

# Heterogeneity in DNA Multiple Alignments: Modeling, Inference, and Applications in Motif Finding <sup>\*</sup>

Gong Chen and Qing Zhou<sup>†</sup>

Department of Statistics, University of California, Los Angeles

## Abstract

Transcription factors bind sequence-specific sites in DNA to regulate gene transcription. Identifying transcription factor binding sites (TFBS's) is an important step for understanding gene regulation. Although sophisticated in modeling TFBS's and their combinatorial patterns, computational methods for TFBS detection and motif finding often make oversimplified homogeneous model assumptions for background sequences. Since nucleotide base composition varies across genomic regions, it is expected to be helpful for motif finding to incorporate the heterogeneity into background modeling. When sequences from multiple species are utilized, variation in evolutionary conservation violates the common assumption of an identical conservation level in multiple alignments. To handle both types of heterogeneity, we propose a generative model in which a segmented Markov chain is used to partition a multiple alignment into regions of homogeneous nucleotide base composition and a hidden Markov model (HMM) is employed to account for different conservation levels. Bayesian inference on the model is developed via Gibbs sampling with dynamic programming recursions. Simulation studies and empirical evidence from biological data sets reveal the dramatic effect of background modeling on motif finding, and demonstrate that the proposed approach is able to achieve substantial improvements over commonly used background models.

**Key Words:** Background modeling; Evolutionary conservation; HMM; Motif finding; Nucleotide base composition; Segmentation; Transcription factor binding site.

## 1 Introduction

Gene transcription is regulated by interactions between transcription factors (TFs) and their binding sites on DNA sequences. Locating transcription factor binding sites (TFBS's) is crucial for understanding how the cell regulates its genes in response to developmental and environmental changes. Due to the time-consuming nature of experimental identification, many computational approaches have been developed to detect TFBS's by utilizing sequence similarity among binding sites of the same TF. Such sequence similarity is usually summarized as a motif, and the detection procedure is referred to as motif finding. As a statistical model, a motif is defined as a sequence of  $w$  independent multinomial

---

<sup>\*</sup>UCLA Statistics Preprints (2009), to appear in *Biometrics*.

<sup>†</sup>Email: zhou@stat.ucla.edu (corresponding author).

distributions on the four nucleotide bases (A, C, G, and T), where each distribution is defined for one position of a motif. Such a model is known as the position-specific weight matrix (PWM) (Stormo and Hartzell, 1989; Lawrence and Reilly, 1990). In what follows, we may call TFBS's motif sites or sites for simplicity. By treating motif sites as signals and surrounding DNA sequences as background, motif finding can be understood as a problem of detecting signals from background. To detect motif sites on a sequence with a given PWM, a null model, usually called the background model, is first estimated from all the bases of the sequence; then a sequential scan is applied to calculate the probability ratio of every word of width  $w$  under the PWM model over the background model.

Substantial efforts have been made recently to enhance probabilistic models for motifs and their combinatorial patterns. Please see Ji and Wong (2006) for a review. On the contrary, less attention has been paid to the modeling of background sequences and its potential effects on motif finding. An i.i.d. multinomial distribution or a homogeneous first order Markov chain is commonly assumed for modeling background (nucleotide) bases, such as in Lawrence et al. (1993), Bailey and Elkan (1994), Liu, Neuwald, and Lawrence (1995), and many other methods reviewed in Ji and Wong (2006). However, base content in a DNA sequence changes from region to region, which obviously violates the homogenous assumption and may have a negative impact on motif finding. For example, in regions of high GC content, base G or C (G/C) appears more frequently than in neutral or low GC regions. Hence, G/C's observed in such regions should be given less credit for being a part of a motif site. A homogenous background model would mix GC rich regions with low GC content regions to give a neutral estimation of GC content. Consequently, it would increase the chance of finding spurious motifs, such as `GGCCGGG` which is likely to appear in a GC rich region. From a statistical point of view, such an inaccurate background model is expected to lower the efficiency in motif finding, resulting in more false positive predictions and a lower power for detecting motif sites.

One natural way to handle this problem is to segment a DNA sequence into homogeneous regions in terms of base composition. There have been many statistical studies on DNA sequence segmentation. We refer to Braun and Müller (1998) for a review. Churchill (1989) introduces a hidden Markov model (HMM) in which each hidden state indicates a different base emission distribution (or segment type) and bases are generated independently given the hidden state. Boys and Henderson (2004) model local dependence between neighboring bases via a Markov chain of an unknown order. They further assume that the number of hidden states is unknown and infer jointly the number of states and the order of neighboring dependence. Liu and Lawrence (1999) propose to segment a sequence into consecutive non-overlapping segments with multiple change points. The number and locations of change points are inferred in a Bayesian setting. In a similar formulation, Braun, Braun, and Müller (2000) use quasi-deviance to measure the model fitting quality and adapt the Schwarz criterion for selecting the number of segments. More recently, it is observed that motif scores with a segmented background model enhance the separation between TF binding sequences and random control sequences (Zhou and Liu, 2008).

Providing another source of information for motif finding, comparative genomics have shown that motif sites can be conserved across multiple species—these sites are bound by transcription factors, and

thus evolve slowly in the evolution. Several recent studies, such as Moses, Chiang, and Eisen (2004), Sinha, Blanchette, and Tompa (2004), Li and Wong (2005), Siddharthan, Siggia, and van Nimwegen (2005), Zhou and Wong (2007), Ray et al. (2008), and Xie et al. (2008), were proposed to utilize evolutionary conservation to facilitate motif finding, with the notion that motif sites tend to be more conserved than background bases. However, when data change from sequences of a single species to sequence alignments of multiple species, heterogeneity in background becomes even more complicated due to evolutionary divergence. Regions under different selective pressures present different levels of conservation. The common assumption of a single evolutionary rate in background modeling as in many of the above studies deserves serious checking. Many conserved regions that contain no TFBS’s exist in multiple alignments of regulatory sequences. False positives from such regions may be produced in motif finding under the homogeneity assumption which pools regions of different conservation levels to estimate a single evolutionary rate for background.

Thus, it is desirable to incorporate heterogeneity in both base composition and evolutionary conservation in multiple alignments into background modeling for motif finding applications. In this study, we propose a generative model to capture these two aspects simultaneously. More specifically, a segmented Markov chain is developed to account for heterogeneity in base composition of aligned DNA sequences, and an HMM is used to model different conservation levels. Empirical evidence from simulated and real world data sets demonstrates that with the proposed heterogeneous background model the performance of motif detection methods improves considerably. The paper is organized as follows. In Section 2 we define statistical models in our framework. The computing strategy and Bayesian inference are developed in Section 3. We define a multiple species motif model for motif detection applications and describe computational methods to be compared in Section 4. We present results from a simulation study in Section 5 and from a case study on two biological data sets in Section 6. We conclude this paper with a discussion in Section 7.

## 2 Statistical models

The input of our model is a (multiple) alignment, which is a set of aligned sequences, each from a species (Figure 1(a)). In addition to the four bases  $\mathcal{D} = \{A, C, G, T\}$ , there exists in an alignment another symbol “-”, which is called a gap in biology. A gap, denoted by  $\phi$ , indicates absence of a base in an aligned position of a sequence (see Web Appendix A for more details on treatment of gaps). We write an alignment  $\mathcal{S} = (s_{ij})_{n \times L}$ ,  $s_{ij} \in \mathcal{D} \cup \{\phi\}$ , as an  $n \times L$  array of symbols, where  $n$  is the number of species (sequences) and  $L$  is the length of the alignment (number of columns). Denote by  $\mathcal{S}_j = (s_{1j}, \dots, s_{nj})$  the  $j$ th column of  $\mathcal{S}$  for  $j = 1, \dots, L$ . The symbols in an alignment column  $\mathcal{S}_j$  are not independent, and their dependence is captured by a binary tree  $\mathcal{T}_j$  (Figure 1(b)), which is referred to as an evolutionary/phylogenetic tree. The tree topology specifies how a common ancestor evolves into its descendants over the course of evolution. The length of a branch is proportional to the amount of evolutionary divergence between the parent node and the child node on the branch. Statistically, an evolutionary tree can be viewed as a graphical model in which every node (vertex)

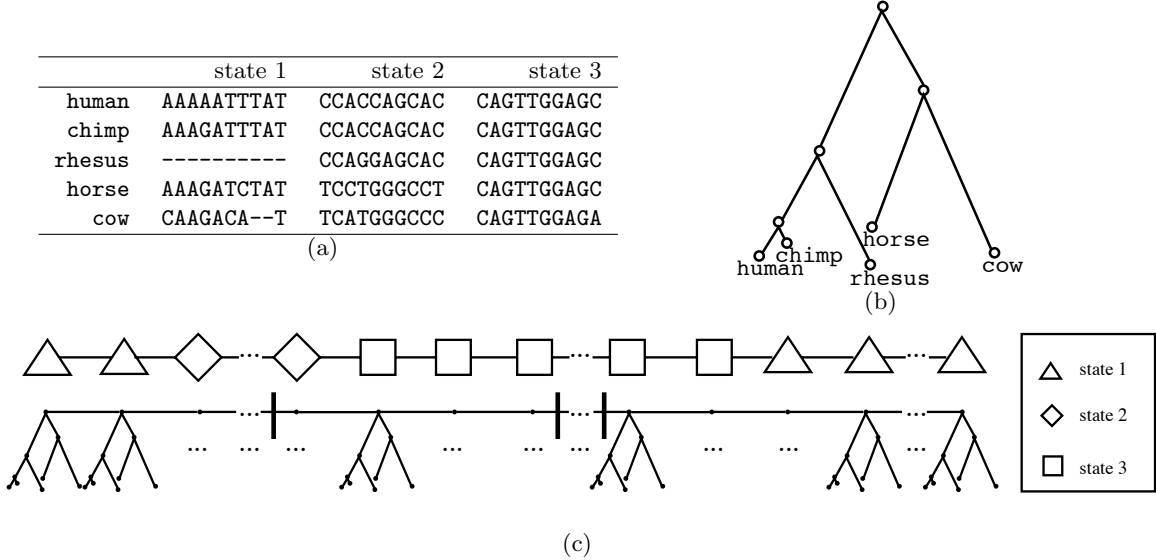


Figure 1: Input data and our model. (a) Alignment columns in three conservation states with a dash indicating a gap. (b) An evolutionary tree of five species. (c) The model schema. The root sequence is segmented into regions of homogeneous base composition, with segment boundaries indicated by solid vertical bars. On the other hand, each alignment column (tree) is associated with a hidden state for evolutionary conservation. The three hidden states are labeled by the symbols of triangle, diamond, and square. Both segmentation and hidden conservation states are to be estimated. Note that although tree topologies are the same, the parameters  $(\beta, \theta)$  are segment-specific and  $(\lambda_D, \lambda_M)$  are state-specific.

represents a random variable taking values on  $\mathcal{D} \cup \{\phi\}$ . The leaf nodes of the tree  $\mathcal{T}_j$  correspond to the observed alignment column  $\mathcal{S}_j$  with one leaf node matching the symbol from one of the  $n$  species. The internal nodes are unobserved. To take into account local dependence in DNA sequences, root nodes are modeled by a first order Markov chain. Under these assumptions, an alignment is augmented to a sequence of trees correlated via their root nodes.

In our framework, a sequence of root nodes is partitioned into consecutive non-overlapping subsequences in terms of base composition as shown in Figure 1(c). Root nodes in the same subsequence or segment share a common preference of the four letters A, C, G, and T. In parallel, a hidden Markov model with three hidden states is used to model evolutionary events on the trees, where the hidden states encode different levels of evolutionary conservation across species (Figure 1(c)). Specifically, state 1 is intended to cover columns involving many gaps, state 2 is for columns with few gaps but many distinct bases from the  $n$  species, and state 3 labels the most conserved columns having almost identical bases. Figure 1(a) illustrates some potential alignment regions in the three states. We note that a tree within one segment may be associated with any of the three conservation states.

In the following, we will first introduce the evolutionary model associated with a tree and then describe the segmentation and the HMM components of our model. In this study, we assume that tree topologies and branch lengths are given. For readers' convenience, Table 1 summarizes notations in the article.

Table 1: Summary of notations

Notation	Definition
$\phi$	Gap
$\mathcal{D}$	{A, C, G, T}
$\mathcal{S}$	Alignment
$L$	Length of an alignment
$K$	Number of segments
$\mathcal{U}$	Hidden conservation states
$\mathbf{V}_D$	Deletion indicators
$\mathbf{V}_M$	Mutation indicators
$\mathbf{V}$	$(\mathbf{V}_D, \mathbf{V}_M)$
$\mathbf{W}$	Segment start positions
$\mathbf{R}$	Root nodes
$\mathbf{Z}$	Internal nodes excluding root nodes
$\mathbf{I}$	$(\mathbf{R}, \mathbf{Z})$
$\mathbf{X}$	$(\mathbf{R}, \mathbf{Z}, \mathcal{S})$
$Z_p$	Parent node
$Z_c$	Child node
$\alpha$	Transition probabilities of the HMM for conservation levels
$\beta$	Transition probabilities of a segmented Markov chain for root sequence
$\theta$	Cell probabilities of segment-specific multinomial distributions for mutated nodes
$\psi$	$(\beta, \theta)$
$\lambda_D$	Deletion rates
$\lambda_M$	Mutation rates
$\lambda$	$(\lambda_D, \lambda_M)$

## 2.1 An evolutionary model

An evolutionary process is assumed to generate symbols for the nodes of a tree. Typical evolutionary events during this process include deletion, insertion, and mutation. The three events correspond to, respectively, three types of transitions from the parent node to the child node on a branch, namely, the transition from  $d$  to  $\phi$ , from  $\phi$  to  $d$ , and from  $d$  to  $d'$  for  $d, d' \in \mathcal{D}$ . In evolutionary biology, deletion and insertion are often regarded as reversible processes. Thus, to simplify our model and to avoid potential non-identifiability problems, we do not consider insertion in this work. Consequently, we assume that there is always a base on a root node. Under this assumption, regions that contain many gaps can be characterized by many deletion events, which gives good background contrast to few deletion events in motif sites. In this sense, deletion alone seems sufficient for the purpose of background modeling in motif finding, as confirmed in our real data analyses.

Let  $t$  be the length of a branch, and denote by  $Z_p$  and  $Z_c$  the parent and child nodes, respectively. For the ease of understanding, we introduce two indicator variables,  $V_D$  and  $V_M$ , to indicate the occurrence of deletion and mutation on a branch, respectively. If  $Z_p \in \mathcal{D}$ , either deletion ( $V_D = 1$ ) or mutation ( $V_M = 1$ ) may happen. We assume that  $P(V_D = 1 | Z_p \in \mathcal{D}) = 1 - e^{-\lambda_D t}$ , where  $\lambda_D$  is referred to as the deletion rate, and  $P(V_M = 1 | V_D = 0, Z_p \in \mathcal{D}) = 1 - e^{-\lambda_M t}$ , where  $\lambda_M$  denotes the mutation rate, i.e., the expected number of mutations per unit branch length (Felsenstein, 1981). Intuitively, the larger the mutation/deletion rate and the longer the branch length, the more likely a

mutation/deletion event will happen. Hereafter, mutation and deletion rates may be called collectively as evolutionary rates. Please note that  $V_D = 1$  automatically implies  $V_M = 0$  since the two events are mutually exclusive. When a mutation event happens, a mutated base is generated independently from a multinomial distribution,

$$[Z_c | Z_p \in \mathcal{D}, V_M = 1] \sim \mathcal{MN}(1, \boldsymbol{\theta}), \quad (1)$$

with cell probabilities  $\boldsymbol{\theta} = (\theta_A, \theta_C, \theta_G, \theta_T)$ . If neither deletion nor mutation happens,  $Z_c$  will be identical to  $Z_p$ . Finally, if  $Z_p = \phi$  then  $Z_c = \phi$ , which implies that the descendants of a gap node are all gaps. In this case we define  $V_D = V_M = 0$ . Obviously, given a root base and the parameters  $\lambda_D$ ,  $\lambda_M$ , and  $\boldsymbol{\theta}$ , evolutionary events and all descendant nodes can be generated according to the above evolutionary model.

## 2.2 A segmentation model

To model heterogeneity in base composition, we assume that distributions for generating bases change along the root sequence. Specifically, a tree sequence is segmented at the root level into  $K$  consecutive non-overlapping segments, where  $K$  is unknown, and each root segment corresponds to a subsequence of trees (Figure 1(c)). Let  $\mathbf{W} = (W_1, \dots, W_K)$  denote the start positions of the segments,  $\mathbf{R}$  the root nodes, and  $\mathbf{Z}$  the other unobserved internal nodes. We use  $\mathbf{R}_{[W_p, W_{p+1}]}$ ,  $\mathbf{Z}_{[W_p, W_{p+1}]}$ , and  $\mathcal{S}_{[W_p, W_{p+1}]}$  to denote the respective subsets of  $\mathbf{R}$ ,  $\mathbf{Z}$ , and  $\mathcal{S}$  in the  $p$ th segment. For each  $p$ ,  $\mathbf{R}_{[W_p, W_{p+1}]}$  is modeled as a homogeneous first order Markov chain with a transition probability matrix  $\boldsymbol{\beta}_p$ . The multinomial distribution, with cell probabilities  $\boldsymbol{\theta}_p$ , for generating mutated bases (Equation (1)) in the evolutionary model is also assumed to be segment-specific. In summary, the segment-specific parameters  $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K)$  and  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)$  capture the base compositional heterogeneity.

## 2.3 An HMM for evolutionary conservation

Several studies have been proposed to model conservation variation in a multiple alignment (Yang, 1995; Felsenstein and Churchill, 1996; Siepel and Haussler, 2004). Felsenstein and Churchill (1996) introduce an HMM to account for variation in mutation rates. Similarly, we utilize an HMM with three hidden states as the underlying structure to model variation in conservation. Each hidden state has its own evolutionary rates. The spatial dependence between states of neighboring trees is captured by a first order Markov chain with a transition matrix  $\boldsymbol{\alpha} = (\alpha_{q'q})_{3 \times 3}$ , where  $\alpha_{q'q}$  is the transition probability from state  $q'$  to state  $q$  for  $q', q \in \{1, 2, 3\}$ . The deletion and mutation rates for the three states are denoted by  $\boldsymbol{\lambda}_D = (\lambda_{1D}, \lambda_{2D}, \lambda_{3D})$  and  $\boldsymbol{\lambda}_M = (\lambda_{1M}, \lambda_{2M}, \lambda_{3M})$ , respectively. Let  $U_i$  denote the hidden conservation state of the  $i$ th tree for  $i = 1, \dots, L$ . Given  $U_i = q$ , the evolutionary events on the  $i$ th tree are generated by the evolutionary model defined in Section 2.1 with the rates  $\lambda_{qD}$  and  $\lambda_{qM}$ .

### 3 Bayesian inference

Let  $\mathbf{U}$  denote conservation states,  $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_D, \boldsymbol{\lambda}_M)$  evolutionary rates, and  $\mathbf{V} = (\mathbf{V}_D, \mathbf{V}_M)$  evolutionary event indicators. Under the model assumptions, the complete data likelihood can be written as (see Table 1 for summary of notations)

$$P(\mathbf{R}, \mathbf{Z}, \mathcal{S}, \mathbf{V}, \mathbf{U} \mid \mathbf{W}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) = P(\mathbf{R} \mid \mathbf{W}, \boldsymbol{\beta})P(\mathbf{U} \mid \boldsymbol{\alpha})P(\mathbf{Z}, \mathbf{V}, \mathcal{S} \mid \mathbf{R}, \mathbf{U}, \mathbf{W}, \boldsymbol{\theta}, \boldsymbol{\lambda}). \quad (2)$$

To conduct Bayesian inference on the parameters of interest, we prescribe a prior distribution

$$\pi(\mathbf{W}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) = \pi(\mathbf{W})\pi(\boldsymbol{\beta} \mid \mathbf{W})\pi(\boldsymbol{\theta} \mid \mathbf{W})\pi(\boldsymbol{\lambda})\pi(\boldsymbol{\alpha})$$

on the unknowns. A priori, we assume that the number of segments  $K = |\mathbf{W}|$  is uniformly distributed on  $\{1, \dots, k_{max}\}$ , where  $k_{max}$  is the maximum possible number of segments, and that the  $(K - 1)$  segment start positions are uniformly placed on the  $(L - 1)$  tree-sequence positions. Note that the start position of the first segment  $W_1 \equiv 1$ . We pre-determine  $k_{max}$  as  $L/l$ , where  $l$  is an expected lower-bound length of a segment (say  $l = 100$  or  $200$ ). Further increase of  $k_{max}$  does not change the results in this work but will cost more computation. The prior distribution of  $\lambda_{qM}$  is assumed to be the Gamma distribution  $\mathcal{G}(a, b)$  with  $a = b = 0.01$  for  $q \in \{1, 2, 3\}$ . This prior is chosen to approximate a non-informative prior. Such a prior setting has little influence on posterior inference when there is at least one mutation event and a reasonable number of non-mutation events (when a child node retains the same base from its parent node). In practice, these conditions are easily satisfied; otherwise, the value 0.01 tends to give a small mutation rate, which may be reasonable given no evidence of mutation from data. The same prior distribution is assumed for  $\lambda_{qD}$ . The prior distributions for other parameters are flat Dirichlet or flat product Dirichlet distributions. We are interested in the joint posterior distribution

$$P(\mathbf{W}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\alpha} \mid \mathcal{S}) \propto \sum_{\mathbf{R}, \mathbf{Z}, \mathbf{V}, \mathbf{U}} \pi(\mathbf{W}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\alpha})P(\mathbf{R}, \mathbf{Z}, \mathcal{S}, \mathbf{V}, \mathbf{U} \mid \mathbf{W}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\alpha}), \quad (3)$$

where  $P(\mathbf{R}, \mathbf{Z}, \mathcal{S}, \mathbf{V}, \mathbf{U} \mid \mathbf{W}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\alpha})$  is the complete data likelihood (Equation (2)).

Since the above summation has no analytical solution, we devise a Gibbs sampling strategy to sample from the joint distribution  $P(\mathbf{R}, \mathbf{Z}, \mathbf{V}, \mathbf{U}, \mathbf{W}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\alpha} \mid \mathcal{S})$  for posterior inference. In one sampling iteration, two threads proceed first, one thread sampling segments and segment-specific parameters and the other sampling conservation states and related parameters, and then internal nodes and evolutionary event indicators are sampled. Let  $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\theta})$  denote segment-specific parameters and  $\mathbf{I} = (\mathbf{R}, \mathbf{Z})$  denote internal nodes. The conditional sampling in one iteration consists of four steps: (1)  $[\mathbf{W}, \boldsymbol{\psi} \mid \mathbf{I}, \mathcal{S}, \mathbf{V}]$ , (2)  $[\mathbf{U} \mid \mathbf{I}, \mathbf{V}, \boldsymbol{\lambda}, \boldsymbol{\alpha}]$ , (3)  $[\boldsymbol{\lambda}, \boldsymbol{\alpha} \mid \mathbf{U}, \mathbf{I}, \mathbf{V}]$ , (4)  $[\mathbf{I}, \mathbf{V} \mid \mathcal{S}, \mathbf{W}, \mathbf{U}, \boldsymbol{\psi}, \boldsymbol{\lambda}]$ .

We derive step 1 here. Steps 2 and 3 follow the classical HMM paradigm with modifications on the emission model. Step 4 is based on bottom-up summation, introduced as the Pulley principle in Felsenstein (1981), and top-down sampling along a tree. Since the adaptation of these steps to our

framework is not trivial, we provide details in Web Appendix B.

In step 1 we first sample the number of segments  $K = |\mathbf{W}|$ . Let  $\mathbf{X} = (\mathbf{R}, \mathbf{Z}, \mathcal{S})$ . With the uniform prior  $\pi(K = k)$ , the marginal posterior probability

$$P(K = k \mid \mathbf{X}, \mathbf{V}) \propto P(\mathbf{X} \mid \mathbf{V}, K = k) = \sum_{\mathbf{W}:|\mathbf{W}|=k} P(\mathbf{X}, \mathbf{W} \mid \mathbf{V}, K = k). \quad (4)$$

To derive the recursion for summing over  $\mathbf{W}$  to compute  $P(\mathbf{X} \mid \mathbf{V}, K = k)$  for  $k = 1, \dots, k_{max}$ , we write  $P(\mathbf{X} \mid \mathbf{V}, K = k) = P(\mathbf{X}_{[1,L]} \mid \mathbf{V}_{[1,L]}, W_{k+1} = L+1)$ , where  $W_{k+1}$  is a dummy variable indicating the end position of the  $k$ th segment is  $L$ , the index for the last tree. A recursive summation (Auger and Lawrence, 1989; Liu and Lawrence, 1999) is employed to compute  $P(\mathbf{X}_{[1,j]} \mid \mathbf{V}_{[1,j]}, W_{p+1} = j+1)$  for  $p = 1, \dots, k_{max}$  and  $j = p, \dots, L$ :

$$\begin{aligned} & P(\mathbf{X}_{[1,j]} \mid \mathbf{V}_{[1,j]}, W_{p+1} = j+1) \\ &= \sum_{i=p}^j P(\mathbf{X}_{[1,i-1]} \mid \mathbf{V}_{[1,i-1]}, W_p = i) \\ & \quad \times P(\mathbf{X}_{[i,j]} \mid \mathbf{V}_{[i,j]}, W_p = i, W_{p+1} = j+1) \\ & \quad \times \pi(W_p = i \mid W_{p+1} = j+1), \end{aligned} \quad (5)$$

where  $\pi(W_p = i \mid W_{p+1} = j+1)$  is the conditional prior probability of placing  $W_p$  on the  $i$ th tree-sequence position given that the next segment starts at the  $(j+1)$ th position, and  $P(\mathbf{X}_{[i,j]} \mid \mathbf{V}_{[i,j]}, W_p = i, W_{p+1} = j+1)$  is the marginal likelihood of observing all the tree nodes in the  $p$ th segment, with parameters  $(\boldsymbol{\beta}_p, \boldsymbol{\theta}_p)$  integrated out (Equation (1) in Web Appendix B).

With  $P(\mathbf{X} \mid \mathbf{V}, K = k)$  computed by Equation (5) for  $k = 1, \dots, k_{max}$ ,  $K$  can be sampled according to the posterior probabilities in Equation (4). Given  $K$ ,  $\mathbf{W}$  is then sampled backwards with probabilities computed in the forward summation. Specifically, given the  $(p+1)$ th segment start position as  $j+1$ , the  $p$ th segment start position is sampled with probability  $P(W_p = i \mid W_{p+1} = j+1, \mathbf{X}_{[1,j]}, \mathbf{V}_{[1,j]})$  proportional to the summation terms in Equation (5) for  $i = p, \dots, j$ . Conditional on the segments, the segment-specific parameters follow Dirichlet or product Dirichlet distributions. Please note that there is no need to employ reversible jump MCMC (Green, 1995), since the parameters are integrated out in the step of sampling segmentation.

We adopt column-wise averages to estimate parameters of interest based on posterior samples. Let  $\hat{\boldsymbol{\Psi}}_i = \frac{1}{N} \sum_{t=1}^N \boldsymbol{\Psi}_i^{(t)}$  be the column-wise estimate of a generic parameter for the  $i$ th column ( $i = 1, \dots, L$ ) where  $\boldsymbol{\Psi}_i^{(t)}$  is the sampled parameter for the column in the  $t$ th iteration after burn-in. For example, if the  $i$ th column is located in the second segment with the first hidden state, then  $\boldsymbol{\Psi}_i^{(t)} = \boldsymbol{\beta}_2^{(t)}$  for calculating  $\hat{\boldsymbol{\beta}}_i$  and  $\boldsymbol{\Psi}_i^{(t)} = \boldsymbol{\lambda}_1^{(t)}$  for  $\hat{\boldsymbol{\lambda}}_i$ . The above estimates will be used in motif scoring comparisons and later analyses.

## 4 Motif detection in a multiple alignment

To demonstrate the importance of modeling heterogeneity in background for motif detection applications, we will compare various methods with different background models in Sections 5 and 6. In this section, we first define a motif model for multiple species, which will be used in scoring a candidate alignment as a motif site; then we describe the motif detection methods to be compared.

Suppose a motif is of width  $w$ . A multiple species motif model defines a generative process for  $w$  consecutive alignment columns  $\mathcal{S}^m = (\mathcal{S}_1^m, \dots, \mathcal{S}_w^m)$ . For the  $w$  positions (columns) of a motif site, let  $\mathbf{R}^m = (R_1^m, \dots, R_w^m)$  denote the corresponding root nodes,  $\mathbf{Z}^m = (\mathbf{Z}_1^m, \dots, \mathbf{Z}_w^m)$  the other internal nodes, and  $\mathbf{V}^m = (\mathbf{V}_1^m, \dots, \mathbf{V}_w^m)$  the evolutionary event indicators. At the root level, a product of independent multinomial distributions (a PWM) is used to model  $\mathbf{R}^m$ : For the  $i$ th position,  $R_i^m \sim \mathcal{MN}(1, \boldsymbol{\theta}_i^m)$ , where  $\boldsymbol{\theta}_i^m = (\theta_{iA}^m, \theta_{iC}^m, \theta_{iG}^m, \theta_{iT}^m)$ , and  $P(\mathbf{R}^m | \boldsymbol{\theta}^m) = \prod_{i=1}^w \theta_{iR_i^m}^m$ , with  $\boldsymbol{\theta}^m = (\boldsymbol{\theta}_1^m, \dots, \boldsymbol{\theta}_w^m)$ . Given  $R_i^m$ , the other nodes and the evolutionary event indicators on the  $i$ th tree,  $\mathbf{Z}_i^m$ ,  $\mathcal{S}_i^m$ , and  $\mathbf{V}_i^m$ , are assumed to be generated according to the evolutionary model in Section 2.1 with a deletion rate  $\lambda_D^m$ , a mutation rate  $\lambda_M^m$ , and cell probabilities  $\boldsymbol{\theta}_i^m$  of the multinomial distribution for generating mutated bases (Equation (1)), for  $i = 1, \dots, w$ . We assume that all the sites of a motif share the same deletion rate  $\lambda_D^m$  and the same mutation rate  $\lambda_M^m$ . Let  $\boldsymbol{\lambda}^m = (\lambda_D^m, \lambda_M^m)$ . The complete data likelihood for a motif site can be expressed as

$$P(\mathbf{R}^m, \mathbf{Z}^m, \mathbf{V}^m, \mathcal{S}^m | \boldsymbol{\theta}^m, \boldsymbol{\lambda}^m) = \prod_{i=1}^w P(R_i^m | \boldsymbol{\theta}_i^m) P(\mathbf{Z}_i^m, \mathbf{V}_i^m, \mathcal{S}_i^m | R_i^m, \boldsymbol{\theta}_i^m, \boldsymbol{\lambda}^m).$$

Suppose there are  $M$  (independent) sites for a motif. Let  $\mathbb{R}^m, \mathbb{Z}^m, \mathbb{V}^m$ , and  $\mathbb{S}^m$  denote root nodes, other internal nodes, evolutionary event indicators, and alignments (leave nodes), respectively, of these sites. A Gibbs sampler that iterates between the conditional sampling of  $[\mathbb{R}^m, \mathbb{Z}^m, \mathbb{V}^m | \boldsymbol{\theta}^m, \boldsymbol{\lambda}^m, \mathbb{S}^m]$  and that of  $[\boldsymbol{\theta}^m, \boldsymbol{\lambda}^m | \mathbb{R}^m, \mathbb{Z}^m, \mathbb{V}^m, \mathbb{S}^m]$  is employed for the Bayesian inference of the parameters  $\boldsymbol{\theta}^m$  and  $\boldsymbol{\lambda}^m$  with the same prior specification as in the background model. The sample averages are taken as the estimates.

A motif score is computed for every candidate alignment (or sequence) of width  $w$  when we scan an alignment (or sequence) of length  $L$ . The score is defined as the probability ratio of the observed data under a motif model over a background model. We refer to such candidate alignments (or sequences) as candidates for simplicity. In the single species case, it is straightforward to compute the probability of observing a candidate under either model. In the multiple species case, the ratio is

$$\frac{P(\mathcal{S}_c | \boldsymbol{\Psi}^m)}{P(\mathcal{S}_c | \boldsymbol{\Psi})} = \frac{\sum_{\mathbf{R}_c, \mathbf{Z}_c, \mathbf{V}_c} P(\mathbf{R}_c, \mathbf{Z}_c, \mathbf{V}_c, \mathcal{S}_c | \boldsymbol{\Psi}^m)}{\sum_{\mathbf{R}_c, \mathbf{Z}_c, \mathbf{V}_c} P(\mathbf{R}_c, \mathbf{Z}_c, \mathbf{V}_c, \mathcal{S}_c | \boldsymbol{\Psi})}, \quad (6)$$

where  $\mathcal{S}_c$  denotes a candidate alignment (such as the alignments of width 10 in Figure 1(a)),  $\boldsymbol{\Psi}^m$  is the parameters of a motif model, and  $\boldsymbol{\Psi}$  is the parameters of a background model. Under either model, the summation over internal nodes ( $\mathbf{R}_c, \mathbf{Z}_c$ ) and evolutionary event indicators ( $\mathbf{V}_c$ ) can be calculated exactly by dynamic programming recursions (Equations (2) and (3) in Web Appendix B).

Table 2: The inputs and assumptions of the compared methods

Methods	Multiple species	Segmentation	HMM for conservation
HomoSingle	N	N	N
HomoMulti	Y	N	N
HeteMulti	Y	Y	N
HeteMultiHMM	Y	Y	Y

Candidates with a motif score greater than a given cutoff value are regarded as predicted sites, and the corresponding false discovery rate can be estimated by scoring a large set of control alignments. In practice, one may choose a cutoff that gives a reasonable false discovery rate, say  $< 40\%$ .

We consider four methods for motif detection and assign them shorthand names for reference. For scanning single species sequences, HomoSingle uses a homogeneous Markov chain background model and a PWM as the motif model. The other three methods all take multiple alignments as input and use the same multiple species motif model defined above, but their background models are different. HomoMulti uses a homogeneous background model without segmentation, whereas HeteMulti employs the segmentation model for handling heterogeneity in base composition. They both assume a single set of evolutionary rates. HeteMultiHMM is the most comprehensive model that uses the segmented background model for base composition and the three-state HMM for conservation levels. Table 2 summarizes the key features of these methods. To estimate parameters of the multiple species motif model, we ran the Gibbs sampler described in this section for 1,000 iterations. For background parameter estimation in HeteMultiHMM, we ran the Gibbs sampling procedure defined in Section 3 for 1,500 iterations. The same procedure was carried out for HeteMulti without sampling conservation states and state-related parameters. A further reduced procedure was used for HomoMulti without sampling segment start positions and segment-specific parameters. For all runs, the first 50% of iterations were used as burn-in periods. The number of iterations and fraction of burn-in were chosen based on empirical efficiency and convergence analysis, with details provided in Web Appendix C.

## 5 A simulation study

We simulated data sets to verify our parameter estimation for HeteMultiHMM and to compare scoring performance of various motif detection methods. The simulation used an evolutionary tree of five species estimated from a multiple alignment in the muscle data set to be introduced later.

An alignment was formed by three segments, each of length 1,000, with different base compositions: A GC rich region, a uniform region, and an AT rich region. Motif sites were simulated from three distinct motifs, a GC rich motif (MGC), a uniform motif (MUN), and an AT rich motif (MAT) (Web Figure 1). Please see Web Tables 1 and 2 for the simulation parameters for background alignments and motif sites. We generated 20 background alignments and 40 sites for each motif according to these parameters. Two strategies were applied to insert motif sites in a background alignment. In strategy A every motif site was uniformly inserted in a background alignment, while in strategy B motif sites were only inserted in a GC rich segment. The first type of insertion should be fair for every method

Table 3: Average false discovery rates (%) for the data sets DUN (top) and DGC (bottom)

Motif	MGC				MUN				MAT			
Sensitivity	25%	50%	75%	100%	25%	50%	75%	100%	25%	50%	75%	100%
Number of sites	10	20	30	40	10	20	30	40	10	20	30	40
HomoSingle	0	4	19	37	0	6	20	35	0	3	16	35
HomoMulti	0	0	8	23	0	2	9	23	0	0	6	23
HeteMulti	0	0	4	17	0	0	6	19	0	0	3	16
HeteMultiHMM	0	0	1	14	0	0	0	12	0	0	0	7

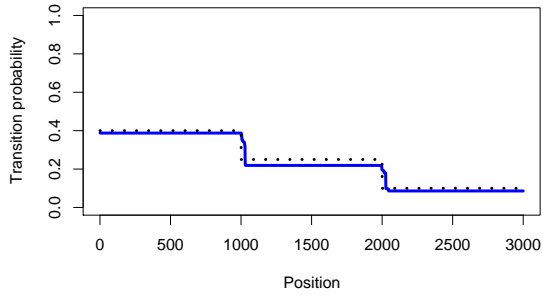
Motif	MGC				MUN				MAT			
Sensitivity	25%	50%	75%	100%	25%	50%	75%	100%	25%	50%	75%	100%
Number of sites	10	20	30	40	10	20	30	40	10	20	30	40
HomoSingle	0	6	22	39	0	6	20	35	0	3	15	33
HomoMulti	0	0	10	25	0	2	9	23	0	0	6	22
HeteMulti	5	13	20	33	0	1	6	18	0	0	0	5
HeteMultiHMM	0	2	8	21	0	0	1	11	0	0	0	2

without any prior knowledge about characteristics of regions surrounding motif sites. However, our analyses of two real world data sets showed that motif sites tend to be located in GC rich regions when more than one segment exists. Therefore, the second type of insertion was considered. These two types of insertion gave rise to two data sets with 20 alignments in each. We refer to the data set with the first type of insertion as DUN and the one with the second type as DGC.

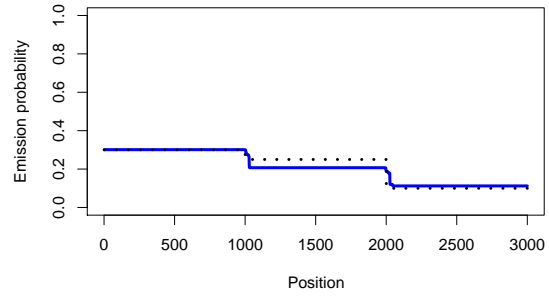
We illustrate our posterior inference with the results of a simulated alignment. The posterior probability  $P(K = 3 | \mathcal{S}) = 1$ , estimated from posterior samples, shows that our method accurately detected the number of segments. The column-wise estimates for typical segment-specific parameters and evolutionary rates are seen to be close to the true parameters (Figure 2). In addition, as is evident from the figure, the segments and the conservation states can be clearly recovered through these estimates.

Cross-validation was performed to evaluate scoring results. Analogous to the scheme used in Barash et al. (2003), the cross-validation treated candidates in one alignment as a test data set and treated sites in the other alignments as a training data set. To compute the motif score of a candidate in a particular alignment (test data), the parameters of the motif model  $\Psi^m$  were estimated from the simulated sites in the other alignments (training data). The background parameters were learned from individual alignments. This procedure was repeated to obtain the motif scores of candidates in every alignment. For each motif, the scores of all candidates in an alignment were computed by Equation (6), and the candidates were ranked in the descending order of their scores. To obtain a global evaluation of motif identification, we calculated the averages of false discovery rates (AFDR) for four sensitivities, 25%, 50%, 75%, and 100% (Table 3). Specifically, for the sensitivity  $\alpha$ ,  $AFDR_\alpha = \frac{1}{M\alpha} \sum_{i=1}^{M\alpha} \frac{FP_i}{FP_i + i}$ , where  $FP_i$  is the number of false positives when  $i$  (simulated) motif sites are detected, and  $M$  is the total number of motif sites. This measure is in essence similar to partial area under the curve (Pepe et al., 2003) for comparing average performance of classification methods.

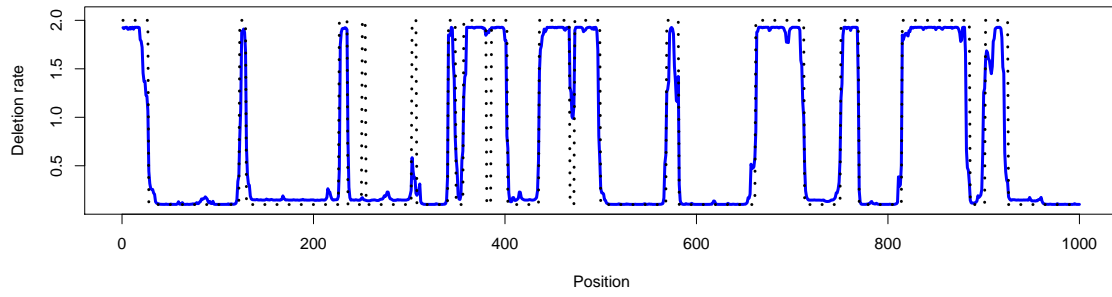
For the data set DUN, Table 3 (top) shows the progressive and substantial improvement of a multiple species motif detection method with more proper background modeling of heterogeneity for all the three motifs. For the data set DGC, the same situation holds for the uniform (MUN) and the



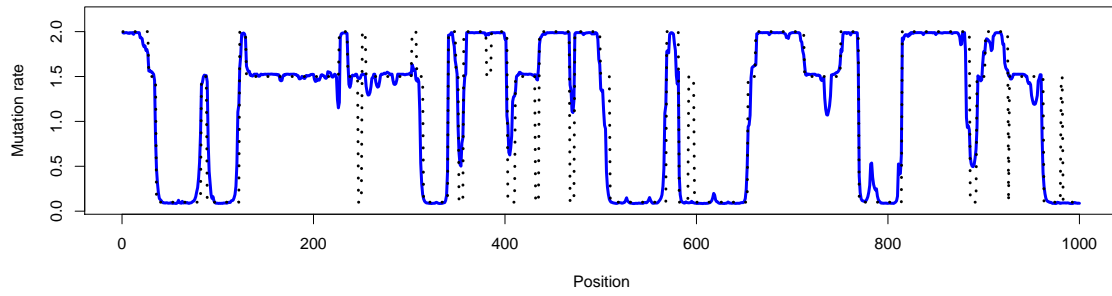
(a)



(b)



(c)



(d)

Figure 2: Estimated parameters: (a) The transition probability  $\beta_{GC}$ , (b) the emission probability  $\theta_G$ , (c) the deletion rate in the first segment, and (d) the mutation rate in the first segment. The dotted and solid lines report the column-wise true parameters and the column-wise estimates, respectively.

AT rich (MAT) motifs. However, when the GC rich motif (MGC) is inserted in GC rich segments, HeteMulti gives more false positives than HomoMulti. In such cases, the background probability for an MGC site is lower under the homogeneous background model than that under the heterogeneous model. The lower background probability leads to a higher score of the MGC site given the same motif model. Therefore, in practice, such behavior of HomoMulti may improve scoring results if it only brings up a negligible amount of false positives whose scores are also increased inappropriately in the same manner as motif sites. On the contrary, when AT rich motif sites are inserted into GC rich regions, HeteMulti provides a much stronger contrast between the motif and the background models than HomoMulti, and thus may improve the scoring results considerably. For example,  $AFDR_{100\%}$ 's for both HeteMulti and HeteMultiHMM when scoring the MAT motif in DGC are much lower than those in DUN.

## 6 A case study on two biological data sets

We applied the four motif detection methods in Table 2 to two biological data sets to further examine the effects of background modeling. Evolutionary trees were estimated from the alignments in the data sets by the PHYLIP package (version 3.67) (Felsenstein, 1989).

The first data set contains human DNA sequences of length between 2 kilobases (kb) and 3kb in the upstream regions of 24 genes with skeletal muscle-specific expression, for which experimentally validated sites of five motifs have been annotated (Thompson et al., 2004). The second data set is composed of 23 human upstream sequences of length 3kb for known  $NF\kappa B$  responsive genes collected from TRANSFAC (Wingender et al., 2000) 9.1 release. For these two data sets, we extracted multiple alignments of 27 vertebrates (Web Figure 2) that are aligned to the human genome (hg18) from the UCSC genome browser database (Karolchik et al., 2008). In an alignment, the top five species in terms of their percent of identity to the human sequence were kept. Then, only species with at least 60 percent of identity were selected from the top five species. Such a stringent selection aims at collecting alignments of relatively high quality so that we may reduce the uncertainty in multiple alignment to a minimal level. Following the common practice of preprocessing data in motif finding applications, we masked out repeat sequences.

### 6.1 Heterogeneity

In order to examine the heterogeneity in base composition and conservation levels in the two data sets, we report some related statistics obtained from HeteMultiHMM, which employs the most comprehensive background model. As shown in Figure 3(a), about half of the alignments show strong empirical evidence for the existence of more than one segment as quantified by the estimated posterior probability  $\hat{P}(K \geq 2 | \mathcal{S}) > 0.9$ . Figure 3(b) displays the estimated evolutionary rates for the three conservation states from the 47 alignments. A state-specific evolutionary rate was estimated by its posterior sample average. The figure clearly illustrates the separation between the three sets of evolutionary rates with their expected meanings: State 1 has the largest deletion rates; state 2 owns

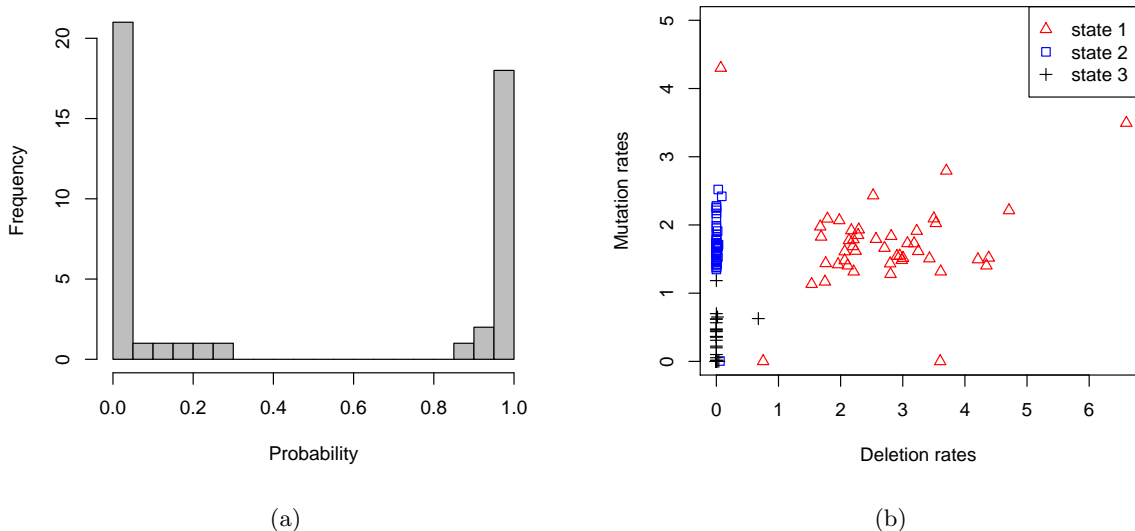


Figure 3: Heterogeneity in the two biological data sets: (a) The histogram of the estimated posterior probability of more than one segment in an alignment and (b) the estimated state-specific evolutionary rates. One case with  $\hat{\lambda}_{1D} = 11.15$  and  $\hat{\lambda}_{1M} = 14.06$  is excluded from (b).

smaller deletion rates than state 1 but higher mutation rates than state 3; and state 3 tends to have the smallest deletion rates and the smallest mutation rates. The above results demonstrate that there exists a large amount of heterogeneity in multiple alignment data, and thus, it is not reasonable to simply make homogeneous assumptions for background modeling.

## 6.2 Motif scoring

Following Thompson et al. (2004), our motif analyses of the muscle data set focused on the well-defined sites for the three TFs, MEF, MYF, and SRF. In summary, the two data sets contain a total of 16, 25, 14, and 29 known binding sites for four transcription factors, MEF, MYF, SRF, and NF $\kappa$ B, respectively. Please see Web Figure 3 for their logo plots. Based on these known binding sites, we employed the same cross-validation scheme to evaluate performance of different methods as we did in the simulation study. Overall, our proposed method HeteMultiHMM is much superior to both HomoSingle, which uses the typical background model in single species motif finding methods, and HomoMulti, whose background model is close to the one commonly used in multiple species methods. The last two rows of Table 4 report the percentage reduction in AFDR from the results of HomoSingle and HomoMulti to those of HeteMultiHMM. To provide another overall performance measure, we highlighted in bold face the methods with the best performance for different sensitivities in Table 4. For 50% sensitivity, HeteMultiHMM shows the best performance in the cases of MEF, MYF, and NF $\kappa$ B. Such performance gain by the heterogeneous models demonstrates the advantage of modeling heterogeneity in background for motif finding. In what follows, we discuss in detail the comparisons

Table 4: Average false discovery rates (%) for the two biological data sets

Motif	MEF			MYF			SRF			NF $\kappa$ B		
	25%	50%	75%	25%	50%	75%	25%	50%	75%	25%	50%	75%
Sensitivity	4	8	12	7	13	19	4	7	11	8	15	22
Number of sites	4	8	12	7	13	19	4	7	11	8	15	22
HomoSingle(HoS)	<b>0</b>	5	16	17	37	48	68	68	72	<b>22</b>	40	<b>47</b>
HomoMulti(HoM)	8	20	30	5	19	32	<b>25</b>	<b>42</b>	<b>57</b>	37	40	51
HeteMulti(HeM)	5	17	29	15	25	36	38	49	62	31	38	49
HeteMultiHMM(HeMH)	<b>0</b>	<b>1</b>	<b>12</b>	<b>4</b>	<b>12</b>	<b>30</b>	42	54	61	25	<b>37</b>	51
(HoS-HeMH)/HoS%	-	71	24	77	69	37	38	22	16	-11	6	-6
(HoM-HeMH)/HoM%	100	93	60	27	38	4	-69	-29	-6	33	7	1

Note: Bold face numbers indicate the methods with the best performance for a particular sensitivity.

between these methods.

First, we compare HomoMulti and HeteMulti to study the segmentation effect. We can see that HeteMulti outperforms HomoMulti for all sensitivities in the cases of MEF and NF $\kappa$ B. For MYF and SRF, however, HeteMulti is not as good as HomoMulti. As in the case of MGC in the simulated data set DGC, some of their sites have a relatively high portion of G/C bases, and they are located in GC rich regions. As we have explained, this may degrade motif scoring results with a heterogeneous background model although segmentation is clearly supported by the data. Second, we compare HeteMulti and HeteMultiHMM to investigate the effect of modeling conservation levels after the base compositional heterogeneity has been taken into account. For all the sensitivities, HeteMultiHMM gives smaller AFDRs than HeteMulti when scoring the motifs MEF and MYF, and the reduction in false positives is mostly more than 50%. For the other two motifs, the two methods seem to have comparable performance across different sensitivities. This confirms the importance of modeling the variation in evolutionary conservation for multiple species motif detection. Finally, our results also confirm the general notion that utilizing sequences from multiple species is helpful for motif detection. However, an important message from this study is that, without proper modeling of background heterogeneity, multiple species methods may be even worse than a single species method. For example, in the case of MEF, whereas HeteMultiHMM shows improvements over HomoSingle, both HomoMulti and HeteMulti show inferior performance. We note that all the multiple species methods give more false positives for 25% and 75% sensitivities in the case of NF $\kappa$ B. The degraded performance is due to the fact that the NF $\kappa$ B sites have very different conservation levels. A more sophisticated motif model that accounts for this additional variation is needed to enhance the detection power of a statistical method.

## 7 Discussion

With increasingly available genomic sequence data, evolutionary conservation across multiple species provides valuable information for detecting TFBS's. However, many existing methods for motif finding ignore background heterogeneity in both base composition and evolutionary conservation present in multiple alignments. In this article, we have proposed a generative model to capture these two types of heterogeneity simultaneously. The empirical evidence from the simulation study and the case study

showed that the model has a great potential to improve motif detection performance.

Although our framework performs segmentation on a first order Markov chain, it can be extended to second or higher-order models because all the segment-specific parameters can be integrated out for sampling segmentation. Despite the usefulness of second order models in some applications suggested by Blaisdell (1985) and Hwang and Green (2004), a first order model seems proper for segmenting background in motif finding applications, where a reasonably large segment of 100 root nucleotide bases often observed in this study may not be sufficient for fitting a model with 64 parameters (as in the case of second order Markov chain) or more. Also, the results in Huang et al. (2004) showed that a Markov chain of a higher order as the background model was not very different from a first order Markov chain for motif scan. Thus, we adopt the simpler model. However, independence assumption on root nucleotide bases should be avoided since it often leads to many very short segments of a few columns in our previous empirical study, which is clearly inappropriate for motif detection. While in our model dependence between nodes of neighboring trees is implicitly taken into account through the inheritance from their root nodes, more elaborate models for such dependence have been considered explicitly by Hwang and Green (2004) and Baele, Van de Peer, and Vansteelandt (2008).

In the context of single species motif finding, several studies have discussed the issue of background modeling. Liu, Brutlag, and Liu (2001) compared an i.i.d. background with a homogeneous third order Markov chain and reported that the later may give more specific predictions. Huang et al. (2004) confirmed the effect of background modeling via the development of a local Markov model. A practical problem in *de novo* motif finding is that single base repeats (e.g., AAAAA ...) or dimer repeats (e.g., CGCGCG ...) are often detected as false positive motifs. To alleviate this problem, Gupta and Liu (2003) proposed to treat low-complexity repeats as a series of adjacent words of the same pattern in a stochastic dictionary model. When turning to multiple species methods, studies such as Thompson et al. (2004), Sinha et al. (2004) and Siddharthan et al. (2005) restrict search space to highly conserved alignment columns, which may reduce heterogeneity in conservation to a certain extent. Nevertheless, it is unclear how much information including motif sites themselves is lost in this way. The approach in this work is expected to utilize more potentially useful information from data by constructing a full model to capture different levels of conservation.

## 8 Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 2-6, and software for background modeling and multiple species motif scan are available at <http://www.stat.ucla.edu/~zhou/htbgscan/>.

## Acknowledgements

We thank the editor, the associated editor, and two referees for their helpful comments and suggestions. This work was supported by NSF grant DMS-0805491.

## References

- Auger, I. E. and Lawrence, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology* **51**, 39–54.
- Baele, G., Van de Peer, Y., and Vansteelandt, S. (2008). A model-based approach to study nearest-neighbor influences reveals complex substitution patterns in non-coding sequences. *Systematic Biology* **57**, 675–692.
- Bailey, T. L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* pages 28–36.
- 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* pages 28–37.
- Blaisdell, B. E. (1985). Markov chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eucaryotic nuclear DNA sequences both protein-coding and noncoding. *Journal of Molecular Evolution* **21**, 278–288.
- Boys, R. J. and Henderson, D. A. (2004). A Bayesian approach to DNA sequence segmentation. *Biometrics* **60**, 573–588.
- Braun, J. V., Braun, R. K., and Müller, H.-G. (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika* **87**, 301–314.
- Braun, J. V. and Müller, H.-G. (1998). Statistical methods for DNA sequence segmentation. *Statistical Science* **13**, 142–162.
- Churchill, G. A. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology* **51**, 79–94.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376.
- Felsenstein, J. (1989). PHYLIP - phylogeny inference package (version 3.2). *Cladistics* **5**, 164–166.
- Felsenstein, J. and Churchill, G. A. (1996). A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution* **13**, 93–104.
- Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.
- Gupta, M. and Liu, J. S. (2003). Discovery of conserved sequence patterns using a stochastic dictionary model. *Journal of the American Statistical Association* **98**, 55–66.

- Huang, H., Kao, M.-C. J., Zhou, X., Liu, J. S., and Wong, W. H. (2004). Determination of local statistical significance of patterns in Markov sequences with application to promoter element identification. *Journal of Computational Biology* **11**, 1–14.
- Hwang, D. G. and Green, P. (2004). Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *The Proceedings of the National Academy of Sciences of the United States of America* **101**, 13994–14001.
- Ji, H. and Wong, W. H. (2006). Computational biology: Towards deciphering gene regulatory information in mammalian genomes. *Biometrics* **62**, 645–663.
- Karolchik, D., Kuhn, R. M., Baertsch, R., Barber, G. P., Clawson, H., Diekhans, M., Giardine, B., Harte, R. A., Hinrichs, A. S., Hsu, F., and et al. (2008). The UCSC genome browser database: 2008 update. *Nucleic Acids Research* **36**, D773–D779.
- Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. (1993). Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science* **262**, 208–214.
- Lawrence, C. E. and Reilly, A. A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins* **7**, 41–51.
- Li, X. and Wong, W. H. (2005). Sampling motifs on phylogenetic trees. *The Proceedings of the National Academy of Sciences of the United States of America* **102**, 9481–9486.
- Liu, J. S. and Lawrence, C. E. (1999). Bayesian inference on biopolymer models. *Bioinformatics* **15**, 38–52.
- Liu, J. S., Neuwald, A. F., and Lawrence, C. E. (1995). Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *Journal of the American Statistical Association* **90**, 1156–1170.
- Liu, X., Brutlag, D., and Liu, J. S. (2001). BioProspector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pacific Symposium on Biocomputing* **6**, 127–138.
- Moses, A. M., Chiang, D. Y., and Eisen, M. B. (2004). Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pacific Symposium on Biocomputing* **9**, 324–335.
- Pepe, M. S., Longton, G., Anderson, G. L., and Schummer, M. (2003). Selecting differentially expressed genes from microarray experiments. *Biometrics* **59**, 133–142.
- Ray, P., Shringarpure, S., Kolar, M., and Xing, E. P. (2008). CSMET: Comparative genomic motif detection via multi-resolution phylogenetic shadowing. *PLoS Computational Biology* **4**, e1000090.

- Siddharthan, R., Siggia, E. D., and van Nimwegen, E. (2005). PhyloGibbs: A Gibbs sampling motif finder that incorporates phylogeny. *PLoS Computational Biology* **1**, 534–556.
- Siepel, A. and Haussler, D. (2004). Combining phylogenetic and hidden Markov models in biosequence analysis. *Journal of Computational Biology* **11**, 413–428.
- Sinha, S., Blanchette, M., and Tompa, M. (2004). PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics* **5**, 170.
- Stormo, G. D. and Hartzell, G. W. (1989). Identifying protein-binding sites from unaligned DNA fragments. *The Proceedings of the National Academy of Sciences of the United States of America* **86**, 1183–1187.
- Thompson, W., Palumbo, M. J., Wasserman, W. W., Liu, J. S., and Lawrence, C. E. (2004). Decoding human regulatory circuits. *Genome Research* **14**, 1967–1974.
- Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Pr, M., Reuter, I., and Schacherer, F. (2000). TRANSFAC: An integrated system for gene expression regulation. *Nucleic Acids Research* **28**, 316–319.
- Xie, D., Cai, J., Chia, N.-Y., Ng, H. H., and Zhong, S. (2008). Cross-species de novo identification of cis-regulatory modules with GibbsModule: Application to gene regulation in embryonic stem cells. *Genome Research* **18**, 1325–1335.
- Yang, Z. (1995). A space-time process model for the evolution of DNA sequences. *Genetics* **139**, 993–1005.
- Zhou, Q. and Liu, J. S. (2008). Extracting sequence features to predict protein-DNA interactions: A comparative study. *Nucleic Acids Research* **36**, 4137–4148.
- Zhou, Q. and Wong, W. H. (2007). Coupling hidden Markov models for the discovery of cis-regulatory modules in multiple species. *Annals of Applied Statistics* **1**, 36–65.