Abstract
Motif discovery, which focuses on locating short sequence patterns associated with the regulation of genes in a species, leads to a class of statistical missing data problems. These problems are discussed first with reference to a hypothetical model, which serves as a point of departure for more realistic versions of the model. Some general results relating to modeling and inference through the Bayesian and/or frequentist perspectives are presented, and specific problems arising out of the underlying biology are discussed.

16.1 Introduction
The goal of motif discovery is to locate short repetitive patterns in DNA that are involved in the regulation of genes of interest. To fix ideas, let us consider the following paragraph modified from Bellhouse [4, Section 3, page 5]:

“Richard Bayes (1596-1675), a great-grandfather of Thomas Bayes, was a successful cutler in Sheffield. In 1643 Richard served in the rotating position of Master of the Company of Cutlers of Hallamshire. Richard was sufficiently well off that he sent one of his sons, Samuel Bayes (1635-1681) to Trinity College Cambridge during the Commonwealth period; Samuel obtained his degree in 1656. Another son, Joshua Bayes (1638-1703) followed in his father’s footsteps in the cutlery industry, also serving as Master of the Company in 1679. Evidence of Joshua Bayes’s wealth comes from the size of his house, the fact that he employed a servant and the size of the taxes that he paid. Joshua Bayes’s influence may be taken from his activities in ... ”
Imagine that a person who has never seen the English language before looks at this paragraph and tries to make sense out of it (this is very much analogous to how we view the genome sequences of various species). Also imagine that all the punctuation marks, capitalizations, and spaces have been taken away from this paragraph so that available to him are collections of sequences like:

\[ \text{richardbayes\text{\textunderscore}15961675agreatgrandfatherofthomasbayeswaswasas \ldots} \]
\[ \text{in\text{\textunderscore}1643richardbayesservedintherotatingpositionofmasterofthe \ldots} \]
\[ \text{richardwassufficientlywell\text{\textunderscore}offthat\thesent\text{\textunderscore}ofhisson\text{\textunderscore}samuelbayes \ldots} \]

How should the non-English speaker proceed? The first natural question to ask is: what might be an internal “linkage” of these sequences? This question leads one to find the most commonly occurring “words” (or, rather, short subsequences of unknown length) that might characterize this paragraph. Indeed, if one tries to list all the possible subsequences of length 5, the word “bayes” pops up as the most frequent or “enriched” one. If one tries this on all segments of length 10 or 11, “joshuabayes” tops the list. After these findings, you may suggest to your collaborators (i.e., biologists) to investigate the properties of “bayes” or “joshuabayes”, which may ultimately lead to the discovery of the great probabilist’s name, although this paragraph *per se* mainly discusses a few relatives of the probabilist. So, it appears that by looking for “significantly enriched” words, one can indeed get some insight on a paragraph written in a completely unknown language.

However, in order to make the above procedure statistically sound, one needs (a) to model in what context a word is “significantly enriched” (thus, a probabilistic structure for generating the observed text is needed); (b) a strategy for determining the length(s) of the enriched word(s) to be discovered; and (c) an efficient computational strategy to find all enriched words. In the genomic context, the problem is even more difficult because the “words” used by the nature are never “exact”, i.e., certain “mis-spellings” can be tolerated. Thus, one also needs (d) a probabilistic model to describe a fuzzy word.

A simplified model leads to the following class of statistical problems. Let $X_{ij}$, $j = 1, \ldots, L_i$, represent the $i$th observed genomic sequence (i.e., each $X_{ij}$ takes 4 different values: A, C, G, and T, instead of the 26 letters in the English alphabet). Our first “null” statistical model is to assume that each $X_{ij}$ is the result of a toss of a 4-sided die.
characterized by the probability vector $\theta_0 = (\theta_{0A}, \theta_{0C}, \theta_{0G}, \theta_{0T})$. The problem of interest is to infer whether there exist subsequences corresponding to one or more enriched words. That is, whether there are subsequences $Y_{ia} = \{Y_{il} : l = a, \ldots, a + w - 1; 1 \leq a \leq L - w + 1\}$ of $\{X_{ij} : 1 \leq j \leq L_i\}$ which are generated from a “run” of tosses from $w$ “special” dice, each characterized by the multinomial probability vector $\theta_m = (\theta_{mA}, \ldots, \theta_{mT})$. Thus, we now use a product-multinomial model, $[\theta_1, \ldots, \theta_w]$, to characterize a fuzzy word.

The values of $w, a, \theta_0$ and $\theta_m (m = 1, \ldots, w)$ are all unknown. The basic set-up above constitutes a missing data problem which somewhat differs from standard missing data problems in two ways: (i) estimating the unknown locations $a$ of the beginning of the “run” is generally considered to be of more interest than the values of the unknown parameters $\theta$ and (ii) at an individual locus level, there exist experimental methods (even though expensive and potentially inaccurate) to verify the computational predictions.

The purpose of this article is (i) to explain in brief the biological background of the preceding problem in relation to gene regulatory binding site discovery, (ii) propose a Bayesian framework for its solution that serves as a point of departure for discussing more realistic versions of the problem, and (iii) describe some alternative models and methods designed to capture the complicating features arising in practice. We consider issues of model selection and robustness of the inference procedures that are especially relevant in the Bayesian context. Some of the problems have close connections in the rich literature on hidden Markov models (HMMs), to which relevant similarities will be discussed.

**16.2 Biology of transcription regulation**

With the completion of many genome sequencing projects, a challenge now facing biologists is to determine which parts of the genome encode for biological functions, and the mechanisms by which sequence information is “translated” into these functions. In transcriptional regulation, sequence signals upstream of each gene provide a target (the promoter region) for an enzyme complex called RNA polymerase (RNAP) to bind and initiate the transcription of the gene into messenger RNA (mRNA). Certain proteins called transcription factors (TFs) can bind to the promoter regions, either interfering with the action of RNAP and inhibiting gene expression, or enhancing gene expression. TFs recognize sequence sites that give a favorable binding energy, which often translates into
a sequence-specific pattern (~8-20 base pairs long). Binding sites thus tend to be relatively well-conserved in composition – such a conserved pattern is termed as a “motif” (corresponding to the “key word” in the example of Section 1). For example, an important TF in *E.coli*, the cyclic AMP receptor protein (CRP), recognizes a pattern of the form TGTGANNNNNNTCACA (‘N’ denotes that any one of the 4 nucleotides may be present) – but a substantial deviation from this pattern may sometimes be tolerated. It is estimated that ~2000 (out of a total of ~30,000) genes in the human genome encode sequence-specific DNA binding TFs [40]. For identifying and understanding the functional role of non-coding sequences in the human and other genomes, it would be valuable to identify all the sequence patterns that can be recognized by these proteins. Experimental detection of TF-binding sites (TFBSs) on a gene-by-gene and site-by-site basis is possible [8], but remains an extremely difficult and expensive task at a genomic level, especially as the amount of sequence to be analyzed increases. Computational methods that assume no prior knowledge of the pattern of the binding sites then become a necessary tool for aiding in their discovery.

16.3 Problem formulation, background and general strategies

One of the first motif-finding approaches was CONSENSUS, an information theory-based progressive alignment procedure [36]. Assuming each sequence contains one motif site of width *w*, the objective was to find the set of sites maximizing “information content”, i.e. Kullback-Leibler entropy distance between the motif site composition and the background distribution: \( \sum_{i=1}^{w} \sum_{j=1}^{J} f_{ij} \log_2 \left( \frac{f_{ij}}{f_{0j}} \right) \), where \( f_{ij} \) is the observed frequency of letter *j* in position *i* of the site, and \( f_{0j} \) denotes the corresponding background letter frequencies. CONSENSUS starts by examining all pairs of *w*-long subsequences (*w*-mers) in the first two sequences, and retains the top-scoring *M* (say, 50) motifs (each consisting of pairs of sites). Each of the *M* motifs is next aligned to all *w*-mers in the third sequence, again the top *M* motifs are retained, and the process is continued for all the sequences.

Other early statistical methods for finding motifs include an EM-algorithm [9] based on a missing-data formulation [23], and a Gibbs sampling algorithm [22]. In both approaches, starting positions of true motif sites were treated as “missing” components of the observed sequence data. Under the assumption that there is exactly one motif site
Motif Discovery

per sequence, an iterative procedure was used to alternately refine the motif description (parameters) and sample sites in the sequences that could represent instances of the motif. Later generalizations that allow for a variable number of motif sites per sequence were a Gibbs sampler [27, 32] and an EM algorithm for finite mixture models [2].

Another class of methods approach the motif discovery problem from a “segmentation” perspective. MobyDick [5] treats the motifs as “words” used by nature to construct the “sentences” of DNA and estimates word frequencies using a Newton-Raphson optimization procedure. The dictionary model was later extended to include “stochastic” words in order to account for variations in the motif sites [14] and a data augmentation (DA) [38] procedure introduced for finding such words.

Recent approaches to motif discovery have improved upon the previous methods in at least two primary ways: (i) improving and sensitizing the basic model to reflect realistic biological phenomena, such as multiple motif types in the same sequence, “gapped” motifs, and clustering of motif sites (cis-regulatory modules) [29, 43, 15] and (ii) using auxiliary data sources, such as gene expression microarrays, phylogenetic information and the physical structure of DNA [7, 21]. Due to limitations of space, in this article we will mainly focus on (i), and indicate ways in which the Bayesian approach has facilitated making significant inroads into this field. We will primarily discuss de-novo methods of discovering uncharacterized motifs in biological sequences, as opposed to scanning sequences with a previously (experimentally) determined motif representation to find probable matches.

16.3.1 Likelihood-based approaches to motif discovery

In Lawrence and Reilly [23], an EM algorithm was developed to estimate the motif pattern and infer the motif site locations. In their formulation, every sequence in the data set is assumed to contain one and only one motif site, and its start position is considered the “missing data” part of the model. In order to model multiple motif sites per sequence, Bailey and Elkan [2] present a simplified model (see Figure 16.1) in which the sequence data set is broken up conceptually into all overlapping subsequences of length $w$ and each of these $w$-mers is assumed to be generated from one of the two classes: “motif” or “background”. More precisely, denoting the set of all $w$-mers by $X = (X_1, X_2, \ldots, X_n)$, each $w$-mer $X_i = (x_{i1}, \ldots, x_{iw})$ is assumed to be generated from a two-component
mixture model indexed by an unobserved group indicator \( Z_{ij} \), where
\[
Z_{ij} = \begin{cases} 
1 & \text{if } X_i \text{ is a motif site of type } j \ (j = 1, 2), \\
0 & \text{otherwise}.
\end{cases}
\]
A similar model is also presented in Liu et al. [27], where \( Z_{ij} \) is allowed to take on \( J + 1 \) possible values to accommodate \( J \) distinct motif types, and a Gibbs sampling strategy is proposed for the inference.

Let us write the set of parameters corresponding to the motif component and background as \( \Theta_1 = (\theta_1, \ldots, \theta_w) \) and \( \Theta_0 = (\theta_0, \ldots, \theta_0) \) (where \( \theta_i = (\theta_{i1}, \ldots, \theta_{i4})^T \)), while \( \pi \) denotes the relative proportion of motif segments (mixing proportion). Given the class indicator \( Z_{ij} = 1 \), \( X_i \) is assumed to be generated from a product-multinomial model characterized by \( \Theta_j \). Under this set-up, considering \( Z_{ij} \) as missing data, it is now possible to set up a standard EM algorithm to maximize the likelihood function \( P(X | \Theta_1, \Theta_0, \pi) \) with respect to \( (\Theta_1, \Theta_0, \pi) \).

A possible way to overcome the limitations of this over-simplified model, as suggested in Liu and Lawrence [26] and explained in more detail in the following section, is to recast the motif finding problem as a problem of segmenting the sequences into two types of contiguous pieces, one described by the block-motif model (of a fixed length \( w \)) and the other by an iid model.

### 16.3.2 Dictionary models for motif discovery

The dictionary model [5] is perhaps one of the first implementations of the aforementioned segmentation idea for motif discovery. In this model, one assumes that nature has a dictionary available, consisting of a list of \( d \) known words \( D = \{M_1, M_2, \ldots, M_d\} \). As a mathematical abstraction, we treat the whole observed data set as a single sequence, \( S \). \( S \) is assumed to be generated by randomly drawing words from the dictionary according to a probability vector \( \rho = (\rho(M_1), \ldots, \rho(M_d)) \) and sequentially concatenating them together.

Since we cannot observe the actual words that are used to compose the data \( S \), we need to sum over all possible segmentations of the sequences to get the likelihood function:

\[
P(S | \rho) = \sum_{\mathcal{H}} \prod_{i=1}^{N(H_i)} \rho(S[H_i]) = \sum_{\mathcal{H}} \prod_{j=1}^{d} [\rho(M_j)]^{N_{M_j}(\mathcal{H})},
\]

where \( \mathcal{H} \) represents a segmentation of the sequence.
where $\mathcal{H} = (H_1, \ldots, H_k)$ is a partition of $S$ so that each part $H_i$ corresponds to a word in the dictionary, $N(\mathcal{H})$ is the total number of words in $\mathcal{H}$, and $N_{M_j}(\mathcal{H})$ is the number of occurrences of word type $M_j$ in the partition. This can be viewed as a missing data problem where the partition $\mathcal{H}$ is missing; the summation over all $\mathcal{H}$ can be achieved recursively [26]. Let $\Phi_{i-1}(\rho)$ be the sum of all legitimate partitions for partial sequence $S_{[1:(i-1)]}$. Then,

$$\Phi_i(\rho) = \sum_{j=1}^{W} \rho(S_{[i-j:i]}) \Phi_{i-j},$$

(16.2)

where $W$ is the length of the longest word in the dictionary. In other words, we check whether the last segment is a word from the dictionary for all possible word lengths $j$. To avoid minor complications, we assume that all the single letters (i.e., A, C, G, and T) are contained in the dictionary; if not, the above recursion needs to be modified slightly.

The maximum likelihood estimate (MLE) of $\rho$ from model (16.1) can be found via a Newton-Raphson algorithm, since one can compute the derivative of the likelihood function (16.1) using a recursive procedure similar to (16.2). One can also employ an EM algorithm or a Gibbs sampler. More precisely, we can derive an estimating equation from (16.1) by taking derivatives with respect to $\rho_i$ [5]; the summations required in the estimating equation being computed recursively as in (16.2).

Bussemaker et al. [5] adopted a progressive strategy to estimate the unknown “dictionary” used by nature for constructing the genome. They start with the simplest dictionary consisting only of the $D = 4$ single-letter words, $D^{(0)} = \{A, C, G, T\}$ and then iterate as follows: For a current dictionary consisting of $D$ words, they find the MLE of the word usage frequencies, $\rho = (\rho_1, \ldots, \rho_D)$ based on model (16.1); then, they consider whether any concatenation of a pair of the estimating words is over-represented compared to what is expected by chance, and these new words are added to the current dictionary. This procedure is carried out iteratively until a stopping criterion is reached. The assumption that longer words are made up of over-represented fragments may not be true, but this defect can be rectified by progressively considering words of increasing lengths. That is, for example, we may let the $(t+1)^{st}$ iteration of the dictionary, $D^{(t+1)}$ be the union of $D^{(t)}$ and all “significant” words of length $t + 1$. After introducing longer words, one can also remove some of the shorter words that appear to be parts of certain long words.

To generalize the model of Bussemaker et al. [5] to “fuzzy” words,
Gupta and Liu [14] and Sabatti and Lange [33] introduce the idea of a *stochastic dictionary*, which consists of a collection of “stochastic words” each represented by a probabilistic word matrix, or exchangeably, a position-specific weight matrix (PWM). Each column of the PWM ($\Theta$) gives the probabilities of finding each letter in that position of the corresponding stochastic word. For example, `ACAGG` and `GCAGA` may be two realizations, with probabilities 0.4328 and 0.0072 respectively, of the stochastic word characterized by the PWM

$$
\Theta = \begin{bmatrix}
A & 0.85 & 0.07 & 0.80 & 0.02 & 0.12 \\
C & 0.05 & 0.78 & 0.07 & 0.01 & 0.01 \\
G & 0.10 & 0.05 & 0.12 & 0.96 & 0.85 \\
T & 0.00 & 0.10 & 0.01 & 0.01 & 0.02
\end{bmatrix}.
$$

In the setting described above, the motif-finding problem reduces to inferring the form of the PWM and the likely locations of the stochastically varying words in the sequence, which can be carried out effectively under a Bayesian framework [14].

### 16.4 A Bayesian approach to motif discovery

In this section, unless otherwise specified, we assume that the data set is a set of $N$ unaligned DNA fragments. Let $S = (S_1, \ldots, S_N)$ denote the $N$ sequences of the data set, where sequence $S_i$ is of length $L_i$, $(i = 1, \ldots, N)$. Multiple instances of the same pattern in the data are referred to as motif *sites* or *elements* while different patterns are termed motifs. Motif type $k$ (of, say, width $w_k$) is characterized by a PWM $\Theta_k = (\theta_{k1}, \ldots, \theta_{kw_k})$, where the $J$-dimensional ($J = 4$ for DNA) vector $\theta_{ki} = (\theta_{ki1}, \ldots, \theta_{kiJ})^T$ represents the probabilities of occurrence of the $J$ letters in column $i$, $(i = 1, \ldots, w_k)$. The corresponding letter occurrence probabilities in the background are denoted by $\theta_0 = (\theta_{01}, \ldots, \theta_{0J})$. Let $\Theta = \{\Theta_1, \ldots, \Theta_K\}$.

We assume for now that the motif widths, $w_k$ ($k = 1, \ldots, K$) are known (this assumption will be relaxed later). The locations of the motif sites are unknown, and are denoted by an array of missing indicator variables $A = (A_{ijk})$, where $A_{ijk} = 1$ if position $j$ ($j = 1, \ldots, L_i$) in sequence $i$ ($i = 1, \ldots, N$) is the starting point of a motif of type $k$ ($k = 1, \ldots, K$). For motif type $k$, we let $A_k = \{A_{ijk} : i = 1, \ldots, N; j = 1, \ldots, L_i\}$, i.e., the indicator matrix for the site locations corresponding
to this motif type, and define the alignment:

\[
S_1^{(A_k)} = \{ S_{ij} : A_{ijk} = 1; i = 1, \ldots, N; j = 1, \ldots, L_i \}, \\
S_2^{(A_k)} = \{ S_{i(j+1)} : A_{ijk} = 1; i = 1, \ldots, N; j = 1, \ldots, L_i \}, \\
\vdots \\
S_{w_k}^{(A_k)} = \{ S_{i(j+w_k-1)} : A_{ijk} = 1; i = 1, \ldots, N; j = 1, \ldots, L_i \}.
\]

In words, \(S_i^{(A_k)}\) is the set of letters occurring at position \(i\) of all the instances of the type-\(k\) motif.

In a similar fashion, we use \(S^{(A^c)}\) to denote the set of all letters occurring in the background, where \(S^{(A^c)} = S \setminus \bigcup_{k=1}^K \bigcup_{l=1}^{w_k} S_l^{(A_k)}\) (For two sets \(A, B\), \(A \subset B\), \(B \setminus A = B \cap A^c\)). Further, let \(C : S \to \mathbb{Z}^A\) denote a "counting" function that gives the frequencies of the \(J\) letters in a specified subset of \(S\). For example, if after taking the set of all instances of motif \(k\), in the first column, we observe a total occurrence of 10 'A's, 50 'T's and no 'C' or 'G's, \(C(S_1^{(A_k)}) = (10, 0, 0, 50)\). Assuming that the motif columns are independent, we have

\[
[C(S_1^{(A_k)}), \ldots, C(S_{w_k}^{(A_k)})] \sim \text{Product-Multinomial}[\theta_k = (\theta_{k1}, \ldots, \theta_{kw_k})],
\]

i.e., the \(i^{th}\) vector of column frequencies for motif \(k\) follows a multinomial distribution parametrized by \(\theta_{ki}\).

We next introduce some general mathematical notation. For vectors \(v = (v_1, \ldots, v_p)^T\), let us define \(|v| = |v_1| + \cdots + |v_p|\), and \(\Gamma(v) = \Gamma(v_1) \cdots \Gamma(v_p)\). Then the normalizing constant for a \(p\)-dimensional Dirichlet distribution with parameters \(\alpha = (\alpha_1, \ldots, \alpha_p)^T\) can be denoted as \(\Gamma(|\alpha|)/\Gamma(|\alpha|)\). For notational convenience, we will denote the inverse of the Dirichlet normalizing constant as \(ID(\alpha) = \Gamma(|\alpha|)/\Gamma(|\alpha|)\). Finally, for vectors \(v\) and \(u = (u_1, \ldots, u_p)\), we use the shorthand \(u^v = \prod_{i=1}^p u_i^{v_i}\).

The probability of observing \(S\) conditional on the indicator matrix \(A\) can then be written as

\[
P(S \mid \Theta, \theta_0, A) \propto \theta_0^{\text{C}(S^{(A^c)})} \prod_{k=1}^K \prod_{i=1}^{w_k} \theta_{ki}^{C(S_i^{(A_k)})}.
\]

For a Bayesian analysis, we assume a conjugate Dirichlet prior distribution for \(\theta_0, \theta_0 \sim \text{Dirichlet}(\beta_0), \beta_0 = (\beta_{01}, \ldots, \beta_{0D})\), and a corresponding product-Dirichlet prior (i.e., independent priors over the columns) \(\text{PD}(B)\) for \(\Theta_k (k = 1, \ldots, K)\), where \(B = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{kw_k})\) is a \(J \times w_k\) matrix with \(\beta_{ki} = (\beta_{k1}, \ldots, \beta_{ki})^T\). Then the conditional posterior dis-
distribution of the parameters given $A$ is:

$$P(\Theta, \theta | S, A) \propto \theta_0^{C(S^{(A)})} + \beta_0 \prod_{k=1}^{K} \prod_{i=1}^{w_i} \theta_{ki}^{C(S^{(A_k)})} + \beta_{ki}.$$  

For the complete joint posterior of all unknowns $(\Theta, \theta, A)$, we further need to prescribe a prior distribution for $A$. In the original model [22], a single motif site per sequence with equal probability to occur anywhere was assumed. However, in the later model [27] that can allow multiple sites, a Bernoulli($\pi$) model is proposed for motif site occurrence. More precisely, assuming that a motif site of width $w$ can occur at any of the sequence positions, $1, 2, \ldots, L^* - w + 1$ in a sequence of length $L^*$, with probability $\pi$, the joint posterior distribution is:

$$P(\Theta, \theta, A | S) \propto \theta_0^{C(S^{(A)})} + \beta_0 \prod_{k=1}^{K} \prod_{i=1}^{w_i} \theta_{ki}^{C(S^{(A_k)})} + \beta_{ki} |A| (1 - \pi)^{L^* - |A|}. \quad (16.3)$$

where $L = \sum_{i=1}^{N} (L_i - w)$ is the adjusted total length of all sequences and $|A| = \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{j=1}^{L_i} A_{ijk}$. If we have reason to believe that motif occurrences are not independent, but occur as clusters (as in regulatory modules), we can instead adopt a prior Markovian model for motif occurrence [15, 39] which is discussed further in Section 16.6.

### 16.4.1 Markov chain Monte Carlo computation

Under the model described in (16.3), it is straightforward to implement a Gibbs sampling (GS) scheme to iteratively update the parameters, i.e., sampling from $[\Theta, \theta_0 | C, A]$, and impute the missing data, i.e., sampling from $[A | C, \Theta, \theta_0]$. However, drawing $\Theta$ from its posterior at every iteration can be computationally inefficient. Liu et al. [27] demonstrated that marginalizing out $(\Theta, \theta_0)$ from the posterior distribution can lead to much faster convergence of the algorithm [28]. In other words, one can use the Gibbs sampler to draw from the marginal distribution

$$p(A | S, \pi) = \int \int p(\Theta, \theta_0 | S, A, \pi) p(A | \Theta, \theta_0) d\Theta d\theta_0, \quad (16.4)$$

which can be easily evaluated analytically.

If $\pi$ is unknown, one can assume a beta prior distribution $Beta(\alpha_1, \alpha_2)$ and marginalize out $\pi$ from the posterior, in which case $p(A | S)$ can be derived from (16.4) by altering the last term in (16.4) to the ratio of normalizing constants for the Beta distribution, $B(|A| + \alpha_1, L - |A| +$
Motif Discovery

Based on (16.4), Liu et al. [27] derived a predictive updating algorithm for $A$, which is to iteratively sample each component of $A$ according to the predictive distribution

$$P(A_{ijk} = 1 \mid S) = \frac{\pi}{1 - \pi} \prod_{l=1}^{w} \left( \frac{\hat{\theta}_{kl}}{\hat{\theta}_0} \right) \theta_{ik}^{C(S_{i,j+k})},$$

(16.5)

where the posterior means are $\hat{\theta}_{kl} = \frac{c(S_{i,j+k}) + \beta_{ik}}{c(S_{i,j+k}) + \beta_{ik}}$ and $\hat{\theta}_0 = \frac{c(S_{i,j+k}) + \beta_0}{c(S_{i,j+k}) + \beta_0}$.

Under the model specified above, it is also possible to implement a “partition-based” data augmentation (DA) approach [14] that is motivated by the recursive algorithm used in Auger and Lawrence [1]. The DA approach samples $A$ jointly according to the conditional distribution

$$P(A \mid \Theta, S) = \prod_{i=1}^{N} P(A_{iL_i} \mid \Theta, S) \prod_{j=1}^{L_i-1} P(A_{ij} \mid A_{ij+1}, \ldots, A_{iL_i}, S, \Theta).$$

At a position $j$, the current knowledge of motif positions is updated using the conditional probability $P(A_{ij} \mid A_{ij+1} \ldots A_{iL_i}, \Theta)$ (backward sampling), with $A_{i, j-1} \ldots A_{i1}$ marginalized out using a forward summation procedure (an example will be given in Section 16.6.1.2). In contrast, at each iteration, GS iteratively draws from the conditional distribution: $P(A_{ijk} \mid A_{ij}, S)$, iteratively visiting each sequence position $i$, updating its motif indicator conditional on the indicators for other positions. The Gibbs approach tends to be “sticky” when the motif sites are abundant. For example, once we have set $A_{ijk} = 1$ (for some $k$), we will not be able to allow segment $S_{[i,j+k]}$ to be a motif site. The DA method corresponds to a grouping scheme (with $A$ sampled together), whereas the GMS corresponds to a collapsing approach (with $\Theta$ integrated out). Both have been shown to improve upon the original scheme [28].

### 16.4.2 Scoring functions and Bayesian optimization

In motif discovery problems, the predictions of interest often correspond to the estimated maximizer $A^*$ of the posterior probability $P(A \mid S)$, rather than the posterior average. In this regard, BioProspector [29] attempts to find a fast approximate estimate of $A$ by slightly altering the Gibbs search strategy. From (16.5), an approximate posterior “scoring” function is derived as

$$\phi(A) = \log(\sum_{i=1}^{w} \sum_{j=1}^{w} \hat{\theta}_{ij} \log \frac{\hat{\theta}_{ij}}{\hat{\theta}_{0j}}).$$
When using the current weight matrix to scan the sequence, all segments whose scores $\phi(\cdot)$ exceed a “high” threshold are automatically called a motif site, while those that are between the high and a “low” threshold are given a chance to be sampled into the set of sites. The low-threshold is started as 0 and increased gradually during iterations to a suitable level. Jensen and Liu [18] present an optimization algorithm that provides (i) a more accurate scoring function approximation of (16.5) and (ii) a simulated annealing procedure to optimize this function.

16.5 Extensions of the product-multinomial motif model

Unknown motif width. In the following discussion, for simplicity of notation, we assume a single motif type $\Theta$ of width $w$. Previously, $w$ was assumed to be known and fixed; we may instead view $w$ as an additional unknown model parameter. Jointly sampling from the posterior distribution of $(A, \Theta, w)$ is difficult as the dimensionality of $\Theta$ changes with $w$. One way to update $(w, \Theta)$ jointly would be through a reversible jump procedure [12]- however, note that we can again integrate out $\Theta$ from the posterior distribution to avoid a dimensionality change during the updating. By placing an appropriate prior distribution $p(w)$ on $w$ (a possible choice is a Poisson($\lambda$)), we can update $w$ using a Metropolis step. Using a Beta($\alpha_1, \alpha_2$) prior on $\pi$, the marginalized posterior distribution of interest is $P(A, w | S)$,

$$P(A, w | S) \propto \prod_{i=1}^{w} \frac{I(D(C(S_i^{(A)})\beta_i) + \beta_i)}{I(D(\beta_i))} \frac{B(|A| + \alpha_1, L - |A| + \alpha_2)}{B(\alpha_1, \alpha_2)} p(w).$$

The product-multinomial model used for $\Theta$ is a first approximation to a realistic model for transcription factor binding sites. In empirical observations, it has been reported that certain specific features often characterize functional binding sites. We mention here a few extensions of the primary motif model that have been recently implemented to improve the performance of motif discovery algorithms.

Variations of the product multinomial assumption. The product multinomial model assumes that all columns of a weight matrix are independent- however, it has been observed that about 25% of experimentally validated motifs show statistically significant positional correlations. Zhou and Liu [42] extend the independent weight matrix model to including one or more correlated column pairs, under the restriction that no two pairs of correlated columns can share a column in common. For example, if columns 1 and 5 are correlated, 2 and 3 can be,
Motif Discovery

but 1 and 2 cannot. A Metropolis-Hastings step is added in the Gibbs sampler [27] that deletes or adds a pair of correlated column at each iteration. Again, the posterior distribution can be collapsed over \( \Theta \) during the Metropolis-Hastings step to avoid a parameter space of varying dimensions for different numbers of correlated columns. Barash et al. [3] proposed a Bayesian tree-like network to model the possible correlation structure among all the positions within a TF model. Zhao et al. [41] described a permuted Markov model–they assume that an unobserved permutation has acted on the positions of all the motif sites and that the original ordered positions can be described by a Markov chain. Thus, mathematically, the model of Zhou and Liu [42] is a sub-case of Zhao et al. [41], which is, in turn, a sub-case of Barash et al. [3].

It has been observed that real TFBSs are not uniformly conserved over all positions–the conserved positions often occur as a group at one or two regions over the motif, since contacts between proteins and the DNA are likely to occur over a few bases at a time (more conservation indicates a higher chance of contact). In the hope that incorporation of this positional trend is more likely to find the correct motif, Kechris et al. [19] use a prior distribution that “penalizes” deviations from a conserved profile. Instead of using a Dirichlet distribution as a prior for the motif column probabilities \( \theta \), they instead use a normal or double exponential prior, e.g. \( p(\theta) \propto e^{-\sum_{i=1}^{\beta_i} |\theta_i-\beta_i|} \). To update parameters of the model, they developed an EM algorithm in which the M-step was slightly modified from Lawrence and Reilly [23] to reflect the change in the prior.

16.6 HMM-type models for regulatory modules

Motif predictions for high eukaryotes (e.g. human, mouse, dog, etc.) are more challenging than that for simpler organisms such as yeast and bacteria. Some of the reasons are: (i) large sections of low-complexity regions (repeat sequences), (ii) weak motif signals, (iii) sparseness of signals compared to entire region under study–binding sites may occur as far as 2000-3000 bases away from the transcription start site, either upstream or downstream, and (iv) motifs occurring in clusters, varying in order or composition between sequences. In complex eukaryotes, regulatory proteins often work in combination to regulate target genes, and their binding sites have often been observed to occur in spatial clusters, or cis-regulatory modules (Figure 16.2). One approach to locating cis-regulatory modules (CRMs) is by predicting novel motifs and looking
for co-occurrences [35]. However, since individual motifs in the cluster may not be well-conserved, such an approach often leads to a large number of false negatives. Our strategy is to first use existing de novo motif finding algorithms and TF databases to compose a list of putative binding motifs, $D = \{\Theta_1, \ldots, \Theta_D\}$, where $D$ is in the range of 50 to 100, and then simultaneously update these motifs and estimate the posterior probability for each of them to be included in the CRM.

Figure 16.2 here

Let $S$ denote the set of $n$ sequences with lengths $L_1, L_2, \ldots, L_n$, respectively, corresponding to the upstream regions of $n$ co-regulated genes. We assume that the CRM consists of $K$ different kinds of TFs with distinctive PWMs. Both the PWMs and $K$ are unknown and need to be inferred from the data. In addition to the indicator variable $A$ defined in Section 16.4, we define a new variable $a_{i,j}$, that denotes the location of the $j$th site (irrespective of motif type) in the $i$th sequence. Let $a = \{a_{ij}; i = 1, \ldots, n; j = 1, \ldots, L_i\}$. Associated with each site is its type indicator $T_{i,j}$, with $T_{i,j}$ taking one of the $K$ values (let $T = (T_{ij})$).

Note that the specification $(a, T)$ is essentially equivalent to $A$.

Next, we model the dependence between $T_{i,j}$ and $T_{i,j+1}$ by a $K \times K$ probability transition matrix $\tau$. The distance between neighboring TF-BSs in a CRM, $d_{ij} = a_{i,j+1} - a_{i,j}$, is assumed to follow $Q(\cdot; \lambda, w)$, a geometric distribution truncated at $w$, i.e. $Q(d; \lambda, w) = (1 - \lambda)^{d-w} \lambda \ (d = w, w+1, \ldots)$. The distribution of nucleotides in the background sequence a multinomial distribution with unknown parameter $\rho = (\rho_A, \ldots, \rho_T)$.

Next, we let $u$ be a binary vector indicating which motifs are included in the module, i.e. $u = (u_1, \ldots, u_D)^T$, where

$$ u_j = \begin{cases} 
1, & \text{if the } j^{th} \text{ motif type is present in the module,} \\
0, & \text{otherwise.}
\end{cases} $$

By construction, $|u| = K$. Thus, the information regarding $K$ is completely encoded by $u$. In light of this notation, the set of PWMs for the CRM is defined as $\Theta = \{\Theta_j: u_j = 1\}$. Since now we restrict our inference of CRM to a subset of $D$, the probability model for the observed sequence data can be written as:

$$ P(S | D, \tau, u, \lambda, \rho) = \sum_a \sum_T P(S | a, T, D, \tau, u, \lambda, \rho) P(a | \lambda) P(T | a, \tau). $$

From the above likelihood formulation, we need to simultaneously estimate the optimal $u$ and the parameters $(D, \tau, \lambda, \rho)$. To achieve this, we
first prescribe a prior distribution on the parameters and missing data:

\[ P(D, \tau, u, \lambda, \rho) = f_1(D \mid u)f_2(\tau \mid u)f_3(\rho)g_1(u)g_2(\lambda). \]

Here the \( f_i(\cdot) \)'s are (product) Dirichlet distributions. Assuming each \( u_i \) takes the value 1 with a prior probability of \( \pi \) (i.e. \( \pi \) is the prior probability of including a motif in the module), \( g_1(u) \) represents a product of \( D \) Bernoulli(\( \pi \)) distributions; and \( g_2(\lambda) \), a generally flat Beta distribution. More precisely, we assume \emph{a priori} that \( \Theta_i \sim \prod_{j=1}^{w} \text{Dirichlet}(\beta_{ij}) \) (for \( i = 1, \ldots, D \)); \( \rho \sim \text{Dirichlet}(\beta_0) \); \( \lambda \sim \text{Beta}(a, b) \). Given \( u \) (with \( |u| = K \)), each row of \( \tau \) is assumed to follow an independent Dirichlet. Let the \( i^{th} \) row \( v_i | u \sim \text{Dirichlet}(\alpha_i) \), where \( i = 1, \ldots, K \).

Let \( \Omega = (D, \tau, \lambda, \rho) \) denote the full parameter set. Then the posterior distribution of \( \Omega \) has the form

\[ P(\Omega, u \mid S) \propto P(S \mid u, \Omega)f_1(D \mid u)f_2(\tau \mid u)f_3(\rho)g_1(u)g_2(\lambda). \]  

(16.6)

A Gibbs sampling approach was developed in Thompson et al. [39] to infer the CRM from a special case of the posterior distribution (16.6) with fixed \( u \). Given the flexibility of the model and the size of the parameter space for an unknown \( u \), it is unlikely that a standard MCMC approach can converge to a good solution in a reasonable amount of time. If we ignore the ordering of sites \( T \) and assume components of \( a \) to be independent, this model is reduced to the original motif model in Section 16.4 which can be updated through the previous Gibbs or DA procedure.

### 16.6.1 A hybrid EMC-DA approach

With a starting set of putative binding motifs \( D \), we simultaneously modify these motifs and estimate the posterior probability for each of them to be included in the CRM through iterations of the following Monte Carlo sampling steps: (i) Given the current collection of motif PWMs (or sites), sample motifs into the CRM by evolutionary Monte Carlo (EMC); (ii) Given the CRM configuration and the PWMs, update the motif site locations through DA; and (iii) Given motif site locations, update the corresponding PWMs and other parameters.

#### 16.6.1.1 Evolutionary Monte Carlo for module selection

It has been demonstrated that the EMC method is effective for sampling and optimization with functions of binary variables [25]. Conceptually, we should be able to apply EMC directly to select motifs comprising the CRM, but a complication here is that there are many continuous
parameters such as the \( \Theta_j \)'s, \( \lambda \), and \( \tau \). We cannot just fix these parameters (as in the usual Gibbs sampler) and update the CRM composition because some of them vary in dimensionality when a putative motif in \( D \) is included or excluded from the CRM. We therefore have to integrate out the continuous parameters \( \Theta \) and \( \tau \) analytically and condition on variables \( a \) and \( T \) when updating the CRM composition. Let \( \Omega^{(u)} = (\Theta, \rho, \tau, \lambda) \) denote the set of all parameters in the model, for a fixed \( u \). Then, the marginalized conditional posterior probability for a module configuration \( u \) is:

\[
P(u | a, T, S) \propto \pi^{[u]} (1 - \pi)^{D - |u|} \int P(S | a, T, \Omega^{(u)}) P(\Omega^{(u)} | u) d\Omega^{(u)},
\]

(16.7)

where only \( \Theta \) and \( \tau \) are dependent on \( u \); and \( a \) and \( T \) are the sets of locations and types, respectively, of all putative motif sites (for all the \( D \) motifs in \( D \)). Thus, only when the indicator \( u_i \) for the weight matrix \( \Theta_i \) is 1, do its site locations and types contribute to the computation of (16.7). When we modify the current \( u \) by excluding a motif type, its site locations and corresponding motif type indicators are removed from the computation of (16.7).

For EMC, we need to prescribe a set of temperatures, \( t_1 > t_2 > \cdots > t_M = 1 \), one for each member in the population. Then, we define

\[
\phi_i(u_i) \propto \exp[\log P(u_i | a, T, S)/t_i],
\]

and \( \phi(U) \propto \prod_{i=1}^{M} \phi_i(u_i) \). The “population” \( U = (u_1, \ldots, u_M) \) is then updated iteratively using two types of moves: mutation and crossover.

In the mutation operation, a unit \( u_k \) is randomly selected from the current population and mutated to a new vector \( v_k \) by changing the values of some of its bits chosen at random. The new member \( v_k \) is accepted to replace \( u_k \) with probability \( \min(1, r_m) \), where

\[
r_m = \phi_k(v_k) / \phi_k(u_k).
\]

In the crossover step, two individuals, \( u_j \) and \( u_k \), are chosen at random from the population. A crossover point \( x \) is chosen randomly over the positions 1 to \( D \), and two new units \( v_j \) and \( v_k \) are formed by switching between the two individuals the segments on the right side of the crossover point. The two “children” are accepted into the population to replace their parents \( u_j \) and \( u_k \) with probability \( \min(1, r_c) \), where

\[
r_c = \frac{\phi_j(v_j) \phi_k(v_k)}{\phi_j(u_j) \phi_k(u_k)},
\]
If rejected, the parents are kept unchanged. On convergence, the samples of \( u_M \) (for temperature \( t_M = 1 \)) follow the target distribution (16.7).

### 16.6.1.2 Sampling motif sites \( A \) through recursive DA

The second part of the algorithm consists of updating the motif sites conditional on a CRM configuration (i.e., with \( u \) fixed). For simplicity, we describe the method for a single sequence \( S = (s_1, \ldots, s_L) \) – the same procedure is repeated for all sequences in the data set. For simplicity of notation, we assume that all motifs are of width \( w \). For fixed \( u \), let

\[
F(i, j, k) = P(s_{[i,j,k]} \mid \Omega^{(u)}, u) = \sum_{l=1}^{K} \sum_{i < j} F(1, i, l) \tau_{l,k} Q(j - i - w; \lambda, w) + P(s_{[i,j-w,0]} | \rho) \times F(j - w + 1, j, k).
\]

By convention, the initial conditions are: \( F(0, 0, k) = 1 \), \( k = 0, 1, \ldots, K \), and \( F(i, j, k) = 0 \) for \( j < i \) and \( k > 0 \). In the backward sampling step, we use Bayes theorem to calculate the probability of motif occurrence at each position, starting from the end of the sequence. If a motif of type \( k \) ends at position \( i \) in the sequence, the probability that the next motif further ahead in the sequence spans position \( (i' - w + 1, i', i' + w) \) and is of type \( k' \), is:

\[
P(A, i', i'-w+1, k') = P(A, i', i'-w+1, k = 1) \frac{F(1, i', k') P(s_{[i'+1,i'-w,0]} | \rho) F(i'-w+1, i, k) Q(i'-i'-w; \lambda, w) \tau_{k',k}}{F(1, i, k)}.
\]

The required expressions have all been calculated in the forward sum.

### 16.6.1.3 Sampling parameters from posterior distributions

Given the motif type indicator \( u \) and the motif position and type vectors \( a \) and \( T \), we now update the parameters \( \Omega = (\Theta, \rho, \Sigma, \lambda) \) by a random
sample from their joint conditional distribution. Since conjugate priors have been assumed for all parameters, their conditional posterior distributions are also of the same form and are straightforward to simulate from. For example, the posterior of $\Theta_i$ will be $\prod_{j=1}^{w} \text{Dirichlet}(\beta_{ij} + n_{ij})$, where $n_{ij}$ is a vector containing the counts of the 4 nucleotides at the $j$th position of all the sites corresponding to motif type $i$. For those motifs that have not been selected by the module (i.e., with $u_i = 0$), the corresponding $\Theta_i$'s still follow their prior distribution. Similarly, the posterior distribution of $\rho$ is $\text{Dirichlet}(\beta_0 + n_0)$, where $n_0$ denotes the frequencies for the 4 nucleotides in the background sequence.

For updating $\tau$, we note that if $m_{ij} \{ i, j \in D : u_i = u_j = 1 \}$ denotes the number of transitions from PWM type $i$ to $j$ (when $i$ and $j$ are both included in the module), then the posterior distribution of $\tau_i$ is $\text{Dirichlet}(\alpha_i + m_i)$. Let the distance between consecutive sites on sequence $i$ ($i = 1, \ldots, n$) be $d_{ij} = a_{i,j+1} - a_{ij}$, where each $d$ follows $Q(\lambda, w)$, a geometric($\lambda$) distribution truncated at $w$. Let $d = \sum_{i=1}^{n} \sum_{j=1}^{A_i} |A_i|^{-1} d_{ij}$ be the total length of sequence covered by the CRMs, where $|A_i|$ is the total number of sites in sequence $i$, and $|A'| = \sum_{i=1}^{n} (|A_i| - 1)$. Then the posterior distribution of $\lambda$ is $\text{Beta}(a + |A'|, b + d - w|A'|)$.

16.6.2 A case-study

We compared the performance of EMC-DA with EM- and Gibbs sampling-based methods in an analysis of mammalian skeletal muscle regulatory sequences [39]. The raw data consist of upstream sequences of lengths up to 5000 bp each corresponding to 24 orthologous pairs of genes in the human and mouse genomes—each of the sequences being known to contain at least one experimentally reported transcription-factor binding site corresponding to one of 5 motif types: MEF, MYF2, SRF, SP1 and TEF. Following the procedure of Thompson et al. [39], we aligned the sequences for each orthologous pair (human and mouse) and retained only the parts that shared a percent identity greater than 65%, cutting down the sequence search space to about 40% of the original sequences.

Using BioProspector, EM (MEME) and AlignAce (Gibbs sampler for independent motifs), we obtained initial sets of 100 motifs including redundant ones. The top-scoring 10 motifs from BioProspector and MEME respectively contained 2 and 3 matches to the true motif set (of 5), whereas AlignAce found none. The Gibbs sampler under a module
model [39] found 2 matches in general, but could find 2 others with a more detailed and precise prior input (the number of sites per motif and motif abundance per sequence), which is generally unavailable in real applications. The best scoring module configuration from EMC-DA contained 3 of the true 5, MYF, MEF2, and SP1, and two uncharacterized motifs. There are few TEF sites matching the reported consensus in these sequences, and they were found by none of the algorithms.

The relative error rates for the algorithms could be compared in this case as we had exact knowledge of each of the experimentally determined TFBSs. Table 16.1 shows that EMC-DA significantly cuts down the percentage of false positives in the output, compared to the methods that do not adjust for positional clustering of motifs. We next tested whether it is beneficial to choose a subset of motifs from the eukaryotic motif database [34] as the starting motif set for EMC-DA. This time, EMC-DA can find four out of five expected motifs. Figure 16.3 shows the posterior probability of site occurrence over the first three aligned sequence pairs, indicating a strong evidence of clustering. For multiple runs of EMC-DA with different starting seeds, there was no noticeable difference in the results over a wide range of prior settings.

### Table 16.1, Figure 16.3 here

#### 16.7 Model selection through a Bayesian approach

One of the basic questions that arise in motif discovery is whether the patterns “discovered” from the sequence data by these algorithms are “real”. Although the biological relevance of such findings generally needs further biological experimentation, we can at least try to assess the significance of the predictions from a statistical viewpoint.

As a Bayesian model selection problem, it is of interest to assess whether the sequence data should be better explained by model $M_1$, which assumes the existence of a nontrivial motif, than by $M_0$, which says that the sequences are generated entirely from a background model (e.g., an i.i.d. or Markov model). The Bayes factor, which is the ratio of the marginal likelihoods under the two models, can be computed as

$$
\frac{p(S \mid M_1)}{p(S \mid M_0)} = \frac{\sum_A \int_{\theta} p(A, S, \theta \mid M_1) d\theta}{\int_{\theta} p(S, \theta \mid M_0) d\theta} = \frac{\sum_A p(A, S \mid M_1)}{p(S \mid M_0)}
$$

(16.8)

The individual additive terms in the numerator of (16.8) consist of ratios of products of gamma functions. To evaluate this sum exhaustively
over all partitions involves prohibitive amounts of computation. A lower bound for (16.8) is $p(A^*, S | M_1)/p(S | M_0)$, where $A^*$ is the maximizer of the ratio. This bound, called the maximum a posteriori score (denoted by $\text{MAP}(A^*)$), can be used as a model selection criterion which can be tracked along with the Gibbs or DA iterations. As a frequentist evaluation of its performance, we have elsewhere [13] shown that the MAP asymptotically attains several desirable properties. For example, with a single motif type (width $w$) having occurrence probability $\pi$, under mild conditions, the MAP score selects the correct model, with the performance of the MAP improving as $w$ and $\pi$ increases (Table 16.2).

Table 16.2 here

16.8 Discussion: motif discovery beyond sequence analysis

TFBS prediction remains an important unsolved problem in molecular biology. The availability of massive amounts of genomic data, such as multi-species genome sequence, gene expression microarray, and the physical structure of DNA has thrown up a huge challenge to computational scientists to develop investigative tools to infer biological function. In this article we have mainly demonstrated how Bayesian statistical models can be used to capture significant aspects of genomic sequence data and lead to more accurate motif predictions. Using the Bayesian approach naturally leads to a host of flexible Monte Carlo-based algorithms that can deal with high-dimensional integration and multi-modality problems effectively. The Bayesian framework also provides us with a basic infrastructure for hierarchically modeling dependence and for dealing with nuisance parameters without leading to overwhelming analytical complexity. Finally, the Bayesian paradigm allows a “learning” capability so that “historical” data can be used in modeling a new, but similar problem. This can be important in building prior distributions based on partial information for known motifs, improving estimation of novel motifs.

It is becoming increasingly clear that sequence information alone is insufficient for accurate motif predictions (being especially true for complex genomes). An important aspect of further development of motif discovery tools is the efficient integration of multiple sources of genomic data, with the aim of refining and improving predictions.
16.8.1 Cross-species phylogenetic information

When multi-species sequence information is available, it is often seen that multiply aligning sequences [10] and using regions having a high sequence similarity improves the specificity of motif search. Thompson et al. [39] use this strategy to simultaneously search for motifs in aligned pairs of regulatory sequences from the human and mouse genomes, significantly improving predictions. It is still a challenge to efficiently incorporate more complete evolutionary information such as a phylogenetic tree of sequence divergence [11] in probabilistic motif models.

16.8.2 Chromatin structure information

The precise control of transcription in eukaryotes depends on the binding of TFs to promoter regions on DNA. Although in general practice, TF-DNA binding is represented as a one-dimensional process; in reality, binding occurs in three dimensional space. DNA is neither static nor one-dimensional—much of DNA is wrapped around certain proteins called histones in a specific pattern, and binding is most likely to occur at the regions of exposed DNA [24]. For a given TF, there may be many potential TFBSs conserved in sequence, scattered throughout the genome of an organism, however, only a subset of these is actually active. As more experimental data become available (e.g. from chromatin immunoprecipitation or ChIP-chip experiments), knowledge of DNA structure holds a huge potential to aid in motif discovery.

16.8.3 Further incorporation of gene expression information

A highly successful tactic for computational motif prediction is to cluster genes based on their expression profiles, and search for motifs in the sequences upstream of tightly clustered genes. When noise is introduced into the cluster through spurious correlations, however, such an approach may result in many false positives. A filtering method based on the specificity of motif occurrences has been shown to effectively eliminate false positives [17]. An iterative procedure for simultaneous gene clustering and motif finding has been suggested [16], but no effective algorithm has been implemented to demonstrate its advantage. Two methods for TFBM discovery via the association of gene expression values with motif abundance have been proposed by Bussemaker et al. [6] and Keles et al. [20]. These first conduct word enumeration and then use
regression to check whether the genes whose upstream sequences contain a set of words have significant changes in their expression. Conlon et al. [7] provide an algorithm to further utilize gene expression or ChIP-chip information to help motif discovery. They first use an algorithm such as BioProspector, MEME, or MDscan [30] to report a large set of putative motifs, and then do a linear stepwise regression to select candidates that correlate with the microarray expression data (Tadesse et al. [37] later present a Bayesian version). However, these methods still face drawbacks such as the inappropriateness of the linearity assumption, high-dimensionality, and difficulties in using multiple microarray data sets simultaneously. Surmounting such challenges and finding effective ways to integrate multiple data sources into motif discovery is likely to hold the future key to accurate inference in biological systems.

References


Motif Discovery


Table 16.1. Error rates for module prediction methods. Total: the total number of sites predicted. There are 154 true sites.

<table>
<thead>
<tr>
<th>Method</th>
<th>MEF</th>
<th>MYF</th>
<th>SP1</th>
<th>SRF</th>
<th>Total</th>
<th>SENS</th>
<th>SPEC</th>
<th>TSpec</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM</td>
<td>0</td>
<td>1</td>
<td>21</td>
<td>0</td>
<td>161</td>
<td>0.14</td>
<td>0.14</td>
<td>0.20</td>
</tr>
<tr>
<td>BioProspector</td>
<td>6</td>
<td>1</td>
<td>8</td>
<td>1</td>
<td>155</td>
<td>0.10</td>
<td>0.10</td>
<td>0.36</td>
</tr>
<tr>
<td>GS</td>
<td>6</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>84</td>
<td>0.10</td>
<td>0.25</td>
<td>0.44</td>
</tr>
<tr>
<td>GS&quot;</td>
<td>14</td>
<td>14</td>
<td>4</td>
<td>6</td>
<td>162</td>
<td>0.25</td>
<td>0.23</td>
<td>0.60</td>
</tr>
<tr>
<td>EMC-DA</td>
<td>12</td>
<td>12</td>
<td>5</td>
<td>7</td>
<td>180</td>
<td>0.23</td>
<td>0.20</td>
<td>0.67</td>
</tr>
<tr>
<td>EMC-DA&quot;</td>
<td>17</td>
<td>13</td>
<td>8</td>
<td>10</td>
<td>108</td>
<td>0.31</td>
<td>0.44</td>
<td>0.80</td>
</tr>
<tr>
<td>True</td>
<td>32</td>
<td>50</td>
<td>44</td>
<td>28</td>
<td>15</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Fig. 16.1. (a) The basic motif model [23] and (b) the mixture model approximation [2].
Table 16.2. **Comparison of model selection criteria for 3 data sets from Bacillus subtilis (BS), yeast (Y), and E. coli (EC).** “Order” represents the order in which the motif was found using the method of Gupta and Liu [14]. Experimentally confirmed motifs are highlighted in boxes. For all data sets, the MAP score decreased after the true motif was found [BIC: Bayes Information Criterion; AIC: Akaike Information Criterion; KLI: Kullback-Leibler Information].

<table>
<thead>
<tr>
<th>TF</th>
<th>Order</th>
<th>logMAP</th>
<th>BIC</th>
<th>AIC</th>
<th>KLI</th>
<th>Motif consensus</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>SigA (BS)</td>
<td>1</td>
<td>-326.27</td>
<td>185.01</td>
<td>258.26</td>
<td>4.10</td>
<td>TTTTTT</td>
<td>186</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-192.02</td>
<td>163.2</td>
<td>309.69</td>
<td>4.07</td>
<td>TGAAAA</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>263.31</td>
<td>148.98</td>
<td>5.70</td>
<td>4.91</td>
<td>TGTGCA</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>90.19</td>
<td>101.46</td>
<td>247.95</td>
<td>4.18</td>
<td>TATAAT</td>
<td>41</td>
</tr>
<tr>
<td>GAL (Y)</td>
<td>1</td>
<td>20.3</td>
<td>-130.74</td>
<td>43.45</td>
<td>11.52</td>
<td>GGAGGACTTTTCTCGA</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-20.26</td>
<td>-259.75</td>
<td>88.62</td>
<td>9.97</td>
<td>AGCCGCAAACCTGGAAA</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-49.26</td>
<td>-341.15</td>
<td>7.23</td>
<td>11.44</td>
<td>GGTGCTGCGATGG</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>-12.53</td>
<td>-398.98</td>
<td>-50.61</td>
<td>9.10</td>
<td>TTTTTTTTTTTTTGAA</td>
<td>16</td>
</tr>
<tr>
<td>CRP (EC)</td>
<td>1</td>
<td>2.27</td>
<td>-167.93</td>
<td>17.8</td>
<td>10.70</td>
<td>TTATTTGAACGAGTCTACCTT</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-7.53</td>
<td>-394.14</td>
<td>-22.67</td>
<td>10.83</td>
<td>GGATCATATGTTGTGAAATA</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-19.33</td>
<td>-401.54</td>
<td>-30.07</td>
<td>9.44</td>
<td>ATTTATAAACATTTAAAAATCGT</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>-1.99</td>
<td>-360.39</td>
<td>11.08</td>
<td>10.63</td>
<td>TGATTTGAGTCTCTACGAAA</td>
<td>5</td>
</tr>
</tbody>
</table>

![Fig. 16.2. Graphical illustration of a CRM.](image1)

![Fig. 16.3. Posterior probability of sampling sites in human-rodent sequence pairs. The light (dotted) and dark (solid) lines correspond to the MEF2 and MYF motif; the horizontal axes denote the site location on sequences.](image2)