Statistical Models for Biological Sequence Motif Discovery

Jun S Liu*, Mayetri Gupta*, Xiaole Liu†, Linda Mayerhofere‡, and Charles E. Lawrence‡

* Department of Statistics, Harvard University, Cambridge, MA 02138
† Medical Informatics, Stanford University, Stanford, CA 94305.
‡ The Wadsworth Center, Albany, NY 12201.

Abstract. With the completion of genomes of many species and the advances of microarray technologies, we begin to possess a tremendous amount of valuable biological data — but these raw products are still far from usable. One of the most challenging problems of this century is to decipher this huge amount of biological information, turning the data into knowledge. The past decade has witnessed a number of successful applications of statistical models in computational biology. This article focuses on one of these success stories: Using Bayesian models and Monte Carlo methods to find short repetitive patterns in a set of DNA or protein sequences, a task often referred to as the *motif discovery*. We review a few probabilistic models that have recently been shown useful for motif discovery and provide a novel framework based on a Bayesian segmentation model to unify these approaches. We show how to combine the dictionary model with the Gibbs sampler and how a segmentation-based data augmentation scheme can be implemented. A few interesting open problems are also discussed.

1 Introduction

The human genome and many other genome sequencing projects have resulted in rapidly growing and publicly available databases of three most important biopolymers, DNA, RNA, and proteins (e.g., the GeneBank). The data in these databases are sequences of letters, without punctuation or space characters, from an alphabet of size d ($d = 4$ for DNA and RNA sequences, and $d = 20$ for protein sequences). Recent advances in gene-chip technologies further enable and entice biologists to generate huge amounts of numerical data measuring the copy numbers of mRNAs in the cell. The task of extracting biological insight from these sequence databases or the microarray data represents one of the grand scientific challenges of the 21st century. The near-completion of the human genome in 2000 makes this interesting question more an urgent task in the scientific community. This task, often termed as "data mining" in the literature, differs from the mining of other types of databases for three reasons (i) many sophisticated structures have been built into the biopolymer databases, (ii) much is already known about the structure and function of these molecules, and (iii) fundamental laws in physics and chemistry apply to them. Consequently, more sophisticated mathematical/statistical models are often critical in developing a "mining" strategy for these datasets.

Our focus here is to find short (7-30 base pairs) recurring patterns in a set of biopolymer sequences. The pattern is called a *motif* and each occurrence of the pattern is called a *motif element*. The problem is analogous to finding a commonly occurring *word* in a text. What makes the task difficult is that there are often "typos" (sometimes very serious ones) in each occurrence of the "word." It is therefore rather natural to employ a probabilistic model to describe the motif pattern (i.e., a stochastic word) and let basic statistical principles (i.e., the Bayesian methodology) guide us in the discovery of these patterns.

The motivation for this motif discovery task is that repetitive patterns in biopolymer sequences often correspond to functionally or structurally important parts of these molecules. For example, motif elements in noncoding regions of DNA sequences are very likely the sites to which certain proteins, called transcription factors (TF), bind so as to control the gene's expression (see Section 2 for more biological background of this problem). Conserved patterns in protein sequences can reveal the proteins' functional roles and may correspond to these molecules' active sites interacting with ligands, a critical piece of information for drug designers. After reviewing the basic biology about gene regulation, this article discusses two related statistical approaches for motif finding: One is based on the product multinomial block-motif model [9] (Section 3), and the other based on a segmentation model (Section 4). A segmentation-based Gibbs sampler is implemented and shown to outperform the original Gibbs motif sampler [9].

2 Regulation of Gene Expression

Every cell has a complete copy of the orgamism's genetic information. There are over 4,000 genes in the E. coli genome and over 30,000 in the human genome. These genes encode the information needed to make the proteins and RNAs. The expression of these genes is tightly regulated so as to meet the requirements of specific cells and for cells to respond to changes in their environment. So they need to continuously monitor external conditions and respond in ways that optimize survival. Cells respond to the external signals by regulating the types and quantities of its proteins. They do this by regulating the genes that encode the information for making the proteins. For example, human skin cells called melanocytes, when exposed to direct sunlight, respond by producing the proteins that produce the pigment melanin. When your skin cells are exposed to direct sunlight, they respond by expressing the genes that code for the proteins involved in producing the pigment melanin. Once you get out of the sun, the cells no longer receive the signal and so the gene expression is turned off.

In bacteria, environmental stimuli play an even greater role in gene regulation because, as unicellular organisms, bacteria are in direct contact with the external environment. Bacterial cells have to be poised to respond to environmental changes quickly so as to optimally match the environmental conditions. For example, bacteria need iron since it is essential for many metabolic processes. However, since free iron is always in short supply, bacteria are equipped with iron scavenging proteins. The genes that code for these proteins are turned off, or repressed, when iron levels are adequate, but activated when iron levels decrease.

During development of a multicellular organism, cells become specialized in a process called differentiation. Differentiated cells acquire a specific complement of proteins suitable to its activities. For example, although every cell contains a full copy of the genetic information, the photoreceptor protein rhodopsin is produced only in the retina. Genes that are not needed are silenced — the photoreceptor protein rhodopsin is only expressed in the retina and other light-sensitive cells (like in the pineal gland, where circadian rhythms are regulated), and you will not be able to detect rhodopsin in you skin cells or anywhere else where it is not needed.

2.1 Transcriptional regulation

While there are several levels of regulation of gene expression, the dominant mechanism is the transcription regulation. Specific sequence signals upstream of each gene provide a target, called the promoter, for RNA polymerase to initiate the polymerization of an exact copy of the gene in the from of a "messenger" RNA. When transcription factors (TFs) bind directly in the promoter region of a gene, they interfere with the function of RNA polymerase and, thus, inhibit its expression. When a TF binds further upstream (sometimes downstream) it can attract RNA polymerase and enhance gene expression. TFs identify the genes they are intended to regulate and their specific binding sites based on sequence specific binding preferences that arise through the energetic interactions of the atoms of the TF and the atoms a sequence of DNA bases. The strength of binding varies with the specific DNA sequence of these sites. Cells capitalize on such variations to quantitatively regulate transcription.

Transcription regulation in multicellular organisms often involves multiple TFs that bind in to form regulatory modules. These modules may involve combinations of several transcription factors and other proteins. During development, the expression of transcription factors is also regulated to produce combinations appropriate to specific cell types. The inference challenge addressed here is to simultaneously identify the parameters of a model describing the base type preference and the specific locations of a transcription factor's binding sites given only sets of sequences upstream of genes that are expected to be regulated together.

The different TFs have different target sequences. These target sites are generally 10-20 nucleotide base pairs long, and many have a palindromic pattern that enables dimeric transcription factors to clamp onto the DNA at the target site and read it from both strands. A sequence is said to be *palindromic* if the reverse of its complementary strand is the same as the original sequence. For example, the "iron box" GATAATGATAATCATTATC is the target site for the transcription factor Fur (ferric uptake regulator). Its complement CTATTACTATTAGTAATAG reads identically as the original sequence from backward.

2.2 Experimental methods for identifying TF binding sites

The initial determination of the binding of a specific TF to a site is achieved using a gel shift assay. In these experiments the DNA fragment and the TF are incubated together under favorable conditions. Then the DNA-protein mixture is placed in a gel. When and electric current is applied to the gel, molecules are sorted by size. In this gel mobility shift assay, DNA fragments with the TF attached run more slowly than do naked fragments. The DNA protection assay, DNAse footprinting, is employed to identify the specific binding site within a fragment. Again the DNA and TF are incubated together under favorable conditions. Then, the DNA is chemically degraded in a controlled manner by DNAse I, and enzyme. The degraded DNA is then sorted on a gel by size. All size fragments are represented on the gel except those that were protected from degradations by the binding of the TF at its cognate site. The specific action stemming from this binding is determined using deletion experiments, in which either the gene that codes for the TF is mutated or deleted and the effect be determined, or the binding site is mutated or deleted and the effect on gene expression assessed.

2.3 Revealing gene regulation information by cross-species comparisons and microarray analyses

Multiple sequences that are likely to be coregulated (i.e., regulated by the same TF or a small number of common TFs) are required for computational identification of transcription factor binding sites. Since microarrays identify genes that are coexpressed (i.e., the sets of genes that behave similarly in response to various treatments), they also provide information on coregulation. However, since multiple regulatory events may be imitated under any given condition, coexpressed genes are not necessarily coregulated. In spite of this difficulty data from microarrays has been used successfully as input for the computational identification of transcription factor binding sites [6, 11, 15].

Cross-species comparison provides another means to identify multiple genes that are likely to be regulated similarly. This approach capitalizes on the fact that the genes that code for the same protein in related species are likely to be similarly regulated. Since this approach obviates the need to assay for coexpression, it can be applied on a genome wide scale using only the sequences from a set of related species. For example, using this approach McCue et.al. [12] showed in a study set of 184 genes that 80% of such predictions correspond with known sites and proceeded to predicted over 2,000 transcription factor binding sites in E. coli.

2.4 Sequence data used for DNA motif discovery

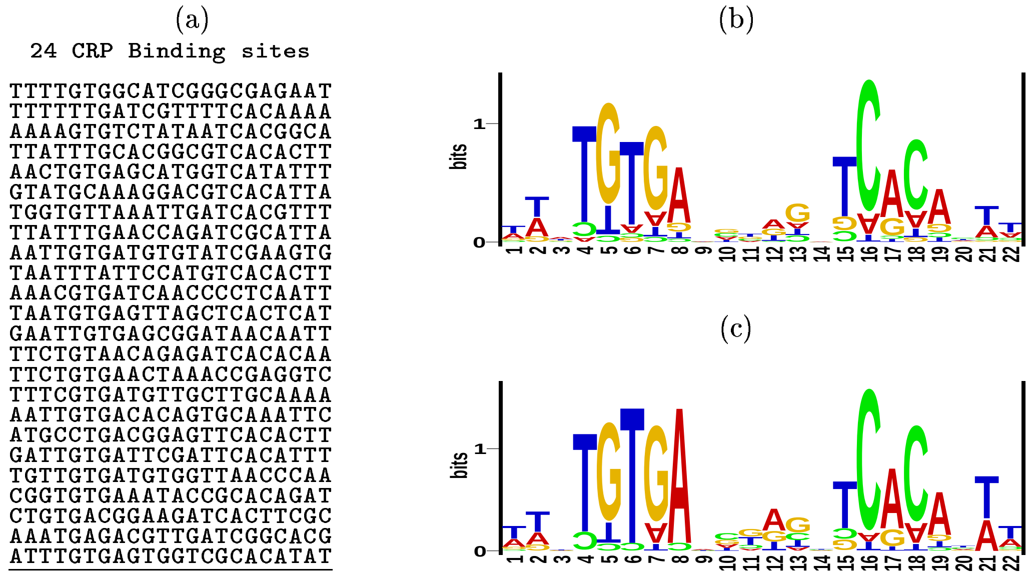
From either the cross-species comparison or the microarray clustering analysis, one can at best end up with a set of DNA sequences of hundreds or thousands of base pairs believed to contain some common TF binding sites. Figure 1 displays a typical dataset of this nature, first given by [18], from which one wishes to infer the locations and patterns of the TF binding sites. It contains 18 UTRs, each of 105 base pairs long, corresponding to 18 genes in E. Coli that are all regulated by the cyclic-AMP receptor protein (CRP), an important TF for controlling metabolism. The CRP-binding sites are about 22 base pairs long, two of which are indicated by the capital letters in Figure 1.

```
taatgtttgtgctggtTTTTGTGGCATCGGGCGAGAATagcgcgtggtgtgaaagactgtttttttgatcgttttcacaaaaatggaagtccacagtcttgacag
gacaaaaacgcgtaacaaaagtgtctataatcacggcagaaaagtccacattgaTTATTTGCACGGCGTCACACTTtgctatgccatagcattttatccataag
acaaatcccaataacttaattattgggatttgttatatataactttataaattcctaaaattacacaaagttaataactgtgagcatggtcatatttttatcaat
cacaaagcgaaagctatgctaaaacagtcaggatgctacagtaatacattgatgtactgcatgtatgcaaaggacgtcacattaccgtgcagtacagttgatagc
acggtgctacacttgtatgtagcgcatctttctttacggtcaatcagcatggtgttaaattgatcacgttttagaccatttttcgtcgtgaaactaaaaaaacc
agtgaattatttgaaccagatcgcattacagtgatgcaaacttgtaagtagatttccttaattgtgatgtgtatcgaagtgtgttgcggagtagatgttagaata
gcgcataaaaaacggctaaattcttgtgtaaacgattccactaatttattccatgtcacacttttcgcatctttgttatgctatggttatttcataccataagcc
gctccggcggggttttttgttatctgcaattcagtacaaaacgtgatcaacccctcaattttccctttgctgaaaaattttccattgtctcccctgtaaagctgt
aacgcaattaatgtgagttagctcactcattaggcaccccaggctttacactttatgcttccggctcgtatgttgtgtggaattgtgagcggataacaatttcac
acattaccgccaattctgtaacagagatcacacaaagcgacggtggggcgtaggggcaaggaggatggaaagaggttgccgtataaagaaactagagtccgttta
ggaggaggcgggaggatgagaacacggcttctgtgaactaaaccgaggtcatgtaaggaatttcgtgatgttgcttgcaaaaatcgtggcgattttatgtgcgca
gatcagcgtcgttttaggtgagttgttaataaagatttggaattgtgacacagtgcaaattcagacacataaaaaaacgtcatcgcttgcattagaaaggtttct
gctgacaaaaaagattaaacataccttatacaagactttttttcatatgcctgacggagttcacacttgtaagttttcaactacgttgtagactttacatcgcc
ttttttaaacattaaaattcttacgtaatttataatctttaaaaaaagcatttaatattgctccccgaacgattgtgattcgattcacatttaaacaatttcaga
cccatgagagtgaaattgttgtgatgtggttaacccaattagaattcgggattgacatgtcttaccaaaaggtagaacttatacgccatctcatccgatgcaagc
ctggcttaactatgcggcatcagagcagattgtactgagagtgcaccatatgcggtgtgaaataccgcacagatgcgtaaggagaaaataccgcatcaggcgctc
ctgtgacggaagatcacttcgcagaataaataaatcctggtgtccctgttgataccgggaagccctgggccaacttttggcgaaaatgagacgttgatcggcacg
gatttttatactttaacttgttgatatttaaaggtatttaattgtaataacgatactctggaaagtattgaaagttaatttgtgagtggtcgcacatatcctgtt
```

Figure 1: The upstream regions of length 105 for 18 genes, all regulated by the transcription factor CRP, in E. Coli. The capitalized segments in the first two sequences are two experimentally determined CRP-binding sites.

Scientists have identified by experimental methods the locations of 24 TF binding sites in these 18 sequences, so this dataset serves as a good testing case for computational motif finding algorithms. An alignment of these sites is shown in Figure 2(a). A sequence logo plot for the alignment is shown in Figure 2(b), which uses a color scheme to display the conservation pattern of each position [17] of the binding motif. As a comparison, Figure 2 (c) displays the alignment logo for the 18 binding sites found by our algorithm described in Section 4.

From the alignment as well as the sequence logos, one can determine not only the consensus sequence but also the relative frequency of bases and the information content (measured in bits) at every position of the motif. It is also seen that the conservation of the pattern at some positions of the motif can be very low, which is often due to some good biological reasons and can be further exploited in statistical modeling. As indicated in the logo plot, a highly conserved pattern is ***TGTGA(6)TCACA***, a perfect palindrome, whereas the 6 middle positions are much less conserved. Further biological analyses reveal that the middle 6 positions are buried in the minor groove of the double strand DNA helix so that CRP cannot "see" them clearly (and



24 CRP Binding sites

```
TTTTGTGGCATCGGGCGAGAAT
TTTTTTGATCGTTTTCACAAAA
AAAAGTGTCTATAATCACGGCA
TTATTTGCACGGCGTCACACTT
AACTGTGAGCATGGTCATATTT
GTATGCAAAGGACGTCACATTA
TGGTGTTAAATTGATCACGTTT
TTATTTGAACCAGATCGCATTA
AATTGTGATGTGTATCGAAGTG
TAATTTATTCCATGTCACACTT
AAACGTGATCAACCCCTCAATT
TAATGTGAGTTAGCTCACTCAT
GAATTGTGAGCGGATAACAATT
TTCTGTAACAGAGATCACACAA
TTCTGTGAACTAAACCGAGGTC
TTTCGTGATGTTGCTTGCAAAA
AATTGTGACACAGTGCAAATTC
ATGCCTGACGGAGTTCACACTT
GATTGTGATTCGATTCACATTT
TGTTGTGATGTGGTTAACCCAA
CGGTGTGAAATACCGCACAGAT
CTGTGACGGAAGATCACTTCGC
AAATGAGACGTTGATCGGCACG
ATTTGTGAGTGGTCGCACATAT
```

Figure 2: (a) The alignment of the 24 CRP-binding sites in the dataset of Figure 1 found by experimental methods; (b) the sequence logo [17] for the alignment in (a); and (c) the sequence logo for computationally predicted CRP-binding sites. The height of each letter is proportional to its frequency, and the letters are sorted so the most common one is on top. The height of the entire stack is then adjusted to signify the information content of the sequences at that position.

does not care much about them either). This feature was taken into consideration in [5] and is now an option for the user to specify in the Gibbs motif sampler and BioProspector [10]

3 The Block-Motif Models

As depicted in Figure 3, this model treats the unknown motif as a contiguous block and describes its pattern by a product-multinomial distribution [9]. This model was first used in [5] via an EM algorithm and is the basis for the popular Gibbs sampler for motif discovery [4, 7]. Compared with the more general HMM as employed by [3], this model does not allow insertions and deletions in the middle of the motif. We show in Section 4.5 that this constraint can be relaxed

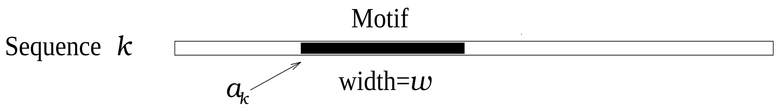


Figure 3: A schematic plot of the block-motif model used for our pattern finding.

easily.

3.1 A single-site model

Here we assume that each of the K DNA (or protein) sequences (of lengths $n_1, \ldots, n_K$, respectively) contains exactly one motif element. More precisely, each sequence is assumed to be generated by i.i.d. draws from the alphabet $\{A, C, G, T\}$ with the frequency vector $\boldsymbol{\theta}_0 = (\theta_{0a}, \ldots, \theta_{0t})$, except for a segment, as illustrated by the blackened region in Figure 3, which is a *motif* element of length w. The *motif* is a frequency matrix $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_w)$, also called a *weight matrix* in the bioinformatics literature [18], where each $\boldsymbol{\theta}_j^T$ is a probability vector of length 4 representing the preference of the nucleotide types at position j of a motif element. Thus, position j of a motif element (or an occurrence of the motif) is a draw from the alphabet with frequency vector $\boldsymbol{\theta}_j$. Both the motif matrix $\boldsymbol{\Theta}$ and the positions of the motif elements are not observed and must be inferred.

Gibbs site sampler. An intuitive solution (for a statistician) to the motif discovery problem just described is to use a missing-data formulation: The position of the motif element, a_j in particular, can be treated as missing data. Then, one can design an EM algorithm [5] or a data augmentation (DA) algorithm [19] to iterate between imputing the a_j and estimating the $\boldsymbol{\theta}_j$. A slight modification [7] of the DA algorithm gives rise to the so-called *site sampler* [4].

In the *site sampler*, the motif locations (sites) are initialized at random; that is, position $a_k^{(0)}$ (for $k = 1, \ldots, K$) is a randomly chosen position of the kth sequence. For $t = 1, \ldots, m$, the algorithm iterates the following steps:

- Select a sequence, say the kth sequence, either deterministically or at random.
- Draw a new motif location a_k according to the predictive distribution

$$P(a_k \mid a_1^{(t)}, \ldots, a_{k-1}^{(t)}, a_{k+1}^{(t)}, \ldots, a_K^{(t)}) \tag{1}$$

  and update the current motif location $a_k^{(t)}$ to $a_k^{(t+1)} = a_k$.
- Let $a_j^{(t+1)} = a_j^{(t)}$ for $j \neq k$.

Although there are many choices of the distribution (1) [e.g., one can let $P(a_k \mid \ldots)$ be proportional to certain fitness measure of the segment indexed by a_k to the current multinomial profile resulting from the $a_j^{(t)}$, $j \neq k$], those that make the foregoing iteration coherent must be derived from a complete Bayesian statistical model.

A theoretical calculation. To understand the nature of the problem, we consider an analogous problem. Consider K sequences of coins (not necessarily fair) of length L each. For each sequence S_k, we assume that starting from an unknown position a_k there is a segment of w special coins, each has probability θ to show head. The remaining coins has a known probability

θ_0 to show head. Our questions are two: (a) Can we estimate θ consistently? (b) How accurately can we predict the locations of these special coins? To answer these questions, we start with the basic likelihood analysis. Let $\boldsymbol{Y} = (S_1, \ldots, S_K)$ be the sequence data. Then

$$P(\boldsymbol{Y} \mid \theta) = \prod_{k=1}^{K} \left(\sum_{i=1}^{L-w+1} \left(\frac{\theta}{\theta_0} \right)^{N_{k,i}} \left(\frac{1-\theta}{1-\theta_0} \right)^{w-N_{k,i}} \right),$$

where $N_{k,i}$ is the number of heads in segment $(s_{k,i}, \ldots, s_{k,i+w-1})$. When $w = 1$, the observed Fisher information is

$$I_{\text{obs}} = \sum_{k=1}^{K} \left(\frac{\frac{N_k}{\theta_0} - \frac{L-N_k}{1-\theta_0}}{\frac{N_k}{\theta_0}\theta + \frac{L-N_k}{1-\theta_0}(1-\theta)} \right)^2 .$$

If both K and L are large, we have

$$I_{\text{exp}} \approx \frac{K}{L\theta_0(1-\theta_0)},$$

implying that K/L needs to go to infinity for the estimation of θ to be consistent.

When $w > 1$, the expression for the Fisher information is too messy to give any insight. We thus resort to a method-of-moment (MOM) approach. More precisely, we solve θ from the moment equation

$$K(w\theta + (L-w)\theta_0) = N_H,$$

where N_H is the total number of heads in the data set, we obtain the method-of-moment estimate of θ as

$$\hat{\theta} = \frac{N_H - (L-w)K\theta_0}{wK}.$$

It is easy to see that

$$\text{var}\left(\hat{\theta}\right) = \frac{\text{var}(N_H)}{w^2K^2} = \frac{(\theta-\theta_0)(1-\theta-\theta_0)}{wK} + \frac{L\theta_0(1-\theta_0)}{Kw^2}.$$

Hence we need $L/(Kw^2) \to 0$ so as to have a consistent estimate for θ. It is interesting to note here that the effect of motif width w is quite substantial, consistent with our observations based on simulations and real applications.

3.2 Repetitive Block-Motif Model

For the purpose of motif discovery, it is also insightful to view the whole sequence dataset $\boldsymbol{S}$ as consisting of a single long sequence of length n, although in this way one will not be able to capture the global evolutionary relationship among the individual sequences. In order to find repetitive motif elements in such a sequence, we propose a simple model as illustrated by Figure 4. Compared with the single-site model in the previous section, the new view here can accommodate the case when some sequences in the dataset contain multiple copies of the motif element whereas others do not contain any. This single-sequence repetitive-motif model was first developed in [9] and has been employed to find subtle repetitive patterns, such as the helix-turn-helix structural motif [13] and gene regulatory binding motifs [15], in both protein and DNA sequences. The repetitive patterns as represented by the shaded rectangle occur irregularly in the dataset. The total number of occurrences of the motif is unknown. A simple first model is to assume that at any sequence position i, there is a small probability p_0 that a motif pattern starts from i. A more sophisticated prior based on biological knowledge (e.g., physical properties of the DNA compositions, the distance to the translation start site, etc.) can be used to improve the model. Similar to the single-site model, we need $w+1$ probability vectors to describe the motif and the background if the motif's width is w: $\boldsymbol{\theta}_0 = (\theta_{0a}, \ldots, \theta_{0t})$ describe the base frequencies in the background; and each $\boldsymbol{\theta}_k$ describes the base frequency at position k of the motif. Thus, the motif weight matrix is $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_w)$.

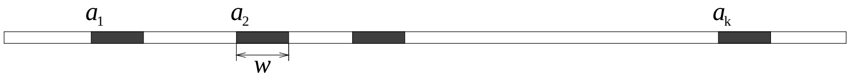


Figure 4: A graphical illustration of the repetitive motif model.

Scoring the motif candidates. A Bayesian solution to the foregoing alignment problem was derived in [9]. With a Dirichlet($\boldsymbol{\alpha}$) prior distribution for all the $\boldsymbol{\theta}_i$, we can obtain the Bayes estimates of the $\boldsymbol{\theta}_i$ very easily *if* we know the positions of the motif. To facilitate analysis, we introduce an indicator vector $\boldsymbol{I} = (I_1, \ldots, I_n)$ and treat it as *missing data.* An $I_i = 1$ means that position i is the start of a motif pattern, and $I_i = 0$ means otherwise. We assume *a priori* each I_i has a small probability p_0 to be equal to 1 independently, with the restriction that the motif elements cannot overlap. With this setup, we can write the joint posterior distribution:

$$p(\boldsymbol{\theta}_0, \boldsymbol{\Theta}, \boldsymbol{I} \mid \boldsymbol{S}) \propto p(\boldsymbol{S} \mid \boldsymbol{I}, \boldsymbol{\theta}_0, \boldsymbol{\Theta}) p(\boldsymbol{I} \mid \boldsymbol{\theta}_0, \boldsymbol{\Theta}) f_0(\boldsymbol{\theta}_0, \boldsymbol{\Theta}) \tag{2}$$

Suppose that the alphabet size is d (=4 for DNA sequences), motif length is w, and the motif pattern can be described by a $d \times w$ weight matrix $\boldsymbol{\Theta}$. Assume also that the background of non-motif regions can be described by a multinomial vector $\boldsymbol{\theta}_0$, which is assumed known in advance (this can be relaxed). Let $|\boldsymbol{I}|$=$I_1 + \cdots + I_n$, i.e., the total number of sites. Then,

$$\begin{aligned} \log p(\boldsymbol{S}, \boldsymbol{I}) &= \log \int p(\boldsymbol{S} \mid \boldsymbol{I}, \boldsymbol{\Theta}, \boldsymbol{\theta}_0) p(\Theta) p(\boldsymbol{I}) d\boldsymbol{\Theta} \\ &\approx \log p(\boldsymbol{I}) + \log p(\boldsymbol{S} \mid \boldsymbol{\theta}_0) + |\boldsymbol{I}| \sum_{j=1}^{w} \left\langle \hat{\boldsymbol{\theta}}_j, \log \frac{\hat{\boldsymbol{\theta}}_j}{\boldsymbol{\theta}_0} \right\rangle . \end{aligned}$$

Since $P(S \mid \boldsymbol{\theta}_0)$ is constant for all $\boldsymbol{I}$, the Gibbs motif sampler developed in [9] optimizes the score function

$$\psi(\boldsymbol{S}) = |\boldsymbol{I}| \left[\log p_0 + \sum_{j=1}^{w} I_{\text{ent}}(\hat{\boldsymbol{\theta}}_j \| \boldsymbol{\theta}_0) \right], \tag{3}$$

where p_0 is the prior probability for any position to be the start of the binding site, and the entropy distance I_{ent} between two discrete distributions $\boldsymbol{p} = (p_1, \dots, p_d)$ and $\boldsymbol{q} = (q_1, \dots, q_d)$, is defined as

$$I(\boldsymbol{p}\|\boldsymbol{q}) = \left\langle \boldsymbol{p}, \log\frac{\boldsymbol{p}}{\boldsymbol{q}} \right\rangle \equiv \sum_{i=1}^{d} p_i \log(p_i/q_i).$$

If we assume that p_0 is unknown and give it a prior distribution $f(p_0)$ (say, a Beta(a_0, b_0) distribution), then

$$\begin{aligned} p(\boldsymbol{I}) &\approx \int p_0^{|\boldsymbol{I}|}(1-p_0)^{N-w|\boldsymbol{I}|} f(p_0)dp_0 \\ &= \frac{\Gamma(a_0+b_0)}{\Gamma(a_0)\Gamma(b_0)} \int p_0^{|\boldsymbol{I}|+a_0-1}(1-p_0)^{N-w|\boldsymbol{I}|+b_0-1}dp_0 \\ &= \frac{\Gamma(a_0+b_0)\Gamma(|\boldsymbol{I}|+a_0)\Gamma(N-w|\boldsymbol{I}|+b_0)}{\Gamma(a_0)\Gamma(b_0)\Gamma(N-(w-1)|\boldsymbol{I}|+a_0+b_0)} \end{aligned}$$

Then the new score function becomes

$$\psi'(\boldsymbol{S}) = \log p(\boldsymbol{I}) + |\boldsymbol{I}| \sum_{j=1}^{w} I_{\text{ent}}(\hat{\boldsymbol{\theta}}_j \| \boldsymbol{\theta}_0). \tag{4}$$

In many biological problems, people focus on finding the *maximum a posteriori* (MAP) alignment $\hat{\boldsymbol{I}}$ according to either (3) or (4). A main reason for this "point-estimate" approach is that the space of $\boldsymbol{I}$ is discrete and enormous, which makes it extremely difficult to characterize the posterior uncertainty. In fact, there are two related difficulties in the analysis: (a) finding the true mode (the mixing issue of the algorithm), and (b) deciding whether the mode we find is a "chancy" one or the one corresponding to the true motif (the posterior uncertainty). The next paragraph discusses some approaches for coping with (a). For question (b), we resort to the frequentist "null" hypothesis analysis. That is, suppose all the data are generated from the background model (no motif elements at all), we find the distribution of the optimal scores by Monte Carlo simulation. Another related question is whether we can find a scoring function that can better discriminate the true motifs from the spurious ones. For example, we may consider

$$\phi(\boldsymbol{S}) = \log(|\boldsymbol{I}|) \sum_{j=1}^{w} I_{\text{ent}}(\hat{\boldsymbol{\theta}}_j \| \boldsymbol{\theta}_0). \tag{5}$$

This scoring function allows the total number of sites $|\boldsymbol{I}|$ to play a role when $|\boldsymbol{I}|$ is relatively small, but its role decreases as $|\boldsymbol{I}|$ increases. Some recent simulations showed that this scoring function performed quite well [10, 11].

3.2.1 Search strategy

Popular motif search strategies include the progressive comparison method employed by CONSENSUS [18], the EM-based method in MEME [5, 1], and the MCMC-based Gibbs Motif Sampler (GMS [9], renamed as AlignACE in [15]). Other word-frequency based approaches do not seem to be as sensitive to detect subtle patterns as these statistical-model based methods. The GMS employs a predictive updating approach: At each iteration, one uses the current weight matrix for the motif pattern to score every segment of width w of all the sequences in the whole dataset and select those "significant" candidates to form a collection of possible motif sites. Then a new weight matrix is computed based on these candidate motif sites. In order to avoid being trapped in a local mode, GMS uses a probabilistic rule to decide whether a sequence segment being examined should be included as a potential motif site or not.

A new search strategy as implemented by BioProspector (BP) [10], called the "threshold sampler," outperformed the GMS in many cases. BP starts by giving each sequence a high threshold and a low threshold. When using the current weight matrix to scan the sequence, all segments whose scores are above the high threshold are automatically called a motif site and all those that are below the high threshold but above the low threshold are given a chance to be sampled into the set of motif sites. The low-threshold is started as 0 and increased gradually to a suitable level. Other major differences between the BP and the GMS are (a) a Markov background model (up to 3rd order) is used, (b) the scoring function is changed to be (5), and (c) motifs consisting of two blocks with flexible gaps in between are allowed.

An important question, then, is how to judge "significance" and how to incorporate additional information revealed by, say, cross-species comparisons, gene expression clusterings, or data from the chromatin-immunoprecipitation and microarray hybridization (ChIP-array, see [6]). A natural route is to build an appropriate statistical model to reflect these information and to construct a search algorithm accordingly. For example, if a sequence segment is located in a region where a cross-species comparison shows that it is highly conserved, then it is highly likely that the segment corresponds to a protein binding site. Otherwise, such a prior probability would be small. In ChIP-array experiments, the intergenic DNA segments with the highest enrichment scores based on differential expression values [14, 6] are most likely to contain the targeted binding sites, and maybe multiple copies of them. Hence, it is essential to direct the GMS to search these few sequences more thoroughly before it wanders off to other less likely sequences. Our new method MDscan [11] was designed to capture these features. This method is especially suitable for discovering novel motifs in a set of intergenic sequence segments that are ranked according to certain scores derived from ChIP-array experiments or comparative hybridization experiments.

4 Motif, Dictionary, and Segmentation

The motif discovery problem as revealed by Figure 4 can also be viewed as a segmentation problem. That is, the problem is equivalent to segmenting the sequences into two types of contiguous pieces, one described by the block-motif model (with a fixed length of w), and the other by an iid model (flexible length). This view can be further generalized into a dictionary model as proposed in [2]. One first assumes the availability of a *dictionary* consisting of a list of d *known* words $\mathcal{D} = \{M_1, M_2, \ldots, M_d\}$. The observed sequence $\boldsymbol{S}$ (as a mathematical abstraction, we treat the whole dataset as a single-sequence) is assumed to be generated by randomly drawing words from the dictionary according to a probability vector $\boldsymbol{\rho} = (\rho(M_1), \ldots, \rho(M_d))$ and sequentially concatenating them together.

We can write down the likelihood of $\boldsymbol{S}$ for this simple model:

$$P(\boldsymbol{S} \mid \boldsymbol{\theta}) = \sum_{\Pi} \prod_{i=1}^{N(\Pi)} \rho(\boldsymbol{S}[P_i]) = \sum_{\Pi} \prod_{j=1}^{d} [\rho(M_j)]^{N_{M_j}(\Pi)}, \tag{6}$$

where $\Pi = (P_1, \ldots, P_k)$ is a partition of the sequence so that each part P_i corresponds to a word in the dictionary, $N(\Pi)$ is the total number of partitions in Π, and $N_{M_j}(\Pi)$ is the number of occurrences of word type M_j in the partition. Clearly, this can be viewed as a missing data problem where the partition Π is missing. The summation over all Π can be achieved recursively, as shown in [8]. Let $L_{i-1}(\boldsymbol{\rho})$ be the sum of all legitimate partitions for partial sequence $\boldsymbol{S}_{[1:(i-1)]}$. Then

$$L_i(\boldsymbol{\rho}) = \sum_{j=1}^{W} \rho(\boldsymbol{S}_{[(i-j):i]}) L_{i-j}, \tag{7}$$

where W is the length of the longest word in the dictionary. In other words, we check whether the last segment of length is a word from the dictionary for all possible word length 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).

Estimating $\boldsymbol{\rho}$ from this model is conceptually simple. There are a few approaches: One can directly optimize (6) via a Newton-type algorithm [2]. Alternatively, one can employ an EM algorithm or a Gibbs sampler. The Gibbs sampler is conceptually simplest, but is perhaps slow in comparison with the Newton-type method. In particular, we can derive an estimating equation from (6) by taking derivative with respect to $\rho(M)$ as

$$\rho(M) = E_{\boldsymbol{\theta}}[N_M(\Pi)/N(\Pi)].$$

This is also derived in [2] from a physics viewpoint.

Since the nature's "dictionary" for the genome construction of a species is unknown to us, the authors in [2] adopted a recursive strategy to estimate it. Briefly, one starts with the simplest dictionary consisting only of the 4 single-letter words, $\mathcal{D}^{(1)} = \{A, C, G, T\}$ and estimates the frequencies of these four words from the dataset (e.g., all the intergenic regions of the yeast genome). Then one considers whether any pair of the letters, AT, say, is over-represented comparing to what is expected by chance. All over-represented pairs are added into the initial dictionary to form a new one $\mathcal{D}^{(2)}$. This strategy can be applied recursively. Based on the current dictionary $\mathcal{D}^{(n)}$, one considers all the concatenations of the pairs of words in $\mathcal{D}^{(n)}$ and forms the new dictionary $\mathcal{D}^{(n+1)}$ by including those new words produced from concatenations that are significantly more abundant than that expected by chance.

4.1 Connection with the block-motif model

Suppose our dictionary consists of only five words, $D = \{A, C, G, T, M_1\}$, where M_1 is a motif sequence of length L. For example, M_1 can be $TGACA$. Then, one can estimate in the dataset the frequency of word M_1, with the consideration of its chance-occurrence. In other words, even if the dataset is generated entirely from the four single letters, there is still a chance of observing M_1. Due to this ambiguity, the frequency estimation of M_1 is not as straightforward as counting.

Now let us assume that M_1 is a fuzzy word in which each position is not a letter, but a probability vector on the four letters (e.g., the second position of the word has 85% chance to be T, 10% to be A, 5% chance to be C and 0% chance to be G). Then, the computation of the segmentation becomes a motif scan algorithm. That is, it is equivalent to scan the whole dataset to see whether there are matches to the postulated pattern represented by M_1. What differs from the usual pattern scanning is that the "threshold" for considering a candidate segment as the occurrence of M_1 is not given in advance, but determined by the dataset.

We can further relax the model by assuming that the *weight matrix* of the word M_1 is completely unknown. Then the dictionary model is equivalent to the repetitive block-motif model subject to minor modifications. A serious statistical question concerning this model is whether and when the parameters can be estimated consistently. Since the motif pattern (weight matrix) is assumed unknown, the only source of information for its inference is its over-abundance in comparison with those "motif-like" patterns occurred by chance under the "null" model.

4.2 Finding the unknown word via Gibbs sampling

As with the previous subsection, we let $\mathcal{D} = \{A, C, G, T, M_1\}$, where M_1 a fuzzy word of length w. Let the usage frequencies of these words be $\boldsymbol{\rho} = \{\rho(\alpha);\ \alpha \in \mathcal{D}\}$ and let the stochastic word matrix (weight matrix) for M_1 be $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_w]$. If the missing partition Π were known, according to (6), we have

$$\boldsymbol{N} = (N_a, \ldots, N_t, N_1 \mid \Pi) \sim \text{Multinom}(N, \boldsymbol{\rho})$$

where $N = \sum_{\alpha \in \mathcal{D}} N_\alpha$. For M_1, we use the same product multinomial model as in Section 3. The prior for Θ is a conjugate product Dirichlet distribution, $PD(\boldsymbol{B})$, where $\boldsymbol{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots \boldsymbol{\beta}_w)$ is a $4 \times w$ matrix with $\boldsymbol{\beta}_j = (\beta_{aj}, \ldots \beta_{tj})^T$. The prior on $\boldsymbol{\rho}$ is $\boldsymbol{\rho} \sim \text{Dir}(\boldsymbol{\gamma})$, $\boldsymbol{\gamma} = (\gamma_a, \ldots, \gamma_t)$. Under these assumptions, the posterior distribution of Θ, conditional on Π, is the product Dirichlet PD($\boldsymbol{B}+\ \boldsymbol{C}$), i.e., the pseudo-counts $\boldsymbol{B}$, updated by the column counts matrix $\boldsymbol{C} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_w)$, where each $\boldsymbol{c}_j$ is the vector of counts of the four nucleotides at the $j$the position of the multiple occurrences of the fuzzy word. The posterior distribution of $\boldsymbol{\rho}$ is $\text{Dir}(\boldsymbol{N} + \gamma)$. In the following, we use the notation $\boldsymbol{\theta}$ to denote the collection of parameters $(\boldsymbol{\rho}, \boldsymbol{\Theta})$, and $\boldsymbol{R}$ to denote $(\boldsymbol{N}, \boldsymbol{C})$.

Under this framework, with $\Pi = (P_1, \ldots, P_k)$ representing the *missing* data relating to the correct partitioning of the sequences, we can write the joint posterior distribution as

$$P(\boldsymbol{\theta}, \Pi \mid \mathbf{R}) = P(\Pi \mid \boldsymbol{\theta}, \mathbf{R}) P(\boldsymbol{\theta} \mid \mathbf{R})$$

Now the Bayes estimate of $\boldsymbol{\theta}$ can be approximated by the Gibbs sampler using the full conditional distributions $P(\Pi \mid \boldsymbol{\theta}, \mathbf{R})$ and $P(\boldsymbol{\theta} \mid \mathbf{R}, \Pi)$. The initial step samples a segmentation of the sequence conditional on the current value of the weight matrix $\boldsymbol{\Theta}$ and word frequencies $\boldsymbol{\rho}$. Then it updates the weight matrix $\boldsymbol{\Theta}$ given the current partition.

Sampling for partitions given the word weight matrix can be done efficiently by using the dynamic programming technique, as in [16]. This initially involves recursive summation of probabilities as given in (7) over all legitimate partitions of all sequences. This is followed by "backward sampling" for words, starting from the end of the sequence and progressing backwards. Let $\mathcal{A}_i$ denote the set of words (sampled partitions) from position i onward to position n, where n is the length of the current sequence. At position i, we sample for a word α_j of size j, $(j = 1, \ldots w)$ according to the conditional probability (given words occurring positions $i+1$ and beyond),

$$P_i(\alpha_j \mid \mathcal{A}_{i+1}) = \frac{\theta(\boldsymbol{S}_{[(i-j):i]}) L_{i-j}}{\sum_{j=1}^{w} \theta(\boldsymbol{S}_{[(i-j):i]}) L_{i-j}} = \frac{\theta(\boldsymbol{S}_{[(i-j):i]}) L_{i-j}}{L_i(\boldsymbol{\theta})}, \qquad j = 1, \ldots w$$

After the step of segmentation sampling it is easy to update the stochastic word matrix from its posterior Dirichlet distribution conditional on the current segmentation. It is not difficult to generalize this idea for the case of more than one "word", with differing lengths. It would mean expanding the dictionary to include $\{M_2, \ldots, M_d\}$ with associated stochastic word matrices $\{\Theta_i : 2 \leq i \leq d\}$. Two unsolved statistical questions are (a) whether the parameters remain distinguishable and (b) whether there are sufficient conditions that guarantee that two words are unconfounded.

Many real biological sequences demonstrate a certain degree of serial correlations and the presence of short nucleotide repeats (of lengths 2 or 3). These are often biologically insignificant but tend to trap motif-searching algorithms. An attempt to get around this problem is by using a Markovian background structure instead of a random background. In the dictionary model, a comparable background structure may be introduced by including over-represented dimers or trimers as *known words* in the dictionary, to account for first and second order dependence. An interesting point to note here is that we are using short polymers to "discount" from the motif signal, whereas in the [2] the dictionary is built up by successive concatenating over-represented short oligomers.

It is useful to note here the that the segmentation model is equivalent to the repetitive block motif model in Section 3. The Gibbs motif sampler (GMS) progressively looks at the sequences and decides whether the segment focused on is more likely to come from the motif model or the background. In this sense its inference is based on the conditional probability $P(I_i \mid \boldsymbol{I}_{[-i]}, \mathbf{M})$ where I_i is the indicator variable for the start of a motif pattern at i, and $\boldsymbol{I}_{[-i]}$ refers to the other motif sites sampled under the motif model $\mathbf{M}$. The segmentation model focuses on the joint likelihood of all motif sites in order to update its knowledge of motif site and composition, i.e., $P(\boldsymbol{I} \mid \mathbf{M})$. It is of interest to study how much the segmentation-based Gibbs sampler can improve over the previous GMS.

As with the original Gibbs sampler [4, 7], the new algorithm just described can be easily trapped in a suboptimal "shifted" mode. For example, let $\mathbf{z} = (z_1, \ldots, z_K)$ be the set of starting positions of the target motif of length w. Then $\mathbf{z} + \delta$, for a small integer δ, are local modes of the distribution, differing from the true mode by a common shift, since $w - \delta$ positions are still correctly aligned. In order to encourage global shifting, we insert a Metropolis step, after a certain degree of stability has been reached in the Gibbs algorithm. A shift-move entails to check either of the two positions flanking the motif element for a higher degree of nucleotide conservation. Similar to [7], we also marginalize out the nuisance parameter $\boldsymbol{\theta}$ in this step and get the unconditional likelihood for the current set of motif start positions $P(\boldsymbol{z} \mid \boldsymbol{S}) \propto \int P(\boldsymbol{S}, \boldsymbol{z} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$. The Metropolis adjustment is carried out in the following steps:

- Choose $\delta = \pm 1$ with probability 1/2 each;
- Update the motif positions to $\boldsymbol{z} + \delta$ with probability $\min \left\{ 1, \frac{P(\boldsymbol{z}+\delta \mid \boldsymbol{S})}{P(\boldsymbol{z} \mid \boldsymbol{S})} \right\}$.

4.3 Influence of the prior

An important question in a Bayesian analysis Bayesian analysis is how to choose a prior distribution so that it would not unduly influence our results yet not fail to use potentially important knowledge we have regarding the data. The prior information in the segmentation model is incorporated in the form of pseudo-counts for the Dirichlet prior over the word probabilities in the starting dictionary. To see the effect of prior parameters on our results, we applied the algorithm on the CRP data set as shown in Figure 1. The degree of correct detection of binding sites, (the total number of experimentally determined sites is 24) under different degrees of prior strength for the presence of motif, can be seen in the table below (TP and FP denote true and false positives occurring more than 20% in the posterior draws, PC denotes pseudo-counts of motif as a fraction of base pseudo-counts):

Motif PC(%)	0.25	1.25	2.5	5	10
Average TP: mean (s.e.)	17 (0.99)	17.3 (1.21)	17.8 (0.35)	18 (0.45)	18.8 (0.37)
Average FP	1.3	2	4.3	7.2	11.2

Even though some of the "FPs" may not be real FPs (i.e., it is not known for sure that these are NOT CRP-binding sites), it appears that there is a trade-off between the percentage of correct sites sampled and the number of false detections. For this example, it seems that a stringent prior is preferred. It is the true motif frequencies for this dataset is about 1.27 per hundred bases.

4.4 Finding gapped motifs through segmentation

Another interesting aspect of this motif finding problem is to deal with motifs that may have one or more insertions of nucleotides, *gaps*, within them. The segmentation model can be extended to allow for this new possibility. A mathematical question of interest is, how far these gaps can extend, beyond which the original pattern becomes indiscernible.

The problem of searching for gapped motifs can be thought of as trying to align a $d \times w$ weight matrix with a segment of length $(w+g)$, where g is the maximum number of gaps allowed in a motif. Thus, in such an alignment we should not penalize the gaps that occur at the beginning or end of the sequence segment. Our probability model for segmentations now is:

$$P(\Pi, \mathcal{G} \mid \boldsymbol{\theta}, \mathbf{R}) = P(\Pi \mid \boldsymbol{\theta}, \mathbf{R}) P(\mathcal{G} \mid \Pi, \boldsymbol{\theta}, \mathbf{R}),$$

where $\mathcal{G}$ denotes the collective set of gap positions within all motifs. Here we need to additionally sample from the full conditional distribution of $\mathcal{G}$, given the current set of sampled partition Π and weight matrix $\boldsymbol{\Theta}$. At this stage, we introduce additional probability parameters, p_m, probability of a match between a nucleotide of the segment with the weight matrix, p_{go}, gap-opening penalty, p_{ge}, gap-extension penalty (typically lower than gap-opening penalty, as a motif is more likely to have few but longer gaps within it than a series of numerous small gaps). At this stage we assume there are no deletions in the motif, hence no gaps in the weight matrix, an assumption which may be later relaxed.

Now we need to deal with an alignment problem, basically sampling for gaps within a segment $\boldsymbol{x} = (x_1, \ldots, x_{w+g})$. This may also be done recursively. Let us assume, without loss of generality, that the motif is aligned exactly with the weight matrix at the end, i.e., the last position is not part of a gap. Let M_j denote the number of matches before position j. Let $P_{\theta_0}(x)$ denote probability of nucleotide x if it lies in a gap (i.e., under the background model), and $P_\Theta(x, i)$ denote the probability of x being realized from the i-th column of the weight matrix. Denote the probabilities of the j-th position being the i-th match, or belonging to a gap after the i-th match, as, respectively,

$$F_{i,j}(M) = P(x_j \in \mathcal{G}^c \mid M_j = i-1), \qquad \text{and} \qquad F_{i,j}(H) = P(x_j \in \mathcal{G} \mid M_j = i)$$

Then we can calculate the above quantities recursively, (with initial conditions $F_{0,0}(H) = 1, F_{1,1}(M) = p_m P_\Theta(x_1, 1)$),

$$\begin{aligned}
F_{0,k}(H) &= F_{0,k-1}(H) P_{\theta_0}(x_k), \quad 2 \le k \le g \\
F_{j,k}(H) &= [F_{j,k-1}(H) p_{ge} + F_{j,k-1}(M) p_{go}] P_{\theta_0}(x_k) \quad j+1 \le k \le g+j, \quad 1 \le j \le w-1 \\
F_{j,k}(M) &= [F_{j,k-1}(H) + F_{j,k-1}(M)] p_m P_\Theta(x_k, j) \quad j \le k \le g+j, \quad 1 \le j \le w-1
\end{aligned}$$

Then, for the segment $\boldsymbol{x}$,

$$P(\boldsymbol{x}) = F_{w,w+g}(M) = [F_{w-1,w+g-1}(H) + F_{w-1,w+g-1}(M)] p_m P_\Theta(x_{w+g}, w)$$

Sampling for gaps can now be accomplished by sampling a two-dimensional pathway from the aligned ends of the segment $\boldsymbol{x}$ and the weight matrix $\boldsymbol{\Theta}$, up to the starting point of the segment.

In summary, it appears that using the segmentation model provides an elegant Bayesian tool for motif-finding that can be modified without excessive complications to suit different biological contingencies. The model makes use of minimal assumptions about the composition and profusion of the motif and updates these aspects based only on the information contained in the sequence data. A potential problem of this approach is the possibility that it might fail to yield significant results in situations where the motif signal is comparatively faint, as the only source of information it uses in inference is over-abundance of patterns in comparison to chance occurrences under the background (null) model, but this requires further study before conclusions can be made. Additional efforts can also be made to refine the characterization of the motif sites, e.g., the mutual distances, relationships among different motifs, and palindromic patterns.

5 Further Topics

As we mentioned in Section 2, an effective strategy for motif discoveries [15] is to (a) use microarray data to find clusters of coregulated genes, (b) extract parts of the upstream regions of these genes, and (c) use the Gibbs sampler or a comparable algorithm to find the binding sites. However, this "stepwise" approach is not as desirable as an approach that combines the motif analysis and the microarray analysis under a unified model. However, there has not been such an approach with demonstrated success. Intuitively, the two analyses should enhance each other in a biologically meaningful way: The gene clusters inferred from the microarray analysis often reveal genes involved in related biological pathway or genes that are regulated by the same TF. If the motif analysis can indeed reveal some significant motifs for these genes, it not only confirms the clustering result, but also suggests future experimental directions.

More recently, a new protocol called the chromatin-immunoprecipitation followed by microarray hybridization (ChIP-array) has been developed for discovering protein-DNA interaction loci *in vivo* [14, 6]. In these experiments, DNA is cross-linked *in vivo* to proteins (i.e., covalently bonded) at sites of DNA-protein interactions, and sheared to 1-2kb fragments. The DNA-protein complexes are precipitated by antibodies specific to the protein of interest. The precipitated protein-bound DNA fragments are amplified, labeled fluorescently, and hybridized to microarrays containing every ORF and intergenic region of the organism's genome. DNA fragments that are consistently enriched by ChIP-array over repeated experiments are identified as containing the protein-DNA interacting loci. The DNA fragments selected by the ChIP-array experiments have an average size of 1-2kb determined by the shearing process, which is still difficult for the biologist to determine the actual binding sites and pattern. The motif discovery methods discussed in this article can all be applied here to further pinpoint the exact sites [6]. After finding the motif weight matrices from these ChIP-enriched segments, [6] plotted the logarithm of the "motif-score" for each segment in the dataset (i.e., how likely the segment contains certain motif) against the median percentile rank of the expression levels and found a significant monotone relationship. This observation confirms that the motif found by the BioProspector, a variation of the Gibbs sampler, is very likely to be authentic.

Acknowledgment

We thank the Workshop organizers, especially Professor Rob Kass, for inviting us to present our work, and Professors Andrew Gelman and Michael Newton for their helpful suggestions. This work was partly supported by NSF grant DMS0094613 to J.S.L. and NIH grants R21RR14036 and R01HG01257 to C.E.L.

References

[1] T. L. Bailey and C. P. Elkan. Fitting a mixture model by expectation-maximization to discover motifs in biopolymers. *ISMB*, pages 28–36, 1994.

[2] H. J. Bussemaker, H. Li, and E. D. Siggia. Building a dictionary for genomes: Identification of presumptive regulatory sites by statistical analysis. *Proc. Nat'l Acad. Sci. USA*, 97(18):10096–10100, 2000.

[3] A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler. Hidden markov-models in computational biology : Applications to protein modeling. *Journal of Molecular Biology*, 235(5):1501–1531, 1994.

[4] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–14, 1993.

[5] C. E. Lawrence and A. A. Reilly. An expectation maximization (em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, 7:41–51, 1990.

[6] J. D. Lieb, X. Liu, D. Botstein, and P. O. Brown. Promoter-specific binding of rap1 revealed by genome-wise maps of protein-dna association. *Nature Genetics*, 28:327–334, 2001.

[7] J. S. Liu. The collapsed gibbs sampler in bayesian computations with applications to a gene-regulation problem. *Journal of the American Statistical Association*, 89(427):958–966, 1994.

[8] J. S. Liu and C. E. Lawrence. Bayesian inference on biopolymer models. *Bioinformatics*, 15(1):38–52, 1999.

[9] J. S. Liu, A. F. Neuwald, and C. E. Lawrence. Bayesian models for multiple local sequence alignment and gibbs sampling strategies. *Journal of the American Statistical Association*, 90(432):1156–1170, 1995.

[10] X. Liu, D. L. Brutlag, and J. S. Liu. Bioprospector: Discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. In *Pacific Symposium on Bioinformatics*, volume 6, pages 127–138, Hawaii, 2001.

[11] X. Liu, D. L. Brutlag, and J. S. Liu. A fast computational method for finding protein-dna interaction sites from chromatin immunoprecipitation microarray experiments. Technical report, Department of Statistics, Harvard University, 2001.

[12] L. A. McCue, W. Thompson, C. S. Carmack, M. P. Ryan, J. S. Liu, V. Derbyshire, and C. E. Lawrence. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Research*, 29(3):774–782, 2001.

[13] A. F. Neuwald, J. S. Liu, and C. E. Lawrence. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci*, 4(8):1618–32, 1995.

[14] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of dna binding proteins. *Science*, 290:2306–2309, 2000.

[15] F. P. Roth, J. D. Hughes, P. W. Estep, and G. M. Church. Finding dna regulatory motifs within unaligned noncoding sequences clustered by whole-genome mrna quantitation. *Nature Biotechnology*, 16(10):939–945, 1998.

[16] S. C. Schmidler, J. S. Liu, and D. L. Brutlag. Bayesian segmentation of protein secondary structure. *Journal of Computational Biology*, 7(1-2):233–248, 2000.

[17] T. D. Schneider Stephens and R. M. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.*, 18:6097–6100, 1990.

[18] G. D. Stormo and G. W. Hartzell III. Identifying protein-binding sites from unaligned dna fragments. *Proceedings of the Nathional Academy of Science, USA*, 86:1183–1187, 1989.

[19] M. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–550, 1987.