2004

Computational Discovery of Gene Regulatory Binding Motifs: A Bayesian Perspective

Shane T. Jensen
University of Pennsylvania

X. Shirley Liu
Harvard University

Qing Zhou
Harvard University

Jun S. Liu
Harvard University

Recommended Citation
Jensen, S. T., Liu, X., Zhou, Q., & Liu, J. S. (2004). Computational Discovery of Gene Regulatory Binding Motifs: A Bayesian Perspective. *Statistical Science, 19* (1), 188-204. http://dx.doi.org/10.1214/088342304000000107

Computational Discovery of Gene Regulatory Binding Motifs: A Bayesian Perspective

Abstract
The Bayesian approach together with Markov chain Monte Carlo techniques has provided an attractive solution to many important bioinformatics problems such as multiple sequence alignment, microarray analysis and the discovery of gene regulatory binding motifs. The employment of such methods and, more broadly, explicit statistical modeling, has revolutionized the field of computational biology. After reviewing several heuristics-based computational methods, this article presents a systematic account of Bayesian formulations and solutions to the motif discovery problem. Generalizations are made to further enhance the Bayesian approach. Motivated by the need of a speedy algorithm, we also provide a perspective of the problem from the viewpoint of optimizing a scoring function. We observe that scoring functions resulting from proper posterior distributions, or approximations to such distributions, showed the best performance and can be used to improve upon existing motif-finding programs. Simulation analyses and a real-data example are used to support our observation.

Keywords
gene regulation, motif discovery, Bayesian models, scoring functions, optimization, Markov chain Monte Carlo

Disciplines
Statistics and Probability

Statistical Science
2004, Vol. 19, No. 1, 188–204
DOI 10.1214/088342304000000107

Computational Discovery of Gene Regulatory Binding Motifs: A Bayesian Perspective

Shane T. Jensen, X. Shirley Liu, Qing Zhou and Jun S. Liu

Abstract. The Bayesian approach together with Markov chain Monte Carlo techniques has provided an attractive solution to many important bioinformatics problems such as multiple sequence alignment, microarray analysis and the discovery of gene regulatory binding motifs. The employment of such methods and, more broadly, explicit statistical modeling, has revolutionized the field of computational biology. After reviewing several heuristics-based computational methods, this article presents a systematic account of Bayesian formulations and solutions to the motif discovery problem. Generalizations are made to further enhance the Bayesian approach. Motivated by the need of a speedy algorithm, we also provide a perspective of the problem from the viewpoint of optimizing a scoring function. We observe that scoring functions resulting from proper posterior distributions, or approximations to such distributions, showed the best performance and can be used to improve upon existing motif-finding programs. Simulation analyses and a real-data example are used to support our observation.

Key words and phrases: Gene regulation, motif discovery, Bayesian models, scoring functions, optimization, Markov chain Monte Carlo.

1. THE BIOLOGY OF TRANSCRIPTION REGULATION

The complete information that defines the characteristics of living cells within an organism is encoded in the form of a moderately simple molecule, deoxyribonucleic acid, or DNA. The building blocks of DNA are four nucleotides, abbreviated by their attached organic bases as A, C, G and T. A–T and C–G are complementary bases between which hydrogen bonds can form. A DNA molecule consists of two long chains of nucleotides that are complementary to each other and joined by hydrogen bonds twisted into a double helix. This structure gives rise to the term "base pair" when describing a DNA sequence. The specific ordering of these nucleotides, the "genetic code," is the means by which information is stored that completely defines all functions within a cell. With the recent development of high-throughput sequencing technology, the National Institutes of Health genetic sequence database, GenBank, has sustained an exponential growth rate since 1982. Right now GenBank contains the complete genomic sequences of over 1,000 organisms (Benson et al., 2002) with approximately 22 billion DNA bases.

The central dogma of molecular biology dictates that certain segments of the DNA (i.e., genes) are transcribed into another molecule, RNA, which serves as a transient template to make the basic building blocks of cellular life, proteins. Although all the cells in the same organism possess exactly the same DNA sequences (i.e., genetic information), they display different physiological characteristics in different tissues,

Shane T. Jensen and Qing Zhou are Ph.D. students and Jun S. Liu is Professor, Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, Massachusetts 02138, USA (e-mail: jensen@stat.harvard.edu, zhou@stat.harvard.edu, jliu@stat.harvard.edu). X. Shirley Liu is Assistant Professor, Department of Biostatistics, Harvard School of Public Health and Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA (e-mail: xsliu@jimmy.harvard.edu).

developmental stages and environmental conditions. This "differentiation" is caused by the differences among the collections of proteins that are synthesized in different cells or at different cell states. If a protein is being synthesized at a certain state, its coding DNA (called a gene) is termed as "active" or "expressed." Thus, a cell in a particular physiological state can be roughly viewed as a mechanical system where each different gene is switched either on (active) or off (inactive).

In many organisms, the DNA that codes for proteins (genes) is only a small portion of the total genomic DNA. For example, genes make up only about 1.5% of the human genome (International Human Genome Sequencing Consortium, 2001). The noncoding components of DNA, which were initially considered as "junk" sequences, actually contain the control mechanisms for activating and deactivating the genes, and thus the synthesis and nonsynthesis of proteins. Most of the control sequences for a gene lie in the *upstream regulatory region*, which is the region of a few thousand base pairs directly before the gene [also called the transcription regulatory region (TRR), or the promoter]. Transcribing or activating a gene requires not only the DNA sequence in the TRR, but also many proteins called transcription factors (TFs). When these TFs are present, they bind to specific DNA patterns in the TRR of genes and either induce or repress the transcription of these genes by recruiting other necessary proteins (Lodish et al., 1995).

One transcription factor can bind to many different upstream regions, thus regulating the transcription of many genes. The binding sites of the same transcription factor show a significant sequence conservation, which is often summarized as a short (5–20 bases long) common pattern called a transcription factor binding motif (TFBM) or binding consensus, although some variability is tolerated. It is the main focus of this paper to discover the locations and common pattern of these motif sites.

In prokaryotes (lower organisms without nuclei), there are fewer TFs, their motifs tend to be relatively long and the strength of regulation for a particular gene often depends on how closely a particular site matches the consensus for the motif. The more mismatches to the consensus in a binding site, the less often the TF will bind and therefore the less control it will exert on the target gene. The variability between sites is sometimes crucial to the regulatory process, since TF binding sites that are perfect matches to the optimal pattern would bind the TF too tightly, preventing the subsequent steps of transcription (Pfahl, 1981).

In eukaryotes (higher organisms with nuclei), many more transcription factors are involved in the regulation of a gene, and their binding motifs tend to be shorter. Eukaryotic upstream regions usually contain regulatory modules, a collection of adjacent binding sites (sometimes multiple binding sites) of several transcription factors. Transcription regulation not only relies on the combination of the TFs involved, but also on the number of site copies in the upstream regions (Werner, 1999).

Characterizing the motifs of TFs and locating TF binding sites are crucial tasks for understanding how the cell regulates its genes in response to developmental and environmental changes. However, the gold standard experimental procedures to determine binding sites are inefficient, sometimes impractical, and it can only discover one transcription factor binding site at a time. With the availability of complete genome sequences, biologists are using techniques such as DNA microarray (Schena, Shalon, Davis and Brown, 1995) or serial analysis of gene expression (SAGE; Velculescu, Zhang, Vogelstein and Kinzler, 1995) to measure the expression level of every gene in an organism in various conditions. A collection of expressions of a gene measured under various conditions is called the expression profile of the gene. A genome can be divided into gene clusters according to similarities in their expression profiles (Eisen, Spellman, Brown and Botstein, 1998). Genes in the same expression cluster respond similarly to environmental and developmental changes and thus may be coregulated by the same TF or the same group of TFs. Therefore, our computational analysis can be focused on the search for TF binding sites in the upstream of genes contained in a particular cluster. Another experimental procedure called chromatin immunoprecipitation followed by microarray (ChIP-array or ChIP-on-chip; Buck and Lieb, 2004) can measure where a particular TF binds to DNA in the whole genome, although at a coarse resolution of 1–2 thousand base pairs. Again, computational analysis is required to pinpoint the short binding sites of a transcription factor from all the long TF binding targets.

With the ever expanding number of whole genomes sequenced and high-throughput gene expression and protein–DNA binding data, motif finding and transcription regulatory network elucidation have become major research topics in computational biology. In Section 2, we describe the basic formulation of the motif finding problem and review discovery methods that

are popular in the field. A formal Bayesian statistical model together with its various extensions is given in Section 3. In particular, we discuss models that allow for unknown motif width and unknown motif abundance ratio. We then investigate the advantages of using scoring functions for motif finding in Section 4. Section 5 discusses the use of a Metropolis-algorithm-based optimization method to improve the results from a Gibbs-sampling-based algorithm, BioProspector, and examines a few simulation studies for comparing different scoring functions. We observed that those scoring functions resulting from a proper Bayes model usually performed the best. Section 7 concludes with a brief discussion.

2. MOTIF FORMULATION AND GENERAL DISCOVERY STRATEGIES

There are two ways of discovering novel binding sites of a TF: *scanning* methods and *de novo* methods. In a scanning method, one uses a motif representation resulting from experimentally determined binding sites to scan the genome sequence to find more matches. In de novo methods, one attempts to find *novel* motifs that are "enriched" in a set of upstream sequences. This article focuses on the latter class of methods. The de novo methods can also be divided into two classes, according roughly to two general data formulations for representing a motif: the consensus sequence or a position-specific weight matrix (PSWM).

2.1 Consensus Sequence Methods

The consensus sequence shows the motif as a string of IUPAC characters (Table 1; see IUPAC, 1986). For example, the Mse motif consensus CRCAAAW suggests that the Mse protein binds to sites starting with a C, followed by A or G, followed by CAAA and followed by A or T. In the following sections, we use *word* and *segment* interchangeably to mean a short DNA sequence being tested by our motif model as a potential binding site. When scanning a set of sequences against a consensus, all words matching the consensus are considered putative binding sites. This sometimes results in many false positive sites, and it may miss some true sites with variability that is not represented by the consensus sequence.

TABLE 1
IUPAC nomenclatures for DNA consensus

A	Adenine	C	Cytosine
G	Guanine	T	Thymine
R	Purines (A, G)	Y	Pyrimidines (C, T)
W	Weak hydrogen bond (A, T)	S	Strong hydrogen bond (C, G)
M	Amino group (A, C)	K	Keto group (G, T)
B	Not A (C, G, T)	D	Not C (A, G, T)
H	Not G (A, C, T)	V	Not T (A, C, G)
N	Any (A, C, G, T)		

Early research on discovering motifs was usually simplified to finding a sequence pattern enriched or overrepresented in the sequence dataset compared to the genome background. Therefore, many computational algorithms for finding motif consensus sequences adopted a "pattern-driven" or "word enumeration" approach by enumerating predefined consensus patterns to see which is significantly enriched in the sequence dataset.

The first consensus sequence enumeration method was developed (Galas, Eggert and Waterman, 1985) to search for a TATA-box motif that appears once in each upstream region. They first align all the upstream sequences at the transcription start site. Then for every aligned position, they search in the nine-base windows centered at that position of all the sequences. In this window, every possible pattern b_i of width 6 is scored according to $S(b_i) = (6/6)q_{i6} + (5/6)q_{i5} + (4/6)q_{i4}$, where q_{ik} is the number of sequences whose best matching 6-mer (subsequence of length 6) to b_i in the nine-base window has k matched positions. The highest scoring pattern is considered as a potential motif and the positions corresponding to this are considered potential binding locations.

In most motif finding problems, the binding site locations are unknown and their distances from the transcription start site vary extensively. Therefore, oligoanalysis (van Helden, Andre and Collado-Vides, 1998) was developed to find sequence patterns enriched in the whole upstream region. This method enumerates every possible pattern b_i of a certain width to determine whether it occurs in the dataset more than expected. Sinha and Tompa (2000) later extended this method to allow for one-base mismatch and to use the IUPAC alphabet to find motifs with more flexible base substitutions. To speed up computation, Sinha and Tompa calculated the mean and variance of the number of occurrences of b_i and determined its significance by a Z-test. Their calculations were based on a third-order Markov model for noncoding sequences in the genome. As shown in Liu, Brutlag and Liu (2001), the Markov model discriminates against meaningless patterns such as AAAA or ATAT that are frequently found in the noncoding sequences and therefore increases the specificity of the discovered motifs.

The time to enumerate all possible consensus patterns increases exponentially as the pattern width increases, so finding longer motif patterns is a challenge. Since many long motifs are more conserved near the two ends, van Helden, Rios and Collado-Vides (2000) proposed to detect long motifs as spaced dyad patterns such as $w_1 \cdot ns \cdot w_2$, where w_1 and w_2 are the dyad motif words with short enough widths, and ns is the s-base spacer of unspecified sequence. The expected occurrences of a spaced dyad can be determined either by calculating from the joint distribution of w_1 and w_2 assuming that w_1 and w_2 are conditionally independent, or by counting $w_1 \cdot ns \cdot w_2$ occurrences in the whole genome noncoding sequences.

Another method encodes nucleotides using a two-bit binary number instead of an eight-bit character and converts the sequence into a much shorter array for quick access (Hampson, Baldi, Kibler and Sandmeyer, 2000). A third method uses a suffix tree to represent all patterns of all widths that exist in the whole genome noncoding regions (Brazma, Jonassen, Vilo and Ukkonen, 1998). Keich and Pevzner (2002) introduce models for more refined consensus pattern searching, which are useful in the case of very subtle motifs. Each node contains a sequence pattern that reflects the path from the root to the node and stores information of the count and location of all the sequences matching that pattern. In addition, each node can branch into A, C, G and T to form patterns one base longer. Although building the full tree is extremely time and memory intensive, one can trim many "rare" nodes to speed up tree-building.

A recent method called MobyDick builds longer motifs from concatenating shorter ones (Bussemaker, Li and Siggia, 2000). MobyDick models the sequence dataset as being generated by concatenations of words drawn independently from a *dictionary* with their respective "usage" frequencies. The initial motif dictionary contains individual bases A, C, G and T, with their frequencies estimated from genome noncoding sequences. Longer patterns are formed by adding into the dictionary those concatenated word pairs that have occurred more than expected (e.g., "CG" would be treated as a new word if its occurrence is significantly more than what is expected from the independent pairing). The frequencies are reestimated for all the words in the new dictionary to maximize the likelihood of generating the sequence dataset. The process is repeated until no new words can be added. This method has recently been generalized to a stochastic dictionary model (Gupta and Liu, 2003).

2.2 Position-Specific Weight Matrix and Statistical Models

An alternative motif formulation is a position-specific weight matrix, or simply *motif matrix*, which measures the desirability of each base at each position of the motif. The simplest matrix is an alignment matrix n_{jk}, which records the occurrence of base k at position j of all the aligned sites for this motif (Table 2). Also shown in Table 2 are the corresponding frequency matrix ($f_{jk} = n_{jk}/N$), where N is the number of motif sites, and weight matrix $\log[f_{jk}/\theta_{0k}]$ (Hertz and Stormo, 1999), where θ_{0k} is the proportion of base k in the nonmotif (background) positions.

A formal statistical model for the weight matrix method was described in Lawrence and Reilly (1990) and a complete Bayesian method was given in Liu (1994) and Liu, Neuwald and Lawrence (1995). In this model, the sequence data is represented as an array **S**, where S_{ij} is the base in position j of sequence i. Each base can take on $K = 4$ different values corresponding to the nucleotides A, C, G and T. To reflect the fact that

TABLE 2
Matrix representation of transcription factor binding motif BCD

	Alignment matrix				Frequency matrix				Weight matrix			
Pos	**A**	**C**	**G**	**T**	**A**	**C**	**G**	**T**	**A**	**C**	**G**	**T**
1	0	4	7	1	0.00	0.33	0.58	0.08	−2.56	0.27	0.80	−0.96
2	2	1	8	1	0.17	0.08	0.67	0.08	−0.37	−0.96	0.93	−0.96
3	0	0	12	0	0.00	0.00	1.00	0.00	−2.56	−2.56	1.33	−2.56
4	12	0	0	0	1.00	0.00	0.00	0.00	1.33	−2.56	−2.56	−2.56
5	0	0	0	12	0.00	0.00	0.00	1.00	−2.56	−2.56	−2.56	1.33
6	0	0	0	12	0.00	0.00	0.00	1.00	−2.56	−2.56	−2.56	1.33
7	12	0	0	0	1.00	0.00	0.00	0.00	1.33	−2.56	−2.56	−2.56
8	6	1	2	3	0.50	0.08	0.17	0.25	0.65	−0.96	−0.37	0.00

the motif sites within **S** are substrings of length w that are conserved relative to each other, we model them as independent realizations from a common *Motif* model. That is,

$$(s_1, \ldots, s_w) \sim \text{ProductMultinomial}(\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_w))$$

if $(s_1, \ldots, s_w)$ is an observed motif site in **S** (a substring of width w), where $\boldsymbol{\theta}_j = (\theta_{jA}, \theta_{jC}, \theta_{jG}, \theta_{jT})$ is a probability vector for the preference of the nucleotide types in position j. This model means that, for example, the motif site "TTACTAA" is generated with probability $\theta_{1T}\theta_{2T}\theta_{3A}\theta_{4C}\theta_{5T}\theta_{6A}\theta_{7A}$. The remainder of the sequences are classified as nonsites, for which the simplest model is the i.i.d. multinomial distribution with the "null" frequency $\boldsymbol{\theta}_0 = (\theta_{0A}, \ldots, \theta_{0T})$. Since the motif sites are only a tiny fraction of the whole sequence data, we can estimate $\boldsymbol{\theta}_0$ first (e.g., direct counting of the four nucleotide types) and subsequently treat it as known. It has been shown recently that using a Markov chain to model the nonsite positions can improve the motif specificity (Liu, Brutlag and Liu, 2001).

From the alignment of a set of binding sites, we can easily derive a frequency matrix f_{jk}, which is the MLE of θ_{jk}, and the weight matrix given in Table 2. These matrices can be used to scan the whole genome sequence, by computing for each segment its likelihood of being generated from the motif model, to discover novel realizations of the binding motif. This strategy tends to be more accurate in capturing the correct sites than using the matching criterion based upon the consensus sequence formulation.

In a majority of gene regulation analysis problems, we know neither the locations of the motif sites nor the motif pattern (i.e., $\boldsymbol{\Theta}$ or an estimate of it). Thus, we need to simultaneously estimate the motif matrix and locate the possible motif sites in the sequence data. A particularly successful class of computational algorithms for this problem adopts a "data-driven" or "matrix update" approach based either on the EM algorithm or Gibbs sampling (Lawrence and Reilly, 1990; Lawrence et al., 1993; Liu, 1994). These methods typically initiate a motif matrix randomly and use the sequence dataset to gradually refine the motif. It is the focus of this article to give an overview and extension of this class of algorithms, providing for them a rigorous Bayesian or likelihood foundation, and to discuss possible improvements.

2.3 Motif Discovery Methods Based on Motif Matrix Updating

The first algorithm for discovering novel motifs was CONSENSUS (Stormo and Hartzell, 1989). Assuming that each sequence contains one motif site, the algorithm starts by examining all possible locations of the motif sites in the first two sequences [a total of $(n_1 - w + 1)(n_2 - w + 1)$ comparisons], and chooses the top X pairs of motif sites according to the relative entropy scores of their corresponding motif matrix, where the score is defined as $\psi_{\text{ENT}} = \sum_{j=1}^{w}\sum_{k=A}^{T} f_{jk} \log f_{jk}/\theta_{0k}$, where f_{jk} is the observed frequency of base type k in the jth position and $\log f_{jk}/\theta_{0k}$ is the weight matrix given in Table 2. Later, another scoring function was deduced to estimate the p-value of each motif, which is the probability of observing a motif from random alignment of the same size that scores equally or higher (Hertz and Stormo, 1999). Only motifs with high information content or low p-value are retained, and each is aligned with every possible w-mer (subsequence of length w) in the third sequence to form a set of new matrices and the top K matrices are retained. The algorithm cycles through all the sequences in the same fashion and the best-scoring motifs are reported at the end as potential TFBMs. When there are more motif sites in the first few sequences in the dataset, especially the first two sequences, CONSENSUS is effective. Otherwise, a number of runs using different sequence orders are needed.

Another matrix motif discovery algorithm is based on a missing data formulation, which will be detailed in the next section, and the EM algorithm (Lawrence and Reilly, 1990). The original algorithm restricts each sequence to contain one TF site. A later method called MEME overcomes this limitation (Bailey and Elkan, 1994; Grundy, Bailey and Elkan, 1996) by introducing a prior probability for every position to be the start of a motif site. The algorithm also uses every existing w-mer in the sequence dataset to initialize the EM iteration, thus improving the convergence properties of the original method of Lawrence and Reilly (1990).

About the same time, a Bayesian method and several related Gibbs sampling algorithms for motif discovery were also developed (Lawrence et al., 1993; Liu, 1994; Liu, Neuwald and Lawrence, 1995), and these Bayesian approaches together with powerful Markov chain Monte Carlo tools demonstrate more modeling and computational flexibilities. For example, many

new methods have been explored to extend the functionality of Gibbs sampling. Gibbs Motif Sampler incorporates a prior probability of motif occurrence in the sampling, thus allowing variable number of motif sites in each input sequence (Liu, Neuwald and Lawrence, 1995). By only considering the k positions out of w in the motif with the richest information content, it allows the motif to contain small gaps. AlignACE continues to improve the Gibbs Motif Sampler by iteratively masking out aligned sites to find multiple different motifs (Roth, Hughes, Estep and Church, 1998). BioProspector uses a Markov model estimated from the whole genome noncoding sequences to represent the nonmotif background in order to improve the motif specificity (Liu, Brutlag and Liu, 2001). It can also find motifs that have two conserved blocks separated by a nonconserved gap of variable length.

Algorithms based on word matches are usually exhaustive in finding motifs, but are limited by the maximum width of the motif that can be enumerated. Algorithms based on matrix update algorithms can find motifs of any specified width, but none can guarantee convergence or a globally optimal motif. To strike a balance of the two, a recent algorithm, MDscan (Liu, Brutlag and Liu, 2002), first uses a word enumeration method to search motifs from the top L sequences that biologists are most confident contain the motif. Using every existing w-mer in these sequences as a seed, MDscan finds all w-mers in the L sequences that are similar to the seed and constructs from them a motif matrix. All the motif matrices are evaluated by a semi-Bayesian scoring function and the best ones are further refined using all the sequences in the dataset. When the motif is weak and the data are noisy, searching for motifs first from sequences with high signal-to-background ratio increases the chance of success.

In the past decade, much effort has been made in the area of regulatory motif analysis and many algorithms have been developed. Although there may still be debate and arguments over which algorithms are "best" in a certain situation, the few most popular motif-finding algorithms (e.g., CONSENSUS, MEME, AlignACE, Gibbs Motif Sampler, BioProspector) are all based on explicit statistical modeling, either fully or partially, in contrast to the word enumeration methods of van Helden and co-workers (van Helden, Andre and Collado-Vides, 1998; van Helden, Rios and Collado-Vides, 2000), Sinha and Tompa (2000), Hampson et al. (2000) and Brazma et al. (1998).

We can comfortably claim that the introduction of the full statistical model and the missing-data formulation has played a pivotal role in revolutionizing this particular research area as well as the field of computational biology in general.

3. A BAYESIAN TREATMENT OF THE BINDING MOTIF MODEL

3.1 A Complete Bayesian Model

As in the previous section, we let $\mathbf{S}$ denote the set of sequences under investigation, where each S_{ij} takes value in an alphabet of size K ($K = 4$ for DNA sequences). Within $\mathbf{S}$ we postulate that there are substrings of length w that are sites of an unknown *motif* model. The locations of these sites are unknown, so we introduce a missing array of indicators $\mathbf{A}$, where A_{ij} is either one or zero indicating whether or not position j in sequence i is the starting point of a motif site. A particular realization of $\mathbf{A}$ gives us a subset of $\mathbf{S}$, denoted as $\mathbf{S}(\mathbf{A})$, which consists only of the bases in the motif sites, and the complementary subset $\mathbf{S}(\mathbf{A}^c)$, which are the remaining background bases. We can further break down $\mathbf{S}(\mathbf{A})$ into $\mathbf{S}(\mathbf{A}_{(1)}), \mathbf{S}(\mathbf{A}_{(2)}), \ldots, \mathbf{S}(\mathbf{A}_{(w)})$, where $\mathbf{S}(\mathbf{A}_{(j)})$ is the set of bases in the jth position of the motif sites.

We let $\mathbf{N}(\mathbf{C}) = (n_1, n_2, \ldots, n_K)$ be a vector of the counts of the different base types in a particular subset $\mathbf{C}$ of $\mathbf{S}$. With a slight abuse of notation, we also let $\mathbf{N}(\mathbf{A}_{(2)})$ be the vector of the base counts in position 2 of all the motif sites, and we let $\mathbf{N}(\mathbf{A}^c)$ be the vector of all base counts that are not part of a motif site. For two vectors $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ and $\mathbf{N} = (n_1, \ldots, n_K)$, we define

$$\boldsymbol{\theta}^{\mathbf{N}} = \prod_{j=1}^{K} \theta_j^{n_j}, \quad \frac{\boldsymbol{\theta}}{\mathbf{N}} = \prod_{j=1}^{K} \frac{\theta_j}{n_j}, \quad \Gamma(\mathbf{N}) = \prod_{j=1}^{K} \Gamma(n_j),$$

where $\Gamma(\cdot)$ is the Gamma function. For the moment we assume that the motif width w is known and we will attempt a generalization to a variable motif width in a later section.

With the statistical model introduced in Section 2.2, we have

$$\{\mathbf{N}(\mathbf{A}_{(1)}), \ldots, \mathbf{N}(\mathbf{A}_{(w)})\} \sim \text{ProductMultinomial}\big(\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_w)\big),$$

which means that each vector of column totals $\mathbf{N}(\mathbf{A}_{(j)})$ follows a multinomial distribution parameterized by $\boldsymbol{\theta}_j$

independent of the other columns. Viewing $\mathbf{A}$ as missing data, we can write the likelihood of $\mathbf{S}$ as

$$p(\mathbf{S}|\boldsymbol{\Theta}, \boldsymbol{\theta}_0, \mathbf{A}) \propto \boldsymbol{\theta}_0^{\mathbf{N}(\mathbf{A}^c)} \times \prod_{j=1}^{w} \boldsymbol{\theta}_j^{\mathbf{N}(\mathbf{A}_{(j)})}.$$

To enable a Bayesian analysis, we employ the following conjugate prior distributions for $\boldsymbol{\Theta}$ and $\boldsymbol{\theta}_0$:

$$\boldsymbol{\Theta} \sim \text{ProductDirichlet}\big(\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_w)\big)$$

and

$$\boldsymbol{\theta}_0 \sim \text{Dirichlet}(\boldsymbol{\beta}_0),$$

where $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jK})$. For a brief review of multinomial models with Dirichlet prior distributions, refer to Gelman, Carlin, Stern and Rubin (1995). With these prior distributions, the conditional posterior distribution is

$$p(\boldsymbol{\Theta}, \boldsymbol{\theta}_0|\mathbf{S}, \mathbf{A}) \propto \boldsymbol{\theta}_0^{\mathbf{N}(\mathbf{A}^c)+\boldsymbol{\beta}_0} \times \prod_{j=1}^{w} \boldsymbol{\theta}_j^{\mathbf{N}(\mathbf{A}_{(j)})+\boldsymbol{\beta}_j}.$$

A nearly trivial but important improvement of this basic general model is to treat the "nonsite" background bases as being generated by an lth-order Markov chain (empirically $l = 3$ works the best). More generally, we can write the above model as

$$p(\boldsymbol{\Theta}, \boldsymbol{\theta}_0|\mathbf{S}, \mathbf{A}) \propto p\big(\mathbf{S}(\mathbf{A}^c)|\boldsymbol{\theta}_0\big) p\big(\mathbf{S}(\mathbf{A})|\boldsymbol{\Theta}\big) p(\boldsymbol{\theta}_0, \boldsymbol{\Theta}),$$

where $\boldsymbol{\theta}_0$ denotes the parameters in the background Markov model. After prescribing a prior distribution $p(\mathbf{A})$ to $\mathbf{A}$ independent of $\boldsymbol{\theta}_0$ and $\boldsymbol{\Theta}$, we have the joint posterior distribution of all unknowns:

$$p(\boldsymbol{\Theta}, \boldsymbol{\theta}_0, \mathbf{A}|\mathbf{S}) \propto p\big(\mathbf{S}(\mathbf{A}^c)|\boldsymbol{\theta}_0\big) p\big(\mathbf{S}(\mathbf{A})|\boldsymbol{\Theta}\big) p(\boldsymbol{\theta}_0, \boldsymbol{\Theta}) p(\mathbf{A}).$$

In the early methods (e.g., Lawrence and Reilly, 1990; Cardon and Stormo, 1992; Lawrence et al., 1993) it has been assumed that each sequence must contain one and only one motif site, which corresponds to assuming that $A_{ij} = 0$ for all but one entry in the ith row. Thus, no explicit prior distribution for $\mathbf{A}$ is needed if we suppose that the motif site can be anywhere in the sequence with equal probabilities. It has been recently demonstrated, however, that incorporating a model that takes into account the location of the motif site relative to the end of each sequence can improve the sensitivity of the algorithm (McCue et al., 2001). Since biological reality calls for a relaxation of the one-site-per-sequence assumption, Liu, Neuwald and Lawrence (1995) propose an explicit prior for $\mathbf{A}$ and propose a set of Markov chain Monte Carlo algorithms for the computation.

3.2 Markov Chain Monte Carlo Algorithms for Motif Discovery

In a typical data-augmentation-based Gibbs sampling algorithm (Tanner and Wong, 1987), the desired posterior distribution $p(\boldsymbol{\Theta}, \boldsymbol{\theta}_0, \mathbf{A}|\mathbf{S})$ can be simulated by starting with arbitrary initial values of the unknown parameters $\boldsymbol{\Theta}^0$ and $\boldsymbol{\theta}_0^0$, and then for $t = 0, 1, \ldots$ iteratively sampling from the two conditional distributions:

1. $p(\mathbf{A}^t|\boldsymbol{\Theta}^t, \boldsymbol{\theta}_0^t, \mathbf{S})$;
2. $p(\boldsymbol{\Theta}^{t+1}, \boldsymbol{\theta}_0^{t+1}|\mathbf{A}^t, \mathbf{S})$.

Given enough time steps, the draws simulated in this fashion will converge to draws from the desired posterior distribution. Typically, we are most interested in the draws from $p(\mathbf{A}|\mathbf{S})$ which would indicate the most likely positions of the unknown conserved sites.

However, since $\boldsymbol{\theta}_0$ and especially $\boldsymbol{\Theta}$ are of rather high dimension, drawing these parameters at every iteration can be both time-consuming and inefficient. As demonstrated by Liu (1994), the algorithm can be improved by integrating over $\boldsymbol{\Theta}$ and $\boldsymbol{\theta}_0$ so that we can simulate draws via Gibbs sampling from the posterior distribution $p(\mathbf{A}|\mathbf{S})$ directly, where

$$p(\mathbf{A}|\mathbf{S}) = \iint p(\boldsymbol{\Theta}, \boldsymbol{\theta}_0|\mathbf{S}, \mathbf{A}) p(\mathbf{A})\, d\boldsymbol{\theta}_0\, d\boldsymbol{\Theta}.$$

We now give variations on the basic motif model under different assumptions and the algorithmic consequences of these assumptions. First, we present the simplest model, a site sampler where the total number of sites is fixed. Then we present an improved model, the Bernoulli sampler, where the total number of sites is allowed to vary. We briefly discuss extending the model to multiple motifs. Finally, we discuss relaxing the assumptions of fixed motif abundance and width.

3.2.1 *The site sampler—one site per sequence.* This algorithm, as described in Lawrence et al. (1993) and Liu (1994), is based on the following assumptions: (a) there is only one type of motif present in the sequence data (with fixed known width w); (b) there is one and only one motif site per sequence. In this case, the missing indicator array $\mathbf{A}$ can be reduced to a vector $\mathbf{A} = (a_1, \ldots, a_N)$, where a_i gives the location of the single site within sequence i. As given in Liu (1994), the collapsed distribution $p(\mathbf{A}|\mathbf{S})$ implies the following conditional distribution for the location a_i of

the single site $\mathbf{A}_i$ in sequence i, conditional on the site locations in the other sequences, $\mathbf{A}^*$:

$$p(a_i|\mathbf{A}^*, \mathbf{S}) \propto \frac{\Gamma[\mathbf{N}((\mathbf{A}^*)^c) - \mathbf{N}(\mathbf{A}_i) + \boldsymbol{\beta}_0]}{\Gamma[\mathbf{N}((\mathbf{A}^*)^c) + \boldsymbol{\beta}_0]} \cdot \prod_{j=1}^{w} \left(\mathbf{N}(\mathbf{A}^*_{(j)}) + \boldsymbol{\beta}_j\right)^{\mathbf{N}(\mathbf{A}_{i(j)})} \approx \prod_{j=1}^{w} \left(\frac{\hat{\boldsymbol{\theta}}^*_j}{\hat{\boldsymbol{\theta}}^*_0}\right)^{\mathbf{N}(S_{i,a_i+j-1})},$$

where $\hat{\boldsymbol{\theta}}^*_j$ are the posterior means of $\boldsymbol{\theta}_j$ conditional on $\mathbf{S}$ and $\mathbf{A}^*$, and $\hat{\boldsymbol{\theta}}^*_0$ are the corresponding means for the background. More precisely, as given in Lawrence et al. (1993),

$$\hat{\theta}^*_{jk} = \frac{C_{jk} + \beta_{jk}}{N - 1 + |\boldsymbol{\beta}_j|},$$

where $|\boldsymbol{\beta}_j| = \sum_{k=1}^{K} \beta_{jk}$ and C_{jk} are the counts of base type k in position j of all sites except for the site in sequence i. Thus, a_i can be randomly drawn from all possible starting points in sequence i with probability proportional to $p(a_i|\mathbf{A}^*, \mathbf{S})$ given above, in either exact or approximate form. To avoid being trapped in a phase-shift mode, they also included a Metropolis step to allow for all the motif sites to move to the left or right by a few positions. That is, a move of the type $A \to \mathbf{A} \pm \delta$ is considered.

3.2.2 *Bernoulli sampler—unknown number of motif sites.* As pointed out in Liu, Neuwald and Lawrence (1995), it is often too restrictive an assumption to hold the total number of unknown sites as fixed and known. If we allow an unknown number of motif sites per sequence, this is equivalent to allowing multiple sites in one long super sequence created by concatenating all the sequences, that is, $\mathbf{S} = (S_1, \ldots, S_{L^*})$, where L^* is the total length of all N the sequences in the dataset. Since the motif site is not allowed to overlap with the endpoints of the original sequences, we let $L = L^* - N(w - 1)$ be the adjusted total sequence length. Thus, our missing data array can be written as a long vector $\mathbf{A} = (a_1, a_2, \ldots, a_L)$ of indicator variables, where each a_i is either 1 (site) or 0 (nonsite) with a priori probability p_0 and $1 - p_0$, respectively, where p_0 is termed as the motif *abundance ratio*. Under this model, the joint posterior distribution is

$$p(\mathbf{A}, \boldsymbol{\Theta}, \boldsymbol{\theta}_0|\mathbf{S}, p_0) \propto \boldsymbol{\theta}_0^{\mathbf{N}(\mathbf{A}^c)+\boldsymbol{\beta}_0} \times \prod_{j=1}^{w} \boldsymbol{\theta}_j^{\mathbf{N}(\mathbf{A}_{(j)})+\boldsymbol{\beta}_j} p_0^{|\mathbf{A}|}(1 - p_0)^{L-|\mathbf{A}|},$$

where $|\mathbf{A}|$ is the total number of sites, now assumed to be unknown. Integrating out $\boldsymbol{\Theta}$ and $\boldsymbol{\theta}_0$, we have

$$p(\mathbf{A}|\mathbf{S}, p_0) \propto \frac{\Gamma(\mathbf{N}(\mathbf{A}^c) + \boldsymbol{\beta}_0)}{\Gamma(L - |\mathbf{A}| + |\boldsymbol{\beta}_0|)} \prod_{j=1}^{w} \frac{\Gamma(\mathbf{N}(\mathbf{A}_{(j)}) + \boldsymbol{\beta}_j)}{\Gamma(|\mathbf{A}| + |\boldsymbol{\beta}_j|)} \cdot p_0^{|\mathbf{A}|}(1 - p_0)^{L-|\mathbf{A}|}. \quad (1)$$

Based on this formula, Liu, Neuwald and Lawrence (1995) constructed a *predictive updating* algorithm based on the conditional distribution

$$\frac{p(a_i = 1|\mathbf{A}^*, \mathbf{S})}{p(a_i = 0|\mathbf{A}^*, \mathbf{S})} \propto \frac{p_0}{1 - p_0} \prod_{j=1}^{w} \left(\frac{\hat{\boldsymbol{\theta}}^*_j}{\hat{\boldsymbol{\theta}}^*_0}\right)^{\mathbf{N}(S_{i,a_i+j-1})},$$

where $\mathbf{A}^*, \hat{\boldsymbol{\theta}}^*_j, \hat{\boldsymbol{\theta}}^*_0$ are the same as in Section 3.2.1.

An immediate next question is how to find a proper abundance ratio p_0. Some earlier literature has let p_0 be in the range of $1/200$ to $1/2000$ (Liu, Neuwald and Lawrence, 1995; Neuwald, Liu and Lawrence, 1995; Roth et al., 1998). However, our empirical studies have found that the choice of p_0 can have a significant effect on the motif discovery results. This issue will be discussed in Section 3.3.

3.2.3 *Dealing with multiple motif types.* Although this situation is not the focus of this paper, it is worth mentioning that the above Bernoulli sampler model can be extended to the situation where we suspect that multiple distinct motif patterns exist in the same set of sequences. The simplest strategy is to introduce more motif matrices, one for each motif type, and to let the variable A_{ij} indicate not only the start of a motif site, but also the motif type (Liu, Neuwald and Lawrence, 1995). Another strategy is to mask out the discovered sites of the first motif and repeat the Bernoulli sampler (Roth et al., 1998).

As pointed out in Lawrence et al. (1993), searching for several patterns simultaneously permits the sharing of information between them to aid in the discovery of unknown sites of each. They present a multiple-motif version of the multinomial sampler, where the multiple motifs are restricted to have the same ordering (collinearity) between different sequences. Potential modeling of the spacing between motifs is also mentioned but not implemented. Liu, Neuwald and Lawrence (1999) mention that this early model for collinearity is computationally inefficient and propose that the models for a single motif be combined with a hidden Markov model (HMM) for insertions and deletions between different motifs. This unified

model, called the *propagation model*, capitalizes on the collinearity properties inherent to hidden Markov models but does not require the large amount of free parameters that a typical HMM would. There is the additional model selection issue (Gelman et al., 1995; Kass and Raftery, 1995) for determining the appropriate total number of different motif patterns.

More recently, Xing, Wu, Jordan and Karp (2003) presented LOGOS, a hidden Markov model for the occurrence of multiple motifs combined with a separate hierarchical Bayesian Markovian model for each different motif. Frith et al. (2003) introduce software, Cluster-Buster, which combines the information from known motif patterns to find dense clusters of motifs in genome-wide searches.

3.3 Flexible Motif Width and p_0

If we assume that the motif abundance ratio p_0 is unknown with a Beta(a, b) prior distribution, then the joint posterior distribution becomes

$$p(\mathbf{A}, \boldsymbol{\Theta}, \boldsymbol{\theta}_0, p_0|\mathbf{S}) \propto \boldsymbol{\theta}_0^{\mathbf{N}(\mathbf{A}^c)+\boldsymbol{\beta}_0} \times \prod_{j=1}^{w} \boldsymbol{\theta}_j^{\mathbf{N}(\mathbf{A}_{(j)})+\boldsymbol{\beta}_j} \cdot p_0^{|\mathbf{A}|+a-1}(1-p_0)^{L-|\mathbf{A}|+b-1}.$$

Integrating out the parameters $\boldsymbol{\Theta}$, $\boldsymbol{\theta}_0$ and p_0, we have

$$p(\mathbf{A}|\mathbf{S}) \propto \frac{\Gamma(\mathbf{N}(\mathbf{A}^c)+\boldsymbol{\beta}_0)}{\Gamma(L-|\mathbf{A}|+|\boldsymbol{\beta}_0|)} \prod_{j=1}^{w} \frac{\Gamma(\mathbf{N}(\mathbf{A}_{(j)})+\boldsymbol{\beta}_j)}{\Gamma(|\mathbf{A}|+|\boldsymbol{\beta}_j|)} \cdot B_{a,b}(|\mathbf{A}|, L-|\mathbf{A}|),$$

where $B_{a,b}(c, d)$ is the Beta function

$$\frac{\int_0^1 x^{a+c-1}(1-x)^{b+d-1}\,dx}{\int_0^1 x^{a-1}(1-x)^{b-1}\,dx}.$$

This expression can be used to construct a predictive updating algorithm similar to that based on (1).

In the Bernoulli sampler model, the assumption of fixed motif width w can be relaxed somewhat to allow so-called *fragmentation* of motifs. In a fragmentation model, only J columns of a motif of width w are selected to form the motif pattern. This is accomplished by positing additional missing indicator variables for whether or not each of the w positions of a motif is considered as part of a conserved motif pattern. This new missing data can be incorporated into a larger model and a Gibbs sampling strategy can again be used for implementation. This fragmentation model is useful for correcting the problem that earlier Gibbs sampling strategies could get stuck in local modes that were phase-shifted versions of the true signal. A slightly different approach to correcting this same phase shift problem is to insert a Metropolis step within the Gibbs sampler that shifts each motif in one direction or the other (Liu, 1994).

If we view w as an unknown variable and treat it directly, then we face a Bayesian model selection problem (Gelman et al., 1995) since, for different widths w the dimensionality of the motif parameter $\boldsymbol{\Theta}$ is different. Lawrence et al. (1993) use an ad hoc *information per parameter* criterion to select the best motif width. This criterion, however, tends to bias in favor of motifs with strong conserved sites on the two ends. Noting that $\boldsymbol{\Theta}$ can be integrated out from the model to avoid the dimensionality change, Gupta and Liu (2003) place a prior distribution on w and use a Metropolis step to update w based on the joint distribution. In summary, with the mutually independent prior distributions $\boldsymbol{\theta}_0 \sim \text{Dirichlet}(\boldsymbol{\beta}_0)$, $w \sim p(w)$, $p_0 \sim \text{Beta}(a, b)$ and $\boldsymbol{\Theta}|w \sim \text{ProductDirichlet}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_w)$, we have

$$p(\mathbf{A}, w|\mathbf{S}) \propto \frac{\Gamma(\mathbf{N}(\mathbf{A}^c)+\boldsymbol{\beta}_0)}{\Gamma(L-|\mathbf{A}|+|\boldsymbol{\beta}_0|)} \cdot \prod_{j=1}^{w} \frac{\Gamma(\mathbf{N}(\mathbf{A}_{(j)})+\boldsymbol{\beta}_j)}{\Gamma(|\mathbf{A}|+|\boldsymbol{\beta}_j|)} \frac{\Gamma(|\boldsymbol{\beta}_j|)}{\Gamma(\boldsymbol{\beta}_j)} \cdot B_{a,b}(|\mathbf{A}|, L-|\mathbf{A}|) \times p(w).$$

4. MUCH ADO ABOUT SCORING FUNCTIONS

In the frequent situation where the single "best" answer to a motif-finding problem is desired (i.e., the "best" set of site predictions or the "best" consensus matrix), our goal is to find the "optimum" of a certain scoring function. In our Bayesian formulation, an appropriate log-posterior distribution can serve our purpose. Although it is still a subject of debate (see Stormo, 2000, for a review) whether the current Bayesian formulation is the "best" one for the motif-finding problem, the methods built based on a statistical model have been shown to be more accurate in many cases than heuristic ones, such as the word enumeration techniques outlined in the second section. Because of the need for a speedy algorithm, it is sensible to seek strategies, such as optimizing a scoring function, instead of a full posterior analysis (via MCMC sampling). Here we examine a few functions that have been used in practice to evaluate a discovered motif and attempt some generalizations of them. Throughout this section we assume that the background parameter $\boldsymbol{\theta}_0$ is known.

4.1 Bayesian Scoring Functions

We begin the discussion assuming that the motif width w and the abundance ratio p_0 are known, as well as the background parameters $\boldsymbol{\theta}_0$. We also assume that the number of prior counts in each column of the motif matrix is constant, that is, $|\boldsymbol{\beta}_j| = |\boldsymbol{\beta}|$ for all j. In each scoring function, we let K be the collection of terms that are constant with respect to the unknown parameters. The first scoring function is the exact log-posterior density for $\mathbf{A}$:

$$\begin{aligned}
&\psi_{\text{exact}}(\mathbf{A}) \\
&\quad = \log p(\mathbf{A}|\boldsymbol{\theta}_0, p_0, w, \mathbf{S}) \\
&\quad = K + |\mathbf{A}| \operatorname{logit}(p_0) - w \log \Gamma(|\mathbf{A}| + |\boldsymbol{\beta}|) \\
&\qquad + \sum_{j=1}^{w} \sum_{k} \log \Gamma(n_{jk} + \beta_{jk}) - n_{jk} \log \theta_{0k}.
\end{aligned} \tag{2}$$

When p_0 is unknown and is assigned a prior distribution Beta(a, b), we have

$$\begin{aligned}
\psi'_{\text{exact}}(\mathbf{A}) &= K + \log B_{a,b}(|\mathbf{A}|, L - |\mathbf{A}|) \\
&\quad - w \log \Gamma(|\mathbf{A}| + |\boldsymbol{\beta}|) \\
&\quad + \sum_{j=1}^{w} \sum_{k} \log \Gamma(n_{jk} + \beta_{jk}) - n_{jk} \log \theta_{0k}.
\end{aligned}$$

Here $L = N - (w-1)m$, where N is the total number of nucleotides and m is the number of sequences. L is the total number of possible site positions, since sites are not allowed to overlap the ends of a sequence. Using Stirling's formula (Stirling, 1730), $\Gamma(x+1) = x! \approx x^x e^{-x} (2\pi x)^{1/2}$, we can approximate ψ_{exact} as

$$\begin{aligned}
&\psi_{\text{Stir}}(\mathbf{A}) \\
&\quad = K + |\mathbf{A}| \operatorname{logit}(p_0) - \frac{3}{2} w \log(|\mathbf{A}| + |\boldsymbol{\beta}| - 1) \\
&\qquad + \sum_{j=1}^{w} \sum_{k} \left(n_{jk} + \beta_{jk} - \frac{1}{2} \right) \\
&\qquad\qquad \times \log \left(\frac{n_{jk} + \beta_{jk} - 1}{|\mathbf{A}| + |\boldsymbol{\beta}| - 1} \right) - n_{jk} \log \theta_{0k} \\
&\quad \approx K + |\mathbf{A}| \left[\operatorname{logit}(p_0) + \sum_{j=1}^{w} \sum_{k} \hat{\theta}_{jk} \log \left(\frac{\hat{\theta}_{jk}}{\theta_{0k}} \right) \right] \\
&\qquad - \frac{3}{2} w \log(|\mathbf{A}| + |\boldsymbol{\beta}| - 1),
\end{aligned} \tag{3}$$

where $\hat{\theta}_{jk} = \frac{n_{jk} + \beta_{jk}}{|\mathbf{A}| + |\boldsymbol{\beta}|}$. Our empirical results showed that the Stirling approximation tracks ψ_{exact} very well for realistic values of $|\mathbf{A}|$ and n_{jk}. When p_0 is assigned a Beta$(1, 1)$ prior, we can again use the Stirling formula to approximate $\log[B_{1,1}(|\mathbf{A}|, L - |\mathbf{A}|)]$ so that

$$\begin{aligned}
\psi'_{\text{Stir}}(\mathbf{A}) &\approx K + |\mathbf{A}| \left[\operatorname{logit}(\hat{p}_0) - 1 \right. \\
&\qquad \left. + \sum_{j=1}^{w} \sum_{k} \hat{\theta}_{jk} \log \left(\frac{\hat{\theta}_{jk}}{\theta_{0k}} \right) \right] \\
&\quad - \frac{3}{2} w \log(|\mathbf{A}| + |\boldsymbol{\beta}| - 1),
\end{aligned}$$

where $\hat{p}_0 = |\mathbf{A}|/L$ is the estimated motif abundance ratio.

Furthermore, we can consider w as unknown with prior $p(w)$, which will give us several extra terms in the scoring function for our exact log-posterior density,

$$\begin{aligned}
&\psi''_{\text{exact}}(\mathbf{A}, w) \\
&\quad = K + \log B_{1,1}(|\mathbf{A}|, L - |\mathbf{A}|) \\
&\qquad + \log p(w) - w \log \left(\frac{\Gamma(|\mathbf{A}| + |\boldsymbol{\beta}|)}{\Gamma(|\boldsymbol{\beta}|)} \right) \\
&\qquad + \sum_{j=1}^{w} \sum_{k} \log \left(\frac{\Gamma(n_{jk} + \beta_{jk})}{\Gamma(\beta_{jk})} \right) - n_{jk} \log \theta_{0k},
\end{aligned}$$

and the corresponding Stirling approximation,

$$\begin{aligned}
&\psi''_{\text{Stir}}(\mathbf{A}, w) \\
&\quad \approx K + \log p(w) \\
&\qquad + |\mathbf{A}| \left[\operatorname{logit}(\hat{p}_0) - 1 + \sum_{j=1}^{w} \sum_{k} \hat{\theta}_{jk} \log \left(\frac{\hat{\theta}_{jk}}{\theta_{0k}} \right) \right] \\
&\qquad - \sum_{j=1}^{w} \sum_{k} \left(\beta_{jk} - \frac{1}{2} \right) \log \left(\frac{\beta_{jk} - 1}{|\boldsymbol{\beta}| - 1} \right) \\
&\qquad - \frac{3}{2} w \log \left(\frac{|\mathbf{A}| + |\boldsymbol{\beta}| - 1}{|\boldsymbol{\beta}| - 1} \right).
\end{aligned}$$

A natural prior distribution for w would be the Poisson(w_0), where w_0 represents our a priori expectation for the motif width. One could also consider other prior distributions for w, such as Geometric(w_0) or Exponential(w_0).

Another scoring function approximation that we can consider is based on the entropy distance between the motif and background parameters (or Kullback–Leibler information),

$$\begin{aligned}
&\psi_{\text{ent}}(\mathbf{A}) \\
&\quad = |\mathbf{A}| \left[\operatorname{logit}(p_0) + \sum_{j} \sum_{k} \hat{\theta}_{jk} \log \left(\frac{\hat{\theta}_{jk}}{\theta_{0k}} \right) \right].
\end{aligned} \tag{4}$$

Compared with this heuristic-based scoring function, ψ_{Stir} has an additional term, which gives an additional penalty to a large number of motif sites.

4.2 Non-Bayesian Scoring Functions

A form similar to the entropy scoring function is mentioned in Lawrence et al. (1993). It is interesting to note that scoring functions related to this entropy approximation have arisen in the motif-finding literature outside the context of a Bayesian formulation. In developing their CONSENSUS algorithm, Stormo and Hartzell (1989) introduced a scoring function very similar to ψ_{ent} which they call the *information content*:

$$\psi_{\text{info}}(\mathbf{A}) = \sum_j \sum_k \hat{\theta}_{jk} \log \frac{\hat{\theta}_{jk}}{\theta_{0k}}, \quad \text{where} \tag{5}$$
$$\hat{\theta}_{jk} = \frac{n_{jk}}{N}.$$

This function is equivalent to all the foregoing scoring functions when the total number of motif sites $|\mathbf{A}|$ and the motif width w are assumed known, which was the case in Stormo and Hartzell (1989), Lawrence and Reilly (1990) and Lawrence et al. (1993). However, when $|\mathbf{A}|$ is unknown, function ψ_{info} cannot be used to find a proper set of motif sites—it will converge to a set of very few motif sites with high conservation and ignore potential sites that are less conserved. A way to remedy this is to give a prior distribution $f(\mathbf{A})$ and then construct

$$\psi'_{\text{info}}(\mathbf{A}) = \log f(\mathbf{A}) + |\mathbf{A}| \sum_j \sum_k \hat{\theta}_{jk} \log \frac{\hat{\theta}_{jk}}{\theta_{0k}}.$$

This scoring function is nearly equivalent to the entropy one we have shown earlier except that a more flexible prior of $\mathbf{A}$ is allowed here. A temptation here is to use a prior on $|\mathbf{A}|$ directly, but this overlooks the "entropy number," that is, the number of different $\mathbf{A}$'s that can give rise to the same value of $|\mathbf{A}|$.

Liu, Brutlag and Liu (2002) present an algorithm called MDScan for motif-finding based not only on sequence data but also on gene expression information from microarray experiments. Since the true p_0 is rarely known in practice, they propose to optimize the following scoring function:

$$\psi_{\text{md}}(\mathbf{A}) = \frac{\log(|\mathbf{A}|)}{w} \sum_j \sum_k \hat{\theta}_{jk} \log \frac{\hat{\theta}_{jk}}{\theta_{0k}}. \tag{6}$$

The functional form again shares some similarities with the entropy approximation given above. Although function ψ_{md} is not intended as an approximation to the posterior distribution $p(\mathbf{A}|\boldsymbol{\theta}_0, p_0, \mathbf{S})$, it can still be used as a scoring function in the optimization algorithm presented below.

4.3 Optimizing a Scoring Function

One disadvantage of the Gibbs sampling method described in Section 3.2 is that it typically takes a longer time than a researcher (especially biologists) can tolerate. It is also much more involved to summarize the results using the posterior samples. Even more seriously, different Gibbs sampling chains with different starting values often get stuck in different modes, due to both the "stickiness" of the posterior distribution surface and the limitation of computing power. Here, we seek to achieve a simpler goal: optimizing one of the scoring functions described above by using a Metropolis-algorithm-based annealing approach.

In the Metropolis steps, we systematically scan through every element of the matrix $\mathbf{A}$ and decide whether the indicator variable at this position should be "changed" to its opposite value. If we denote $\mathbf{A}'$ as $\mathbf{A}$ with this change made, then we calculate the following Metropolis ratio:

$$r = \min\{1, \exp\{\psi(\mathbf{A}') - \psi(\mathbf{A})\}/T\}.$$

The decision to accept the change or to keep $\mathbf{A}$ unchanged is made with probability r and $1 - r$, respectively. The scoring function ψ can be taken as any of the scores discussed earlier in this section. The parameter T is called the *temperature* of the algorithm, with low temperatures restricting the algorithm to accept only small jumps and high temperatures allowing for more freedom to move around the parameter space. We consider the following optimization strategies.

The *Temperature* $= 0$ strategy forces the algorithm to accept only changes that immediately improve the score, since forcing T to approach 0 then forces r to equal 0 if $\psi(\mathbf{A}') < \psi(\mathbf{A})$ or r to equal 1 if $\psi(\mathbf{A}') \geq \psi(\mathbf{A})$. With this type of deterministic strategy, it is important that we start the algorithm in an area near the mode of the density, or else our simple hill-climbing algorithm is guaranteed to get stuck in an inferior local mode. Therefore, one would first want to run the dataset through a sensitive program such as BioProspector (Liu, Brutlag and Liu, 2001), which would give a set of predicted sites that is near the area of high posterior density, and then use these predicted sites as the starting point of a $T = 0$ optimization algorithm. In this scenario, our optimization strategy is intended to "clean up" the output produced by

a stochastic-based algorithm such as BioProspector, Consensus or AlignACE.

The *Temperature* $= 1$ strategy is equivalent to sampling from the posterior distribution, if the score function is the exact log-posterior. However, for other types of score functions this approach imposes a target density on the parameter space, which may or may not be desirable. One can run this algorithm over many iterations and analyze the Monte Carlo samples thus obtained. We did not implement this strategy because of an overlap of the effort with previous approaches such as Gibbs Motif Sampler, AlignACE and BioProspector. A *simulated annealing* (Kirkpatrick, Gelatt and Vecchi, 1983) strategy combines deterministic and stochastic strategies by starting the algorithm at a high temperature such as $T = 4$ and then slowly decreasing the temperature to $T = 0$ as the algorithm continues through many iterations through all positions of $\mathbf{A}$. For the current exposition, we restrict ourselves to the modest goal of the $T = 0$ strategy, that is, deterministic improvement upon the output from Gibbs sampling algorithms such as BioProspector.

5. EMPIRICAL STUDIES

5.1 Effect of Putting a Prior on p_0

Earlier methods such as the Gibbs Motif Sampler and AlignACE (Liu, Neuwald and Lawrence, 1995; Liu, Brutlag and Liu, 2001; Roth et al., 1998) use a fixed motif abundance ratio p_0. However, some of our recent studies (Liu, Brutlag and Liu, 2002) suggest that this abundance ratio, if not given properly, may have adverse effect on the accuracy (in terms of finding true sites) of the findings. We also confirmed this finding by some simple simulation experiments. To circumvent this problem, Liu, Brutlag and Liu (2001, 2002) proposed to optimize a slightly different scoring function ψ_{md} as shown in (6). Here we investigate the advantage of treating p_0 as an unknown parameter in a full Bayesian formulation.

Table 3 shows the results of a simulation in which 20 sequences of 500 base pairs each were generated according to a first-order Markov model. In each sequence a motif site of width 10 was inserted with motif strength 0.9 (i.e., the most frequent letter is 90% conserved). Different starting values for p_0 ranging from $1/100$ to $1/2000$ were tested and p_0 was then updated in the Gibbs sampler iterations. The results are compared to those from methods using fixed p_0.

It is seen from the table that, when the fixed p_0 is large, we tend to get many false positive sites; whereas when p_0 is small, we tend to pick up fewer true sites, leading to more false negatives. The optimal value of p_0 is around 500 to 1,000. However, in the results where we treat p_0 as an unknown variable and update it along with the Gibbs sampling iterations, the performance was quite stable and invariant to the starting values. With different starting values for p_0, we ended up with approximately the same number of false positive and false negative sites, comparable to the results from using a fixed p_0 at its optimal value. The best draw of p_0, in the sense of maximizing the joint posterior distribution, is close to the true value of p_0 $(1/500)$.

TABLE 3
FN is the number of false negative sites; FP is the number of false positive sites; $K = (1 - p_0)/p_0$ and K^* is the best draw of K

	Starting values for K				
Strategy	**100**	**200**	**500**	**1,000**	**2,000**
Fixed p_0					
FN	3	4	4	5	7
FP	49	18	5	3	3
Sample p_0					
K^*	452	405	550	399	497
FN	3	4	4	4	5
FP	4	6	4	5	4

5.2 Comparison of Scoring Functions

In Section 4 we outlined a few scoring functions that could be used in a motif-finding algorithm: the exact log-posterior as in (2), its Stirling approximation as in (3), its entropy approximation as in (4), the scoring function (6) used by the MDscan (Liu, Brutlag and Liu, 2002) and the information-content function (5) used by CONSENSUS. We designed the following simulation study to investigate the relative ability of each scoring function to find unknown motif sites under various sequence conditions.

Since ψ_{info} is only suitable for the case in which the number of sites is known, we only compared the effectiveness of the first four scoring functions. We include the MDscan scoring function here since we are interested in evaluating its performance against the other scoring functions, though it not an approximation to our posterior distribution.

Each simulated dataset consisted of 20 sequences of 200 base pairs each, with each sequence containing exactly one true motif. Datasets were generated multiple (200) times under each combination of several conditions. The first condition was the length of the hidden

motif, either 8 or 16 base pairs. The second condition was the degree of conservation of the hidden motif signal, either high conservation or low conservation. High conservation means that each motif position had a dominant nucleotide with 91% probability (all others 3% equally). Low conservation means that each motif position had a dominant nucleotide with 70% probability (all others 10% equally).

We tested the effect of the $T = 0$ strategy for improving the results from BioProspector. BioProspector was run on each dataset and the best motif result was retained. We then applied our optimization algorithm, based upon each of the four scoring functions mentioned above, to this best BioProspector result. The motif result from each optimization algorithm was also retained after the optimization algorithm had converged.

We also compared the effects of the prior distribution on $\boldsymbol{\Theta}$ by using two different sizes of pseudocounts, $\beta_{jk} = 2$ versus $\beta_{jk} = 1.1$. This comparison will affect the three scoring functions derived from our complete Bayesian model, but will not affect ψ_{md} since no prior distribution was involved in its derivation.

Table 4 gives the accuracy of the results from algorithms using each of the four scoring functions. Accuracy is measured by two statistics, the percentage of correct sites found and how close the motif consensus found matches the true motif consensus.

The first conclusion we can reach is that the $T = 0$ strategy seems to improve the accuracy of the predicted sites in comparison with the BioProspector result. Regardless of motif width or conservation, the "accuracy of predicted sites" is higher for each scoring function compared to the BioProspector output, except in the case of a short motif and low conservation, where no method seems to work. The results are not as dramatic for the consensus match, suggesting that the scoring function optimization is primarily refining the signal that has already been found by the Gibbs sampling-based BioProspector. Thus, it seems that this $T = 0$ strategy has accomplished its intended goal of "cleaning up" the BioProspector output.

In general, the algorithms do not do nearly as well for low conservation as for high conservation, especially in the case of the shorter motif. This is partly due to the fact that the $T = 0$ strategy is deterministically restricted to stay in the same local

TABLE 4
Simulation results for $T = 0$ strategy

Prior	Motif width	Conservation	BioProspector results	Optimization results using scoring function			
				Exact	Stirling	Entropy	MDscan
			Accuracy of predicted sites (average |A|)				
1.1	8	91	79 (18)	80 (18)	81 (19)	81 (20)	80 (18)
2	8	91	79 (18)	80 (18)	80 (18)	67 (15)	80 (18)
1.1	8	70	9 (15)	8 (8)	10 (11)	3 (2)	12 (19)
2	8	70	9 (15)	1 (0)	1 (0)	0 (0)	12 (19)
1.1	16	91	85 (17)	91 (19)	91 (20)	91 (23)	80 (16)
2	16	91	84 (17)	91 (20)	91 (20)	91 (24)	80 (16)
1.1	16	70	41 (11)	51 (14)	59 (17)	62 (20)	43 (11)
2	16	70	41 (11)	51 (13)	54 (14)	41 (10)	43 (11)
			Consensus match (average |A|)				
1.1	8	91	98 (18)	98 (18)	98 (19)	98 (20)	98 (18)
2	8	91	98 (18)	98 (18)	98 (18)	82 (15)	98 (18)
1.1	8	70	22 (15)	18 (8)	22 (11)	10 (2)	26 (19)
2	8	70	22 (15)	6 (0)	6 (0)	2 (0)	26 (19)
1.1	16	91	100 (17)	100 (19)	100 (20)	100 (23)	100 (16)
2	16	91	100 (17)	100 (20)	100 (20)	100 (24)	100 (16)
1.1	16	70	86 (11)	88 (14)	90 (17)	88 (20)	88 (11)
2	16	70	86 (11)	86 (13)	88 (14)	62 (10)	88 (11)

NOTES: "Accuracy of predicted sites" is the percentage of true sites found in each simulated dataset, averaged over all simulated datasets. Shifting of up to 3 base pairs was allowed. "Consensus match" is the proportion of datasets where the consensus found matches the true consensus (up to 2 mismatched or shifted letters allowed when $w = 8$, and 4 allowed when $w = 16$). The average number of predicted sites is given in parentheses.

mode that the BioProspector output is stuck in, and so these algorithms do not have the freedom to correct a poor starting point.

For the low conservation datasets, performance is much better for a longer motif than for a shorter motif, suggesting that a certain threshold of information is needed for the Gibbs sampling algorithm BioProspector, and consequently our optimization algorithm, to be successful. If conservation is reduced, one needs a longer motif for the algorithms to do well. In the case of a short motif and low conservation, extra information (such as prior information about the motif locations or $\boldsymbol{\Theta}$) is clearly needed.

The exact, Stirling and entropy scoring functions display similar performance in most situations, although the entropy scoring function appears to do noticeably worse in some cases with larger prior pseudocounts and is in general most affected by a change in prior pseudocounts.

MDscan in general does not perform as well as the three Bayesian scores, except in the case where the signal is very weak (low conservation and short motif). This may be because in the case of a really weak signal, the priors used for the Bayesian scores swamp the weak signal so that it cannot be detected. This is also shown by the slightly improved performance in Table 4 when the prior pseudo-counts are smaller. However, in situations where prior information is actually available, the formal use of a prior distribution will allow us to incorporate that information properly.

Overall, these simulation results for the predicted sites suggest that there is almost always a benefit associated with using a deterministic optimization algorithm to further improve the output from a stochastic algorithm such as BioProspector, and that this benefit seems generally to be the greatest when using the exact scoring function or one of its approximations, in terms of a reasonable number of predicted sites and the accuracy of those sites. The additional computational cost of the optimization algorithm is small (≈ 2 minutes for each simulated dataset).

5.3 Application to Cyclic-AMP Receptor Protein Motif Sites

We examine a dataset consisting of 18 sequences that contain cyclic-AMP receptor protein (CRP) binding sites. Each sequence is 105 base pairs long and each contains at least one 22-base-pair motif site that has been experimentally determined via the footprinting method (Lawrence and Reilly, 1990). This dataset has been previously analyzed by Lawrence and Reilly (1990) using an EM algorithm and by Liu (1994) using a Gibbs sampler.

Similar to our strategy with the simulated datasets, we first used the program BioProspector to find a set of initial motif sites and then used our $T = 0$ optimization strategy with one of the four scoring functions to further improve the BioProspector result. For the first three scoring functions prior pseudocounts of $\beta_{jk} = 1.1$ were used.

Table 5 shows the results from these optimization algorithms, in terms of the consensus sequence for the motif, the number of sites predicted and the number of predicted sites that corresponded to one of the 24 experimentally established ("correct") positions of the CRP binding sites. These results are similar to the ones from our simulation study. For each scoring function the optimization algorithm improved upon the original BioProspector signal in terms of the number of correct sites predicted.

As shown in Table 5, the consensus sequences of the motifs found by using different scoring functions are similar. The three scoring functions (exact, Stirling and entropy) that are closely related to the complete Bayesian model seem to perform noticeably better than the MDscan score, with the Stirling scoring function performing the best in this example. As a comparison, the "true" motif based on the alignment of the 24

TABLE 5
Results from optimization algorithms with each scoring function applied to the BioProspector output; the number of experimentally determined binding sites is 24

Scoring function	Consensus sequence	Number of predicted sites	Number of correct sites
BioProspector	`ttatttgatcgaggtcacactt`	9	9
Exact	`ttatgtgaacgagttcacattt`	15	15
Stirling	`ttttgtgatcgagctcacattt`	18	18
Entropy	`taatgtgatcgaggtcacattt`	20	17
MDscan	`ttatgtgaacgaggtcacactt`	11	11

FIG. 1. *Sequence logo of the CRP binding motif based on the alignment of* 24 *experimentally determined sites. The height of each position is equal to its information content and the size of each letter is proportional to the letter's relative frequency.*

experimental sites is displayed in Figure 1 in the form of a sequence logo. It is seen that the differing positions of the five consensus sequences in Table 5 correspond to the information-weak or ambiguous positions shown in the sequence logo.

6. DISCUSSION

Motif discovery is an important problem in computational biology since the binding of transcription factors to upstream region motifs is crucial to the mechanism of gene regulation. We have presented various techniques used in the past for motif discovery, a set of Bayesian models useful for developing motif-finding tools and generalizations of these models that allow for unknown motif width w and unknown motif abundance ratio p_0. We have also discussed the use of scoring functions for motif finding. Viewing Bayesian models in terms of scoring functions has provided insight to the similarities between the full Bayesian model-based approaches and some non-Bayesian methods, such as CONSENSUS (Stormo and Hartzell, 1989). We observed that an annealing optimization process can further improve the results obtained from the usual Gibbs sampling implementation, such as the program BioProspector, and the best results were obtained from the scoring functions that most closely approximated a true posterior distribution.

There are still many interesting open problems in this field. The vast majority of motif-finding research has assumed that all information about the interaction between transcription factors and their DNA binding motifs can be summarized just by looking at the one-dimensional nucleotide sequence. Benos, Lapedes and Stormo (2002, b) discuss one-dimensional nucleotide models and conclude that, although their fit is not perfect, they do provide a very good approximation to the true nature of protein–DNA interactions. However, in actuality this interaction is occurring in three-dimensional space, so ideally motif models should incorporate characteristics of DNA morphology. As an example, in eukaryotic organisms, DNA is stored in the form of tightly compacted chromosomes where substantial portions of the DNA sequence are wrapped around proteins called histones. This is important information to include in future models, since portions of the sequence that are wrapped around histones are less free to interact with DNA-binding proteins such as transcription factors.

Another interesting problem is to establish a model-based approach for incorporating gene expression information, such as microarray results, into the motif discovery problem. The MDscan program mentioned above gives one approach to this problem, since the upstream regions that are examined for motifs are updated in an iterative fashion, based on microarray information. A more recent method, Motif Regressor (Conlon, Liu, Lieb and Liu, 2003), directly uses the microarray expression values to help screen out false positive findings of MDscan. However, model-based approaches may still be desirable since these models may provide us a principled way to tune relevant parameters and guide us to achieve the optimal combination of the two sources of information (i.e., genome sequences and microarray values).

ACKNOWLEDGMENT

Research supported in part by NSF Grant DMS-02-04674 and National Institutes of Health Grant R01 HG02518-01.

REFERENCES

BAILEY, T. L. and ELKAN, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proc. Second International Conference on Intelligent Systems for Molecular Biology* 28–36. AAAI Press, Menlo Park, CA.

BENOS, P. V., BULYK, M. L. and STORMO, G. D. (2002). Additivity in protein–DNA interactions: How good an approximation is it? *Nucleic Acids Res.* **30** 4442–4451.

BENOS, P. V., LAPEDES, A. S. and STORMO, G. D. (2002). Probabilistic code for DNA recognition by proteins of the EGR family. *J. Molecular Biol.* **323** 701–727.

BENSON, D. A., KARSCH-MIZRACHI, I., LIPMAN, D. J., OSTELL, J., RAPP, B. A. and WHEELER, D. L. (2002). GenBank. *Nucleic Acids Res.* **30** 17–20.

BRAZMA, A., JONASSEN, I., VILO, J. and UKKONEN, E. (1998). Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.* **8** 1202–1215.

BUCK, M. J. and LIEB, J. D. (2004). ChIP-chip: Considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **83** 349–360.

BUSSEMAKER, H. J., LI, H. and SIGGIA, E. D. (2000). Building a dictionary for genomes: Identification of presumptive regulatory sites by statistical analysis. *Proc. Natl. Acad. Sci. U.S.A.* **97** 10,096–10,100.

CARDON, L. R. and STORMO, G. D. (1992). Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. *J. Molecular Biol.* **223** 159–170.

CONLON, E. M., LIU, X. S., LIEB, J. D. and LIU, J. S. (2003). Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl. Acad. Sci. U.S.A.* **100** 3339–3344.

EISEN, M. B., SPELLMAN, P. T., BROWN, P. O. and BOTSTEIN, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* **95** 14,863–14,868.

FRITH, M. C., LI, M. C. and WENG, Z. (2003). Cluster–Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.* **31** 3666–3668.

GALAS, D. J., EGGERT, M. and WATERMAN, M. S. (1985). Rigorous pattern-recognition methods for DNA sequences. Analysis of promoter sequences from *Escherichia coli*. *J. Molecular Biol.* **186** 117–128.

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (1995). *Bayesian Data Analysis*. CRC, Boca Raton, FL.

GRUNDY, W. N., BAILEY, T. L. and ELKAN, C. P. (1996). ParaMEME: A parallel implementation and a web interface for a DNA and protein motif discovery tool. *Computer Applications in the Biosciences* **12** 303–310.

GUPTA, M. and LIU, J. S. (2003). Discovery of conserved sequence patterns using a stochastic dictionary model. *J. Amer. Statist. Assoc.* **98** 55–66.

HAMPSON, S., BALDI, P., KIBLER, D. and SANDMEYER, S. B. (2000). Analysis of yeast's ORF upstream regions by parallel processing, microarrays, and computational methods. In *Proc. Eighth International Conference on Intelligent Systems for Molecular Biology* 190–201. AAAI Press, Menlo Park, CA.

HERTZ, G. Z. and STORMO, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15** 563–577.

INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM (2001). Initial sequencing and analysis of the human genome. *Nature* **409** 860–921.

IUPAC, Nomenclature Committee of the International Union of Biochemistry (NC-IUB) (1986). Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. *Proc. Natl. Acad. Sci. U.S.A.* **83** 4–8.

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.

KEICH, U. and PEVZNER, P. A. (2002). Finding motifs in the twilight zone. *Bioinformatics* **18** 1374–1381.

KIRKPATRICK, S., GELATT, C. D. and VECCHI, M. P. (1983). Optimization by simulated annealing. *Science* **220** 671–680.

LAWRENCE, C. E., ALTSCHUL, S. F., BOGUSKI, M. S., LIU, J. S., NEUWALD, A. F. and WOOTTON, J. C. (1993). Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science* **262** 208–214.

LAWRENCE, C. E. and REILLY, A. A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins* **7** 41–51.

LIU, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Amer. Statist. Assoc.* **89** 958–966.

LIU, J. S., NEUWALD, A. F. and LAWRENCE, C. E. (1995). Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Amer. Statist. Assoc.* **90** 1156–1170.

LIU, J. S., NEUWALD, A. F. and LAWRENCE, C. E. (1999). Markovian structures in biological sequence alignments. *J. Amer. Statist. Assoc.* **94** 1–15.

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

LIU, X. S., BRUTLAG, D. L. and LIU, J. S. (2002). An algorithm for finding protein–DNA binding sites with applications to chromatin immunoprecipitation microarray experiments. *Nature Biotechnology* **20** 835–839.

LODISH, H., BALTIMORE, D., BERK, A., ZIPURSKY, S. L., MATSUDAIRA, P. and DARNELL, J. (1995). Regulation of transcription initiation. In *Molecular Cell Biology*, 3rd ed. (J. Darnell, H. Lodish and D. Baltimore, eds.) 405–481. Scientific American Books, New York.

MCCUE, L. A., THOMPSON, W., CARMACK, C. S., RYAN, M. P., LIU, J. S., DERBYSHIRE, V. and LAWRENCE, C. E. (2001). Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.* **29** 774–782.

NEUWALD, A. F., LIU, J. S. and LAWRENCE, C. E. (1995). Gibbs motif sampling: Detection of bacterial outer membrane protein repeats. *Protein Science* **4** 1618–1632.

PFAHL, M. (1981). Characteristics of tight-binding repressors of the lac operon. *J. Molecular Biol.* **147** 1–10.

ROTH, F. P., HUGHES, J. D., ESTEP, P. W. and CHURCH, G. M. (1998). Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology* **16** 939–945.

SCHENA, M., SHALON, D., DAVIS, R. W. and BROWN, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270** 467–470.

SINHA, S. and TOMPA, M. (2000). A statistical method for finding transcription factor binding sites. In *Proc. Eighth International Conference on Intelligent Systems for Molecular Biology* 344–354. AAAI Press, Menlo Park, CA.

STIRLING, J. (1730). *Methodus Differentialis*. London.

STORMO, G. D. (2000). DNA binding sites: Representation and discovery. *Bioinformatics* **16** 16–23.

STORMO, G. D. and HARTZELL, G. W. (1989). Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. U.S.A.* **86** 1183–1187.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82** 528–550.

VAN HELDEN, J., ANDRE, B. and COLLADO-VIDES, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Molecular Biol.* **281** 827–842.

VAN HELDEN, J., RIOS, A. F. and COLLADO-VIDES, J. (2000). Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.* **28** 1808–1818.

VELCULESCU, V. E., ZHANG, L., VOGELSTEIN, B. and KINZLER, K. W. (1995). Serial analysis of gene expression. *Science* **270** 484–487.

WERNER, T. (1999). Models for prediction and recognition of eukaryotic promoters. *Mammalian Genome* **10** 168–175.

XING, E. P., WU, W., JORDAN, M. I. and KARP, R. M. (2003). LOGOS: A modular Bayesian model for de novo motif detection. IEEE Computer Society Bioinformatics Conference, CSB2003.