Generalized hierarchical Markov models for the discovery of length-constrained sequence features from genome tiling arrays

Mayetri Gupta
Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.

SUMMARY. A generalized hierarchical Markov model for sequences that contain length-restricted features is introduced. This model is motivated by the recent development of high density tiling array data for determining genomic elements of functional importance. Due to length constraints on certain features of interest, as well as variability in probe behavior, usual hidden Markov-type models are not always applicable. A robust Bayesian framework that can incorporate length constraints, probe variability and bias is developed. Moreover, a novel recursion-based Monte Carlo algorithm is proposed to estimate the parameters and impute hidden states under length constraints. Application of this methodology to yeast chromosomal arrays demonstrate substantial improvement over currently existing methods in terms of sensitivity as well as biological interpretability.

KEY WORDS: Bayesian hierarchical model; Chromatin; Data augmentation; Gene regulation; Nucleosome.

data: gupta@bios.unc.edu

1
1. Introduction

With the advent of technology to study genomic features at a molecular level, recent studies prove that the regulation of biological processes are influenced not just by the information contained in the DNA sequence, but the actual physical conformation of chromatin that constitutes DNA (Lee et al., 2004). DNA is tightly packed into the nucleus of the cell, with stretches of DNA wrapped around bead-like histone proteins (nucleosomes) at approximately regular intervals. Strong biological evidence (Luger, 2006) shows that about 147 bp of DNA is wrapped around nucleosomes, which are separated by short stretches of DNA called linkers. DNA structure plays an important role in successful protein-DNA interactions, leading to transcription regulation and formation of essential protein products. Nucleosome-free regions are also expected to be more susceptible to damage from environmental agents. For instance, mutations in nucleosome-free areas of regulatory regions of oncogenes could lead to the development of cancerous cells.

Although it is difficult to determine directly how chromatin “packing” takes place by observing the living cell, recent microarray-based techniques (Lee et al., 2004) can measure certain related DNA characteristics. Whole genome high density tiling arrays are microarrays involving short overlapping probes, covering the entire genome. These arrays have been used to experimentally measure genomic positions of nucleosomes, in which the measurement “intensities” indicate how likely that locus is to have been wrapped around a histone. To complicate matters, various intracellular processes are intermittently involved in unwrapping and wrapping DNA. This results in certain nucleosomal regions appearing to be “delocalized”, that is, having in-
tensities more variable in length and measurement magnitude than expected. Differentiating the nucleosomal and nucleosome-free regions in the DNA is further hindered by the facts that (i) the measured intensities are susceptible to inherent probe-specific bias and variability and (ii) the effect of noise is magnified by the short probe lengths, and the measurement error.

Compared to mRNA microarrays, whole genome tiled arrays pose severe challenges for data analysis. Tiling induces a spatial data structure, invalidating independence assumptions across loci. Hidden Markov models, or HMMs (Rabiner, 1989), can accommodate such spatial structure. HMMs are based on a doubly stochastic process where the underlying (hidden) state develops as a Markov chain that can be inferred only through observations from another set of stochastic processes. But as HMMs induce an exponentially decaying distribution of state lengths, they are not directly appropriate for assessing structural features such as nucleosomes that have restrictions in physical dimension. Profile HMMs (PHMMs) (Durbin et al., 1998) allow states with length restrictions; however, they can only model a single sequence at a time. Since the parameter space increases linearly with the segment length, PHMMs are infeasible for modeling long sequences. HMMs with state durations have been proposed for discrete data (Rabiner, 1989), but parameter estimation using likelihood-based procedures is often complicated and computationally inefficient. Related Bayesian sequence “segmentation” approaches have been successfully used for discrete data models, for example, protein secondary structure prediction (Schmidler et al., 2000).

We introduce a generalized Bayesian framework for drawing statistical inference about a latent process with length-restricted characteristics. A
hierarchical model robust to various sources of probe variability and measurement error (Section 3) is presented and an explicit state duration model is used to model state lengths. For efficient estimation of states and the model parameters, a novel Monte Carlo method is developed that adapts recursive likelihood computation techniques to reduce computational complexity (Section 4). Estimation for the length-restricted parameters is designed for arbitrary families of distributions, not limited to the exponential family, making it applicable for a variety of situations. Application of this methodology to yeast tiling arrays demonstrates significant gains in terms of both sensitivity and biological interpretability (Section 5).

2. Biological background and motivation

Determining how gene expression is regulated is an essential step in understanding biological pathways, including the development and progression of disease. Functional elements such as transcription factor binding sites (TFBSs) initiate gene regulation by providing signals to proteins, called transcription factors (TFs) to bind to the DNA at specific locations and activate transcription of genes. TFBSs often occur as short, conserved patterns in the promoter regions of genes they regulate. Powerful computational methods for TFBS discovery from DNA sequence have been developed (e.g. Lawrence et al. (1993), Liu et al. (2001), Gupta and Liu (2003), Thompson et al. (2004)). However, in complex genomes, they often yield a large number of false positives, making direct site-by-site biological validation expensive and infeasible. Recent in-vivo biological assays (Lee et al., 2004) show that nucleosome positioning is closely associated with binding of TFs to the DNA–linker regions in general are significantly enriched for TFBSs. Accurate deter-
mination of nucleosome positions thus gives a powerful estimate of genomic loci that could be targeted for the discovery of functional TFBSs.

The data set motivating the methodology is now described. Yuan et al. (2005) developed a DNA microarray to locate nucleosome positions for 30 selected regions of yeast chromosome III, covering a total of 270 kilobases (Kbp). Nucleosomal DNA was isolated from linkers by treatment with micrococcal nuclease, labeled with Cy3 (green) fluorescent dye, and mixed with Cy5-labeled (red) genomic DNA. The mixture was hybridized to a whole genome microarray which consisted of overlapping 50-mer oligonucleotide probes tiled every 20 base pairs, with 8 replicate measurements for each probe. A graph of green-to-red ratio values for spots along the chromosome is expected to show nucleosomes as peaks of about 140 nucleotides wide (equivalent to 6 to 8 microarray spots), surrounded by lower ratio values representing linker regions. The data was normalized in a pre-processing step to filter out potentially cross-hybridizing probes. To discover nucleosome positions, Yuan et al. (2005) used a profile HMM model for the median value over replicates, restricting nucleosomal states to being within 6 to 8 probes. However, as PHMMs are only feasible for short sequences, the sequence was first divided into shorter “windows” of about 40 probes, and post-processed to ensure that the window boundaries did not cut across nucleosomes.

Our proposed generalized HMM incorporates length restrictions, without the necessity of imposing artificial “windows”. Additionally, the hierarchical Bayesian framework can accommodate: (i) probe variability by using measurements from all replicates and (ii) probe-specific biases by hierarchically modeling probe-specific distributions, avoiding model over-parametrization.
3. Bayesian framework for length-restricted array features

The general model is now described. First, the case where the length range for each state is known is considered. Next is discussed the generalization of the framework to the case where the length duration of certain states may be partially unknown. In the case where the range of state lengths is completely unknown, it may be more appropriate to use a usual HMM.

3.1 Hierarchical generalized hidden Markov model (HGHMM)

Assume that the model consists of $K$ ($\geq 2$) states. The possible length duration in state $k$, ($k = 1, \ldots, K$) is given by the set $D_k = \{r_k, \ldots, s_k\} \subset \mathbb{N}$ (i.e. $\mathbb{N}$ denotes the set of positive integers). The generative model for the observed data is described below.

(i) The initial distribution of states is characterized by the probability vector $\pi = (\pi_1, \ldots, \pi_K)$.

(ii) The probability of spending time $d$ in state $k$ is given by the distribution $p_k(d|\phi)$, $d \in D_k$ ($1 \leq k \leq K$), characterized by the parameter $\phi = (\phi_1, \ldots, \phi_K)$. For the motivating application, $p_k(d)$ is chosen to be a truncated negative binomial distribution, between the range specified by each $D_k$. More precisely,

$$p_k(d) = c_k \left( \frac{d-1}{r_k-1} \right) (1-\phi_k)^{r_k-1} \phi_k^d, \quad d \in D_k = \{r_k, r_k + 1, \ldots, s_k\} \quad (1)$$

where the normalizing constant $c_k = \left[ \sum_{d=r_k}^{s_k} \frac{(d-1)^{r_k-1}}{r_k^d} \phi_k^d \right]^{-1}$. This can equivalently be written in terms of a binomial sum (Morris, 1963), viz. $c_k^{-1} = \sum_{d=0}^{s_k-r_k} \binom{s_k}{d} (1-\phi_k)^d \phi_k^{s_k-d}$. A conjugate Beta($\gamma_k, \delta_k$) prior is assumed for $\phi_k$. 

6
(iii) The latent state for probe \(i\) is denoted by the variable \(Z_i\) \((i = 1, \ldots, N)\).

(iv) Logarithms of spot measurement ratios are denoted by \(y_{ij}\) \((1 \leq i \leq N; 1 \leq j \leq r)\) for \(N\) spots and \(r\) replicates each. Assume that given the (unobservable) state \(Z_i\), \(y_{ij}'s\) are independent, with \(y_{ij}|Z_i = k \sim g_k(\cdot; \xi_{ik}, \sigma_{ik}^2)\). For specifying \(g_k\), a hierarchical model is developed that allows robust estimation of the parameters. Let \(\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)\) and \(\Sigma = \{\sigma_{ik}^2; 1 \leq i \leq N; 1 \leq k \leq K\}\). Assume

\[
y_{ij}|Z_i = k, \xi_{ik}, \sigma_{ik}^2 \sim N(\xi_{ik}, \sigma_{ik}^2)
\]

\[
\xi_{ik}|\mu_k, \sigma_{ik}^2 \sim N(\mu_k, \tau_0 \sigma_{ik}^2)
\]

\[
\sigma_{ik}^2 \sim \text{Inv-Gamma}(\rho_k, \alpha_k), \text{ i.e. } p(\sigma_{ik}^2) \propto (\sigma_{ik}^2)^{-\frac{\alpha_k}{2} - 1} e^{-\frac{\alpha_k}{2} \cdot \frac{\sigma_{ik}^2}{\rho_k}}
\]

where at the top level, \(\mu_k \propto \text{constant}\), and \(\rho_k, \alpha_k,\) and \(\tau_0\) are hyperparameters discussed in more detail in Section 5.1. Then, the marginal distribution of \(y_{ij}\) is \(y_{ij}|Z_i = k, \mu_k \sim t\left(\mu_k, (1 + \tau_0)\frac{\alpha_k}{\rho_k}, 2\rho_k\right), \text{ i.e. a non-central t-distribution with non-centrality and scale parameters } \mu_k \text{ and } (1 + \tau_0)\frac{\alpha_k}{\rho_k} \text{ respectively, with } 2\rho_k \text{ degrees of freedom (df).}

(v) The transition probabilities between the states are given by the matrix \(\boldsymbol{\tau} = (\tau_{jk}), (1 \leq j, k \leq K)\). Assume a Dirichlet prior for state transition probabilities, i.e. \(\tau_{k1}, \ldots, \tau_{k,k-1}, \tau_{k,k+1}, \ldots, \tau_{k,K} \sim \text{Dirichlet}(\boldsymbol{\eta})\), where \(\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{k-1}, \eta_{k+1}, \ldots, \eta_K)\). Since the duration in a state is being modeled explicitly, no transition back to the same state can occur, i.e. there is a restriction \(\tau_{kk} = 0\) for all \(1 \leq k \leq K\).

4. Model fitting and parameter estimation

For notational simplicity, assume a single long sequence of length \(N\), \(\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}\), with \(r\) replicate observations for each \(\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ir})\). If
there are gaps, each separated segment of the sequence should be taken separately, and the same procedure repeated for each segment. Let the set of all parameters be generically denoted by $\theta = (\mu, \tau, \phi, \pi, \Sigma)$, and let the latent variables $Z = (Z_1, \ldots, Z_N)$ and $L = (L_1, \ldots, L_N)$ denote the state identity and state lengths, where $L_i = l$ is a non-zero number denoting the state length if it is a point where a run of states ends. Mathematically,

$$L_i = \begin{cases} l & \text{if } Z_{i+1} \neq Z_i = Z_{i-1} = \cdots = Z_{i-l+1} = k \neq Z_{i-l} \text{ for some } k \in \{1, \ldots, K\}, \\ 0 & \text{otherwise}. \end{cases}$$

The observed data likelihood then may be written as:

$$L(\theta; Y) = \sum_{Z} \sum_{L} p(Y|Z, L, \theta)P(L|Z, \theta)P(Z|\theta) \tag{2}$$

The likelihood computation (2) is analytically intractable, involving a sum over all possible partitions of the sequence $Y$ with different state conformations, and different state lengths (under the state restrictions). Further, direct computation is infeasible for any realistic data set as it involves an order of computation exponential in the length of the sequence. The following section describes a Monte Carlo-based method which utilizes a recursive data augmentation technique to efficiently sample from the posterior distributions of interest. An alternative methodology would be a Gibbs sampler for iteratively sampling from the full conditional distributions. However, the proposed data augmentation scheme utilizes a grouped sampling scheme, essentially sampling from the joint posterior distribution of the state variables, thus being more efficient in terms of MCMC convergence (Liu et al., 1994).

4.1 Recursive data augmentation.

In the data augmentation algorithm, the key is to update the states and state length durations in an recursive manner, after calculating the required
probability expressions through a \textit{forward summation} step. Let an indicator variable $I_t$ take the value 1 if a segment boundary is present at position $t$ of the sequence, meaning that a state run ends at $t$ ($I_t = 1 \Leftrightarrow L_t \neq 0$). In the following, the notation $y_{[1:t]}$ is used to denote the vector \{ $y_1, y_2, \ldots, y_t$ \}.

Define the partial likelihood of the first $t$ probes, with the state $Z_t = k$ ending at $t$ after a state run length of $L_t = l$, by the \textit{forward} probability:

\[
\alpha_t(k, l) = P(Z_t = k, L_t = l, I_t = 1, y_{[1:t]}).
\]

Also, let the state probability marginalized over all state lengths be given by

\[
\beta_t(k) = \sum_{l=r_k}^{s_k} \alpha_t(k, l) \tag{3}
\]

Let $d_{(1)} = \min\{ D_1, \ldots, D_K \}$ and $d_{(K)} = \max\{ D_1, \ldots, D_K \}$. Then, assuming that the length spent in a state and the transition to that state are independent, i.e. $P(l, k|l', k') = P(L_t = l|Z_t = k) \tau_{k', k} = p_k(l) \tau_{k', k}$, we have

\[
\alpha_t(k, l) = \sum_{k' \neq k} \sum_{l' \in \mathcal{D}_{k'}} \alpha_{t-l}(k', l') P(l, k|l', k') P(y_{[t-l+1:t]}|Z_t = k)
\]

\[
= P(y_{[t-l+1:t]}|Z_t = k) p_k(l) \sum_{k' \neq k} \tau_{k', k} \sum_{l'} \alpha_{t-l}(k', l')
\]

\[
= P(y_{[t-l+1:t]}|Z_t = k) p_k(l) \sum_{k' \neq k} \tau_{k', k} \beta_{t-l}(k'), \tag{4}
\]

for $2 \leq t \leq N$; $1 \leq k \leq K$; $l \in \{ d_{(1)}, d_{(1)}+1, \ldots, \min[d_{(k)}, t] \}$. To complete the calculation, the boundary conditions needed are: $\alpha_t(k, l) = 0$ for $t < l < d_{(1)}$, and $\alpha_t(k, l) = \pi_k P(y_{[1:t]}|Z_t = k) p_k(l)$ for $d_{(1)} \leq l \leq d_{(k)}$, $k = 1, \ldots, K$. $p_k(\cdot)$ denotes the $k$-th truncated negative binomial distribution given in (1).

The set of equations (4) can alternatively be concisely written as a matrix recursion. Define $f^{(t)}_{kl} = P(y_{[t-l+1:t]}|Z_t = k) p_k(l)$, and $a^{(t)}_{kl} = \alpha_t(k, l)$,
and define the $K \times d_{(K)}$ matrices $F^{(t)} = (f_{kl}^{(t)})$ and $A^{(t)} = (a_{kl}^{(t)})$. Then, $A^{(t)} = F^{(t)} \tau [A^{(t-1)} A^{(t-2)} \ldots A^{(t-M)}] (1_{d_M} \otimes I_M)$, where $M = \min[d_{(K)}, t]$, $1_n$ denotes an $n$-dimensional vector of ones, $I_n$ denotes an identity matrix of dimension $n$, and $\otimes$ denotes the Kronecker product. The notation $\tau$ is used to denote the elementwise multiplication of matrices, i.e. for two $m \times n$ matrices $B = (b_{jk})$ and $C = (c_{jk})$, $B \tau C = (d_{jk})$ where $d_{jk} = b_{jk} c_{jk}$.

The states and state duration lengths $(Z_t, L_t)$ $(1 \leq t \leq N)$ can now be updated, for current values of the parameters $\theta = (\mu, \tau, \phi, \pi, \Sigma)$, using a backward sampling-based imputation step.

Algorithm: Recursive state and duration sampling.

(i) Set $i = N$. Update $Z_N|y, \theta$ using

$$P(Z_N = k|y, \theta) = \frac{\beta_N(k)}{\sum_k \beta_N(k)}.$$

(ii) Next, update $L_N|Z_N = k, y, \theta$ using

$$P(L_N = l|Z_N = k, y, \theta) = \frac{P(L_N = l, Z_N = k|y, \theta)}{P(Z_N = k|y, \theta)} = \frac{\alpha_N(k, l)}{\beta_N(k)}.$$

(iii) Next, set $i = i - L_N$, and let $LS(i) = L_N$. Let $D(2)$ be the second smallest value in the set \{$\min D_1, \ldots, \min D_K$\}. While $i > D(2)$, repeat the following three steps:

- Draw $Z_i, y, \theta, Z_{i+LS(i)}, L_{i+LS(i)}$ using

$$P(Z_i = k|y, \theta, Z_{i+LS(i)}, L_{i+LS(i)}) = \frac{P(Z_i, Z_{i+LS(i)}|L_{i+LS(i)}, y, \theta)}{P(Z_{i+LS(i)}|L_{i+LS(i)}, y, \theta)} \frac{\beta_i(k) \tau_{kZ_{i+LS(i)}}}{\sum_k \beta_i(k) \tau_{kZ_{i+LS(i)}}}.\]
where \( k \in \{1, \ldots, K\} \setminus Z_{i+LS(i)} \), the simplification resulting from the assumption that the duration in the previous state and the next state transition are independent events.

- Draw \( L_i | Z_i, y, \theta \) using

\[
P(L_i = l | Z_i, y, \theta) = \frac{\alpha_i(Z_i, l)}{\beta_i(Z_i)}.
\]

- Set \( LS(i - L_i) = L_i, \ i = i - L_i \).

Note that the proposed sampling algorithm is generally applicable to any length restricted HMM and not limited to the forms of the state-specific distributions used here. Once the states and state duration lengths \((Z_i, L_i)\) \((1 \leq i \leq N)\) have been updated, updating the parameters from their posterior distributions is standard (details are in the Supplementary material).

4.2 Modified algorithm for partially unknown state lengths

In the case that only partial information is available on the range of state lengths \((r_k, s_k)\) for one or more states, the above approach cannot be directly applied. However, one can assign a prior distribution on the boundaries of the range and update them from their posterior distributions. Assume that only an approximate range of lengths is known for state \( k' \), and choose prior distributions \( r_k' \sim h_{r_k'}(\cdot) \) and \( s_k' \sim h_{s_k'}(\cdot) \) (e.g. \( h_{r_k'}(\cdot) \sim \text{Poisson}(\lambda_k) \)). A Metropolis-Hastings step can be used to adjust the values of \((r_{k'}, s_{k'})\) at intervals within the data augmentation steps. Assume, at iteration \( j \), that the current specification of \((r_{k'}, s_{k'}) = (r_{k'}^{(j)}, s_{k'}^{(j)})\). Perform the following update:

**Metropolis-Hastings update for \((r_{k'}, s_{k'})\).**

- Choose to update \( r_{k'} \) or \( s_{k'} \). If \( r_{k'} = 1, s_{k'} = 2 \), set \( \epsilon_1 = 1, a = s_{k'}, b = \)
Otherwise, draw $\epsilon_1 \sim Bernoulli(1/2)$; if $\epsilon_1 = 0$, set $a = r_k$, $b = s_kr'$, else set $a = s_k$, $b = r_kr'$. $a$ will be updated while $b$ remains fixed.

- **Choose an increase or decrease in $a$.** If $\epsilon_1 = 0$ and $r_k = 1$, set $\delta = 1$.
  If $\epsilon_1 = 0$ and $r_k + 1 = s_k$, set $\delta = -1$. If $\epsilon_1 = 1$, and $r_k + 1 = s_k$, set $\delta = 1$. Otherwise, draw $\epsilon_2 \sim Bernoulli(1/2)$ and set $\delta = 2\epsilon_2 - 1$.

- Now update $a \leftarrow a + \delta$ with probability $\min\{1, r\}$, where the Metropolis ratio $r$ is given by $r = \frac{P(y|\theta, a + \delta, b)h_a(a + \delta)}{P(y|\theta, a, b)h_a(a)}$.

The likelihood $P(y|\theta, a, b)$ is already calculated in the data augmentation step, given by $\sum_s \beta_N(s)$ (Eqn. 3). For the numerator, this likelihood must be calculated in a similar forward summation step (Section 4.1).

5. **Applications**

5.1 **Hyperparameter choice and sensitivity analysis**

Hyperparameters $\eta, \gamma_k, \delta_k$ for the Dirichlet and Beta densities are chosen to be non-informative— (i) $\eta_{ij} = 1$ (ii) $\gamma_k = \delta_k = 2$. To provide an intuition for the choice of the inverse gamma (IG) hyperparameters $\rho_k, \alpha_k$, we re-scale the inverse gamma into an equivalent inverse chi-square ($I_{\chi^2}$) distribution, using the relation $\frac{1}{2\alpha}IG(\rho, \alpha) = I_{\chi^2}(2\rho)$. With the scale parameter $\alpha_k = \frac{1}{2}$, the prior for $\sigma^2_{ik}$ is then $I_{\chi^2}(2\rho_k)$. Putting a uniform density on $\log(\sigma_{ik})$, i.e. $p(\sigma^2_{ik}) \propto \sigma_{ik}^{-2}$ (equivalent to $I_{\chi^2}(0)$) leads to a non-informative (but improper) prior. For testing sensitivity of our prior specifications, we first set $\alpha_k = \frac{1}{2}$ and varied values of $\rho (\equiv (\rho_1, \ldots, \rho_K)'$) over a large range starting from small values ($> 2$), thus keeping the priors proper and increasingly informative.

Data sets were generated from $t$-distributions with $\rho$ as $(5, 10, 20, 30)$, $\mu = (-0.8, 0.2, 0.9)$ and states simulated under the length-constrained HMM.
Using the HGHMM, the maximum MSE for all parameters was close to zero for all hyperprior specifications and data sets (Figure 1). Misclassification rates are also low (< 0.02), robust to hyperparameter specification and demonstrate overall similar behaviour. The variation of estimates is highest and the performance is worst when the true $\rho = 5$. This simulation study, together with Figure 2, suggests that although the robust model improves the fit near the tails (as compared to a Gaussian distribution), the extra benefit may be counter-acted upon by the extra variability of estimation if too stringent a df is chosen for the $t$, leading to slightly higher (though minor) misclassification rates and MSEs. The argument for keeping a $t$ in the model, instead of a Gaussian, however, is that it allows thicker tails and hence is more robust in the presence of chance outliers. For example, we can see that the df set at 80 does not perform best when the true $\rho = 30$, indicating a more robust model is better. A moderately high value of $\rho$ (between 10 and 20) appeared to perform best for all simulated data sets. Sensitivity analyses for $\alpha$ indicated that the results were robust to variations over a 10-fold range, and hence we kept $\alpha = \frac{1}{2}$ for all further analyses.

[Figure 1 about here.]

The remaining hyperparameter $\tau_0$ measures within replicate variability. A general recommendation would be to take a large value only if the experimental variation is high (this can be assessed using ANOVA-based comparisons) or estimate it from similar existing data. In our applications, both within- and between-state variability was moderate, so we chose $\tau_0 = 1$. Posterior inference was insensitive to values between 0.1 to 10. To determine the effect of misspecification of the lengths for the different states,
HGHMM was run with two additional choices of $D$ for states 2 and 3:

(i) Set 2: $D_2 = \{10, \ldots, 25\}$, $D_3 = \{5, 6, 7, 8, 9\}$ and

(ii) Set 3: $D_2 = \{8, \ldots, 25\}$, $D_3 = \{5, 6, 7\}$.

Figure 1 shows that the posterior estimates for both sets of data ($\rho = 5$ and $\rho = 20$) are robust to misspecifications of $D_k$’s, and the fraction of misclassified observations is also virtually constant.

In summary, it is recommended to carry out sensitivity analyses and goodness-of-fit analyses balancing lack-of-fit and consistent parameter estimation before fixing hyperparameter values. Goodness-of-fit tests may show adequate fitting using a normal distribution. However, if a lack of fit is demonstrated near the tails, sensitivity of results can be assessed by choosing different values of $\rho$, starting from a moderate value (say 5) and choosing a value within a range that gives consistent posterior inferences (for the yeast data we chose $\rho = 20$). With sufficient data, an alternative strategy is to treat $\rho$ as an unknown parameter and update it from its posterior distribution.

5.2 Yeast data set

The HGHMM algorithm was implemented using the R software. As proof of principle, it was applied to the longest contiguous mapped region, corresponding to about 61 Kbp (chromosomal coordinates 12921 to 73970), of yeast chromosome III (Yuan et al., 2005). Cleaned and normalized data was used from all 8 replicates. The length ranges for the three states were: (1) linker: $D_1 = \{1, 2, 3, \ldots\}$, (2) delocalized nucleosome: $D_2 = \{9, \ldots, 30\}$, and (3) well-positioned nucleosome: $D_3 = \{6, 7, 8\}$. Although the model assumes that delocalized nucleosomes occupy within the range of 9-30 probes, extensions beyond 30 did not lead to any significant differences in the predictions, so the upper restriction was used to maximize computational efficiency.
5.2.1 Model fitting. Performing 10,000 MCMC iterations of the algorithm required about 48 hours on an AMD Athlon 1600+ 1.4GHz MP processor with 2GB DDR RAM. Based on the posterior autocorrelations and convergence diagnostics using the R CODA package, it was found that half the number of iterations, with a burn-in of about 10% was sufficient (see Supplementary material), making it favorable to apply this method in whole genome studies. The most computationally expensive step in the algorithm corresponds to state updating, which scales up linearly with the number of regions under study. QQ plots indicated that the state-specific hierarchical t-distribution used in the HGHMM fitted the data well for each predicted state (Figure 2) especially near the tails, which are not fitted well using Gaussian distributions. The posterior estimates for the state-specific mean parameters $\mu$ indicate that the distributions are well-separated, and the posterior expectation of the state lengths are significantly different (Table 1). The modified algorithm allowing the nucleosome state length to vary gave approximately equivalent results, with a slight increase in the variance of estimates. The algorithm was most successful when the length sampling step was introduced after the algorithm had attained a moderate degree of convergence, with the state lengths converging to be the correct ones, even if the initial values are chosen to be vastly different. MCMC convergence diagnostics show that this procedure is computationally feasible for a data set of this size; Gelman’s $\hat{R}$ statistic $\leq 1.01$ (Gelman and Rubin, 1992) (Figures S1 and S2 in Supplementary material).

[Figure 2 about here.]

[Table 1 about here.]
5.2.2 Biological validation of predictions. The model was next assessed by judging how well it supports biological facts relating to this data, in particular whether nucleosome-free state predictions correlate with the location of TFBSs. Harbison et al. (2004) used genomewide location analysis (ChIP-chip) to determine occupancy of DNA-binding transcription regulators under a variety of conditions. The ChIP-chip data give locations of binding sites to only a 1Kb resolution, making further analysis necessary to determine the location of binding sites at a single nucleotide level.

**TFBS overlap with nucleosome regions.** The challenge was to determine whether known functional binding sites could be computationally discovered based only on the knowledge of predicted state. For the HGHMM algorithm, the probabilities of state membership for each probe were estimated from the posterior frequencies of visiting each state (Figure 3) in $M$ iterations (excluding burn-in). Each region was assigned to the occupancy state $k$, for which the estimated posterior state probability $\hat{P}(Z_i = k|Y) = \sum_{j=1}^{M} I(Z_i^{(j)} = k)/M$ was maximum. For all probes, this probability ranged from 0.5 to 0.9 (Figure S3 in the Supplementary material), and hence there was no necessity for resolving ties. An alternative strategy for state prediction (for cases where the decision may not be as clear cut) may be to use the states corresponding to the maximum observed posterior probability (MAP). For this application, however, both approaches yielded identical results. The motif sites predicted using two algorithms was compared for overlap with the bound regions detected by ChIP-chip experiments (Harbison et al, 2004). The binding data were compiled by combining results of all 310 experiments for the region of chromosome III relevant to
the analysis. Binding was considered significant if the multiple comparison-adjusted p-value was ≤ 0.005 in 3 yeast species.

**Results.** Two motif discovery methods SDDA (Gupta and Liu, 2003) and BioProspector (Liu et al., 2001) were used to analyze the sequences. SDDA predicts motifs based on over-represented patterns in the data, through a recursive data augmentation procedure in a Bayesian framework. It uses a stopping rule based on the maximum a posteriori score to prevent over-fitting by inclusion of too many motifs. Bioprospector uses an approximate Gibbs sampling technique. Motif searches were run separately on linker (L), nucleosomal (N) and delocalized nucleosomal (D) regions predicted by the HGHMM procedure, as well as states predicted from the profile HMM (PHMM) of Yuan et al. (2005). All probes in a single run of states were concatenated into “regions”. Under the HGHMM procedure, merging led to a total of 410 contiguous regions, from 599 (L), 1612 (D) and 840 (N). For the PHMM procedure, there were a total of 478 regions (L:1257, D:624, N:1170).

Motif discovery using BioProspector (BP) and SDDA was conducted for motif lengths of 8 to 10 and a maximum of 20 motifs per set. Motif predictions were compared by (i) their match to known motifs in the SCPD database (Zhu and Zhang, 1999) and literature referenced in Harbison et al. (2004), and (ii) positional overlap with factor-bound regions (from the Harbison ChIP-chip data). The entire procedure was repeated for the regions predicted by the PHMM classification of Yuan et al. (2005).

[Table 2 about here.]

The highest specificity (proportion of regions containing motif sites corresponding to high binding propensities in the Harbison data) was for the
linker regions predicted by HGHMM: 61% by SDDA and 40% by BP (Table 2). The linker regions predicted by PHMM overlapped to a lower extent with the binding data (38% by SDDA and 23% by BP). Sensitivity is defined as the proportion of highly TF-bound regions found when regions were classified according to specific state predictions. The highest overall specificity and sensitivity was observed for the linker regions predicted with HGHMM. The PHMM predicts a much higher number of linker regions, so it is not surprising that almost all highly bound regions overlap with PHMM-predicted linker regions, leading to comparable sensitivity but low specificity. Figure 3 is a snapshot of a segment of the genomic region indicating predicted binding site “hotspots” in regions with high posterior linker state probabilities.

[Figure 3 about here.]

The motif profiles predicted were compared to the annotated motifs in the SCPD database and Harbison et al. (2004). Among the top 20 motifs predicted by HGHMM-SDDA, there were 9 partial or full matches to known motifs, of which 8 were predicted in linker regions and 1 in the delocalized nucleosomal regions. HGHMM-BP found approximately 4 of these motifs. PHMM-SDDA and PHMM-BP each found 3 motifs (Table 3); however, one of these (FKH1) was in a predicted nucleosomal region and was a weaker match than the one predicted by HGHMM, indicating that this may have been a misclassified state. Motifs were determined to “match” the specificity in the literature if they were equivalent in at least 80% of the positions that had an information content of higher than 1 bits. The strongest motifs found by HGHMM were STE12, RAP1 and FKH1. Surprisingly, one of the
motifs predicted by HGHMM-SDDA, FKH1, had an even better match to the binding specificity from the literature than the one found by Harbison et al. (2004) in the location experiment.

[Table 3 about here.]

6. Discussion

This paper proposes a novel model and methodology for analyzing high density tiling array data with explicit modeling of state-specific lengths, in a Bayesian framework. The hierarchical structure of the model allows robust modeling of the data adjusting for heavy tails and can incorporate probe-specific biases and variability without over-parametrization. An efficient method for state updating and parameter estimation based on recursive data augmentation techniques is presented. The method can use general truncated range state duration distributions beyond the exponential family, hence making the model attractive for a variety of applications, not necessarily limited to nucleosomal positioning arrays, but any experiments designed to determine length-restricted features in sequential data. Another attractive feature of using length-restricted HMMs where applicable is the resulting stability of estimates, avoiding label-switching problems typical to HMMs. Although the present data set consisted of independent replicates at every point, it is straightforward to extend the method to multivariate observations.

The application on yeast nucleosomal positioning array data shows several favorable results, including highly significant overlaps of predicted nucleosome-free regions with known functional transcription factor (TF) binding sites. Using predicted linker regions without any prior knowledge of binding site locations or specificities, led to the de-novo discovery of 9 out of 20 binding
motifs that are known to be functional in those regions. Results indicate a promising avenue for future research into how nucleosome occupancy relates to TF binding, and the development of appropriate statistical models that may capture this relationship.

7. Supplementary materials
These may be accessed at the Biometrics website.

ACKNOWLEDGEMENTS
The author would like to thank Jason Lieb and Paul Giresi for many helpful discussions on chromatin structure and tiling array technology, and also Jun Liu for helpful comments and suggestions. This work was partially supported by an IBM junior faculty research award from the University of North Carolina at Chapel Hill.

REFERENCES


Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F.


21
List of Figures

1 Sensitivity analysis for hyperparameter $\rho$ in a simulation study. Top panel: maximal posterior MSEs (over all parameters) and misclassification rates for HGHMM for different settings of hyperparameter $\rho$ (solid, dotted and dashed lines of varying widths) for data generated under different choices of $\rho$ (horizontal axis). Lower panel: MSEs and misclassification rates under misspecification of values of truncated negative binomial range for two data sets generated under different hyperparameter specifications (solid and dotted lines), and with a hyperparameter setting $\rho = 20$. The horizontal axis depicts the settings for the range of the truncated negative binomials- Range 1: $(r_2, s_2) = (9,30), (r_3, s_3) = (6,8)$; Range 2: $(r_2, s_2) = (10,25), (r_3, s_3) = (5,9)$; and Range 3: $(r_2, s_2) = (8,25), (r_3, s_3) = (5,7)$. This figure appears in color in the electronic version of this article. .................. 24

2 QQ plots of the three predicted states in nucleosomal array data using a HGHMM with the hierarchical t-distribution (upper panel) and with a Gaussian distribution (lower panel). This figure appears in color in the electronic version of this article. .................. 25
Region of chromosome III with posterior probability of states from HGHMM and corresponding predicted motif sites from SDDA, showing certain binding site “hotspots” in predicted linker regions (blue crosses) which are almost completely devoid of well-positioned nucleosomes. This figure appears in color in the electronic version of this article.
Figure 1.
Figure 2.
Figure 3.
Table 1

Overall posterior summaries for 3 states for $\mu$ (expectation of probe level measurement) and $r/\phi$ (approximate expectation of state length) for the HGHMM with (a) fixed state lengths and (b) variable nucleosome state length.

<table>
<thead>
<tr>
<th>Method</th>
<th>Param</th>
<th>Linker</th>
<th>Deloc</th>
<th>Nucleo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>(a) Fixed</td>
<td>$\mu$</td>
<td>-0.833</td>
<td>0.011</td>
<td>0.165</td>
</tr>
<tr>
<td>Length</td>
<td>$r/\phi$</td>
<td>3.969</td>
<td>0.281</td>
<td>12.266</td>
</tr>
<tr>
<td>(b) Variable</td>
<td>$\mu$</td>
<td>-0.811</td>
<td>0.008</td>
<td>0.165</td>
</tr>
<tr>
<td>Length</td>
<td>$r/\phi$</td>
<td>3.952</td>
<td>0.304</td>
<td>12.056</td>
</tr>
</tbody>
</table>
Table 2

Specificity (Spec): Proportion of regions predicted in a state overlapping with TF-bound regions from Harbison et al (2004), and Sensitivity (Sens): % of bound regions not missed. Higher proportion of linker regions than nucleosomal overlap with TF-bound regions from Harbison et al.

<table>
<thead>
<tr>
<th></th>
<th>HGHMM</th>
<th></th>
<th></th>
<th>PHMM</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SDDA</td>
<td>BP</td>
<td>SDDA</td>
<td>BP</td>
<td>SDDA</td>
<td>BP</td>
</tr>
<tr>
<td>Linker</td>
<td>Spec</td>
<td>Sens</td>
<td>Spec</td>
<td>Sens</td>
<td>Spec</td>
<td>Sens</td>
</tr>
<tr>
<td></td>
<td>0.61</td>
<td>0.7</td>
<td>0.40</td>
<td>0.87</td>
<td>0.38</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>0.19</td>
<td>0.8</td>
<td>0.15</td>
<td>0.63</td>
<td>0.16</td>
<td>0.37</td>
</tr>
<tr>
<td>Nucleosomal</td>
<td>0.16</td>
<td>0.5</td>
<td>0.09</td>
<td>0.43</td>
<td>0.20</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3
Sequence logos (Schneider and Stephens, 1990) of motifs in linker regions predicted by HGHMM and SDDA. Sequence logos are a graphical representation of a multiple sequence alignment. Each logo consists of stacks of symbols, one stack for each position in the sequence. The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each letter at that position. “ExptMatch” denotes matches in SCPD database and from other previous literature (“Source”), cited in Harbison et al, 2004 (“n” denotes none of the letters A, C, G or T had a significantly higher frequency). “PHMM” denotes the predicted state of the region if the motif was also found on running PHMM (Yuan et al., 2005) and SDDA (L, D and N for predicted linker, delocalized nucleosomal, or nucleosomal). All predicted motifs in the table are from SDDA (BP gave fewer matching predictions for both cases - HGHMM and PHMM).

<table>
<thead>
<tr>
<th>HGHMM motif</th>
<th>TF match</th>
<th>ExptMatch</th>
<th>Source</th>
<th>PHMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATTTA</td>
<td>BAS1</td>
<td>SCPD</td>
<td>L</td>
<td></td>
</tr>
<tr>
<td>TTGTTTACC</td>
<td>FKH1</td>
<td>Literature</td>
<td>L</td>
<td></td>
</tr>
<tr>
<td>GATAA</td>
<td>GAT1</td>
<td>Literature</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>TTCTAGAA</td>
<td>HSF1</td>
<td>Literature</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>CACCCAnACAn</td>
<td>RAP1</td>
<td>SCPD</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>TGAAACA</td>
<td>STE12</td>
<td>SCPD</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>TTTCG</td>
<td>SW16</td>
<td>SCPD</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>ATTTCA</td>
<td>LNO2</td>
<td>Literature</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>