Bayesian Models for Multiple Local Sequence Alignment and Gibbs Sampling Strategies

Jun S. LIU, Andrew F. NEUWALD, and Charles E. LAWRENCE

A wealth of data concerning life's basic molecules, proteins and nucleic acids, has emerged from the biotechnology revolution. The human genome project has accelerated the growth of these data. Multiple observations of homologous protein or nucleic acid sequences from different organisms are often available. But because mutations and sequence errors misalign these data, multiple sequence alignment has become an essential and valuable tool for understanding structures and functions of these molecules. A recently developed Gibbs sampling algorithm has been applied with substantial advantage in this setting. In this article we develop a full Bayesian foundation for this algorithm and present extensions that permit relaxation of two important restrictions. We also present a rank test for the assessment of the significance of multiple sequence alignment. As an example, we study the set of dinucleotide binding proteins and predict binding segments for dozens of its members.

KEY WORDS: Bernoulli sampling; Dinucleotide binding; Dirichlet distribution; Fragmentation; Gibbs sampling; Metropolis algorithm; Product Multinomial; Ranks test.

1. INTRODUCTION

1.1 Background

The linear biopolymers—DNA, RNA, and proteins—are the three central molecular building blocks of life. DNA is an information storage molecule. RNA has a wide variety of roles, including a small but important set of functions. Because RNA can play so many roles, it is believed to be the molecule from which life began. Proteins are the action molecules of life, responsible for nearly all of the functions of all living beings and forming many of life's structures. The diversity of functions performed by proteins is extraordinarily broad. Although the methods we describe here have been successfully applied to all the three biopolymers, to date they have been most widely used for the characterization of proteins, the focus of the application described here.

A protein is an unbranched, linear, heterogeneous polymer composed of a sequence of 20 types of amino acids. An amino acid is composed of a peptide and a side chain residue. The peptide is identical in all but one of the amino acids, proline; however, all 20 side chain residues are unique. As shown in Figure 1, the peptides link together to form the peptide backbone chain, with the side chain residues attached at regular intervals. Thus a protein may be specified in a one-dimensional representation by giving the sequence of its residues. Table 1 gives the sequence of the first four of the 91 proteins in our example.

Most proteins fold to a unique densely packed three-dimensional shape. Because of the dense packing, a figure that includes all the atoms in a protein is extremely cluttered. In Figure 2 we present only the trace of the backbone chain to illustrate the three-dimensional fold of the first protein in Table 1: horse alcohol dehydrogenase (ADHE).

Probably the most important and impressive job performed by proteins is the catalysis of biochemical reactions. Enzymes, which are complexes of one or more proteins, are extremely efficient catalysts, with efficiencies several orders of magnitude greater than the best chemical catalysts. The fact that chains of 20 simple molecules can carry out such a broad spectrum of catalytic function so efficiently makes them arguably the most impressive system in all of the natural world. In large part enzymes owe their impressive catalytic efficiency to the exacting three-dimensional structures to which they spontaneously fold.

Enzymes often recruit other molecules, called cofactors, to assist in their catalytic action and perform their catalytic function through the precise interaction of the enzyme and perhaps a cofactor with the molecule to be chemically transformed, its substrate. This interaction requires the enzyme to bind both the substrate and the cofactor. Many cofactors are derivatives of vitamins and typically contain several chemical groups. In the example used here, three cofactors are involved: nicotinamide adenine dinucleotide (NAD), nicotinamide adenine dinucleotide phosphate (NADP), and flavin adenine dinucleotide (FAD), which are derived from niacin, niacin, and riboflavin. All three cofactors contain two phosphate groups, composed of a phosphorous atom and four oxygen atoms. In the example we focus on the binding of the enzyme to the phosphate groups in these cofactors. In Figure 2 the bound cofactor, NAD, is shown in orange. A magnified view in the vicinity of NAD is shown in Figure 3.

The critical interactions between an enzyme and a substrate, cofactor, or other bond molecule involve only a limited subset of its residues. These critical residues occur in a few segments of the protein chain. In the example,

Jun S. Liu is Assistant Professor, Department of Statistics, Stanford University, Stanford, CA 94305. Andrew F. Neuwald is Postdoctoral Fellow, National Center for Biotechnology Information, NLM-NIH, Bethesda, MD 20894. Charles E. Lawrence is Chief, Biometrics Lab, Wadsworth Center for Laboratories and Research, NYS-DOH, Albany, NY 12201, and a visiting Senior Scientist, National Center for Biotechnology Information, NLM-NIH. Part of the manuscript was prepared when the first author was Assistant Professor, Department of Statistics, Harvard University. This work was partially supported by National Institute of Health Grant R01 HG01257-01, National Science Foundation Grants DMS-9404344 and DMS9501570, and the William F. Milton Fund. The authors thank the editor, the associate editor, and two referees for numerous constructive comments that helped improve the article.

Journal of the American Statistical Association
December 1995, Vol. 90, No. 432, Applications & Case Studies

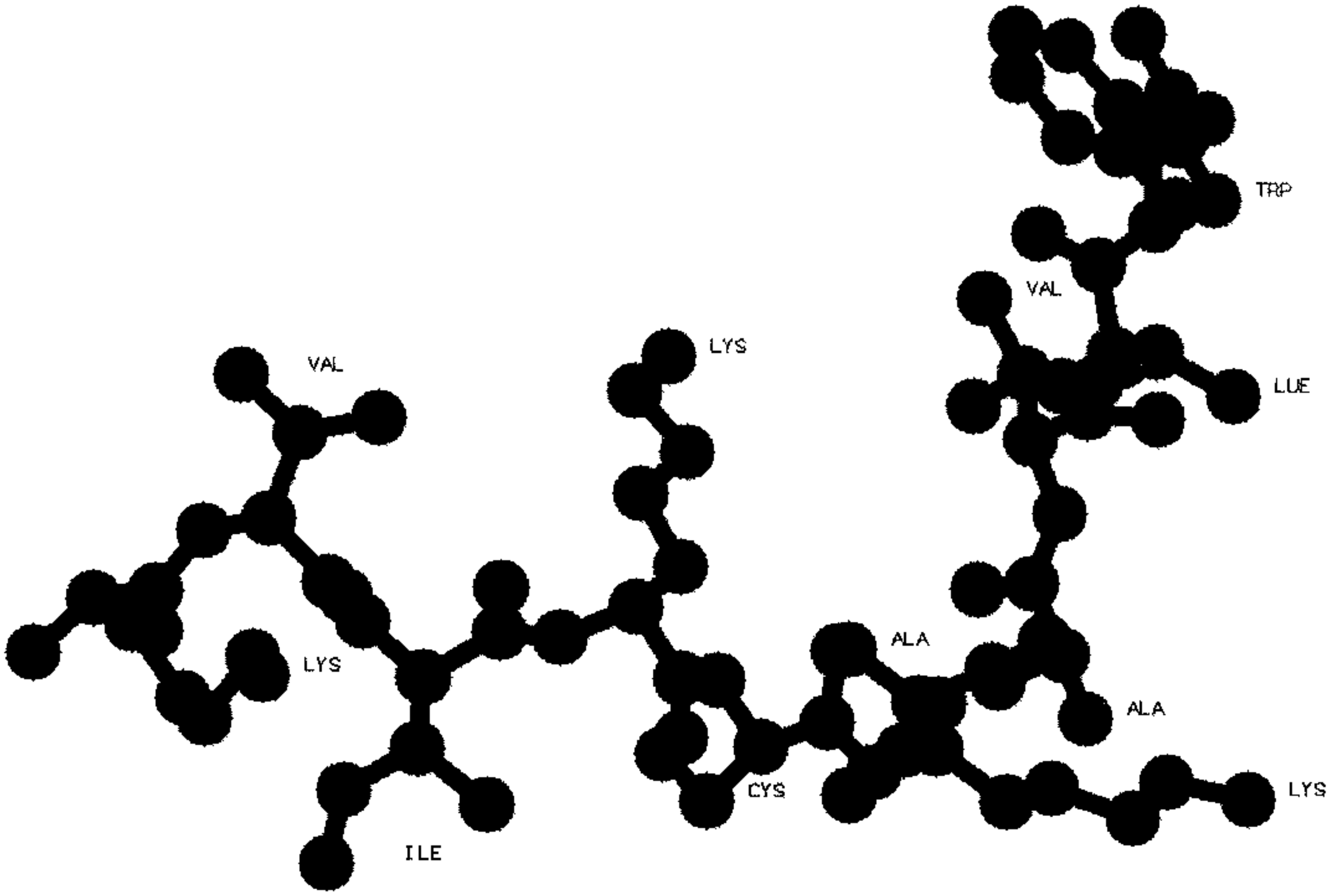

Figure 1. A Short Segment (Residues 5–15) From the Protein ADHE Illustrating the Backbone Chain in Black as an Unbranched Polymer With Side Chain Residues in Gray Attached at Regular Intervals Along the Backbone. Starting at the left, the sequence of side chain residue (three letter code) is as follows: lysine (LYS), valine (VAL), isoleucine (ILE), lysine (LYS), cystine (CYS), lysine (LYS), alanine (ALA), alanine (ALA), valine (VAL), leucine (LUE), and tryptophane (TRP). For comparison with the first sequence in Table 1, residues 5–15, the sequence in one-letter code for this segment is KVIKCKAAVLW.

a few segments are involved in cofactor binding, but only one such segment in each binding site is involved in binding to the important phosphate groups common to all three cofactors. This segment, shown in green in Figures 2 and 3, corresponds to a specific subsequences in ADHE, which is underlined in Table 1. Experimentally verified binding segment in the second sequence of Table 1 is also underlined. The distinctive strand-helix-strand shape of the phosphate binding segment of ADHE in Figures 2 and 3 is critical to its ability to bind to this group. The sequence of residues in the binding segments of these proteins is constrained by the requirement that the residues fold to this distinctive shape and interact properly with the phosphate groups. Surprisingly, however, there is considerable variability in the sequence of residues that can meet these constraints.

1.2 Variations and Commonalities Among Functionally Related Proteins

Variability in the biopolymer sequences arises from mutations that occur during evolution. In most cases, today's sequences are believed to be the progeny of common ancestral sequences and thus are not independent. The mutations happen at a relatively slow rate, so species that are not too far apart in evolutionary time (e.g., man and monkey) are strongly correlated. Homologous proteins within a single organism arise from a duplication of a common ancestral sequence and are passed on in parallel. When biopolymers have been subjected to a limited amount of evolutionary change, their commonality stems primarily from their mutational history. Such closely related sequences are relatively easy to align. The focus here is on the more difficult case that arises when the sequences have been subjected to extensive change and common patterns are subtle.

The other force at work during evolution is natural selection. Intuitively, natural selection limits which mutations survive to be passed on and, therefore, defines the common model for structurally and functionally related proteins. If a mutation is deleterious, then the mutant protein is eliminated from the population. The force of selection on a protein thus arises from its ability to function properly, and function depends on structure. Among distantly related sequences, common features stem primarily from the common set of structural or functional constraints rather than from evolutionary history. Thus in the limit, the sequences may be treated as independent observations. Statistically, the model described later and some other statistical methods that have successfully aligned large numbers of sequences

Table 1. Selected Sequences of Proteins Involved in Binding of NAD, NADP, or FAD

ADHE_HORSE: Alcohol_Dehydrogenase E Chain (EC 1.1.1.1). 2ohx

```
STAGKVIKCKAAVLWEEKKPFSIEEVEVAPPKAHEVRIKMVATGICRSDDHVVSGTLVTP
LPVIAGHEAAGIVESIGEGVTTVRPGDKVIPLFTPQCGKCRVCKHPEGNFCLKNDLSMPR
GTMQDGTSRFTCRGKPIHHFLGTSTFSQYTVVDEISVAKIDAASPLEKVCLIGCGFSTGY
GSAVKVAKVTQGSTCAVFGLGGVGLSVIMGCKAAGAARIIGVDINKDKFAKAKEVGATEC
VNPQDYKKPIQEVLTEMSNGGVDFSFEVIGRLDTMVTALSCCQEAYGVSVIVGVPPDSQN
LSMNPMLLLSGRTWKGAIFGGFKSKDSVPKLVADFMAKKFALDPLITHVLPFEKINEGFD
LLRSGESIRTILTF
```

LDHM_SQUAC: L-Lactate Dehydrogenase M Chain (EC 1.1.1.27) 1ldh

```
ATLKDKLIGHLATSQEPRSYNKITVVGVGAVGMACAISILMKDLADEVALVDVMEDKLKG
EMMDLQHGSLFLHTAKIVSGKDYSVSAGSKLVVITAGARQQEGESRLNLVQRNVNIFKFI
IPNIVKHSPDCIILVVSNPVDVLTYVAWKLSGLPMHRIIGSGCNLDSARFRYLMGERLGV
HSCSCHGWVIGEHGDSVPSVWSGMNVASIKLHPELGTNKDKQDWKKLHKDVVDSAYEVIK
LKGYTSWAIGLSVADLAETIMKNLCRVHPVSTMVKDFYGIKDNVFLSLPCVLNDHGISNI
VKMKLKPDEEQQLQKSATTLWDIQKDLKF
```

LEU3_LEPIN: 3-Isopropylmalate Dehydrogenase (EC 1.1.1.85)

```
MKNVAVLSGDGIGPEVMEIAISVLKKALGAKVSEFQFKEGFVGGIAIDKTGHPLPPETLK
LCEESSAILFGSVGGPKWETLPPEKQPERGALLPLRKHFDLFANLRPAIIYPELKNASPV
RSDIIGNGLDILILRELTGGIYFGQPKGREGSGQEEFAYDTMKYSRREIERIAKVAFQAA
RKRNNKVTSIDKANVLTTSVFWKEVVIELHKKEFSDVQLNHLYVDNAAMQLIVNPKQFDV
VLCENMFGDILSDEASIITGSIGMLPSASLSESGFGLYEPSGGSAPDIAGKGVANPIAQV
LSAALMLRYSFSMEEEANKIETAVRKTIASGKRTRDIAEVGSTIVGTKEIGQLIESFL
```

3BHS_BOVIN: 3-Beta Hydroxy-5-Ene Steroid Dehydrogenase

```
AGWSCLVTGGGGFLGQRIICLLVEEKDLQEIRVLDKVFRPEVREEFSKLQSKIKLTLLEG
DILDEQCLKGACQGTSVVIHTASVIDVRNAVPRETIMNVNVKGTQLLLEACVQASVPVFI
HTSTIEVAGPNSYREIIQDGREEEHHESAWSSPYPYSKKLAEKAVLGANGWALKNGGTLY
TCALRPMYIYGEGSPFLSAYMHGALNNNGILTNHCKFSRVNPVYVGNVAWAHILALRALR
DPKKVPNIQGQFYYISDDTPHQSYDDLNYTLSKEWGFCLDSRMSLPISLQYWLAFLLEIV
SFLLSPIYKYNPCFNRHLVTLSNSVFTFSYKKAQRDLGYEPLYTWEEAKQKTKEWIGSLV
KQHKETLKTKIH
```

NOTE: This table shows the first 4 of the 91 sequences presented by Neuwald and Green (1994). In the first two sequences, dinucleotide-binding sites (underlined) have been experimentally determined and are also correctly predicted by the sampler. Reliable information on binding of cofactor is unavailable for the third and fourth sequences. No element is predicted by the sampler in the third sequence, and one is predicted in the forth sequence (underlined). The full data set is available at anonymous ftp site NCBI.NLM.NIH.GOV, directory pub/gibbs.

(Baldi, Chauvin, McClure, and Hunkapiller 1994; Haussler, Krogh, Mian, and Sjolander 1993) assume that sequences under analysis are *independent* given the common model. In many cases this assumption can be very closely achieved through careful selection of the data set. In practice we have found that the methods we propose here often work well even with substantial departures from this assumption.

There are four classes of mutations: point mutations, insertions/deletions, transpositions, and duplications. A point mutation occurs when a residue at a given point in a sequence is mutated to a new residue type. In our example we focus on proteins that bind either NAD, NADP, or FAD as a cofactor. To achieve binding there must be a force based on the energetic requirement of the protein to hold the cofactor in place. This requirement imposes constraints on point mutations in the binding site. The relationship between energetic constraints and frequencies forms the basis of statistical mechanics, pioneered by Gibbs and Boltzmann. There is an analogous relationship for residue frequencies subject to random point mutations (Berg and von Hipple 1987; Bryant and Lawrence 1991; Pohl 1971), which forms the foundation of the models used here. From a statistical modeling perspective, this relationship is extremely valuable, because it permits us to translate from the language of physics and chemistry that governs molecular behavior to the language of statistics.

Multinomial models have been used successfully to capture the variability and the limitations imposed on these sequences. Because the most important interactions of residues in a binding site area with the cofactor, substrate, or the environment rather than with one another (Bryant and Lawrence 1993), motif models that assume independence of the positions provide a good first approximation. Thus we use *product multinomial models* to describe the binding sites. The remainder of the enzyme forms a scaffold that supports the substrate binding site. Because in distantly related proteins these scaffolds can vary greatly, there is often little in common in the remainder of the protein of interest to us. Consequently, they can usually be described by a simple iid or low-order Markov sequence model.

The other three classes of mutations—insertions/deletions, transpositions, and duplications—result in changes in the length of the sequence or in reordering of the sequence. Let us consider in more detail the effect of deletions with specific attention on the segments involved in cofactor binding. A deletion mutation will remove a segment of a protein sequence, and the resulting two adjacent fragments of the protein will be shifted to form a continuous chain. If the deletion is "upstream" of the cofactor binding segment, then this segment will be shifted to the left and thus *misaligned* with respect to its predecessor. Insertions operate in an analogous manner but add sequence segments. Transpositions move a segment to a new location in the sequence. Duplications replicate segments and then insert them in new locations. On the other hand, the stringent geometric requirements on the binding segments can rarely tolerate internal length changes, and thus in most cases can be modeled well as segments with no internal misalignment.

1.3 Missing-data Methodology for Alignment Problem

The variations of the locations of the phosphate-binding segments in Table 1 provide a good illustrative example of the effects of these mutational processes. The underlined segments, each of which corresponds to a substrate binding segment, occur at various different locations in each sequence. The data in Table 2 show these segments aligned. We seek to find the multinomial model that describes the data in Table 2. Unfortunately, direct data that specify the locations (i.e., the indices of these binding segments in each sequence) are not available. The data available for analysis are like those in Table 1 without underlines or any other indication of the positions of the common segments in these sequences. This type of misalignment of sequences is the distinguishing characteristic in the analysis of biopolymer sequence data.

Lawrence and Reilly (1990) recognized that it is constructive to consider alignment variables as missing data and presented an expectation maximization (EM) algorithm for the alignment of gene regulatory sites in DNA sequences. Cardon and Stormo (1991) extended this approach to allow for one small gap in the middle of each segment. Liu (1994) showed that an iterative sampling approach could be used to advantage in the identification of gene regulatory sites, and described a Gibbs sampling algorithm (Gelfand and Smith 1990) for its solution. The collapsing theorem developed there has been found very useful for these problems. Lawrence et al. (1993) have described a Gibbs sampling algorithm for local multiple alignment of protein sequences. Alignment methods based on hidden Markov models (HMM) have recently been described (Baldi et al. 1994; Haussler et al. 1993). Other statistical approaches for sequence alignment have also been described (Allison, Wallace, and Yee 1992; Bishop and Thompson 1986, Thorne, Kishino, and Felsenstein 1991). More comments on these methods with comparison to ours will be provided in Section 7.

In Sections 2–4 we describe statistical methods for analyzing variants of the basic multiple alignment problem. In Section 5 we present a nonparametric method for assessing the significance of our findings. In Section 6 we describe the application of these methods to dinucleotide binding proteins, and in Section 7 we conclude with a discussion.

2. DATA STRUCTURE AND MULTINOMIAL MODELS

2.1 Defining the Statistical Problem

Given a set of sequences, $R_1, R_2, \ldots, R_K$, such as those in Table 1 without underlining, our goals are twofold: (1) to identify the substrate binding segments, which we call "elements," and (2) to estimate the parameters of the product multinomial model that describes the collection of aligned elements. Here we call this common model a "motif." We call position a "site" if it is described by the motif and "nonsite" otherwise.

In general, the sequence data can be represented as

$$\text{Data } \mathbf{R}: \begin{array}{lcccc} \text{sequence } R_1: & r_{1,1} & r_{1,2} & \cdots & r_{1,L_1} \\ \text{sequence } R_2: & r_{2,1} & r_{2,2} & \cdots & r_{2,L_2} \\ \vdots & \vdots & & \ddots & \vdots \\ \text{sequence } R_K: & r_{K,1} & r_{K,2} & \cdots & r_{K,L_K}, \end{array}$$

where the residue r_{kl} among the sequences take values from an alphabet with p (=20 for protein) different letters, and the L_k are the respective lengths of the sequences. Let the collection of indices be denoted by $I = \{(k,l)\colon l = 1,\ldots,L_K; k = 1,\ldots,K\}$. For any set $S \subseteq \mathbf{I}$, we define $\mathbf{R}_S = \{r_{k,l}\colon (k,l) \in S\}$. We also write $S^c = I \setminus S$. We seek within each sequence mutually similar segments (called elements) of specific length J. These elements are assumed to be *independent* observations from a *product-multinomial* model (called a *motif*) that describes residue frequencies for each position j within an element and consists of parameters $\boldsymbol{\theta}_j = (\theta_{1,j},\ldots,\theta_{p,j})^T, j = 1,\ldots,J$. The background parameter $\boldsymbol{\theta}_0 = (\theta_{1,0},\ldots,\theta_{p,0})^T$ describes the frequencies of those nonsite positions.

As discussed in Section 1, three classes of mutations result in misalignment of common elements. Another data structure, constituting the alignment, is a set of positions a_k for k from 1 to K, for the starting positions of the elements within each sequence. Regarding the a's as missing observations, we can effectively apply missing-data methodology. One of our goals will be to identify the "best," defined as the most probable, motif and alignment.

To help fix the idea for the material that follows, consider the following coin-tossing analogy. The game is played with L coins, say 50. $(L - J)$ of these coins, say 40 (plain coins), all have the same probability of heads, not necessarily a half. The remaining $J = 10$ special coins have probabilities of heads different from the plain coins and different from one another. On each of K independent trials, corresponding to K observed sequences in alignment problems, the coins are shaken in a tumbler and laid out in a row. It is known that the 10 special coins are positioned in a contiguous block of 10, with the special coins always in the same order is each trial. But their actual positions vary from trial to trial and are not observed. The player is challenged to estimate the probabilities of the 11 types of coins and to specify the locations of the special coins in each trial.

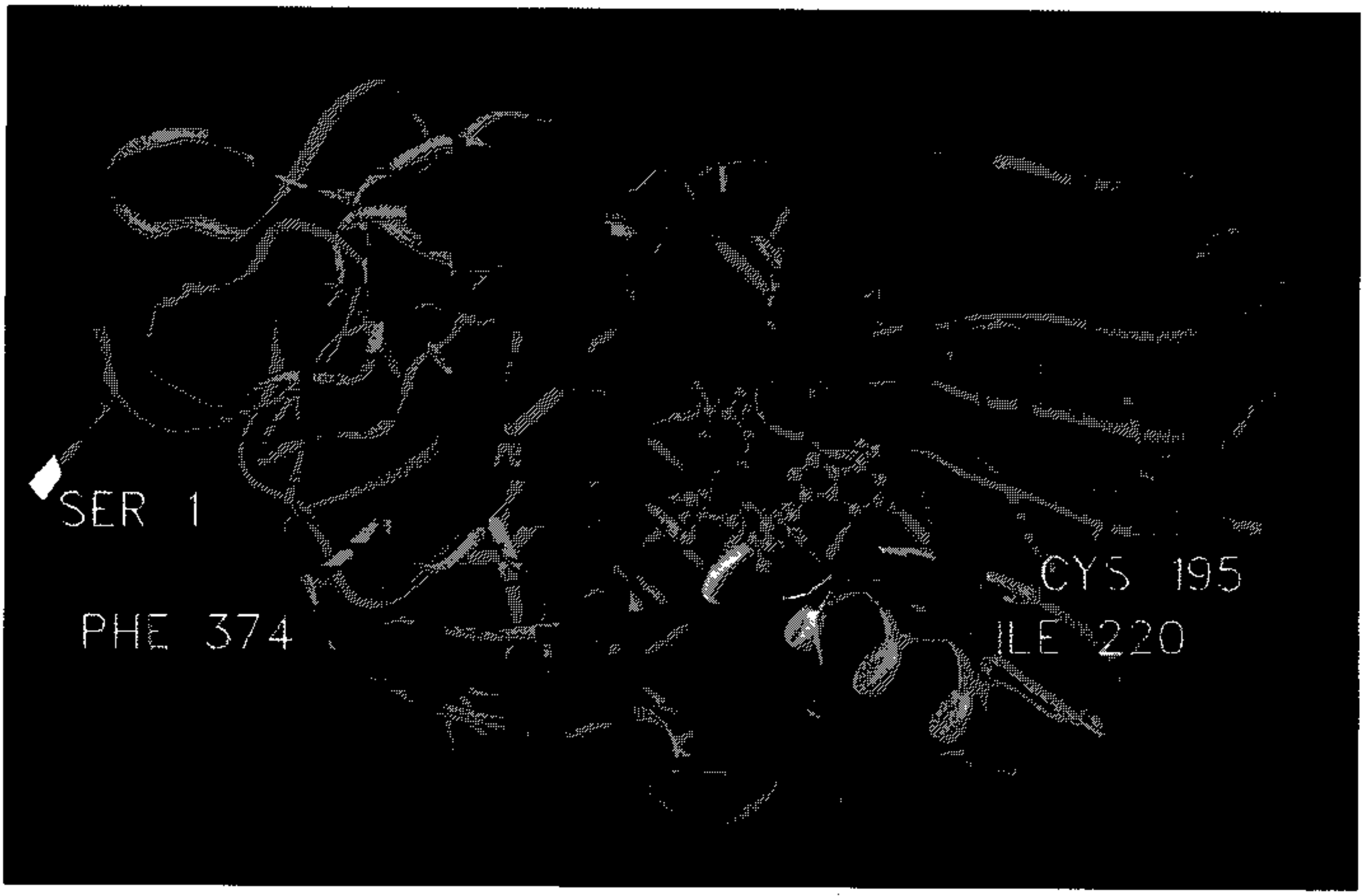

Figure 2. The Trace of the Backbone Chain of ADHE With the Segment (Residues 195–220; see Table 2), Identified by the Sampler in Green and the Cofactor (NAD) in Orange. The dinucleotide binding element in green forms a strand-helix-strand structure. Glycine residues at positions 199, 201, 204 in red, and the valine at position 197 in purple.

This game corresponds to an alignment problem with $p = 2$ (two-letter alphabet), $J = 10$, and $L_1 = \cdots = L_K = 50$. The associated parameter $\boldsymbol{\theta}$'s describe the unknown probabilities of showing heads for the $J+1$ types of coins.

The material we present in Sections 3.1 and 3.2 provides a sampling strategy for the solution of this game assuming that we know there is exactly one block of length $J = 10$ special coins in each trial. In Section 3.3 we extend the model to permit multiple blocks of special coins in each trial, and in Section 3.4 we relax the requirement that the coins occur as contiguous blocks of length J. In Section 4 we relax the requirement that there is a block of special coins for each sequence. We discuss a ranks method to test the hypothesis that there are special coins in the data in Section 5.

2.2 Product Multinomial Model With Dirichlet Priors

For a set of categorical data $R = \{r_1, \ldots, r_m\}$, where each r_i takes values from a p-letter alphabet, we define a counting function $\mathbf{h}$ such that $\mathbf{h}(R) = (m_1, \ldots, m_p)^T$, where m_k is the total number of kth type letter observed in R. For protein sequences, the $\mathbf{h}$ function counts the numbers of different types of amino acids in set R and results in a 20-dimensional column vector. It is noted that the function $\mathbf{h}$ has a nice additive property. If another data set $R' = (r'_1, \ldots, r'_n)$ is given, for example, then $\mathbf{h}(R) + \mathbf{h}(R') = \mathbf{h}(R \oplus R')$, where the left side is just the ordinary addition for vectors and $R \oplus R'$ indicates combining the two categorical data sets R and R'.

For vectors $\mathbf{v} = (v_1, \ldots, v_p)^T$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$, we define that $|\mathbf{v}| = |v_1| + \cdots + |v_p|$, $\mathbf{v} + \boldsymbol{\theta} = (v_1 + \theta_1, \ldots, v_p + \theta_p)^T$, $\mathbf{v}/\boldsymbol{\theta} = (v_1/\theta_1, \ldots, v_p/\theta_p)^T$, $\boldsymbol{\theta}^{\mathbf{v}} = \theta_1^{v_1} \cdots \theta_p^{v_p}$, and $\Gamma(\mathbf{v}) = \Gamma(v_1) \cdots \Gamma(v_p)$. If the v_i's are integers, then we denote $\mathbf{v}! = v_1! \cdots v_p!$. Using these notations, for example, we can denote the norming constant for a p-dimensional Dirichlet distribution $\text{Dir}(\boldsymbol{\alpha})$ as $\Gamma(|\boldsymbol{\alpha}|)/\Gamma(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$. Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J)$ be a $p \times J$ matrix in which each $\boldsymbol{\theta}_j$ is a probability vector of length p; that is, $\boldsymbol{\theta}_j = (\theta_{1_j}, \ldots, \theta_{pj})^T$ satisfies $\theta_{ij} \geq 0$ and $|\boldsymbol{\theta}_j| = 1$ for all j.

A $p \times J$ integer matrix $\mathbf{M} = (\mathbf{m}_1, \ldots, \mathbf{m}_J)$, where $\mathbf{m}_j = (m_{1j}, \ldots, m_{pj})^T$, is said to follow a product multinomial distribution with parameter $\boldsymbol{\Theta}$ (i.e., $\mathbf{M} \sim \text{PM}(\boldsymbol{\Theta}; |\mathbf{m}_1|, \ldots, |\mathbf{m}_J|)$), if each $\mathbf{m}_j$ follows the multinomial distribution $\text{mul}(\boldsymbol{\theta}_j, |\mathbf{m}_j|)$ and the $\mathbf{m}_j$ are mutually independent. The *product Dirichlet* (PD) distribution can be similarly introduced as a conjugate prior for $\boldsymbol{\Theta}$. That is, we say $\boldsymbol{\Theta} \sim \text{PD}(\mathbf{B})$, where $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)$ is a $p \times J$ matrix and $\boldsymbol{\beta}_j = (\beta_{1j}, \ldots \beta_{pj})^T$, if the $\boldsymbol{\theta}_j$ are independent and p-dimensional Dirichlet random variables with distributions $\text{Dir}(\boldsymbol{\beta}_j)$, $j = 1, \ldots, J$.

Suppose that we observe random vectors $R_i = (r_{i1}, \ldots, r_{iJ})$, for $i = 1, \ldots, n$, where the r_{ij} are mutually in-

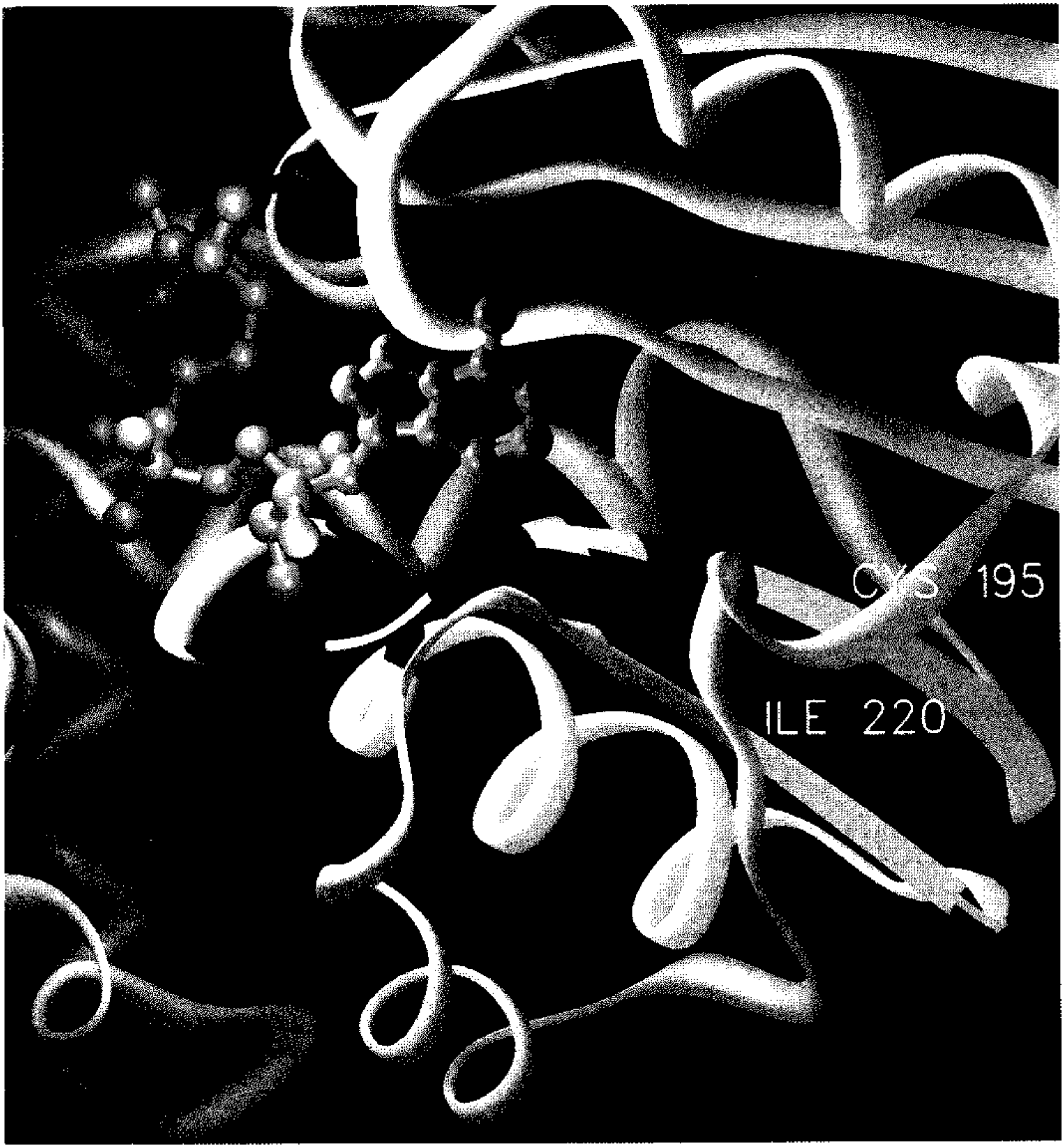

Figure 3. A Magnified View of the Dinucleotide Binding Element of ADHE. Three side chain carbon atoms of the valine 197 residue, in purple, are illustrated. A glycine residue has only a hydrogen atom in its side chain. Hydrogen atoms are generally not visible in the X-ray structure and thus are not illustrated here.

dependent, and with probability θ_{kj} take the kth letter in the alphabet. Let the observations be arranged as

$$\begin{array}{ccc} r_{11} & \cdots & r_{1J_1} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nJ_1} \end{array}$$

and let $R_{\cdot j} = \{r_{1j}, \ldots, r_{nj}\}$. Then the likelihood of Θ can be written as

$$\pi(R_1, \ldots, R_n | \Theta) \propto \prod_{j=1}^{J} \boldsymbol{\theta}_j^{\mathbf{h}(R_{\cdot j})}.$$

Thus $\mathbf{h}(R_{\cdot j})$ represents the sufficient statistics of $\boldsymbol{\theta}_j$ for all j, and $(\mathbf{h}(R_{\cdot 1}), \ldots, \mathbf{h}(R_{\cdot j})) \sim \mathbf{PM}(\Theta; n, \ldots, n)$. If the prior distribution for Θ is PD($\mathbf{B}$), then the posterior distribution of Θ is PD($\mathbf{B}+H$), where $\mathbf{H} = (\mathbf{h}(R_{\cdot 1}), \ldots, \mathbf{h}(R_{\cdot J}))$. The predictive distribution for the next observation R_{new} or, equivalently, $(\mathbf{h}(r_{\text{new},1}), \ldots, \mathbf{h}(r_{\text{new},J}))$ is PM($\tilde{\Theta}; 1, \ldots, 1$), where $\tilde{\Theta} \propto (\mathbf{B}+\mathbf{H})$ is the posterior mean of Θ.

3. MULTINOMIAL SAMPLING: ALIGNMENT WHEN THE NUMBER OF ELEMENTS IS GIVEN

In Sections 3.1 and 3.2 we describe a sampling-based alignment procedure when there is known to be one element in each sequence from a single motif. In Section 3.3 we generalize this to the case of multiple elements and multiple motifs. In Section 3.4 we use a fragmentation model to relax the requirement that an element has to be a contiguous block.

3.1 Model and Posterior Distributions for One Element Per Sequence

Let the starting position of the element in the kth sequence be denoted as a_k, where $k = 1, \ldots, K$, and a_k can take values from 1 to $L_k - J + 1$. Let $A = \{(1, a_1), \ldots, (K, a_K)\}$, which is the unobserved alignment. We define $\{A\} = \{(k, a_k + j - 1): k = 1, \ldots, K, j = 1, \ldots, J\}$, which is the set of indices occupied by the elements with starting positions indicated in A. Let $\mathbf{R}_{\{A\}}$ denote the collection of the residues in $\{A\}$; that is, $\mathbf{R}_{\{A\}} = \{r_{k,a_k+j-1}$: for $j = 1, \ldots, J; k = 1, \ldots, K\}$. We further use $\{a_k\}$ to denote the set of indices occupied by the element in the kth sequence, (i.e., $\{a_k\} = \{(k, a_k + j - 1)$, for $j = 1, \ldots, J\}$ is a collection of J consecutive positions in the kth sequence and use $A(j)$ to denote the set of the jth positions of all elements (i.e., $A(j) = \{(k, a_k + j - 1)$, for $k = 1, \ldots, K\}$).

We assume that all residues outside the motif region are independently drawn from a common multinomial model characterized by a vector $\boldsymbol{\theta}_0 = (\theta_{1,0}, \ldots, \theta_{p,0})^T$, where $|\boldsymbol{\theta}_0| = 1$ and $\theta_{i,0} \geq 0$, for all i. But the residue frequencies for the motifs are modeled by PM(Θ) defined in Section 2, where $\Theta = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J)$. Therefore, $J+1$ p-dimensional parameter vectors are required to fully describe the data.

Treating the alignment data A as missing, we can write the complete-data likelihood of the parameters as

$$\pi(\mathbf{R}, A|\boldsymbol{\theta}_0, \Theta) \propto \boldsymbol{\theta}_0^{\mathbf{h}(\mathbf{R}_{\{A\}^c})} \prod_{j=1}^{J} \boldsymbol{\theta}_j^{\mathbf{h}(\mathbf{R}_{A(j)})}. \qquad (1)$$

As was noted by Tanner and Wong (1987), this simple form of the complete-data likelihood (posterior distribution) is the key to carrying out their data augmentation algorithm. We can further use the collapsing technique of Liu (1994) to integrate out the parameter vectors $\boldsymbol{\theta}_0$ and Θ to obtain a *predictive update* version of the Gibbs sampler. Consequently, the resulting program, by skipping the step of drawing from product Dirichlet distributions, only involves sampling from a product multinomial distribution iteratively (whereas a standard Gibbs sampler would necessarily include the step of sampling from product Dirichlet distributions).

The use of iterative sampling for Bayesian missing-data problems was first undertaken by Tanner and Wong (1987) and Li (1988). This sampling approach and its extensions have recently become a topic of great interest in Bayesian statistics. Gelfand and Smith (1990) illustrated its connection with a more general method—the Gibbs sampler—and greatly popularized the idea. Some theoretical guidances, such as judging convergence and choosing efficient samplers, have been obtained (Gelman and Rubin 1992 and its companion discussions; Liu, Wong, and Kong 1994, 1995; Tierney 1995); see the *Journal of the Royal Statistical Society*, Ser. B, 1993, Vol. 55, pp. 3–102, for an overview.

3.2 Predictive Update Formula

Let the prior for $\boldsymbol{\theta}_0, f(\boldsymbol{\theta}_0)$, be a Dirichlet distribution, Dir($\boldsymbol{\alpha}$), where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$, and let the prior for $\Theta, g(\Theta)$, be a product Dirichlet distribution, PD($\mathbf{B}$), with $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)$ and $\boldsymbol{\beta}_j = (\beta_{1j}, \ldots, \beta_{pj})^T$. Then by using the Bayes theorem and integrating out the $\boldsymbol{\theta}$, we obtain the following predictive distribution for A:

$$\begin{aligned}\pi(A|\mathbf{R}) &\propto \pi(\mathbf{R}, A) \\ &= \int\int \pi(\mathbf{R}, A|\boldsymbol{\theta}_0, \Theta) f(\boldsymbol{\theta}_0) g(\Theta)\, d\boldsymbol{\theta}_0\, d\Theta \qquad (2) \\ &\propto \Gamma\{\mathbf{h}(\mathbf{R}_{\{A\}^c}) + \boldsymbol{\alpha}\} \prod_{j=1}^{J} \Gamma\{\mathbf{h}(\mathbf{R}_{A(j)}) + \boldsymbol{\beta}_j\}. \qquad (3)\end{aligned}$$

Let $A_{[-k]}$ denote the set $\{(l, a_l), l \neq k\}$; that is, the set of starting positions of elements in all sequences but sequence k. Now we derive the predictive distribution $\pi(a_k|A_{[-k]}, \mathbf{R})$ of the starting position of the element in sequence k conditional on $A_{[-k]}$.

Note that $\mathbf{h}(\mathbf{R}_{A_1}) + \mathbf{h}(\mathbf{R}_{A_2}) = \mathbf{h}(\mathbf{R}_{A_1 \cup A_2})$ if $A_1 \cap A_2 = \phi$, which provides us with

$$\mathbf{h}(\mathbf{R}_{\{A\}^c}) = \mathbf{h}(\mathbf{R}_{\{A_{[-k]}\}^c}) - \mathbf{h}(\mathbf{R}_{\{a_k\}})$$

and

$$\mathbf{h}(R_{A(j)}) = \mathbf{h}(\mathbf{R}_{A_{[-k]}(j)}) + \mathbf{h}(r_{k,a_k+j-1}).$$

Here $\mathbf{h}(r_{k,a_k+j-1})$ is a vector of $p-1$ zeros and a 1 to indicate the residue type observed at position $a_k + j - 1$ in sequence k. Using the fact that $\pi(a_k|A_{[-k]}, \mathbf{R}) \propto \pi(A|\mathbf{R})$ and treating functions of $A_{[-k]}$ as constants, we obtain the following expression for $\pi(a_k|A_{[-k]}, \mathbf{R})$ by dividing the right side of (3) by two constants, $\Gamma\{\mathbf{h}(\mathbf{R}_{\{A_{[-k]}\}^c}) + \boldsymbol{\alpha}\}$ and $\Gamma\{\mathbf{h}(\mathbf{R}_{A_{[-k]}}(j) + \boldsymbol{\beta}_j)\}$:

$$\pi(a_k|A_{[-k]}, \mathbf{R}) \propto \frac{\Gamma\{\mathbf{h}(\mathbf{R}_{\{A_{[-k]}\}^c}) + \boldsymbol{\alpha} - \mathbf{h}(\mathbf{R}_{\{a_k\}})\}}{\Gamma\{\mathbf{h}(\mathbf{R}_{\{A_{[-k]}\}^c}) + \boldsymbol{\alpha}\}} \times \prod_{j=1}^{J} \{\mathbf{h}(\mathbf{R}_{A_{[-k]}(j)}) + \boldsymbol{\beta}_j\}^{\mathbf{h}(r_{k,a_k+j-1})}. \qquad (4)$$

We use this relationship to iteratively sample through $a_1, \ldots, a_K$ to obtain what we call the *predictive update version* of the Gibbs sampler. The formula works well, but computing the ratio of two gamma functions can be slow.

When the size of R_2 is small compared to R_1 and the composition of R_2 is relatively diverse, we can use the following approximation:

$$\Gamma\{\mathbf{h}(R_1 \oplus R_2)\}/\Gamma\{\mathbf{h}(R_1)\} \approx \mathbf{h}(R_1)^{\mathbf{h}(R_2)}.$$

Hence when $\mathbf{R}_{\{a_k\}}$, whose size is J, is relatively diverse compared to $\mathbf{R}_{\{A_{[-k]}\}^c}$, the collection of all nonsite residues, we have

$$\frac{\Gamma\{\mathbf{h}(\mathbf{R}_{\{A_{[-k]}\}^c}) + \boldsymbol{\alpha} - \mathbf{h}(\mathbf{R}_{\{a_k\}})\}}{\Gamma\{\mathbf{h}\{\mathbf{R}_{\{A_{[-k]}\}^c}) + \boldsymbol{\alpha}\}} \approx \frac{1}{\mathbf{h}(\mathbf{R}_{\{A\}^c})^{\mathbf{h}(\mathbf{R}_{\{a_k\}})}}.$$

Using this approximation, we arrive at a simple formula for the predictive distribution:

$$\pi(a_k = i|A_{[-k]}, \mathbf{R}) \propto \prod_{j=1}^{J} \left(\frac{\hat{\boldsymbol{\theta}}_{j[k]}}{\hat{\boldsymbol{\theta}}_{0[k]}}\right)^{\mathbf{h}(r_{k,i+j-1})}, \qquad (5)$$

where the $\hat{\boldsymbol{\theta}}_{j[k]}$ are the posterior means of the $\boldsymbol{\theta}_j$ conditioned on the observation $\mathbf{R}$ and the current alignment $A_{[-k]}$, and $\hat{\boldsymbol{\theta}}_{0[k]}$ is the posterior mean of $\boldsymbol{\theta}_0$ based on the current nonsite positions $\mathbf{R}_{\{A_{[-k]}\}^c}$. The formula implies that conditional on the fixed sites of the elements in the rest of the sequences, the probability that the element in sequence k starts at position i is proportional to the likelihood ratio of its being a site to its being a nonsite. We tested both the exact formula (4) and the approximation (5) in our algorithm. We found no observable discrepancies between the results obtained from using the two different formulas.

3.3 Generalizations to Multiple Elements and Multiple Motifs

The foregoing method can be generalized to the case where there is more than one copy of the common element in each sequence. For example, suppose that there is only one sequence (i.e., $K = 1$) but there are m copies of the common element to be aligned. Let $A = \{a_{1,1}, \ldots, a_{1,m}\}$ be the set of starting positions of the m elements. We can similarly write down the complete-data likelihood,

$$\pi(\mathbf{R}, A|\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J) \propto \boldsymbol{\theta}_0^{\mathbf{h}(\mathbf{R}_{\{A\}^c})} \prod_{j=1}^{J} \boldsymbol{\theta}_j^{\mathbf{h}(\mathbf{R}_{A(j)})}.$$

Integrating out the $\boldsymbol{\theta}$'s, we arrive at a form similar to (3):

$$\pi(A|\mathbf{R}) \propto \Gamma\{\mathbf{h}(\mathbf{R}_{\{A\}^c}) + \boldsymbol{\alpha}\} \prod_{j=1}^{J} \Gamma\{\mathbf{h}(\mathbf{R}_{A(j)}) + \boldsymbol{\beta}_j\}.$$

Hence the conditional distributions are also similar to (4), with an added restriction that the elements not overlap. For multiple copies in multiple sequences, we can simply combine the case presented in Section 3.2 and the foregoing to arrive at a formula similar to (5).

3.4 Fragmentation Model

Previously, we modeled the motif as a contiguous block of width J residues. But usually not all the positions within the contiguous block of an element are important for protein structure and function. To accommodate this feature, we may wish to select $J < W$ positions in an aligned block of width W residues in each of the K sequences to form the motif model. This model automatically takes care of the phase-shift problem noted by Liu (1994).

Let $\boldsymbol{\Delta} = (\delta_1, \ldots, \delta_W)$, where $\delta_w = 0$ or 1, and $\sum_{w=1}^{W} \delta_w = J$. The δ's indicate which positions (or columns) within the aligned block are included in the motif model. The vector $\boldsymbol{\Delta}$ is in fact another high-dimensional parameter and can be treated as missing data as well. Treating $\boldsymbol{\Delta}$ as missing data implicitly assumes a flat prior on all its possible values; that is, a priori $\boldsymbol{\Delta}$ takes any 0–1 vector satisfying the constraints with equal probability $1/\binom{W}{J}$. The complete-data likelihood in this case can be written as

$$\pi(\mathbf{R}, A, \boldsymbol{\Delta}|\theta_0, \boldsymbol{\Theta}) \propto \boldsymbol{\theta}_0^{\mathbf{h}(\mathbf{R}_{\Delta\{A\}^c})} \prod_{w=1}^{W} \boldsymbol{\theta}_w^{\delta_w \mathbf{h}(\mathbf{R}_{A(w)})},$$

where $\boldsymbol{\Delta}\{A\}$ denotes those positions with $\delta_w = 1$ and $\boldsymbol{\Delta}\{A\}^c \equiv (\boldsymbol{\Delta}\{A\})^c$. With the same Dirichlet priors on the $\boldsymbol{\theta}$'s, we can obtain the joint posterior distribution of A and $\boldsymbol{\Delta}$; that is,

$$\pi(A, \boldsymbol{\Delta}|\mathbf{R}) \propto \Gamma\{\mathbf{h}(\mathbf{R}_{\boldsymbol{\Delta}\{A\}^c}) + \boldsymbol{\alpha}\} \times \prod_{w=1}^{W} \Gamma[\delta_j\{\mathbf{h}(\mathbf{R}_{A(w)}) + \boldsymbol{\beta}_w\}]. \qquad (6)$$

To sample from this distribution, we proceed in two steps. First, we sample to find a contiguous block of J residues as in Section 3.1. This is equivalent to assuming that $\delta_1 = \cdots = \delta_J = 1$ and all the rest are zeros. Second, we use a Gibbs sampling strategy for identifying the best positions. The way to achieve this is to permit the exchange of one of the J positions in the motif model for one of the $(W - J)$ nonsite positions within the aligned block. We draw one of the J positions to take out of the model completely at random, followed by the sampling of a replacement position from one of the $(W - J + 1)$ unoccupied positions where sampling is now in proportion to the ratio of the site to nonsite gammas in Equation (6).

It is typical in our applications that each position column has only tens of residues, whereas all the nonsite residues add to several thousands. Hence the variation of the nonsite residue frequencies caused by including or excluding a column can usually be ignored, and thus the conditional distribution $\pi(\boldsymbol{\Delta}|\mathbf{R}, A)$ resulting from (6) can be approximated by

$$\pi^*(\boldsymbol{\Delta}|\mathbf{R}, A) \propto \prod_{w:\delta_w=1} g_w, \qquad (7)$$

where

$$g_w = \frac{\Gamma[\delta_w\{\mathbf{h}(\mathbf{R}_{A(w)}) + \boldsymbol{\beta}_w\}]}{\hat{\boldsymbol{\theta}}_0^{\mathbf{h}(\mathbf{R}_{A(w)})}}$$

can be regarded as a weight assigned to position w. This approximation works well in all of our examples. Several sampling strategies for this model and their connections with weighted finite population sampling are developed in the Appendix.

3.5 Weighted Prior For Fragmentation

In previous subsection we modeled a priori that $\mathbf{\Delta}$ takes any W-long 0–1 vector satisfying $\delta_1 + \cdots + \delta_W = J$ with equal probability $1/\binom{W}{J}$. But real protein sequences favor short motif and thus suggest that the "span" of the element should be small. Let $w_0 = \min\{w\colon \delta_w = 1\}$, and let $w_1 = \max\{w\colon \delta_w = 1\}$. The span of an element is defined as $w_1 - w_0 + 1$ whose smallest possible value is J. An alternative prior distribution for $\mathbf{\Delta}$ that accommodates the protein reality is

$$\pi(\mathbf{\Delta}) \propto \binom{w_1 - w_0 - 1}{J-2}^{-1}.$$

The rationale is that given the length of the span, there are $\binom{w_1 - w_0 - 1}{J-2}$ ways of assigning zeros and 1s for positions within w_0 and w_1. In other words, this prior assigns equal probability on all possible (w_0, w_1), the span of the element.

We have tried both prior distributions in many test examples and found that the latter is significantly better than the former in the sense that the sampler identifies the correct locations of the elements more often. This weighted prior for fragmentation is used in the dinucleotide-binding example.

4. BERNOULLI SAMPLING STRATEGY

It is often difficult or impossible to specify the number of elements in a sequence. In this section we show how this requirement may be relaxed.

4.1 Model and Likelihood

To make our discussion simple and precise, we take the observed multiple sequences as one long sequence with total length L^*, with restrictions to exclude sampling of elements that overlap the ends of sequences. Let the long sequence be denoted by

$$\mathbf{R} = (r_1, \ldots, r_{L^*}),$$

and let $L = L^* - J + 1$ be the total number of possible element positions of length J that can occur. An unobserved indicator vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_L)$, where $\xi_l = 0$ or 1, is introduced to indicate the starting positions of the elements. If $\xi_l = 1$, then an element occurs from l to $l + J - 1$, assuming that the width of an element is known to be J. Our task is then to classify each position l of the sequence into two "types": starter of an element or not. Let $|\boldsymbol{\xi}| = \sum_{l=1}^{L} \xi_l$ be the total number of 1s, and let $A_{\boldsymbol{\xi}}$ be the set of those l's such that $\xi_l = 1$. Again, let $A_{\boldsymbol{\xi}}(j)$ be the collection of the jth positions of all elements. Because we do not permit elements to overlap, there is some slight dependency among the ξ's. In our case, because probability ε is very small, ignoring this slight dependency, as in the iid model that follows, has little effect. Therefore, we assume a priori that $\xi_l = 1$ with probability ε independently for all possible l's. Then, by treating $\boldsymbol{\xi}$ as missing data, we have

$$\pi(\mathbf{R}, \boldsymbol{\xi} | \boldsymbol{\Theta}, \boldsymbol{\theta}_0, \varepsilon) = \boldsymbol{\theta}^{\mathbf{h}(\mathbf{R}_{\{A_\xi\}^c})} \prod_{j=1}^{J} \boldsymbol{\theta}_j^{\mathbf{h}(\mathbf{R}_{A_\xi(j)})} \varepsilon^{|\boldsymbol{\xi}|}(1-\varepsilon)^{L-|\boldsymbol{\xi}|}. \tag{8}$$

With Dirichlet conjugate priors $D(\boldsymbol{\alpha})$ on $\boldsymbol{\theta}_0$,PD($\mathbf{B}$) on $\boldsymbol{\Theta}$, and Beta(a, b) on ε, the joint distribution of $\boldsymbol{\xi}, \boldsymbol{\Theta}, \boldsymbol{\theta}_0$, and ε can be written explicitly as

$$\pi(\boldsymbol{\xi}, \boldsymbol{\Theta}, \boldsymbol{\theta}_0, \varepsilon | \mathbf{R}) \propto \boldsymbol{\theta}_0^{\mathbf{h}(\mathbf{R}_{\{A_\xi\}^c}) + \boldsymbol{\alpha} - 1} \times \prod_{j=1}^{J} \boldsymbol{\theta}_j^{\mathbf{h}(\mathbf{R}_{A_\xi(j)}) + \boldsymbol{\beta}_j - 1} \varepsilon^{|\boldsymbol{\xi}|+a-1}(1-\varepsilon)^{L-|\boldsymbol{\xi}|+b-1}. \tag{9}$$

Typically, $a/(a+b)$ is very small, ranging from .001 to .01.

4.2 Predictive Sampling Distribution

Integrating out the parameters except $\boldsymbol{\xi}$ provides us with a more concise formula and a much faster program for doing the Gibbs sampler, as has been implemented by Chen and Liu (1993) for a switch regression problem. Based on (9), we have

$$\begin{aligned}\pi(\boldsymbol{\xi}|\mathbf{R}) \;&\propto\; \frac{\Gamma(|\boldsymbol{\alpha}|)\Gamma\{\mathbf{h}(\mathbf{R}_{\{A_\xi\}^c}) + \boldsymbol{\alpha}\}}{\Gamma(\boldsymbol{\alpha})\Gamma\{|\mathbf{h}(\mathbf{R}_{\{A_\xi\}^c})| + |\boldsymbol{\alpha}|\}} \\ &\times \prod_{j=1}^{J} \frac{\Gamma(|\boldsymbol{\beta}_j|)\Gamma\{\mathbf{h}(R_{A_\xi(j)}) + \boldsymbol{\beta}_j\}}{\Gamma(\boldsymbol{\beta}_j)\Gamma\{|\mathbf{h}(\mathbf{R}_{A_\xi(j)})| + |\boldsymbol{\beta}_j|\}} \\ &\times B_{a,b}(|\boldsymbol{\xi}|, L - |\boldsymbol{\xi}|),\end{aligned} \tag{10}$$

where $B_{a,b}(c,d) = \int_0^1 x^{a+c-1}(1-x)^{b+d-1}\,dx / \int_0^1 x^{a-1}(1-x)^{b-1}\,dx$ is the beta function. Sampling $\boldsymbol{\xi}$ directly from such a complicated distribution is prohibitive. Gibbs sampling strategy can be applied effectively, because of the very simple form of the conditional distribution of any ξ_k given all the rest ξ's, $\boldsymbol{\xi}[-k]$. More precisely,

$$\frac{\pi(\xi_k = 1 | \boldsymbol{\xi}[-k], \mathbf{R})}{\pi(\xi_k = 0 | \boldsymbol{\xi}[-k], \mathbf{R})} = \frac{\hat{\varepsilon}}{1-\hat{\varepsilon}} \prod_{j=1}^{J} \left(\frac{\hat{\boldsymbol{\theta}}_j}{\hat{\boldsymbol{\theta}}_0}\right)^{\mathbf{h}(r_{k+j-1})}, \tag{11}$$

where $\hat{\varepsilon} = (|\boldsymbol{\xi}[-k]| + a)/(L + a + b - 1)$ and $\hat{\boldsymbol{\theta}}_j = \{\mathbf{h}(\mathbf{R}_{A_{\xi[-k]}+j-1}) + \boldsymbol{\beta}_j\}/(|\boldsymbol{\xi}[-k]| + |\boldsymbol{\beta}_j|)$ is the predictive probability for the jth position of the site based on current known sites. The $\hat{\boldsymbol{\theta}}_0$ is understood to be the current estimate of the model for all the nonsites and is approximated by $\hat{\boldsymbol{\theta}}_0 \approx \mathbf{h}(\mathbf{R}_{\{A_{\xi[-k]}\}^c})/(L - |\boldsymbol{\xi}[-k]|J)$. An accurate formula for $\hat{\boldsymbol{\theta}}_0$, however, is expressed in terms of gamma functions as follows:

$$\hat{\boldsymbol{\theta}}_0 = \left[\frac{\Gamma\{\mathbf{h}(\mathbf{R}_{\{A_{\xi[-k]}\}^c}) - \mathbf{h}(\mathbf{R}_{\{k\}}) + \boldsymbol{\alpha}\}}{\Gamma\{\mathbf{h}(\mathbf{R}_{\{A_{\xi[-k]}\}^c}) + \boldsymbol{\alpha}\}} \times \frac{\Gamma\{|\mathbf{h}(\mathbf{R}_{\{A_{\xi[-k]}\}^c})| + |\boldsymbol{\alpha}|\}}{\Gamma\{|\mathbf{h}(\mathbf{R}_{\{A_{\xi[-k]}\}^c})| - |\mathbf{h}(\mathbf{R}_{\{k\}})| + |\boldsymbol{\alpha}|\}}\right]^{1/J},$$

where $|\mathbf{h}(\mathbf{R}_{\{A_{\xi[-k]}\}^c})| = L - |\boldsymbol{\xi}[-k]|J$ and $|\mathbf{h}(\mathbf{R}_{\{k\}})| = J$. The approximation formula is satisfactory when $|\boldsymbol{\xi}|$ is small and $\mathbf{h}(\mathbf{R}_{\{k\}})$ is diverse.

4.3 Motif Sampler

We can take the idea illustrated in Section 4.1 one step further to derive the so-called *motif* sampler that detects and aligns more than one motif. Consider m different types of motifs of lengths $J_1, \ldots, J_m$, each occurring an unknown number of times, in a long sequence containing a total of L^* residues. Let $L = L^* - \min\{J_1, \ldots, J_m\} + 1$. We can similarly introduce the indicator vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_L)$, where $\xi_l = i$ if an element from motif i starts from position l and $\xi_l = 0$ if no elements start from this position l. Suppose that $P\ (\xi_l = i) = \varepsilon_i$, where $\varepsilon_0 + \cdots + \varepsilon_m = 1$. Given what is known about the biology of the sequences being analyzed, a crude guess n_i for the number of elements for motif i is usually possible. Let $n_0 = L - n_1 - \cdots - n_m$. We can represent this prior opinion about the number of occurrences of each type of element by a Dirichlet distribution on $(\varepsilon_0, \ldots, \varepsilon_m)$, which has the form $\text{Dir}(b_0, \ldots, b_m)$ with $b_i = n_i(L_0/L)$, where L_0 represents the "weight" (or "pseudocounts") to be put on this prior belief. Then the same predictive updating approach as illustrated in Section 4.2 can be applied. Precisely, the update formula (11) is changed to

$$\frac{\pi(\xi_k = i|\boldsymbol{\xi}[-k], A)}{\pi(\xi_k = 0|\boldsymbol{\xi}[-k], A)} = \frac{\hat{\varepsilon}_i}{\hat{\varepsilon}_0} \prod_{j=1}^{J_i} \left(\frac{\hat{\boldsymbol{\theta}}_{i,j}}{\hat{\boldsymbol{\theta}}_0} \right)^{\mathbf{h}(r_{k+j-1})},$$

where $\hat{\varepsilon}_i$ is the posterior mean estimate of the rate of occurrence of motif i excluding position k, $\hat{\boldsymbol{\theta}}_{i,j}$ is the current estimate of the residue frequency at position j of motif i, and $\hat{\boldsymbol{\theta}}_0$ has the same meaning as that in (11). More details have been provided by Neuwald, Liu, and Lawrence (1995).

5. THE RANK TEST OF SIGNIFICANCE

The assessment of significance for a sequence alignment is a known difficult problem. It is particularly difficult when the residue frequency parameters must be estimated from the data. The problem is perhaps easily addressed for standard problems where a full Bayesian analysis can provide not only a point estimate, but also the whole posterior distribution. But in our problems, the sample space of the joint alignment A is discrete with an astronomical size. In the example, the size of the space is $2^{36,000}$. The resulting posterior distribution of A is typically spiky with many local modes, even when the sequence data are randomly generated. When likelihood methods are used, the likelihood ratio test is useful only when sample sizes are larger than that frequently available for protein alignment problems. Comparison of results with multiple (typically 100) random shuffles of the data is frequently the only available tool for the assessment of significance. Here we describe a ranks test that compares the sequence to a single control data set. The only restriction on the selection of the control data set is that the lengths of the sequences used in the control be equal to those in the test data set. Next we describe the procedure for Bernoulli sampling. For multinomial sampling, the procedure is the same, except that a control sequence is appended to each test sequence.

Let $\mathbf{R} = (r_1, \ldots, r_{L^*})$ be the sequence to be analyzed, and let $\mathbf{R}' = (r'_1, \ldots, r'_{L^*})$ be a control data set of $\mathbf{R}$. For example, we may use a random shuffle of $\mathbf{R}$.

Step 1. Append the control sequence to the test sequence to form a new sequence,

$$\mathbf{R}^* = (\mathbf{R}\mathbf{R}') = \{r_1, \ldots, r_{L^*}, r'_1, \ldots, r'_{L^*}\},$$

and perform Bernoulli sampling to locate a best set of repeats; that is, the most probable $\boldsymbol{\xi}^*$ in this extended data set.

Step 2. Calculate the following log odds score (los) at each element selected by the sampler; that is, the elements selected by the set A_ξ in the extended data set. For every $k \in A_{\boldsymbol{\xi}}$,

$$\text{los}(k) = \sum_{j=1}^{J} \langle \mathbf{h}(r_{k+j-1}), \qquad \log(\hat{\boldsymbol{\theta}}_j/\hat{\boldsymbol{\theta}}_0)\rangle.$$

Here for a vector $\mathbf{v} = (v_1, \ldots, v_p), \log(\mathbf{v}) = \{\log(v_1), \ldots, \log(v_p)\}$. The notation "$\langle\cdot,\cdot\rangle$" is the usual inner product between two vectors. This score is an immediate output of the sampler. As was revealed in (11), the score is equivalent to the logarithm of the ratio of the probability of a candidate segment being an element to that of it not being an element, treating the estimated $\boldsymbol{\theta}$'s as true ones.

Step 3. Suppose that N_0 such positions are obtained. Rank these positions by decreasing los score; that is, the ith largest score; is ranked as $N_0 - i + 1$. Each rank grade is assigned a positive sign if the corresponding position is in the original data $\mathbf{R}$ and assigned a negative sign if the corresponding position is in the shuffled data $\mathbf{R}'$.

Step 4. Do a Wilcoxon signed rank test treating the foregoing signed ranks as being obtained from two paired samples. That is, calculate the mean rank and obtain a *reference p value* based on a normal approximation or on an exact table derived by Wilcoxon (1945).

The rationale behind the test is that under the null, the elements are equally likely to be solicited from either the study or the control sequences. The circumstance of our test is not the same as that of the classical Wilcoxon test, where typically two paired samples are involved. But conditioned on the total number of elements found, the same exchangeability argument as that of Wilcoxon's can be ap plied. In the case that no elements are found in the control data set, for example, the significance will be simply $2^{-(N_0-1)}$. Because the Bernoulli sampler excludes any overlap of two elements, our test is actually slightly conservative. (The sampler will be pushed towards the control data if the elements are too crowded in the test data.)

The generation of negative control data sets is a well-recognized issue in the field of computational biology. Lipman, Wilbur, Smith, and Waterman (1984) showed that some local dependence of a sequence can affect the statistical significance of a similarity. Thus one may want to use negative control data sets other than the randomly shuffled set to guard against possible artifacts. Our test method is not restricted to random shuffled sequences and can be carried through for any carefully chosen negative control data sets that accurately reflect one's null hypothesis. In this sense, our method is also open to possible

Table 2. Aligned Binding Segments Identified by the Sampler

I	*II*		*III*		*IV*	*V*
1-1	195	kvakvtgqst	CAVFGLGGVGLSVIMGCKAAGAARII	gvdinkdkfa	220	(.9600)
2-1	23	tsqeprsynk	ITVVGVGAVGMACAISILMKDLADEV	alvdvmedkl	48	(.9680)
4-1	6	agwsc	LVTGGGGFLGQRIICLLVEEKDLQEI	rvldkvfrpe	31	(.9180)
5-1	6	malqq	FGLIGLAVMGENLALNIERNGFSLTV	ynrtaektea	31	(.9840)
6-1	8	aiankni	IFVAGLGGIGFDTSREIVKSGPKNLV	ildrienpaa	33	(.9960)
8-1	41	hfstqektpq	ICVVGSGPAGFYTAQHLLKHPQAHVD	iyekqpvpfg	66	(1.0000)
8-2	179	elepdlscdt	AVILGQGNVALDVARILLTPPEHLEA	lllcqrtdit	204	(.9760)
11-1	468	mlfntdqvie	VFVIGVGGVGGALIEQIYRQQPWLKQ	khidlrvcgi	493	(.7640)
12-1	3	mk	IGIVGATGYGGTELVRILSHHPHAEE	cilysssgeg	28	(.5740)
14-1	5	mqfd	YIIIGAGSAGNVLATRLTEDPNTSVL	lleaggpdyr	30	(.9980)
21-1	298	geavarllvq	VVVLGPGRIAGPVGVEVDIVRDVEGA	hdpvqttvvv	323	(.5400)
21-2	348	hfqrglvqrl	IGIEARGRIGRAVAVALHRASRALDV	adhrqiqvia	373	(.6740)
22-1	11	mskntegmgr	AVVIGAGLGGLAAAMRLGAKGYKVTV	vdrldrpggr	36	(1.0000)
22-2	222	sqlekkfgvh	YAIGGVQAIADAMAKVITDQGGEMRL	ntevdeilvs	247	(.9840)
24-1	6	mtnir	VAIVGYGNLGRSVEKLIAKQPDMDLV	gifsrratld	31	(1.0000)
27-1	215	klgidmkkak	IAVQGIGNVGSYTVLNCEKLGGTVVA	maewcksegs	240	(.9240)
28-1	24	aadhhplplt	VGVLGSGHAGTALAAWFASRHVPTAL	wapadhpgsi	49	(.9820)
29-1	148	ngaehfkgkp	ALIVGGGGTARTAIYVLRKWLGVSKI	yivnrdakev	173	(.8240)
41-1	162	ynidtfglna	VVIGASNIVGRPMSMELLLAGCTTTV	thrftknlrh	187	(.9700)
41-2	204	nlrhhlenad	LLIVAVGKPGFIPGDWIKEGAIVIDV	ginrlengkv	229	(.5660)
42-1	7	qtfqad	LAIVGAGGAGLRAAIAAAQANPNAKI	aliskvypmr	32	(.9980)
42-2	383	glfavgecss	VGLHGANRLGSNSLAELVVFGRLAGE	qateraatag	408	(.8020)
44-1	6	msrak	VGINGFGRIGRLVLRAAFLKNTVDVV	svndpfidle	31	(.9620)
46-1	6	mqpir	LGLVGYGKIAQDQHVPAINANPAFTL	vsvatqgkpc	31	(.8000)
47-1	756	taqcfsthhf	ACLIGYGASAVCPYLALETCRQWRLS	nktlnlmrng	781	(.5500)
48-1	7	madkvn	VCIVGSGNWGSAIAKIVGANAAALPE	feervtmfvy	32	(.9200)
49-1	6	mnqva	VVIGGGQTLGAFLCHGLAAEGYRVAV	vdiqsdkaan	31	(.9700)
50-1	260	enrapsvtvc	VGHLGGLDIAERDIARLRGLGRTVSD	siavrsydev	285	(.7920)
51-1	184	hrqvdlssqk	TVIIGAGKMACLLVKHLLAKGATDIT	ivnrsqrrsq	209	(.9920)
55-1	297	knydgdvqsd	IVAQGFGSLGLMTSILVTPDGKTFES	eaahgtvtrh	322	(.7260)
56-1	5	mslr	IGVIGTGAIGKEHINRITNKLSGAEI	vavtdvnqea	30	(1.0000)
57-1	80	klldyfkndt	FALIGYGSQGYGQGLNLRDNGLNVII	gvrkdgaswk	105	(.9860)
63-1	286	itknrlsdht	VLFQGAGEAALGIANLIVMAMEKEGV	skeaavkriw	311	(.9540)
70-1	4	mdt	IAFLGLGNMGGPMAANLLKAGHRVNV	fdlqpkavLG	29	(1.0000)
70-2	38	NVfdlqpkav	LGLVEQGAQGADSALQCCEGAEVVIS	mlpagqhves	63	(.5220)
71-1	3	mk	ALHFGAGNIGRGFIGKLLADAGIQLT	fadvnqvvld	28	(.9360)
76-1	5	msnt	IVVVGAGVIGLTSALLLSKNKGNKIT	vvakhmpgdy	30	(1.0000)
77-1	119	rdgfslydrt	VGIVGVGNVGRRLQARLEALGIKTLL	cdppradrgd	144	(.9960)
78-1	5	mktq	VAIIGAGPSGLLLGQLLHKAGIDNVI	lerqtpdyvl	30	(1.0000)
79-1	167	taagkvppak	VMVIGAGVAGLAAIGAANSLGAIVRA	fdtrpevkeq	192	(1.0000)
82-1	13	ifpipaesyt	LGFIGAGKMAESIARGAVRSGVLPPS	rirtavhfnl	38	(1.0000)
83-1	290	rrlslelngr	LPIIGVGGIDSVIAAREKIAAGASLV	qiysgfifkg	315	(.5960)
85-1	9	mtflkeyv	IVSGASGFIGKHLLEALKKSGISVVA	itrdviknns	34	(.9380)
86-1	218	atdvmlagkv	AVVAGYGDVGKGSAASLKAFGSRVIV	teidpinalq	243	(1.0000)
89-1	16	kieqwkatkv	IGIIGLGDMGLLYANKFTDAGWSVIC	cdreeyydel	41	(1.0000)
sites:			******* ** * * * *			
			5 10 15 20 25			

NOTE: Column I gives the sequence number (e.g., 1, ADHE_HORSE, etc.) and the element number in each sequence. Columns II and IV provide starting and ending positions of each element. Columns III and V display the sequence segment for each element and its frequency of being sampled by the sampler. The sites found by the weighted fragmentation model are indicated by "*" at the bottom of the table. The signed ranks test was applied; the result is highly significant, with a p value 2.37×10^{-6}.

abuses. Another issue concerns the implication of our reference p value, which would perhaps have been different had another control data set been generated. Note that this kind of fluctuation is the characteristic of all Monte Carlo tests. Following Hope (1968), our reference p value based on one shuffled control set is a valid p value. That is, if we calculate that the reference p value is α, then the probability of making a type I error by rejecting the null hypothesis at level α is equal to α. Hope (1968) also pointed out that generating more reference sets can increase the power of the test.

6. EXAMPLE: DINUCLEOTIDE-BINDING PROTEINS

NAD, NADP, and FAD are cofactors used by several different enzymes. Neuwald and Green (1994) described a set of 91 distantly related proteins that are components of enzymes that bind one of these cofactors. This set contains no pair of sequences to be found significantly related using the BLAST algorithm (Altschul, Gish, Miller, Myers, and Lipman 1990). We replaced two of the proteins in this collection with two other dinucleotide-binding proteins whose three-dimensional structure is known. We made this replacement in a manner that assured there were no significantly related pairs. Several of the proteins in this set are components of enzymes complexes of several proteins. Consequently, not all of these 91 proteins are expected to directly bind one of the cofactors. Furthermore, some of these sequences may contain multiple dinucleotide binding sites. This example is thus suited to the Bernoulli sampling strategy described in Section 4.

The prior Dirichlet distributions for the frequencies of the 20 residue types for all site and nonsite positions were chosen so that the expected values of the frequencies were proportional to the composition of the entire data set. The pseudocounts (i.e., $\alpha_1+\cdots+\alpha_p$ or $\beta_{1j}+\cdots+\beta_{pj}$) were chosen as 9, 10% of the total number of sequences. Because we expected some proportion of the sequences to contain an element, we chose the parameters of the prior beta distribution in such a way that the expected number of elements was 45, corresponding to half an element per sequence, and with a standard deviation of our default value of 6, $\frac{2}{3}$% of the mean. This gave us $a = 180$ and $b = 144,984$. In this example the results do not change remarkably for different choices of (a, b) ranging from expected $\frac{1}{3}$ element per sequence to one element per sequence. Because in this example the width of the motif is unknown, we applied fragmentation using the weighted prior with the defaults of $J = 13$ and $W = 130$. We applied the signed ranks test as described in Section 5 with the same prior distributions as those described earlier, except that we modified the beta parameters to correspond to a doubling of the expected number of elements. The sampler applied to the data set including the controls reported 38 elements that were sampled 50% of the iterations after convergence. Seven of these came from the control data set with a total rank of 55, which corresponds to a Z value of -4.6 using normal approximation with correction. In all cases we judged convergence in accordance with the procedures described by Lawrence et al. (1993).

As shown in Table 2, 45 aligned elements from these 91 sequences were identified by the sampler. Table 3 gives the expected values of the a posteriori Dirichlet distributions associated with these aligned elements. Notice that the fragmentation model has selected a motif that spans 26 residues, and that this motif corresponds quite well with the strand–helix–strand motif of ADHE, as shown in Figure 3. It also corresponds to the locations of the known dinucleotide-binding strand–helix–strand structure in the two other proteins in this set for which such structural data are available. Furthermore, this motif corresponds nearly perfectly to a sequence fingerprint model for dinucleotide-binding segments derived by superimposing the structures of several of these proteins (Wierenga, DeMaeyer, and Hol 1985).

In protein motifs, not all of the positions are equally important. The last column in Table 3 gives the relative entropy (Cover and Thomas 1991) in bits (i.e., the Kullback–Leibler distance) between the site and nonsite models of the 13 sampled positions found for the motif. As these columns indicate, positions 5, 7, and 10 are the most distant from the background model. All three positions are most frequently glycines, the smallest and most flexible of the 20 amino acids. In this motif, these three positions, which are colored red in Figure 3, play important distinct roles in the binding of NAD. The flexibility of a glycine residue at position 199 of ADHE (motif position 5) permits the protein chain to make the sharp bend shown in Figure 3. The small size of a glycine at position 201 (motif position 7) permits the motif to closely approach the phosphate groups of NAD to which it binds. The small size of the glycine at position 204 (motif position 10) permits the first strand of the strand–helix–strand structure to closely approach the helix. Position 3 of the motif (position 197 in ADHE), colored purple in Figure 3, is the fourth most distant position. As indicated in Table 3, one of three residues (valine, isoleucine, or leucine) is required at this position. These three residue types are often found in the interior of a protein and help to stabilize a protein's structure. In this case this position helps to hold the first strand of this strand–helix–strand motif in position against the helix.

7. DISCUSSION

In most cases sequences available for analysis do not arise from a process that mimics independent samples from a common model, but rather they emerge by evolving from common ancestors. Some methods for sequence alignment that incorporate evolutionary history have been described (Allison et al. 1992; Bishop and Thompson 1986; Thorne et al. 1991, 1992), but the results from these in multiple sequence alignment have so far been limited. When sufficient diverse data are available, as in our dinucleotide-binding protein example, the independence assumption can be very closely achieved by removing highly correlated observations from the data set. In cases where there are insufficient data to take this approach, a method that can simultaneously align a large set of sequences (especially sequences that include subtle relationships) and account for the correlations stemming from evolutionary history awaits future development. Although theory addressing these correlations is important, in practice we have found that the methods we propose here often work well even with substantial departures from the independence assumption. To date, our experience and that of others using these methods (S. Henikoff, personal communication) indicate that departures from the independence assumptions have nearly no impact on the methods described in Section 3. Although we have less experience with the method described in Section 4 (i.e., Bernoulli sampling methods), preliminary results indicate that this assumption is somewhat more important in this case.

Some basic similarities of this work with previously reported methods should be noted. Existing Gibbs sampling and EM algorithms for multiple sequence alignment share the treatment of alignment data as missing. Two EM approaches have been developed: one that finds alignments of ungapped elements (Cardon and Stormo 1992; Lawrence and Reilly 1990), the other in the form of hidden Markov models (HMM) that permit gaps between any two neighboring residues in a sequence (Baldi et al. 1994; Haussler et al. 1993). Because the joint distribution of multiple elements in each sequence must be explicitly estimated, the first approach has been limited to problems with only one or two motifs. The sampling methods permit us to escape this restriction by exploring the joint distribution through sampling the complete set of conditionals.

When sequence evolution has not been subject to duplications or transpositions, the order of the residues in the sequences will not be altered, and the sequences are said to be collinear. This collinearity induces a Markov relationship on the alignment, which forms the basis of a popular

Table 3. Posterior Mean (× 100) of the Product Multinomial Parameter

POS	*C*	*G*	*A*	*S*	*T*	*N*	*D*	*E*	*Q*	*K*	*R*	*H*	*W*	*Y*	*F*	*V*	*I*	*L*	*M*	*P*	*Inf*
1	2	1	13	1	2	0	0	1	0	0	0	0	0	4	4	29	25	15	0	0	1.1
2	6	27	17	1	2	0	0	1	0	0	0	0	0	0	4	23	3	9	2	2	.9
3	0	1	3	3	2	0	0	1	0	0	0	4	0	0	6	27	37	13	0	0	1.4
4	0	11	5	1	0	2	0	3	6	0	0	2	0	0	4	25	27	11	0	0	.9
5	0	82	9	1	0	0	0	3	0	0	0	0	0	0	0	1	1	1	0	0	2.5
6	0	9	25	11	2	0	0	1	4	0	2	0	0	10	4	13	3	11	0	2	.6
7	0	78	3	1	2	4	0	1	4	0	0	0	0	0	0	1	1	3	0	0	2.3
9	0	3	13	5	2	0	0	1	4	0	0	0	2	2	0	17	25	11	10	2	.8
10	0	70	21	1	0	0	3	1	0	0	0	0	0	0	0	1	1	1	0	0	2.2
14	2	9	39	7	0	0	0	3	2	0	0	2	0	2	0	5	17	9	0	0	.9
17	6	1	11	1	0	0	0	3	0	0	0	0	0	0	4	7	21	43	0	0	1.3
21	0	43	11	3	0	6	3	1	2	5	5	4	0	0	0	1	1	5	2	9	1.0
26	2	1	11	9	7	0	5	7	2	0	0	0	0	0	0	29	15	11	0	0	.8
Nonsite:	1	7	8	5	5	4	5	6	3	5	5	2	1	3	3	7	5	9	2	4	
Site:	1	27	14	2	1	1	0	1	2	0	0	1	0	1	2	14	14	11	1	1	

NOTE: This table shows the posterior expectation of the residue frequency parameter conditioned on the best alignment. The information number provided in the last column is the Kullback–Leibler distance between the estimated frequency $\hat{\theta}_j$ for each position of the motif and the estimated background frequency $\hat{\theta}_0$, calculated using base-2 logarithm. By the standard theory that -2 log of likelihood ratio is asymptotically χ^2 distributed, the null distribution (i.e., if the 45 elements were randomly generated from the background model) of this information number is asymptotically $\chi^2(19)/90 \log(2)$, whose mean and standard deviation are about .3 and .1.

dynamic programming algorithm for the alignment of a pair of sequences. HMM's exploit this same characteristic to explore the joint alignment distribution, which allows gaps anywhere in the sequence. While increasing their modeling flexibility, the unrestricted HMM's have to pay the price of a loss in specificity and an increase in computational complexity. Of course, their use of collinearity restricts them to problems where the orders of the motif elements have not been transposed. The sampling methods discussed here do not have this restriction. All currently available HMM algorithms are based on the EM approach, which, because of its deterministic nature, can become trapped in a local mode. The ideas presented in this article can be useful for building full Bayesian models and Gibbs sampling strategies for the HMM's.

Choosing the pattern width J in our analysis corresponds to the well-known problem of model selection when there are changes in dimensionality of the freely adjustable parameters. The celebrated Akaike information criterion (AIC) (Akaike 1973) and Bayesian information criterion (BIC) (Schwartz 1978) have been useful in many applications. Unfortunately, these criteria did not perform well in selecting those pattern widths that identified correct alignments in data sets with experimentally determined "known solutions." When fragmentation is not used, the "information-per-parameter" criterion proposed by Lawrence et al. (1993) still seems to perform well, although this criterion biases toward elements with strong ends. When fragmentation is used, our experiences suggest that a single default setting of $J = 13$ and $W = 130$ works well for most protein alignment problems. Clearly, this single default is able to identify most motifs because (a) 13 columns are sufficient to identify most protein motifs, (b) the weighting scheme in Section 3.5 effectively eliminates the need to constrain W, and (c) the single default focuses the sampler's attention on the J positions that are farthest in relative entropy from the background frequencies. Consequently, only those positions in a motif with relatively small entropy distance from the background frequencies are excluded from the model. In some circumstances a biologist will have prior information to choose a value for J that differs from the default.

There are several biopolymer sequence data bases. Records in these data bases begin with the entry of the biopolymer sequence. Additional information about the biopolymer is sometimes added to the record as annotations. Among other things, these annotations can conclude experimental evidence that a protein binds a specific ligand and evidence of the location of the binding site in the sequence, which usually requires a known three-dimensional structure. Because these data bases are now in a state of rapid growth, the state of characterization for many sequences is very limited. Thus, as is typical, it is not possible to verify the binding of a cofactor for most of the 91 sequences in our example, much less to identify the specific sites of binding. In this example the best evidence for convergence to the "correct" alignment is the excellent agreement of our motif with the fingerprint model reported by Wierenga et al. (1985).

There are a few approximation formulas throughout Sections 3 and 4. They are not necessary, and the corresponding exact formulas are provided. These approximations speed computation somewhat, but more important, they provide statistical insights into the meanings of iterative sampling procedures.

The models discussed in this article can be viewed as mixture models, and the problem addressed by the Bernoulli sampler can be regarded as a classification problem. More precisely, we can treat each position in every biological sequence as sampling from a mixture of "element" and "nonelement." Our experience demonstrates that this simple viewpoint is especially constructive. Gibbs sampling methods have been shown to be useful for these models and have been particularly constructive for the identification of subtle relationships (Lawrence et al. 1993) among protein sequences. Here we provide the rigorous Bayesian foundations for the application of these sampling methods for multiple sequence alignment. In addition, we show how to extend these methods by relaxing the specification of the number and size of motifs. Furthermore, we develop a ranks test to judge the statistical significance of a multiple alignment.

APPENDIX: WEIGHTED SAMPLING FROM FINITE POPULATION

The conditional distribution (7) can be regarded as a generalized Bernoulli–Laplace model; that is, the one for drawing a sample of size J from a pool of W weighted balls without replacement, such that the joint probability of the J balls is proportional to the product of their respective weights. Chen, Dempster, and Liu (1994) explained how this is a maximum entropy model and presented efficient algorithms to generate the sample sequentially. Because our use of the model is nested in a general Gibbs sampling procedure, an iterative scheme is sufficient.

The sampling strategy that we designed is equivalent to a Markov chain evolving in the following way. Let $\Gamma_t = \{i_1, \ldots, i_J\}$ denote a sample of J balls at step t. At step $t+1$, a ball (say, i_1) is first drawn from Γ_t with equal probability and excluded from the sample Γ_t; then a ball (say, i^*) is picked from the pool $\Gamma_t^c \cup \{i_1\}$ with probability proportional to its weight. So $\Gamma_{t+1} = \Gamma_t \setminus \{i_1\} \cup \{i^*\}$. This Markov chain is reversible with π^* as its equilibrium distribution and is related to the Metropolis scheme of Chen et al. (1994).

Instead of excluding a ball from the current sample Γ_t completely at random, we may want to draw one in proportional to g_i^{-1} for $i \in \Gamma_t$. We call this strategy *doubly proportional sampling.* The equilibrium distribution of this new strategy is found to be

$$\pi^\dagger(\boldsymbol{\Delta}) \propto \prod_{i:\delta_i=1} g_i^2 \left(\sum_{w=1}^{W} \delta_w g_w^{-1} \right).$$

To see why, consider a configuration $\mathbf{Y} = \{1, \ldots, J\}$ and let $\mathbf{X}_{ik} = \mathbf{Y} \setminus \{i\} \cup \{k\}$, where $i \le J$ and $k > J$. So $\mathbf{X}_{ik}$ differs from $\mathbf{Y}$ by one element. The transition functions are

$$T(\mathbf{Y}, \mathbf{Y}) = \sum_{i=1}^{J} \left\{ \frac{g_i^{-1}}{\sum_{j=1}^{J} g_j^{-1}} \right\} \left\{ \frac{g_i}{\sum_{k=J+1}^{W} g_k + g_i} \right\}$$

and

$$T(\mathbf{X}_{ik}, \mathbf{Y}) = \left\{ \frac{g_k^{-1}}{\sum_{j=1}^{J} g_j^{-1} - g_i^{-1} + g_k^{-1}} \right\} \left\{ \frac{g_i}{\sum_{l=J+1}^{W} g_l + g_i} \right\}.$$

Then it is easy to check

$$\sum_{i=1}^{J} \sum_{k=J+1}^{W} \pi^\dagger(\mathbf{X}_{ik}) T(\mathbf{X}_{ik}, \mathbf{Y}) + \pi^\dagger(\mathbf{Y}) T(\mathbf{Y}, \mathbf{Y}) = \pi^\dagger(\mathbf{Y}),$$

so that the $\pi^\dagger$ is indeed invariant under the foregoing transition.

A general doubly proportional sampling chain $M_{\alpha,\beta}$ can be defined as follows: At step $t+1$, a ball (say, i_k) from $\Gamma_t = \{i_1, \ldots, i_J\}$ is chosen to be excluded with probability proportional to the $g_{i_k}^{\alpha}$; then a ball is drawn from the pool $\Gamma_t^c \cup \{i_k\}$ with probability proportional to the g_j^{β}, for $j \in \Gamma_t^c \cup \{i_k\}$. The equilibrium distribution of this procedure is

$$\pi^\ddagger(\Gamma) \propto \prod_{j\in\Gamma} \{g_j^{\beta-\alpha}\} \left\{ \sum_{j\in\Gamma} g_j^{\alpha} \right\},$$

as can be verified by examination of the equilibrium equation as shown previously. By choosing different α and β, we were able to obtain different equilibrium distributions. A further generalization can be easily made to the case when two sets of weights, $\{a_1, \ldots, a_J\}$ and $\{b_1, \ldots, b_J\}$, are used for the two proportional sampling steps.

[Received January 1994. Revised April 1995.]

REFERENCES

Akaike, H. (1973), "Information Theory and the Maximum Likelihood Principle," in *2nd International Symposium on Information Theory,* eds. B. N. Petrov and F. Csàki, Budapest: Akademiai Kiàdo.

Allison, L., Wallace, C. S., and Yee, C. N. (1992), "Minimum Message Length Encoding Evolutionary Trees and Multiple Alignment," in *Proceedings of 25th Hawaii International Conference on System Science,* pp. 663–674.

Altschul, S. F., Gish, W., Miller, M., Myers, E. W., and Lipman, D. J. (1990), "Basic Local Alignment Search Tool," *Journal of Molecular Biology,* 215, 403–410.

Baldi, P., Chauvin, Y., McClure, M., and Hunkapiller, T. (1994), "Hidden Markov Models of Biological Primary Sequence Information," in *Proceedings of the National Academy of Science of USA,* 91, 1059–1063.

Berg, O. G., and von Hipple, P. H. (1987), "Selection of DNA Binding Sites by Regulatory Proteins: Statistical-Mechanical Theory and Application to Operators and Promoters," *Journal of Molecular Biology,* 193, 153–161.

Bishop, M. J., and Thompson, E. A. (1986), "Maximum Likelihood Alignment of DNA Sequences," *Journal of Molecular Biology,* 190, 159–165.

Bryant, S. H., and Lawrence, C. E. (1991), "The Frequencies of Ion Pair Substructures in Proteins Is Quantitatively Related to Electrostatic Potentials: A Statistical Model for Nonbonded Interactions," *PROTEINS: Structure, Function, and Genetics,* 9, 108–119.

——— (1993), "An Empirical Energy Function for Threading Protein Sequence Through the Folding Motif," *PROTEINS: Structure, Function, and Genetics,* 16, 92–112.

Cardon, L. R., and Stormo, G. D. (1992), "An Expectation Maximization Algorithms for Identifying Protein Binding Sites With Variable Gaps From Unaligned DNA Fragments," *Journal of Molecular Biology,* 223, 159–170.

Chen, R., and Liu, J. S. (1996), "Predictive Updating Methods With Applications to Bayesian Classification," *Journal of the Royal Statistical Society,* Ser. B, to appear.

Chen, X., Dempster, A. P., and Liu, J. S. (1994), "Weighted Finite Population Sampling to Maximize Entropy," *Biometrika,* 457–469.

Cover, T. M., and Thomas, J. A. (1991), *Elements of Information Theory,* New York: John Wiley.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association,* 85, 398–409.

Gelman, A., and Rubin, D. B. (1992), "Inference From Iterative Simulation Using Multiple Sequences," *Statistical Science,* 7, 457–472.

Haussler, D., Krogh, A., Mian, S., and Sjolander, K. (1993), "Protein Modeling Using Hidden Markov Models: Analysis of Globins," in *Proceedings of the Hawaii International Conference on System Sciences,* Los Alamitos, CA: IEEE Computer Society Press.

Hope, A. C. A. (1968), "A Simplified Monte Carlo Significance Test Procedure," *Journal of the Royal Statistical Society,* Ser. B, 30, 582–598.

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, 208–314.

Lawrence, C. E., and Reilly, A. A. (1990), "An Expectation Maximization Algorithm for the Identification and Characterization of Common Sites in Unaligned Biopolymer Sequences," *PROTEINS: Structure, Function, and Genetics,* 7, 41–51.

——— (1992), "Likelihood Inferences With Uncertain Indices With Application to Gene Regulation," Technical Report 121, Biometrics Laboratory, Wadsworth Center for Laboratories and Research, New York State Department of Health.

Li, K.-H. (1988), "Imputation Using Markov Chains," *Journal of Statistical Computation and Simulation,* 30, 57–79.

Lipman, D. J., Wilbur, W. J., Smith, T. F., and Waterman, M. S. (1984), "On the Statistical Significance of Nucleic Acid Similarities," *Nucleic Acids Research,* 12, 215–326.

Liu, J. S. (1994), "The Collapsed Gibbs Sampler in Bayesian Computations With Applications to a Gene Regulation Problem," *Journal of the American Statistical Association*, 89, 958–966.

Liu, J. S., Wong, W. H., and Kong, A. (1994), "Covariance Structure of the Gibbs Sampler With Applications to the Comparisons of Estimators and Augmentation Schemes," *Biometrika*, 81, 27–40.

——— (1995), "Covariance Structure and Convergence Rate of the Gibbs Sampler With Various Scans," *Journal of the Royal Statistical Society*, Ser. B, 56, 157–169.

Neuwald, A. F., and Green, P. (1994), "Detecting Patterns in Protein Sequences," *Journal of Molecular Biology*, 239, 698–712.

Neuwald, A. F., Liu, J. S., and Lawrence, C. E. (1995), "Sequence Analysis Using a Motif Sampling Strategy," submitted to *Protein Science*.

Pohl, F. M. (1971), "Empirical Protein Energy Maps," *Nature: New Biology*, 234, 277–379.

Schwartz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distribution by Data Augmentation," (with discussion), *Journal of the American Statistical Association*, 82, 528–550.

Thorne, J. L., Kishino, H., and Felsenstein, J. (1991), "An Evolutionary Model for Maximum Likelihood Alignment of DNA Sequences," *Journal of Molecular Evolution*, 33, 114–124.

——— (1992), "Inching Toward Reality: An Improved Likelihood Model of Sequence Evolution," *Journal of Molecular Evolution*, 34, 3–16.

Tierney, L. (1995), "Markov Chains for Exploring Posterior Distributions (with discussion)," *The Annals of Statistics*, 22, 1701–1762.

Wierenga, R. K., DeMaeyer, M. C. H., and Hol, W. G. J. (1985), "Interaction of Pyrophosphate Moieties With Alpha-Helixes in Dinucleotide Binding Proteins," *Biochemistry*, 24, 1346–1357.

Wilcoxon, F. (1945), "Individual Comparisons by Ranking Methods," *Biometrics*, 1, 80–83.