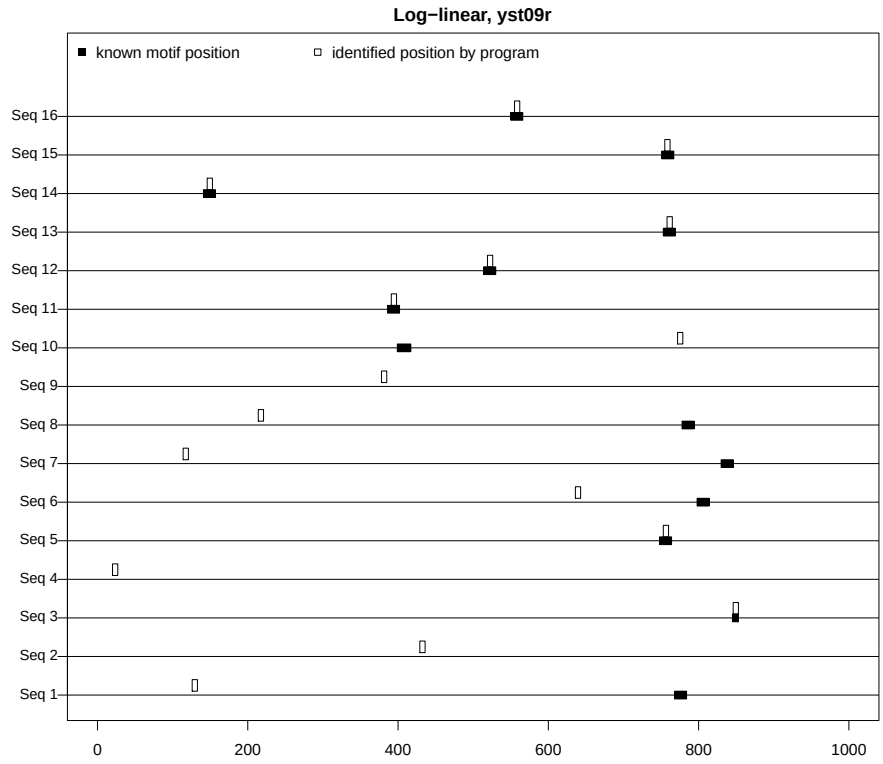

Figure 3.6: The result of the log-linear model for the yst09r dataset. We assume the motif width is 8 and the OOPS motif model on the given strand.

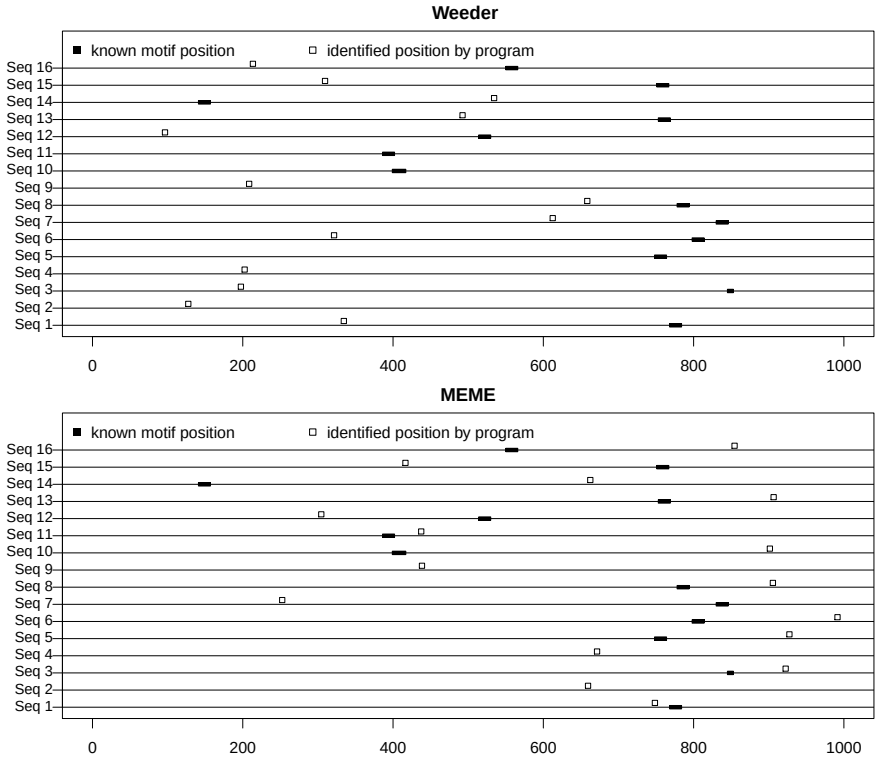

Figure 3.7: The results of Weeder and MEME for the yst09r dataset. We assume the motif width is 8 and the OOPS motif model on the given strand.

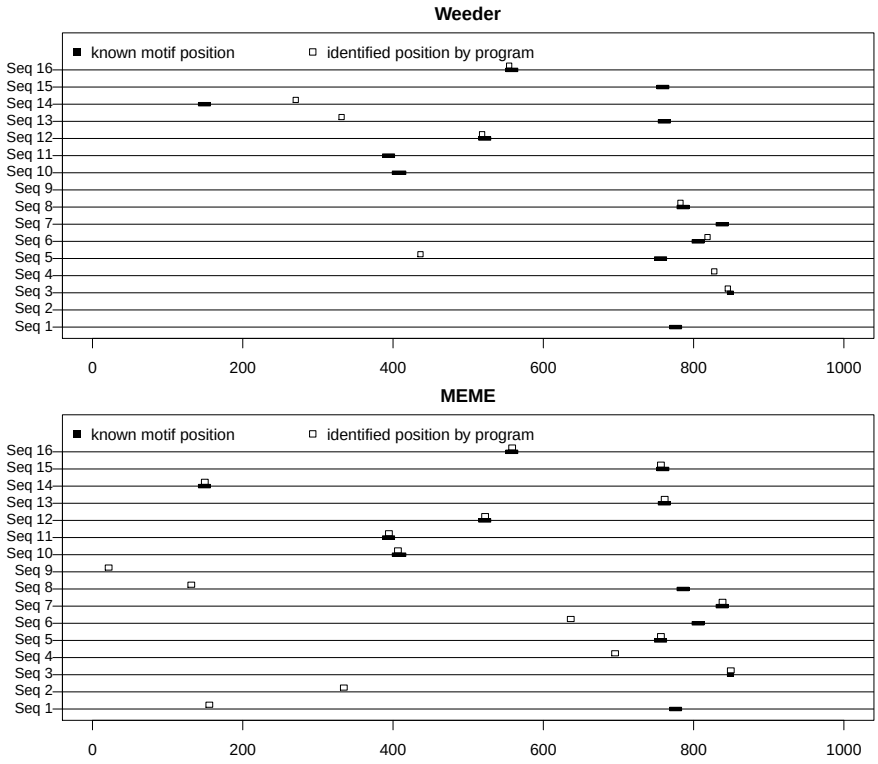

Figure 3.8: The results of Weeder and MEME for the yst09r dataset. We ran the Weeder program assuming the ZOOPS motif model and looked for motifs of length 6, 8, 10 and 12. We ran the MEME program assuming the ZOOPS motif model, allowed instances on both the given DNA strand and the reverse complement strand, and allowed the motif width to be up to 50.

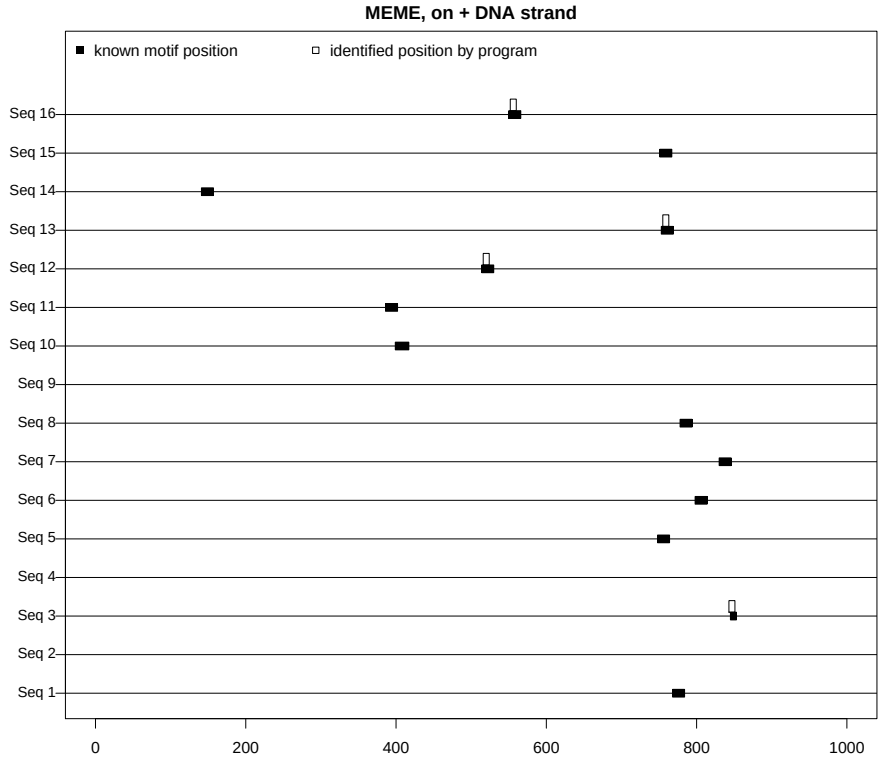

Figure 3.9: The result of MEME for the yst09r dataset. We ran MEME using the OOPS motif model, allowed instances on the given DNA strand and allowed the motif width to be up to 50.

interactions fall in this category. The five main effects are $\beta_{2(C)}(65\%)$, $\beta_{3(C)}(67\%)$, $\beta_{4(G)}(65\%)$, $\beta_{5(C)}(62\%)$, and $\beta_{6(C)}(65\%)$, which is the 5-mer from Position 2 to Position 6 of the posterior mode pattern (GCCGCCGA), while 57% of the posterior mass of $\beta_{1(G)}$ is positive, and 52% of the posterior mass of $\beta_{7(G)}$ is positive. We also examine the set of β_is having more than 70% of their posterior mass greater than 0. While no main effects fall in this category there are 9 two-way interaction effects falling in this category as listed in Table 3.9. For example, Position 1 has an interaction with Position 2 as indicated by our finding that 72% of the posterior mass of $\beta_{1(G)2(C)}$ is positive. We also found that $P(\beta_{1(G)} + \beta_{2(C)} + \beta_{1(G)2(C)} > 0|R) = 0.78$, and we note that $P(\beta_{1(G)} > 0) < 0.78$, $P(\beta_{2(C)}) < 0.78$, and $P(\beta_{1(G)2(C)}) < 0.78$, which indicates that Position 1 has an interaction with Position 2.

3.4.5.3 The log-linear based model with 2-way interactions vs. without interactions

It's interesting to look at the result of the log-linear based model without 2-way interactions as compared to our results with interactions. The result is shown in Figure 3.10 based on 1000000 iterations. Note that the known motif instances found here are almost the same as those found by MEME if instances are only allowed on the given DNA strand (please compare with Figure 3.9). This isn't surprising as the model used by MEME is the same as our log-linear based method with no interactions. Thus modeling the interactions is crucial to successfully identify motif instances here.

	1	2	3	4	5	6	7	8	% posterior samples > 0
$\beta_{1(G)2(C)}$	G	C							72
$\beta_{1(G)3(C)}$	G		C						83
$\beta_{1(G)5(C)}$	G				C				75
$\beta_{2(C)3(C)}$		C	C						73
$\beta_{2(G)3(G)}$		G	G						73
$\beta_{2(C)5(G)}$		C			G				72
$\beta_{3(C)5(C)}$			C		C				70
$\beta_{3(C)7(G)}$			C				G		72
$\beta_{4(G)5(C)}$				G	C				73

Table 3.9: Coefficients with at least 70 % of posterior samples being positive

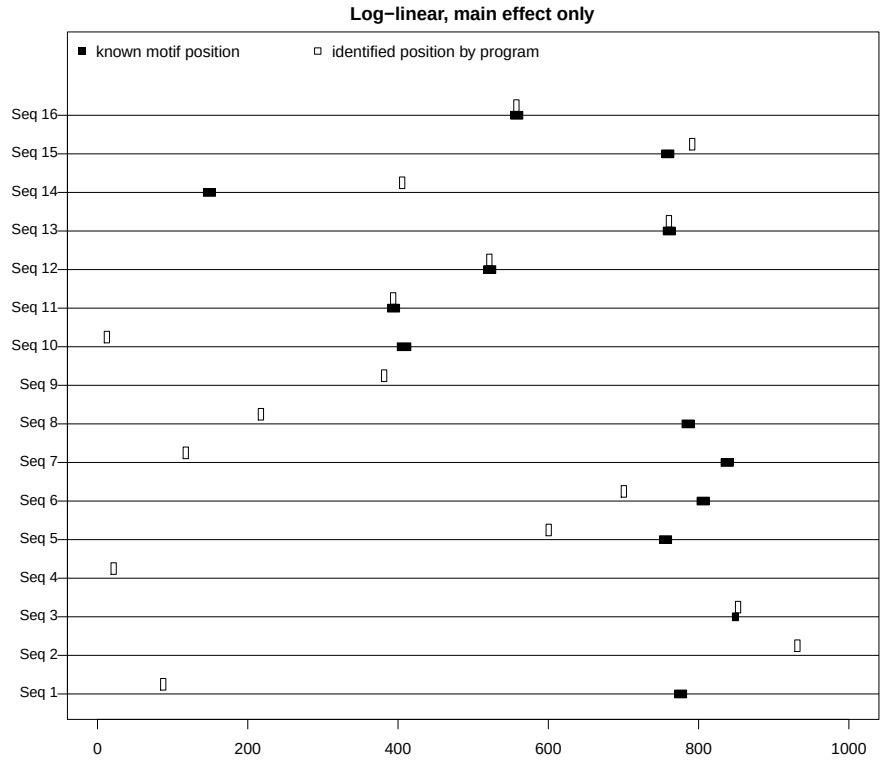

Figure 3.10: The result of the log-linear based model without interactions for the yst09r dataset.

3.5 Discussion

Despite considerable efforts to date and the development of numerous motif finding algorithms over the past decade, DNA TFBSs discovery remains a complex challenge. The performance of the current algorithms is quite low. In one investigation using experimentally confirmed TFBS, the site sensitivity was at most 0.22 (Tompa *et al.* (2005)). Most importantly, the underlying biology of regulatory mechanisms is very incompletely understood. The product-multinomial motif model, used in most of the model-based methods, is limited in its ability to discover the regulatory mechanisms due to the unrealistic assumption that the distribution of letters within the motif are independent across different positions. Similarly non-model based methods, like Weeder, provide no information about the motif structure at the end of analysis. As discussed in Li *et al.* (2006), most current methods do not accurately capture the nature of the binding sites.

Most existing methods require the user to make subjective choices regarding parameter values. For example, in Weeder, how many e mutations should be allowed for a width m motif is arbitrary and subjective. In MEME, how the pre- and postprocessing are conducted is error-prone and very subjective. In contrast, the approach developed here doesn't require the user to specify anything other than the motif width and the structure of the design matrix. Thus we can use the tools developed for linear models to aid in motif discovery.

Here is an attempt to systematically capture the nature of the binding sites through

modeling dependency structure within a motif. Simulations showed that the proposed method can outperform existing methods when there is dependence in the motif and is comparable in cases where there is no dependence. Real data analysis showed the crucial role of modeling the dependency structure within a motif in motif discovery. Theoretically, no matter how complex the structure is, the log-linear model can unveil it through modeling all possible way interactions. In practice, modeling 2-way interactions performed very well as we have shown using some datasets that are commonly used for testing motif discovery methods.

Due to the current memory limitations, we could only model up to 10-mers with two-way interactions. While this may seem to be a limitation, many (but not all) transcription factors binding sites share a common core, consisting of a set of only 5-10 contiguous residues. Due to the presence of this common core, the proposed algorithm is able to identify TFBSs whose length exceeds the limitations imposed by current resource constraints.

Here we focus on highlighting the novel aspects of our proposed approach, modeling the dependency structure within a motif through a log-linear model. While we could further incorporate extensions into the approach developed here, like allowing zero or one instance in each sequence, we see these as further refinements that we will conduct in the future.

Chapter 4

Motif Discovery by Generalizations of the Lasso

4.1 Introduction

Though the lasso has been a widely used technique for simultaneous estimation and variable selection, it has some limitations. If we are interested in finding important explanatory factors in predicting the response variable, where each explanatory factor is represented by a group of derived input variables, the lasso tends to select individual derived input variables from the grouped variables corresponding to each explanatory factor. Analysis of a multiway-contingency table is this kind of problem, in that each factor may have several levels and be expressed through a group of indicator variables. For grouped variables, Yuan and Lin (2006) proposed a generalized lasso that is called

the group lasso. The group lasso is defined as

$$\hat{\boldsymbol{\beta}}_G = \arg\min_{\boldsymbol{\beta}} (\mathbf{y} - \beta_0 \mathbf{1}_n - \sum_{g=1}^{G} \mathbf{X}_g \boldsymbol{\beta}_g)^T (\mathbf{y} - \beta_0 \mathbf{1}_n - \sum_{g=1}^{G} \mathbf{X}_g \boldsymbol{\beta}_g) + \lambda \sum_{g=1}^{G} \|\boldsymbol{\beta}_g\|_{G_g},$$

where G is the number of groups, $\boldsymbol{\beta}_g$ is the vector of βs in group g and $\mathbf{X}_g$ is the design matrix for group g. The G_g's are given positive definite matrices and $\|\boldsymbol{\beta}_g\|_{G_g} = (\boldsymbol{\beta}_g^T G_g \boldsymbol{\beta}_g)^{1/2}$. In general, $G_g = I_{m_g}$, where m_g is the size of the coefficient vector in group g. Yuan and Lin (2006) argued that it does variable selection at the group level.

The lasso shrinkage also produces biased estimates for the nonzero coefficients, and thus it could be suboptimal in terms of estimation risk. Fan and Li (2001) conjectured that the oracle property does not hold for the lasso. We say a procedure has an oracle property if this procedure identifies the right subset model and the estimator based on this procedure has the optimal estimation rate. To overcome the bias issue and obtain the oracle property, Zou (2006) proposed using the adaptively weighted $L1$ penalty to replace the $L1$ penalty in the lasso, and this modified lasso was named the adaptive lasso. The adaptive lasso is defined as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{k=1}^{q} \hat{w}_j |\beta_j|,$$

where the weight vector $\hat{\boldsymbol{w}} = 1/|\hat{\boldsymbol{\beta}}^*|$. Here $\hat{\boldsymbol{\beta}}^*$ is a root-n-consistent estimator of $\boldsymbol{\beta}$ such as $\hat{\boldsymbol{\beta}}(\text{ols})$, the ordinary least squares estimate and $\lambda > 0$. The adaptive lasso enjoys the oracle properties and in generalized linear models the oracle properties still hold under mild regularity conditions (Zou (2006)).

To overcome these shortcomings of the lasso, here we consider methods to incorpo-

rate the Bayesian version of the above extended lasso methods into the log-linear motif discovery algorithm.

4.2 The Bayesian group lasso

In our use of the group lasso, indicator variables derived for the main effect of the same factor are grouped together. For example, $(x_2^j, \ldots, x_p^j)$ for $j = 1, \ldots, w$ are $(p-1)$ indicator variables for the residue in the jth position in our motif discovery context that are grouped together. Similarly indicator variables derived for the interaction effects between two factors are grouped together.

4.2.1 Hierarchical models and full conditionals

Compared with the previous lasso prior, the prior for regression coefficients is now

$$p(\boldsymbol{\beta}) \propto \prod_{g=1}^{G} \lambda^{m_g} \exp^{-\lambda \|\boldsymbol{\beta}_g\|}, \tag{4.1}$$

where m_g is the number of variables in the gth group (i.e. $m_g = p-1$). This prior can be represented by the following hierarchy

$$p(\boldsymbol{\beta}|\tau_1^2, \ldots, \tau_G^2) \propto \prod_{g=1}^{G} \tau_g^{2^{-\frac{m_g}{2}}} \exp\left(-\frac{\|\boldsymbol{\beta}_g\|^2}{2\tau_g^2}\right)$$

$$\tau_g^2 \sim \text{Gamma}(\frac{m_g+1}{2}, \frac{2}{\lambda^2}),$$

where $g = 1, 2, \ldots, G$. To see that this hierarchical specification corresponds to the prior in (4.1) note that

$$\begin{aligned}
&\int p(\boldsymbol{\beta}|\tau_1^2 \ldots \tau_G^2) \prod_{g=1}^{G} \pi(\tau_g^2) \mathrm{d}\tau_1^2 \ldots \mathrm{d}\tau_G^2 \\
\propto &\int \prod_{g=1}^{G} \tau_g^{2^{-\frac{m_g}{2}}} \exp\left(-\frac{\|\boldsymbol{\beta}_g\|^2}{2\tau_g^2}\right) \prod_{g=1}^{G} \lambda^{2^{\frac{m_g+1}{2}}} \tau_g^{2^{(\frac{m_g+1}{2}-1)}} \exp\left(-\frac{\lambda^2\tau_g^2}{2}\right) \mathrm{d}\tau_1^2 \ldots \mathrm{d}\tau_G^2 \\
\propto &\prod_{g=1}^{G} \int \lambda^{2^{\frac{m_g+1}{2}}} \tau_g^{2^{-\frac{1}{2}}} \exp\left(-\frac{\|\boldsymbol{\beta}_g\|^2}{2\tau_g^2} - \frac{\lambda^2\tau_g^2}{2}\right) \mathrm{d}\tau_g^2 \\
\propto &\prod_{g=1}^{G} \lambda^{m_g} \int \lambda \tau_g^{2^{-\frac{1}{2}}} \exp\left(-\frac{\|\boldsymbol{\beta}_g\|^2}{2\tau_g^2} - \frac{\lambda^2\tau_g^2}{2}\right) \mathrm{d}\tau_g^2 \\
\propto &\prod_{g=1}^{G} \lambda^{m_g} \exp^{-\lambda\|\boldsymbol{\beta}_g\|},
\end{aligned}$$

since $\frac{1}{2}e^{-a|z|} = \int_0^\infty \frac{1}{\sqrt{2\pi s}} \frac{a}{2} \exp\left(-\frac{z^2}{2s} - \frac{a^2 s}{2}\right) \mathrm{d}s$ (Andrews and Mallows (1974)). We selected a conjugate prior for λ^2 for computational convenience, which is

$$\lambda^2 \sim \text{Gamma}(\delta, \gamma).$$

Therefore the full conditional distributions of $\boldsymbol{\beta}$ and λ are straightforward to obtain from the joint posterior. For $\boldsymbol{\beta}$ we find that

$$\begin{aligned}
p(\boldsymbol{\beta}|\mathbf{R}, \alpha, \tau_1^2, \ldots, \tau_G^2, \lambda^2) &\propto \prod_{i=1}^{n} p(y_i(\mathbf{R}_{\{\alpha\}}|\boldsymbol{\beta}) \prod_{g=1}^{G} \tau_g^{2^{-\frac{m_g}{2}}} \exp\left(-\frac{\|\boldsymbol{\beta}_g\|^2}{2\tau_g^2}\right) \\
&\propto \prod_{i=1}^{n} \frac{e^{x_i^T\boldsymbol{\beta} y_i} e^{-e^{x_i^T\boldsymbol{\beta}_i}}}{y_i!} \prod_{g=1}^{G} \tau_g^{2^{-\frac{m_g}{2}}} \exp\left(-\frac{\|\boldsymbol{\beta}_g\|^2}{2\tau_g^2}\right)
\end{aligned}$$

where x_i^T is the ith row of the design matrix $\mathbf{X}$. We use the Metropolis algorithm to sample $\boldsymbol{\beta}$ with a normal proposal distribution whose covariance matrix is a scaled identity matrix, with the scalar selected to achieve an acceptance rate around 0.4.

For τ_g^2 we find that

$$\begin{aligned} p(\tau_g^2|\mathbf{R},\alpha,\boldsymbol{\beta},\lambda^2) &\propto p(\boldsymbol{\beta}_g|\tau_g^2)\pi(\tau_g^2|m_g,\lambda^2) \\ &\propto {\tau_g^2}^{-\frac{m_g}{2}} \exp\left(-\frac{\|\boldsymbol{\beta}_g\|^2}{2\tau_g^2}\right) {\tau_g^2}^{(\frac{m_g+1}{2}-1)} \exp\left(-\frac{\lambda^2\tau_g^2}{2}\right) \\ &\propto {\tau_g^2}^{-\frac{1}{2}} \exp\left(-\frac{\|\boldsymbol{\beta}_g\|^2}{2\tau_g^2} - \frac{\lambda^2\tau_g^2}{2}\right), \end{aligned}$$

which is known as the generalized inverse Gaussian distribution (Jorgensen (1982)). We will use GIG $\left(\frac{1}{2},\lambda^2,\|\boldsymbol{\beta}_g\|^2\right)$ to represent this distribution. We let $r_g = \frac{1}{\tau_g^2}$, then

$$\begin{aligned} p(r_g|\mathbf{R},\alpha,\boldsymbol{\beta},\lambda^2) &\propto {r_g}^{\frac{1}{2}} \exp\left(-\frac{\|\boldsymbol{\beta}_g\|^2 r_g}{2} - \frac{\lambda^2}{2r_g}\right) \left|\frac{\mathrm{d}\tau_g^2}{\mathrm{d}r_g}\right| \\ &\propto {r_g}^{\frac{1}{2}} \exp\left(-\frac{\|\boldsymbol{\beta}_g\|^2 r_g}{2} - \frac{\lambda^2}{2r_g}\right) \frac{1}{r_g^2} \\ &\propto {r_g}^{-\frac{3}{2}} \exp\left(-\frac{\|\boldsymbol{\beta}_g\|^2 r_g}{2} - \frac{\lambda^2}{2r_g}\right) \\ &\propto \text{Inverse Gaussian}(\sqrt{\frac{\lambda^2}{\|\boldsymbol{\beta}_g\|^2}},\lambda^2). \end{aligned}$$

Here we choose to sample r_g instead of τ_g^2, since generating random numbers from an inverse Gaussian distribution is much easier than from an generalized inverse Gaussian distribution. To be specific, we can generate random numbers from an inverse Gaussian distribution using normal and uniform random number generators (Michael *et al.* (1976)).

For λ^2 we find that

$$\begin{aligned} p(\lambda^2|\mathbf{R},\alpha,\gamma,\delta,\tau_g^2,\boldsymbol{\beta}) &\propto \prod_{g=1}^{G} p(\tau_g^2|m_g,\lambda^2)\pi(\lambda^2|\gamma,\delta) \\ &\propto \prod_{g=1}^{G}\left[(\frac{1}{\lambda^2})^{-\frac{m_g+1}{2}}\exp(-\frac{\tau_g^2\lambda^2}{2})\right](\lambda^2)^{\gamma-1}\exp(-\frac{\lambda^2}{\delta}) \\ &\propto (\lambda^2)^{\frac{(P-1)+G}{2}+\gamma-1}\exp\left[-\left(\frac{1}{2}\sum_{g=1}^{G}\tau_g^2+\frac{1}{\delta}\right)\lambda^2\right] \\ &\propto \text{Gamma}\left(\frac{P+G-1}{2}+\gamma,\left(\frac{1}{2}\sum_{g=1}^{G}\tau_g^2+\frac{1}{\delta}\right)^{-1}\right), \end{aligned}$$

where P is the total number of the variables in the log-linear model. Hence the full conditionals of τ_g^2 and λ^2 can be sampled using standard routines. We set the parameters (δ,γ) with values such as $(1,1)$. Because of the parameterization, a Gamma(1,1) has large enough support to accommodate non-negative λ^2.

4.3 The Bayesian adaptive lasso

The problem of finding a maximum a posteriori (MAP) estimate of $\boldsymbol{\beta}$ can be expressed as a penalized likelihood problem where $\boldsymbol{\beta}$ is chosen to find a minimum of the function

$$L = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \sum_{i=1}^{q} p_i(|\beta_i|),$$

where the penalty function $\sum_{i=1}^{q} p_i(|\beta_i|)$ corresponds to the negative log prior in the Bayesian model setting.

If we assume a prior for β_i of the form $\pi(\beta_i|\lambda_i) \propto \lambda_i e^{-\lambda_i|\beta_i|}$, assume all the β_i are independent a priori, and $\lambda_i \sim \text{Gamma}(a,b)$, then each β_i has its own shrinkage

parameter λ_i. The oracle property can be achieved by this prior choice. This follows since Fan and Li (2001) showed that the oracle property is achieved if the derivative of the penalty function tends to zero as $|\beta_i|$ tends to infinity. The unconditional prior of β_i is

$$\begin{aligned}\pi(\beta_i) &\propto \int_0^\infty \lambda_i e^{-\lambda_i|\beta_i|}\lambda_i^{a-1}e^{-\frac{\lambda_i}{b}}d\lambda_i \\ &\propto \left(\frac{1}{|\beta_i|+\frac{1}{b}}\right)^{a+1},\end{aligned}$$

the negative log of which corresponds to the penalty function, $p_i(|\beta_i|) = -\log(\pi(\beta_i)) = (a+1)\log(|\beta_i|+\frac{1}{b})$. The derivative of this penalty function for $\beta_i \neq 0$, $p_i'(|\beta_i|)$, equals $\frac{a+1}{|\beta_i|+\frac{1}{b}}$ and this tends to zero as $|\beta_i| \to \infty$.

4.3.1 Hierarchical models and full conditionals

For the Bayesian adaptive lasso, the priors for regression coefficients are

$$\pi(\boldsymbol{\beta}_j|\lambda_j) = \frac{\lambda_j}{2}e^{-\lambda_j|\beta_j|}$$

where $j = 1, 2, \ldots, q$,

$$\lambda_j \sim \text{Gamma}(\delta, \gamma).$$

The full conditional distributions of $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$, δ and γ are straightforward to obtain from the joint posterior. For $\boldsymbol{\beta}$ we find that

$$\begin{aligned}p(\boldsymbol{\beta}|\mathbf{R},\alpha,\lambda) &\propto \prod_{i=1}^{n} p(y_i(\mathbf{R}_{\{\alpha\}})|\boldsymbol{\beta})\prod_{j=1}^{q}p(\boldsymbol{\beta}_j) \\ &\propto \prod_{i=1}^{n}\frac{e^{x_i^T\boldsymbol{\beta}\boldsymbol{y}_i}e^{-e^{x_i^T\boldsymbol{\beta}_i}}}{y_i!}\prod_{j=1}^{q}\frac{\lambda_j}{2}e^{-\lambda_j|\beta_j|}.\end{aligned}$$

We can use the Metropolis algorithm to sample $\boldsymbol{\beta}$. To this end we use a normal distribution with a scaled identity covariance matrix as the proposal distribution. The scalar is selected to achieve an acceptance rate around 0.4.

For $\boldsymbol{\lambda}$ we find that its full conditional is

$$\begin{aligned} p(\lambda_j|\mathbf{R},\alpha,\boldsymbol{\beta},\delta,\gamma) &\propto p(\boldsymbol{\beta}_j|\lambda_j)p(\lambda_j|\delta,\gamma) \\ &\propto \frac{\lambda_j}{2}e^{-\lambda_j|\beta_j|}(\lambda_j)^{\delta-1}\exp(-\frac{\lambda_j}{\gamma}) \\ &\propto (\lambda_j)^{\delta}\exp(-\lambda_j(\frac{1}{\gamma}+|\beta_j|)) \\ &\propto \text{Gamma}(\delta+1,(\frac{1}{\gamma}+|\beta_j|)^{-1}), \end{aligned}$$

hence the full conditional of $\boldsymbol{\lambda}$ can be sampled using a standard routine. We set the parameters (δ,γ) to be $(1,1)$, which has large enough support to accommodate non-negative λ_j.

4.4 The Bayesian group adaptive lasso

In order to select predictors at the group level and maintain the oracle property, we propose the Bayesian group adaptive lasso.

4.4.1 Hierarchical models and full conditionals

We can sample from the joint posterior using the Gibbs sampler as above. To this end we first describe our prior, then derive the full conditional distributions we use for drawing samples. The priors for regression coefficients corresponding to the Bayesian

group adaptive lasso are

$$\boldsymbol{\beta}_g|\tau_g^2 \sim N_{m_g}(\mathbf{0}_{m_g}, \tau_g^2\mathbf{I}_{m_g})$$

where $g = 1, 2, \ldots, G$

$$\tau_g^2 \sim \text{Gamma}(\frac{m_g+1}{2}, \frac{2}{\lambda_g^2})$$

We selected a conjugate prior for λ_g^2 for computational convenience, which is

$$\lambda_g^2 \sim \text{Gamma}(\gamma, \delta).$$

The full conditional distributions of $\boldsymbol{\beta}$ and $(\lambda_1^2, \ldots, \lambda_G^2)$ are straightforward to obtain from the joint posterior. For $\boldsymbol{\beta}$ we find that

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{R}, \alpha, \tau_1^2, \ldots, \tau_G^2) &\propto \prod_{i=1}^{n} p(y_i(\mathbf{R}_{\{\alpha\}})|\boldsymbol{\beta}) \prod_{g=1}^{G} \exp\left(-\frac{\|\boldsymbol{\beta}_g\|^2}{2\tau_g^2}\right) \\ &\propto \prod_{i=1}^{n} \frac{e^{x_i^T\boldsymbol{\beta} y_i} e^{-e^{x_i^T\boldsymbol{\beta}_i}}}{y_i!} \prod_{g=1}^{G} \exp\left(-\frac{\|\boldsymbol{\beta}_g\|^2}{2\tau_g^2}\right). \end{aligned}$$

Again, we can use the Metropolis algorithm to sample $\boldsymbol{\beta}$ as a block and we use the same strategy outlined in our discussion of the Bayesian lasso.

For τ_g^2 we find that

$$\begin{aligned} p(\tau_g^2|\mathbf{R}, \alpha, \boldsymbol{\beta}_g, \lambda_g^2) &\propto p(\boldsymbol{\beta}_g|\tau_g^2)\pi(\tau_g^2|m_g, \lambda_g^2) \\ &\propto \tau_g^{2-\frac{m_g}{2}} \exp\left(-\frac{\|\boldsymbol{\beta}_g\|^2}{2\tau_g^2}\right) \tau_g^{2(\frac{m_g+1}{2}-1)} \exp\left(-\frac{\lambda_g^2\tau_g^2}{2}\right) \\ &\propto \tau_g^{2-\frac{1}{2}} \exp\left(-\frac{\|\boldsymbol{\beta}_g\|^2}{2\tau_g^2} - \frac{\lambda_g^2\tau_g^2}{2}\right) \\ &\propto \text{GIG}\left(\frac{1}{2}, \lambda_g^2, \|\boldsymbol{\beta}_g\|^2\right). \end{aligned}$$

If we let $r_g = \frac{1}{\tau_g^2}$, then

$$\begin{aligned} p(r_g|\mathbf{R}, \alpha, \boldsymbol{\beta}_g, \lambda_g^2) &\propto r_g{}^{\frac{1}{2}} \exp\left(-\frac{\|\boldsymbol{\beta}_g\|^2 r_g}{2} - \frac{\lambda_g^2}{2r_g}\right) \left|\frac{\mathrm{d}\tau_g^2}{\mathrm{d}r_g}\right| \\ &\propto r_g{}^{\frac{1}{2}} \exp\left(-\frac{\|\boldsymbol{\beta}_g\|^2 r_g}{2} - \frac{\lambda_g^2}{2r_g}\right) \frac{1}{r_g^2} \\ &\propto r_g{}^{-\frac{3}{2}} \exp\left(-\frac{\|\boldsymbol{\beta}_g\|^2 r_g}{2} - \frac{\lambda_g^2}{2r_g}\right) \\ &\propto \text{Inverse Gaussian}(\sqrt{\frac{\lambda_g^2}{\|\boldsymbol{\beta}_g\|^2}}, \lambda_g^2) \end{aligned}$$

Again, here we prefer to sample r_g rather than τ_g^2 as discussed above.

For λ_g^2 we find that

$$\begin{aligned} p(\lambda_g^2|\mathbf{R}, \boldsymbol{\beta}, \alpha, \gamma, \delta, \tau_g^2) &\propto p(\tau_g^2|m_g, \lambda_g^2)\pi(\lambda_g^2|\gamma, \delta) \\ &\propto (\frac{1}{\lambda_g^2})^{-\frac{m_g+1}{2}} \exp(-\frac{\tau_g^2\lambda_g^2}{2})(\lambda^2)^{\gamma-1} \exp(-\frac{\lambda_g^2}{\delta}) \\ &\propto (\lambda_g^2)^{\frac{m_g+1}{2}+\gamma-1} \exp\left[-\left(\frac{1}{2}\tau_g^2 + \frac{1}{\delta}\right)\lambda_g^2\right] \\ &\propto \text{Gamma}\left(\frac{m_g+1}{2} + \gamma, \left(\frac{1}{2}\tau_g^2 + \frac{1}{\delta}\right)^{-1}\right), \end{aligned}$$

hence the full conditional of τ_g^2 and λ_g^2 can be sampled using a standard routine. We set parameters (δ, γ) to the values $(1, 1)$, which has large enough support to accommodate non-negative λ_g^2.

4.5 Discussion

Here we propose to incorporate three generalizations of the lasso into the Bayesian log-linear based motif discovery algorithm developed in Chapter 3 to overcome the difficulties

facing the lasso as documented in the literature. The group lasso does variable selection at the group level. Thus in our motif discovery context, the group lasso is expected to do variable selection at the factor or interactions of factors level (factor represents position in the motif here) instead of at the indicator variable level, which will help to select strongly conserved positions and position pairs with strong interactions. Meanwhile by using priors corresponding to the adaptive lasso penalty, larger effect coefficients will be shrunk less. Thus we expect to see a clearer dependency structure, and we expect improved performance of the algorithm by using the Bayesian group lasso, the Bayesian adaptive lasso and the Bayesian group adaptive lasso. Thus we expect a more easily interpreted motif (due to the group lasso) that can outperform our previous approach (due to the oracle property).

Chapter 5

References

5.1 Bibliography

Andrews, D. and Mallows, C. (1974) Scale Mixtures of Normal Distributions. *Journal of the Royal Statistical Society: Series B*, **36** 99-102.

Baker, R.J. (1985) Zero entries in contingency tables. *Computational Statistics and Data Analysis*, **3**: 33-45.

Barash, Y., Elidan, G., Friedman, N. and Kaplan, T. (2003) Modeling dependencies in protein-DNA binding sites. Proceedings of the seventh annual International Conference on Research in Computational Molecular Biology, Berlin, Germany, 28-37

Benjamini, Y. and 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**(1): 289-300.

Bulyk, M.L., Johnson, P.L. and Church, G.M. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Research*, **30**: 1255-1261.

Burge, C. and Karlin, S. (1997) Prediction of complete gene stuctures in human genomic DNA. *Journal of Molecular Biology*, **268**: 78-94.

Chumbley, J.R. and Friston, K.J. (2009) False discovery rate revisited: FDR and topological inference using Gaussian random fields. *Neuroimage*, **44(1)**: 62-70.

Coombes, K.R., Kooman, J.M., Baggerly, K.A., Morris, J.S., Kobayashi, R. (2005) Understanding the characteristics of mass spectrometry data through the use of simulation. *Cancer Informatics*, **1**: 41-42.

Datta, S., DePadilla, L. (2006) Feature selection and machine learning with mass spectrometry data for distinguishing cancer and noncancer samples. *Statistical Methodology*, **3**: 79-92.

Donoho D, Johnstone IM, Kerkyacharian G, Picard D (1995) Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society: Series B*, **57(2)**: 301-337.

Du, P., Kibbe, W.A., and Lin S.M. (2006) Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, **22(17)**: 2059-2065.

Eilers, P.H. (2004) Parametric time warping. *Analytical Chemistry*, **76(2)**: 404-411.

Fan, J., and Li, R. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**: 1348-1360.

Gelman, A., and Rubin, D.B. (1992) Inference From Iterative Simulation Using Multiple Sequences (with discussion). *Statistical Science*, **7**: 457-511.

Haberman, S. (1974) The Analysis of Frequency Data. *University of Chicago Press.*

Heller, R., Stanely, D., Yekutieli, D. and Rubin, N. (2006) Cluster based analysis of fMRI data. *Neuroimage*, **33(2)**: 599-608.

House, L., Clyde, M., and Wolpert, R. (2006) Nonparametric models for peak identification and quantification in mass spectroscopy, with application to MALDI-TOF. Discussion Paper 2006-24, Duke University Department of Statistical Science.

Hughes, J., Estep, P., Tavazoie, S. and Church, G. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *Journal of Molecular Biology*, **296(5)**: 1205-1214.

Jorgensen, B. (1982) Statistical Properties of the Generalized Inverse Gaussian Distribution. *Lecture Notes in Statistics*, **9**. New York-Berlin: Springer-Verlag.

Lawrence, C, Reilly, A. (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, **7**: 41-51.

Li, N. and Tompa, M. (2006) Analysis of computational approaches for motif discovery. *Algorithms for Molecular Biology*, **1**: 8.

Listgarten, J., Neal, R.M., Roweis, S.T., and Emili, A. (2005) Multiple alignment of continuous time series, *In Advances in Neural Information Processing Systems*, Vol.17, MIT Press, Cambridge, MA.

Liu, J. (1994) The collapsed Gibbs sampler with applications to a gene regulation problem. *Journal of the American Statistical Association*, **89**: 958-966.

Liu,J., Neuwald, A. and Lawrence, C. (1995) Bayesian Models for Multiple Local Sequence Alignment and Gibbs Sampling Strategies *Journal of the American Statistical Association*, **90**: 1156-1170.

Liu, X., Britlag, D., and Liu, J.(2001) BioProspector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pacific Symposium on Biocomputing*, **6**: 127-138.

Man, T.K. and Stormo, G.D. (2001) Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Research*, **29**: 2471-2478.

Michael, J., Schucany, W and Haas, R. (1976) Generating Random Variates Using Transformations with Multiple Roots. *American Statistician*, **30(2)**: 88-90.

Morris, J., Brown, P., Herrick, R., Baggerly, K., Coombes, K. (2007) Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics*, **64(2)**: 479-489.

Morris, J.S., Coombes, K.R., Koomen, J., Baggerly, K.A. and Kobayashi, R. (2005) Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, **21(9)**: 1764-1775.

Park,T. and Casella, G. (2008) The Bayesian lasso. *Journal of the American Statistical Association*, **103(482)**: 681-686.

Pavesi, G., Mereghetti, P., Mauri, G., Pesole, G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Research*, **32**: W199-W203.

Reilly, C., Price, P., Gelman, A., Sandgrathe, S. (2004) Using image and curve registration for measuring the goodness of fit of spatial and temporal predictions. *Biometrics*, **60**: 954-964.

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

Sakoe, H., Chiba, S. (1978) Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **26**: 43-49.

Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, **58(1)**: 267-288.

Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A., and Le, Q. (2004) Sample classification from protein mass spectrometry, 'by peak probability contrasts'. *Bioinformatics*, **20(17)**: 3034-3044.

Tompa, M., et al. (2005) Assessing Computational Tools for the Discovery of Transcription Factor Binding Sites. *Nature Biotechnology*, **23(1)**: 137-144.

Van Helden, J. (2003) Regulatory sequence analysis tools. *Nucleic Acids Research*, **31**: 3593-3596.

Vandenbogaert, V. (2008) Alignment of LC-MS images, with applications to biomarker discovery and protein identification. *Proteomics*, **8**: 650-674.

Wu, B. (2004) *Statistical Methods in Analyzing Mass Spectrometry Dataset. PhD Dissertation*, Yale, CT.

Yasui, Y., Randolph, T., Feng, Z. (2006) Profiling high-dimensional protein expression using MALDI-TOF mass spectrometry for biomarker discovery. *Handbook of Statistics in Clinical Oncology*, 2nd edn. Chapman-Hall/CRC, New York.

Yu, W., Li, X., Liu, J., Wu, B., Williams, K., Zhao, H. (2006) Multiple peak alignment in sequential data analysis: A scale-space based approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **3(3)**: 208-219.

Yuan, M. and Lin, Y. (2006) Model Selection and Estimaton in Regression with Grouped Variables. *Journal of the Royal Statistical Society: Series B*, **68**: 49-67.

Zhang, Y., Wroblewski,M., Hertz, M., Wendt, C., Cervenka, T., and Nelsestuen, G.L. (2005) Analysis of chronic lung transplant rejection by MALDI-TOF profiles of bronchoalveolar lavage fluid. *Proteomics*, **6**: 1001-1012.

Zhao, P. and Yu, B. (2006) On model selection consistency of Lasso. *Journal of Machine Learning Research*, **7**: 2541-2563.

Zhou,Q. and Liu,J. (2004), Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, **20 (6)**: 909-916.

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

Zou, H. and Hastie, T. (2005) Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society: Series B*, **67**: 301-320.

Appendix A

Proof

$$
\begin{aligned}
& p(\boldsymbol{\beta}, \lambda, \alpha, \theta_0 | \mathbf{R}) \\
\propto \quad & p(\mathbf{R}_{\{\alpha\}} | \boldsymbol{\beta}, \alpha) p(\mathbf{R}_{\{\alpha\}^c} | \theta_0, \alpha) p(\boldsymbol{\beta} | \lambda, f_i^w) p(\lambda | a, b) p(\theta_0 | \gamma_0) p(\alpha) \\
\propto \quad & \prod_{i=1}^{n} \frac{e^{x_i^T \boldsymbol{\beta} y_i} e^{-e^{x_i^T \boldsymbol{\beta}}}}{y_i!} \prod_{j=1}^{q} \frac{\lambda}{2} e^{-\lambda|\beta_j|} \prod_{i=1}^{K} \frac{e^{(X\boldsymbol{\beta})_{(i)}}}{f_{(i)}^w} p(\mathbf{R}_{\{\alpha\}^c} | \theta_0, \alpha) p(\lambda | a, b) p(\theta_0 | \gamma_0) p(\alpha) \\
= \quad & \prod_{i=1}^{n} \frac{e^{(X\boldsymbol{\beta})_{(i)} y_{(i)}} e^{-e^{(X\boldsymbol{\beta})_{(i)}}}}{y_{(i)}!} \prod_{j=1}^{q} \frac{\lambda}{2} e^{-\lambda|\beta_j|} \prod_{i=1}^{K} \frac{e^{(X\boldsymbol{\beta})_{(i)}}}{f_{(i)}^w} p(\mathbf{R}_{\{\alpha\}^c} | \theta_0, \alpha) p(\lambda | a, b) p(\theta_0 | \gamma_0) p(\alpha) \\
= \quad & \prod_{i=1}^{K} \frac{1}{f_{(i)}^w} \prod_{i=1}^{K} \frac{e^{(X\boldsymbol{\beta})_{(i)} (y_{(i)}+1)} e^{-e^{(X\boldsymbol{\beta})_{(i)}}}}{y_{(i)}!} \prod_{i=K+1}^{n} \frac{e^{(X\boldsymbol{\beta})_{(i)} y_{(i)}} e^{-e^{(X\boldsymbol{\beta})_{(i)}}}}{y_{(i)}!} \\
& \cdot \prod_{j=1}^{q} \frac{\lambda}{2} e^{-\lambda|\beta_j|} p(\mathbf{R}_{\{\alpha\}^c} | \theta_0, \alpha) p(\lambda | a, b) p(\theta_0 | \gamma_0) p(\alpha) \\
= \quad & \prod_{i=1}^{K} \frac{y_{(i)}+1}{f_{(i)}^w} \prod_{i=1}^{K} \frac{e^{(X\boldsymbol{\beta})_{(i)} (y_{(i)}+1)} e^{-e^{(X\boldsymbol{\beta})_{(i)}}}}{(y_{(i)}+1)!} \prod_{i=K+1}^{n} \frac{e^{(X\boldsymbol{\beta})_{(i)} y_{(i)}} e^{-e^{(X\boldsymbol{\beta})_{(i)}}}}{y_{(i)}!} \\
& \cdot \prod_{j=1}^{q} \frac{\lambda}{2} e^{-\lambda|\beta_j|} p(\mathbf{R}_{\{\alpha\}^c} | \theta_0, \alpha) p(\lambda | a, b) p(\theta_0 | \gamma_0) p(\alpha)
\end{aligned}
$$

Since $\sum_{i=1}^{n} y_{(i)} = N$, we have

$$
\prod_{i=1}^{K} \frac{y_{(i)}+1}{f_{(i)}^w} \quad \leq \quad \left(\frac{N+1}{\min(f_i^w)} \right)^K .
$$

Since $\frac{e^{(X\boldsymbol{\beta})_{(i)} (y_{(i)}+1)} e^{-e^{(X\boldsymbol{\beta})_{(i)}}}}{(y_{(i)}+1)!}$ is in the form of a Poisson mass function, no matter what value $y_{(i)}$ takes

$$\frac{e^{(X\boldsymbol{\beta})_{(i)}(y_{(i)}+1)}e^{-e^{(X\boldsymbol{\beta})_{(i)}}}}{(y_{(i)}+1)!} \leq 1$$

Since $\frac{e^{(X\boldsymbol{\beta})_{(i)}y_{(i)}}e^{-e^{(X\boldsymbol{\beta})_{(i)}}}}{y_{(i)}!}$ is is in the form of a Poisson mass function, no matter what value $y_{(i)}$ takes,

$$\frac{e^{(X\boldsymbol{\beta})_{(i)}y_{(i)}}e^{-e^{(X\boldsymbol{\beta})_{(i)}}}}{y_{(i)}!} \leq 1$$

Therefore,

$$p(\boldsymbol{\beta},\lambda,\alpha,\theta_0|\mathbf{R}) \leq \left(\frac{N+1}{\min(f_i^w)}\right)^K \prod_{j=1}^{q}\frac{\lambda}{2}e^{-\lambda|\beta_j|}p(\mathbf{R}_{\{\alpha\}^c}|\theta_0,\alpha)p(\lambda|a,b)p(\theta_0|\gamma_0)p(\alpha)$$

As long as we could show that $\prod_{j=1}^{q}\frac{\lambda}{2}e^{-\lambda|\beta_j|}p(\mathbf{R}_{\{\alpha\}^c}|\theta_0,\alpha)p(\lambda|a,b)p(\theta_0|\gamma_0)p(\alpha)$ is integrable, $p(\boldsymbol{\beta},\lambda,\alpha,\theta_0|\mathbf{R})$ is a proper posterior. And we have

$$\begin{aligned}
& \int\prod_{j=1}^{q}\frac{\lambda}{2}e^{-\lambda|\beta_j|}p(\mathbf{R}_{\{\alpha\}^c}|\theta_0,\alpha)p(\lambda|a,b)p(\theta_0|\gamma_0)p(\alpha)\mathrm{d}\boldsymbol{\beta}\mathrm{d}\alpha\mathrm{d}\lambda\mathrm{d}\theta_0 \\
= & \int\left(\int\prod_{j=1}^{q}\frac{\lambda}{2}e^{-\lambda|\beta_j|}\mathrm{d}\boldsymbol{\beta}\right)p(\mathbf{R}_{\{\alpha\}^c}|\theta_0,\alpha)p(\lambda|a,b)p(\theta_0|\gamma_0)p(\alpha)\mathrm{d}\alpha\mathrm{d}\lambda\mathrm{d}\theta_0 \\
= & \int p(\mathbf{R}_{\{\alpha\}^c}|\theta_0,\alpha)p(\lambda|a,b)p(\theta_0|\gamma_0)p(\alpha)\mathrm{d}\alpha\mathrm{d}\lambda\mathrm{d}\theta_0 \\
= & \int p(\mathbf{R}_{\{\alpha\}^c}|\theta_0,\alpha)p(\theta_0|\gamma_0)p(\alpha)\mathrm{d}\alpha\mathrm{d}\theta_0 \\
\propto & \int\frac{\Gamma(h(\mathbf{R}_{\{\alpha\}^c})+\gamma_0)}{\Gamma(h(\mathbf{R}_{\{\alpha\}^c}))}p(\alpha)\mathrm{d}\alpha \\
\propto & \sum_{\alpha_1=1}^{n_1-w+1}\cdots\sum_{\alpha_N=1}^{n_N-w+1}\frac{\Gamma(h(\mathbf{R}_{\{\alpha\}^c})+\gamma_0)}{\Gamma(h(\mathbf{R}_{\{\alpha\}^c})} \\
< & \infty
\end{aligned}$$

Thus we could conclude that $p(\boldsymbol{\beta},\lambda,\alpha,\theta_0|\mathbf{R})$ is integrable, which means that $p(\boldsymbol{\beta},\lambda,\alpha,\theta_0|\mathbf{R})$ is proper even though prior for $\boldsymbol{\beta}$ is not proper.