BIOINFORMATICS

Vol. 14 no. 1 1998
Pages 25–39

Bayesian adaptive sequence alignment algorithms

Jun Zhu[1], Jun S. Liu[2] and Charles E. Lawrence[1,3]

[1]Wadsworth Center for Laboratories and Research, Albany, NY and [2]Department of Statistics, Stanford University, Stanford, CA, USA

Received on July 10, 1997; revised on October 2, 1997; accepted on October 28, 1997

Abstract

The selection of a scoring matrix and gap penalty parameters continues to be an important problem in sequence alignment. We describe here an algorithm, the 'Bayes block aligner, which bypasses this requirement. Instead of requiring a fixed set of parameter settings, this algorithm returns the Bayesian posterior probability for the number of gaps and for the scoring matrices in any series of interest. Furthermore, instead of returning the single best alignment for the chosen parameter settings, this algorithm returns the posterior distribution of all alignments considering the full range of gapping and scoring matrices selected, weighing each in proportion to its probability based on the data. We compared the Bayes aligner with the popular Smith–Waterman algorithm with parameter settings from the literature which had been optimized for the identification of structural neighbors, and found that the Bayes aligner correctly identified more structural neighbors. In a detailed examination of the alignment of a pair of kinase and a pair of GTPase sequences, we illustrate the algorithm's potential to identify subsequences that are conserved to different degrees. In addition, this example shows that the Bayes aligner returns an alignment-free assessment of the distance between a pair of sequences.

Availability: *Software is available at http://www.wadsworth.org/res&res/bioinfo/*

Contact: *junzhu, lawrence@wadsworth.org, jliu@stat.stanford.edu*

1. Introduction

Biopolymer sequence alignment is playing an increasingly important role in biomedical research. For example, the alignment of the product of a putative human colon cancer gene with a yeast mismatch repair gene played a valuable role in its identification and characterization (Bronner *et al.*, 1994; Papadopoulos *et al.*, 1994). Numerous exact algorithms for the alignment of pairs of sequences have been developed. As the name implies, global alignment algorithms find the best alignment of the entire lengths of a pair of sequences (Needleman and Wunsch, 1970). However, it is often the case that the two biopolymers share only a common substructure. Thus, the development of local alignment algorithms which identify and align the best common subsequence, and thereby exclude unrelated termini, was an important advance (Smith and Waterman, 1981; Goad and Kanehisa, 1982; Sellers, 1984; Lipman and Pearson, 1985; Pearson and Lipman, 1988; Altschul *et al.*, 1990). Further progress along these lines is warranted since, in distantly related sequences, internal segments in one protein may have no conserved counterparts in the others.

The need for specification of gap penalty parameters and scoring matrices is a serious limitation of most current alignment algorithms. These input parameter settings can strongly influence the alignment. Several authors have addressed the issue of using multiple scoring matrices (Schwartz and Dayhoff, 1977; Collins *et al.*, 1988). Altschul developed an information theoretic approach to the selection of scoring matrices (Altschul, 1991), and an alignment scoring system sensitive at all evolutionary distances (Altschul, 1993). Methods for improving DNA alignment using application-specific scoring matrices have been described (States *et al.*, 1993). A Bayesian model for measuring evolutionary distance using optimal ungapped DNA alignments has been presented by Agarwal and States (1996). Several studies have addressed the choice of gap penalty parameters (Waterman *et al.*, 1992; Pearson, 1995). For DNA sequence comparisons, systematic search procedures for finding the best gap penalty parameters have been developed (Waterman *et al.*, 1992; Waterman, 1994). Statistics-based iterative algorithms which find at least locally optimal gap and mismatch penalties for DNA alignments have also been developed (Thorne *et al.*, 1991, 1992; Allison *et al.*, 1992). However, the large size of the scoring matrices makes these approaches difficult for protein sequence alignment. Among the approaches taken to address gaps, the algorithm of Sankoff (1972) is of particular interest. This algorithm bypasses the need to specify gap penalties through the use of constrained optimization. Specifically, this algorithm finds the optimal

[3] *To whom correspondence should be addressed at: Biometrics Lab, Wadsworth Center for Laboratories and Research, Albany, USA*

alignment subject to the constraint that there are no more than k aligned blocks (or, equivalently, $k + 1$ total gaps). It has the additional feature that those portions of the sequences that are not included in the aligned blocks are completely ignored, thereby extending the concept of local alignment. However, it does so at the price of an even more vexing problem: the requirement for specification of the number of gaps. Furthermore, like other alignment algorithms, it requires the specification of a scoring matrix. Practical methods that simultaneously address gapping and the selection of scoring matrices for protein sequences require further investigation.

Bayesian inference methods provide the means to overcome the requirement of setting parameters and for making inferences on all unknown variables. A preliminary report describing this approach has been presented by Zhu *et al.* (1997). Here we present a far more complete description which includes new algorithms, a comparison with existing methods, new inferences, and a delineation of distinctive features of these Bayesian alignments. Bayesian statistics rests on the premise that all the variables (observed data and the unknowns) in an inference problem are random variables. A statistical model which seeks to approximate reality is then specified in the form of the joint distribution of all of the variables. The standard procedure is to specify this distribution as the product of the likelihood function, the probability of the data given the unknowns, and a prior distribution for the unknowns as follows:

$$\text{joint} = \text{likelihood} * \text{prior}$$
$$P(data,unknown) = P(data|unknown)\ P(unknown)$$

Since our interest is focused on the unknowns, it is useful to rewrite the joint distribution another way:

$$P(unknown|data)\ P(data) = P(data,unknown).$$

Now inferences about the unknowns after considering the data are described by the conditional distribution of the unknowns given the data, obtained by using Bayes rule:

$$P(unknown|data) = \frac{P(data,unknown)\ P(unknown)}{P(data)}$$

where $P(data) = \int P(data|unknowns)\ P(unknowns)\ d(unknowns)$. When there are several unknowns, we integrate the posterior distribution further to obtain:

$$P(one\ unknown|data).$$

The more interesting and less rigorous aspect of this problem is in the specification of likelihood and prior functions. This modeling aspect requires the selection of functions which we believe will reasonably approximate the underlying reality. While both the likelihood and the priors must be modeled in this way, the specification of the priors is often more controversial. Several measures are available to address the uncertainties in this modeling process. Priors which incorporate little, if any, modeling preferences can be employed. Such priors are often referred to as uninformed priors. More than one model can be specified to describe reality. The selection of one of these models from among several alternatives is achieved by embedding the alternatives into one unified model and finding the posterior distribution for these alternatives (Box, 1980; Gelman *et al.*, 1995; Lawrence, 1997).

While it is often easy to write down expressions for the desired posterior distributions, the computation of the high dimensional integrations or summations (for discrete variables) to make inferences is often difficult or impossible. The major technical difficulty arises from the computational complexity of this task. Often approximations are required. In fact, exact solutions or good approximations were so rare until recently that Bayesian statistics was a field of interest only to specialists. Bayesian statistics has become much more popular in the last decade with the advance of sampling methods, such as the Gibbs sampler, which can often yield good approximations efficiently. On the rare occasions when the needed summation or integrations can be completed without employing approximations, as with the present case, the resulting posterior distributions are said to be exact. These exact solutions are, of course, exact solutions to approximate models, just as dynamic programming methods yield a guaranteed optimum for approximate models.

Here we show how pairwise alignment can be formulated as a Bayesian inference problem. By employing a modification of the optimal alignment algorithm of Sankoff (1972), we show how to complete the large summation over all possible alignments with a time complexity of $O(N^2)$. Using this sum, we show how Bayesian statistics can be used to overcome the need to set gap penalty parameters or to choose a scoring matrix. Furthermore, the 'evidence' against the null hypothesis, the Bayesian analog of the classical P value, is computed exactly. We compare this method with a popular Smith–Waterman-based local alignment procedure, SSEARCH (Pearson, 1991), and illustrate the distinctive features of our method with two applications.

2. Methods

2.1. Likelihood

Consider a pair of sequences $R^{(1)} = \{R_1^{(1)} \ldots R_I^{(1)}\}$ and $R^{(2)} = \{R_1^{(2)} \ldots R_J^{(2)}\}$, their alignment can be characterized by a matrix of indicator variables $\boldsymbol{A}_{i,j}$. If $R_i^{(1)}$ is aligned with $R_j^{(2)}$, $\boldsymbol{A}_{i,j} = 1$, otherwise $\boldsymbol{A}_{i,j} = 0$. As each residue in one sequence can relate to at most one residue in the other sequence, we require $\sum_i \boldsymbol{A}_{i,j} \leq 1$ and $\sum_j \boldsymbol{A}_{i,j} \leq 1$. The logarithm of the likelihood for the joint distribution realization of the pair of sequences is:

$$\log P(R_1^{(1)}, R_j^{(2)}|\boldsymbol{\Theta}, \boldsymbol{\Psi}, \boldsymbol{A}) = \boldsymbol{\Theta}_{R_i^1} + \boldsymbol{\Theta}_{R_j^2} + \boldsymbol{A}_{i,j}\boldsymbol{\Psi}_{R_i^{(1)}, ,R_j^{(2)}} \quad (1)$$

where $\boldsymbol{\Psi}_{R_i^{(1)},R_j^{(2)}}$ is a matrix of the logarithm of residue interactions, i.e., an alignment score matrix, such as PAM (Dayhoff *et al.*, 1972) or BLOSUM (Henikoff and Henikoff, 1992); $\boldsymbol{\theta}_{R_j}$ is the log marginal probability of observing residue type R_j associated with the selected score matrices. These marginal terms are usually ignored in alignment algorithms since they are constants with respect to the alignment.

Two simplifying assumptions are generally made by alignment algorithms: (i) the parameters θ and ψ are known and fixed; (ii) transpositions are not allowed. The second assumption adds the condition of colinearity which requires that if $\boldsymbol{A}_{i,j} = 1$, then:

$$\boldsymbol{A}_{i+\Delta,j-\delta} = \boldsymbol{A}_{i-\Delta,j+\delta} = 0 \qquad (2)$$

where for all $\Delta, \delta > 0$.

An alignment is composed of aligned segments, called blocks, in both sequences interspersed with unrelated subsequences from one or both sequences, called gaps. Since a gap may span an unrelated subsequence from not only either sequences, but both sequences, its meaning is somewhat broader than is common in this field. A block of length m is determined by three indices i, j and m, satisfying the condition:

$$\begin{aligned} &\boldsymbol{A}_{i,j} = 0, \boldsymbol{A}_{i+m+1,j+m+1} = 0 \\ &\boldsymbol{A}_{i+l,j+l} = 1 \text{ for } l = 1, 2, \ldots, m \end{aligned} \qquad (3)$$

If no additional constraints are given, alignments based on this likelihood yield biologically unrealistic solutions which contain many short alignment segments with far too many gaps. The most popular alignment algorithms (Needleman and Wunsch, 1970; Smith and Waterman, 1981) address this difficulty by adding a gap penalty term to the model given in equation (1). Here we take the alternative path first described by Sankoff (1972) and seek alignments with at most $(k - 1)$ internal gaps.

2.2. Joint probabilities and priors

The joint distribution, which as described above is central to a Bayesian approach, is specified as follows:

$$\text{Joint} = \text{likelihood} * \text{priors}$$
$$P(R^{(1)},R^{(2)},\boldsymbol{A},k,\psi) = P(R^{(1)},R^{(2)} \mid \boldsymbol{A},k,\psi)P(\boldsymbol{A} \mid k)P(k)P(\psi)$$

where k is the number of alignment blocks and we assume that ψ is independent of k or $\boldsymbol{A}$ *a priori*.

As described above, the priors can be used to aid in modeling through the incorporation of previous experience or biological information. The use of such informed priors would be analogous to setting the parameters of an optimal alignment algorithm based on general experience or experience specifically tuned to the problem at hand. Although there have been a number of papers reporting results of investigations to identify ‘good’ parameter values, we have little guide for choosing such an informed prior specification because of the novelty of the Bayesian approach.

Lacking *a priori* information, we employ uninformed priors. Specifically, we assume all possible score matrices are equally likely, i.e. $P(\psi) = \frac{1}{N_\psi}$ where N_ψ is the number of scoring matrices in the series. We further assume, except when calculating Bayesian evidence as described below, that all possible numbers of matching blocks are equally likely, $P(K = k) = \frac{1}{\boldsymbol{\kappa} + 1}$, $k = 0, 1, \ldots \kappa$, where κ is the maximum number of blocks. As there are only a limited number of common motifs in distantly related sequences, by default we set:

$$\boldsymbol{\kappa} = \min\{\frac{L_s}{10}, 20\}$$

where L_s is the length of the shorter sequence. Finally, we assume that all alignments with blocks are equally likely, i.e. $P(\boldsymbol{A} \mid k) = \frac{1}{N_k}$, where N_k is the number of alignments with k blocks.

2.3. Posteriors

As described above, inferences on the number of gaps and on the scoring matrices are made by examining the conditional posterior distributions:

$$P(k|R^{(1)}, R^{(2)}) = \frac{\sum_{\Psi}\sum_{A} P(R^{(1)}, R^{(2)}|A, k, \Psi)P(A|k)P(k)P(\Psi)}{\sum_{k}\sum_{\Psi}\sum_{A} P(R^{(1)}, R^{(2)}|A, k, \Psi)P(A|k)P(k)P(\Psi)} \qquad (4)$$

and

$$P(\boldsymbol{\Psi}|R^{(1)}, R^{(2)}) = \frac{\sum_{k}\sum_{A} P(R^{(1)}, R^{(2)}|A, k, \Psi)P(A|k)P(k)P(\Psi)}{\sum_{k}\sum_{\Psi}\sum_{A} P(R^{(1)}, R^{(2)}|A, k, \Psi)P(A|k)P(k)P(\Psi)} \qquad (5)$$

Here, ψ takes values on a finite set of scoring matrices, e.g. the PAM or BLOSUM series.

2.4. Bayesian evidence

In Bayesian statistics, the ‘evidence’ for the null hypotheses is obtained by examining the posterior probabilities of all alternative models. If there are no blocks, $k = 0$, then $\boldsymbol{A}_{i,j} = 0$ for $i = 1, 2, \ldots I$; $j = 1, 2, \ldots, J$. This result indicates that the two sequences have no pairs of residues in common and thus are unrelated to one another. The alternative is the set of all alignments which have at least one block of related residues. However, in a ‘hypothesis testing setting’, one is not interested in the evidence for the null compared with a specific prior specification for the alternative (even an uninformed specification for the prior as above), but rather against a diffuse alternative. Thus, Bayesian statisticians commonly define the Bayesian evidence considering all possible priors, over alternative models. In accordance with this convention, the Bayesian evidence that the two sequences are not related is:

$$1 - \sup_{P(K)} \{P(K > 0 \mid R^{(1)}, R^{(2)})\} \qquad (6)$$

where the supremum is taken over all prior distributions on K such that $P(K = 0) = 1 - P(K) > 0) = \pi_0$. Typically, if an investigator is aligning two specific sequences of interest, then he or she probably expects that there is at least a 50% chance they are related, accordingly a value of $\pi_0 = 0.5$ is conservative. When employed in a database search framework $1 - \pi_0$ is set to reflect the *a priori* chance that the query sequence is similar to a sequence taken at random from the database. It is well known that Bayesian evidence tends to be conservative compared to the classical P value for some classes of univariate distributions (Berger and Sellke, 1987). Zhu *et al.* (1997) show that when applied to randomly shuffled sequences, Bayesian evidence is very conservative compared with classical P values.

3. Algorithms

Here we present an algorithm for completing the sums in equations (4) and (5) and an algorithm for counting the number of alignments. Also, we present algorithms for direct sampling from the joint posterior alignment distribution and an algorithm which finds the exact marginal posterior alignment distribution. We also give an algorithm for identifying all alignment within δ of the optimal alignment.

3.1. Completing the sums

To obtain the posterior distributions shown in equations (4) and (5), we need to sum out variables k, ψ and $\boldsymbol{A}$. Sums over the small number of blocks, k, and the small number of scoring matrices in a series, ψ, can be completed by direct enumeration. A recursive algorithm for completing sums over the large number of alignments is given below. To complete these sums, this algorithm recursively builds a series of partial sums considering one residue or matched pair of residues at a time. At each iteration of this recursion, the partial sum, up to residue i in sequence 1 and residue j in sequence 2 with t blocks, contains three components. These components correspond to the following last steps: (i) a match of residue $R_i^{(1)}$, $R_j^{(2)}$, denoted by ↘ as the last step, yields the partial sum $PC_{i,j}^{(t)}$; (ii) an insertion in sequence 1 (deletion in sequence 2), denoted by ↓ as the last step, yields the partial sum $PD_{i,j}^{(t)}$; and (iii) an insertion in sequence 2 (deletion in sequence 1) denoted by → as the last step, yields the partial sum $PC_{i,j}^{(t)}$. In order to count each distinct assignment of the variables in the alignment matrix $\boldsymbol{A}$ only once, we impose the rule that a deletion in sequence 1, a → move, cannot be followed by an insertion in sequence 1, a ↓ move. We complete the desired sums recursively as follows:

(i) If the last step is a match ($\boldsymbol{A}_{i,j} = 1$), then $PC_{i,j}^{(t)}$ is dependent only on the partial sums with indices $(i - 1, j - 1)$. It extends the matching block without introducing any new matching blocks if the previous step is ↖. If the previous step is → or ↓, the last move introduces a new matching block. Accordingly,

$$PC_{i,j}^{(t)} = [PC_{i-1,j-1}^{(t)} + PD_{i-1,j-1}^{(t-1)} + PR_{i-1,j-1}^{(t-1)}] * \exp(\Psi_{R_i^{(1)},R_j^{(2)}})$$

(ii) The partial sums with indices $(i - 1, j)$ are required for a ↓ move as the last move. The counting rule requires no → move can be followed by a ↓ move. Furthermore, the last move does not introduce any new matching block, and $\boldsymbol{A}_{i,j} = 0$, so:

$$PD_{i,j}^{(t)} = PC_{i-1,j}^{(t)} + PD_{i-1,j}^{(t)}$$

(iii) The partial sums with indices $(i - 1, j)$ are required for a → move, and the last move does not introduce any matching block, and $\boldsymbol{A}_{i,j} = 0$, so:

$$PR_{i,j}^{(t)} = PC_{i,j-1}^{(t)} + PD_{i,j-1}^{(t)} + PR_{i,j-1}^{(t)}$$

The boundary values are set as following $PC_{i,j}^{(0)} = 0$, $PD_{i,j}^{(0)} = 0$ $PR_{i,j}^{(0)} = 1$, for $j \neq 0$; $PC_{i,0}^{(0)} = 0$, $PD_{i,0}^{(0)} = 1$ and $PR_{i,0}^{(0)} = 0$ for all i; $PC_{i,0}^{(t)} = 0$, $PD_{i,0}^{(t)} = 0$, $PR_{i,0}^{(t)} = 0$, $PC_{0,j}^{(t)} = 0$, $PD_{0,j}^{(t)} = 0$ and $PR_{0,j}^{(t)} = 0$ for $t \neq 0$. The time complexity for completing these sums is $O(IJ)$.

3.2. Counting alignments

The total number of alignments $N_{I,J}^{(k)}$ may also be obtained recursively. Let $NC_{i,j}^{(t)}$ be the number of alignments from pair (1,1) to pair (i,j) with t matching blocks and ↘ as the last step, $ND_{i,j}^{(t)}$ be the number alignments with ↓ as the last step, $NR_{i,j}^{(t)}$ be the number of alignments with → as the last step. Using argument analogous to those in Section 3.1, the following recursive relationships hold:

$$NC_{i,j}^{(t)} = NC_{i-1,j-1}^{(t)} + ND_{i-1,j-1}^{(t-1)} + NR_{i-1,j-1}^{(t-1)}$$
$$ND_{i,j}^{(t)} = NC_{i-1,j}^{(t)} + ND_{i-1,j}^{(t)}$$
$$NR_{i,j}^{(t)} = NC_{i,j-1}^{(t)} + ND_{i,j-1}^{(t)} + NR_{i,j-1}^{(t)}$$

The boundary conditions are the same as those in Section 3.1. The conditional probabilities described in Section 2 can now be calculated using the algorithms of Sections 3.1 and 3.2. These, in turn, yield the inferences on k, the number of blocks, ψ, the selection of scoring matrices, and the quantity from equation (6), the Bayesian evidence for the pair to be related, as described in Section 2.4. However, to examine alignments themselves, a backward recursion is required.

3.3. Posterior alignment distribution

Many traditional methods focus only on the single best alignment and do not address alignment uncertainty. The Bayesian posterior alignment distribution characterizes the complete alignment space and thus explicitly quantifies alignment uncertainty. In Section 3.3.1, we give the formula to calculate the posterior probability for any given alignment, but this procedure does not specify how to obtain probable alignments constructively. Because of the constraints on alignment described in Section 2.1, the alignment variables $\boldsymbol{A}_{i,j}$ are not independent of one another. In Section 3.3.2, we describe an algorithm to sample alignments from the exact posterior joint alignment distribution. While insight can be gained by examining the joint distribution of the alignment based on these samples, it is difficult to visualize the high dimensional space. Thus, the marginal posterior alignment distribution, $P(\boldsymbol{A}_{i,j} = 1 \mid R^{(1)},R^{(2)})$, which is represented by a matrix, is desirable. We show in Section 3.3.2 how to obtain an estimate of this distribution from the sample alignments and in Section 3.3.3 how to obtain this distribution exactly.

3.3.1. Probability of an alignment. When the sequences are subtly related, their alignment will be uncertain. The posterior probability of any given alignment, say $\boldsymbol{A}^*$, is:

$$P(A * |R^{(1)}, R^{(2)}) = \frac{\sum_{\boldsymbol{\Psi}} P(R^{(1)}, R^{(2)}, \boldsymbol{A} *, \boldsymbol{\Psi})}{\sum_{k} \sum_{\boldsymbol{\Psi}} \sum_{\boldsymbol{A}} p(R^{(1)}, R^{(2)}, \boldsymbol{A}, k, \boldsymbol{\Psi})} \tag{7}$$

Note that in the above equation a given alignment $\boldsymbol{A}^*$ specifies both $\boldsymbol{A}$ and k in the numerator, where the sums are obtained as shown in Section 3.1.

3.3.2. Samples from the exact joint posterior alignment distribution. The full characterization of an alignment describes the simultaneous assignment of all the variables of the matrix $\boldsymbol{A}$. The colinearity constraint and the constraint which requires that each residue align at most with one other residue induce dependence on the alignment variables. The joint alignment distribution specifies the probability of the joint occurrence of a full set of variables from the matrix $\boldsymbol{A}$. In this section, we describe how to sample alignments in proportion to their joint posterior probability.

A backward recursion, which is similar to those used in dynamic programming algorithms (Needleman and Wunsch, 1970; Smith and Waterman, 1981), is described here for obtaining a sample alignment. First, we draw a scoring matrix from ψ from $P(\psi \mid R^{(1)},R^{(2)})$, equation (4). Then conditional on ψ, we sample the number of blocks, k, from $P(k \mid R^{(1)},R^{(2)},\psi)$, similar to equation (5). Finally, conditional on both ψ and k, the alignment is drawn recursively as follows: starting from (I,J), there are three choices of moves ↖, ↑ and ←, which relate to forward step choices ↘, ↓ and →, respectively. We draw a sample of these steps according to the following posterior probabilities $\frac{PC_{I,J}^{(k)}}{PC_{I,J}^{(k)} + PD_{I,J}^{(k)} + PR_{I,J}^{(k)}}$, $\frac{PD_{I,J}^{(k)}}{PC_{I,J}^{(k)} + PD_{I,J}^{(k)} + PR_{I,J}^{(k)}}$ and $\frac{PR_{I,J}^{(k)}}{PC_{I,J}^{(k)} + PD_{I,J}^{(k)} + PR_{I,J}^{(k)}}$, respectively. If ↑ is chosen, we move to position $(I - 1,J)$; if ↖ is chosen, we move to $(I - 1,J - 1)$; if ← is chosen, then we move to position $(I,J - 1)$. For any position (i,j), with at most t matching blocks we proceed as follows until i and j are both equal to 1: (i) if the last step is ↖, there are three choices for this move, ↖, ↑, or ←, the probability of each choice is $\frac{PC_{i,j}^{(t)}}{PC_{i,j}^{(t)} + PD_{i,j}^{(t-1)} + PR_{i,j}^{(t-1)}}$, $\frac{PD_{i,j}^{(t-1)}}{PC_{i,j}^{(t)} + PD_{i,j}^{(t-1)} + PR_{i,j}^{(t-1)}}$ and $\frac{PR_{i,j}^{(t-1)}}{PC_{i,j}^{(t)} + PD_{i,j}^{(t-1)} + \mathrm{Pr}_{i,j}^{(t-1)}}$; (ii) if the last sampled move was ↑, we again permit just two moves, ↖, or ↑. The moves ↖ or ↑ have probabilities $\frac{PC_{i,j}^{(t)}}{PC_{i,j}^{(t)} + PD_{i,j}^{(t)}}$ and $\frac{PD_{i,j}^{(t)}}{PC_{i,j}^{(t)} + PD_{i,j}^{(t)}}$, respectively; (iii) if the last move is ←, there are three choices for this move, ↖, ↑ or ←, the probability of each choice is $\frac{PC_{i,j}^{(t)}}{PC_{i,j}^{(t)} + PD_{i,j}^{(t)} + PR_{i,j}^{(t)}}$, $\frac{PD_{i,j}^{(t)}}{PC_{i,j}^{(t)} + PD_{i,j}^{(t)} + PR_{i,j}^{(t)}}$ and $\frac{PR_{i,j}^{(t)}}{PC_{i,j}^{(t)} + PD_{i,j}^{(t)} + PR_{i,j}^{(t)}}$. This back-sampling step has a time complexity $O(\max(I,J))$.

Each of the sampled alignments fully respects the constraints described above. Samples from Markov Chain Monte Carlo (MCMC) methods, like the Gibbs sampler, cannot be guaranteed to be from the full joint posterior distribution. Here the guarantee holds and the samples are said to be exact. A careful examination of these may reveal interesting correlated patterns in the alignment variables, such as those that occur when there are internal repeats in one of the sequences. There is no easy way to view the high dimensional space from which these samples are drawn. However, we can view the marginal distribution of all alignments. This marginal distribution gives the probability that each pair of residues will align, i.e. $P(\boldsymbol{A}_{i,j} = 1 \mid R)$. The marginal distribution takes into account all the other alignment variables, but provides no information on the simultaneous, joint, realization of the variables.

The marginal alignment distribution can be obtained empirically from the samples, and displayed by a two-dimensional histogram $\frac{\boldsymbol{A}^S}{N_S}$, where $\boldsymbol{A}^S$ are the sampled observations from the posterior distribution of the alignment, $P(\boldsymbol{A} \mid R^{(1)},R^{(2)})$. The ratio $\frac{\boldsymbol{A}^S}{N_S}$ will approach marginal posterior alignment distribution as N_S increases.

While the marginal posterior alignment distribution describes the alignment of the two sequences, a more traditional view of the alignment of two sequences is represented by the 'average' alignment. To represent an 'average' alignment with k blocks based on the marginal posterior alignment distribution, saved as a matrix $\{\boldsymbol{MP}_{i,j}\}$, we first find the highest peak of $\boldsymbol{MP}$, and extend the aligned block to the positions whose marginal posterior probability is just above 1/4 of the height of highest peak. This yields an aligned block $(R^{(1)}_{i1} \ldots R^{(1)}_{i2}, R^{(2)}_{j1} \ldots R^{(2)}_{j2})$. We assign zeros to those parts of the alignment matrix implied by the constraints, so that:

$$\boldsymbol{MP}_{i,j} = 0 \text{ for } 1 \le i \le i2 \text{ and } j \ge j1;$$
$$\boldsymbol{MP}_{i,j} = 0 \text{ for } i \ge i1 \text{ and } 1 \le j \le j2$$

Then we repeat the process with the next highest peak in $\{\boldsymbol{MP}_{i,j}\}$, continuing until we have obtained k blocks. If the current aligned block is the extension of an existing block, we concatenate them.

3.3.3. Exact marginal posterior alignment distribution. The marginal posterior alignment distribution of a specific pair of aligned residues can be calculated exactly without sampling as:

$$P(\boldsymbol{A}_{i,j} = 1 | R^{(1)}, R^{(2)}) = \frac{\sum\limits_{k}\sum\limits_{\Psi}\sum\limits_{\boldsymbol{A}_{i,j}=1} P(R^{(1)}, R^{(2)} | A, k, \Psi) P(A|k) P(k) P(\Psi)}{\sum\limits_{k}\sum\limits_{\Psi}\sum\limits_{A} P(R^{(1)}, R^{(2)} | A, k, \Psi) P(A|k) P(k) P(\Psi)} \tag{8}$$

Using the direct implication of colinearity $\boldsymbol{A}_{i,j} = 1$ as stated in equation (2), the numerator of equation (8) can be rewritten as:

$$\sum_{\Psi}\sum_{k}\sum_{0 \le t \le k} P(R^{(1)}_1 \ldots R^{(1)}_{i-1}, R^{(2)}_1 \ldots R^{(2)}_{j-1} | t, \Psi) * P(R^{(1)}_{i+1} \ldots R^{(1)}_{I}, R^{(2)}_{j+1} \ldots R^{(2)}_{j} | k-t, \Psi) * \exp(\Psi_{R^{(1)}_i, R^{(2)}_j}) P(A|k) P(k) P(\Psi) \tag{9}$$

The first two terms in this expression can be obtained by summing over all alignments from (1,1) to $(i-1, j-1)$ and from $(i+1, j+1)$ to (I,J), respectively, i.e. summing in upper left and lower right regions of the alignment matrix. The sum of alignments from (1,1) to $(i-1, j-1)$ is obtained in the forward step of Section 3.1. In this symmetric alignment procedure, the sum of all alignments starting from one end is equivalent to the sum of all alignments starting from the other end. By summing 'backward', we can complete the sums required for the second term. Let $BPC^{(t)}_{i,j}$ be the partial sum of alignments from pair (I,J) to pair (i,j) with at most t matching blocks and ↖ as the last step, $BPU^{(t)}_{i,j}$ the partial sum with ↑ as the last step, $BPL^{(t)}_{i,j}$ be the partial sum with as ← the last step. With the boundary conditions and arguments analogous to those in Section 3.1, except that the rule now is ← cannot be followed by ↑, the following recursive relationships hold:

$$BPC^{(t)}_{i,j} = [BPC^{(t)}_{i+1,j+1} + BPU^{(t-1)}_{i+1,j+1} + BPL^{(t-1)}_{i+1,j+1} * \exp(\Psi_{R^{(1)}_i, R^{(2)}_j})$$
$$BPU^{(t)}_{i,j} = BPC^{(t)}_{i+1,j} + BPU^{(t)}_{i+1,j} + BPL^{(t)}_{i+1,j}$$
$$BPL^{(t)}_{i,j} = BPC^{(t)}_{i,j+1} + BPL^{(t)}_{i,j+1}$$

Then, the sum of all alignments with a pair of residues aligned can be expressed using the forward partial sum and the backward partial sum:

$$\sum_{A_{i,j}=1} P(R^{(1)}, R^{(2)}, A | \Psi, k) = \sum_{0 \le t \le k} PC^{(t)}_{i,j} * (BPC^{(k-t+1)}_{i+1,j+1} + BPU^{(k-t)}_{i+1,j} + BPL^{(k-t)}_{i,j+1} \tag{10}$$

And the exact posterior probability can be calculated by summing over possible matrices and number of matching blocks. The backward sum will take $O(N^2)$ extra time to compute and take extra $O(N^2)$ space to store, it is much slower than the sampling method described in Section 3.3.2. In practice, we always use the backward sampling method to calculate the marginal posterior alignment distribution.

3.4. Optimal alignment

Optimal scores with less than or equal to k blocks for submatrices of the first $i = 1, 2, \ldots, I$ rows and the first $j = 1, 2, \ldots, J$ columns are returned by Sankoff's algorithm and saved as matrix with elements

$$W^{(k)}_{i,j}(\Psi) = Max_A\ (\log_2(P(R^{(1)}, R^{(2)} | A, \Psi)))$$

To obtain a maximum for the joint probability

$$\begin{aligned} P(R^{(1)}, R^{(2)}, \boldsymbol{A}, k, \boldsymbol{\Psi}) &= P(R^{(1)}, R^{(2)} | \boldsymbol{A}, \boldsymbol{\Psi}) * P(\boldsymbol{A}|k) * P(k) * P(\boldsymbol{\Psi}) \\ &= P(R^{(1)}, R^{(2)} | \boldsymbol{A}, \boldsymbol{\Psi}) * (\frac{1}{N^k_{i,j}}) * (\frac{1}{k_{\max}}) * (\frac{1}{N_{\boldsymbol{\Psi}}}) \end{aligned}$$

we must find the optimum for exactly k blocks. However, the alignment returned by the Sankoff algorithm may have less than k blocks. If the optimal alignment corresponding to $W^{(k)}_{I,J}(\boldsymbol{\Psi})$ contains v blocks and $v < k$, then $W^{(k)}_{I,J}(\boldsymbol{\Psi}) = W^{(v)}_{I,J}(\boldsymbol{\Psi})$. Since $N^{(k)}_{i,j}$ increases monotonically in k, $\frac{W^{(k)}_{I,J}(\boldsymbol{\Psi})}{N^{(k)}_{I,J}} < \frac{w^{(v)}_{I,J}(\boldsymbol{\Psi})}{N^{(v)}_{N,J}}$. Accordingly, the maximum over k for $\frac{W^{(k)}_{I,J}(\boldsymbol{\Psi})}{N^{(k)}_{I,J}}$, and thus for the joint, can occur only for values of k such that $k = v$, and

$$\begin{aligned} \log_2(p_{best}) &= Max_{k,\boldsymbol{A},\boldsymbol{\Psi}}\{ \log_2(P(R^{(1)}, R^{(2)}, \boldsymbol{A}, k, \boldsymbol{\Psi},))\} \\ &= Max_{k,\boldsymbol{\Psi}}\{ W^{(k)}_{I,J}(\boldsymbol{\Psi}) - \log_2(N^{(k)}_{I,J} * k_{\max} * N_{\boldsymbol{\Psi}})\} \end{aligned}$$

3.5. Near-optimal alignments

Here we describe an algorithm which identifies a near optimum, i.e. all alignments such that $P(R^{(1)},R^{(2)},\boldsymbol{A},\psi) \geq P_{best} - \delta = P_{cutoff}\ (R^{(1)},R^{(2)},\boldsymbol{A},\ \psi)$. We obtain this set by finding all the alignments for each value of k and ψ which exceed this cutoff value. Because $P(A \mid k)$ varies with k, we re-normalize the cutoff for exactly k blocks as follows:

$$P_{cutoff}(R^{(1)},R^{(2)},\boldsymbol{A},k,\psi) = \frac{P_{cutoff}(R^{(1)},R^{(2)},\boldsymbol{A},\boldsymbol{\Psi})}{P(\boldsymbol{A}|k)P(k)P(\boldsymbol{\Psi})}$$

with corresponding cutoff score,
$S_{cutoff}(k,\psi) = \log_2 (P_{cutoff}(R^{(1)},R^{(2)},\boldsymbol{A},k,\psi))$.

For each value of k and ψ, we employ a branch and bound algorithm with the bound fixed at $S_{cutoff}(k,\psi)$ to identify the near-optimal solutions with exactly k blocks. Our basic strategy is to employ a depth-first trace-back procedure using $W^{(k)}_{i,j}$ to prune a large number of branches, and then remove any additional branches which do not have exactly k blocks.

All possible solutions may be seen as leaves of a tree rooted at the terminal, (I,J). Each node of this tree is characterized by the following items:

(i) $S^t(i,j)$, an alignment score from the ends of the sequences to the current position (i,j), where t is the number of matching blocks used from the ends (I,J) to the current position (i,j);

(ii) an indicator of the last step (↖, ↑, ←).

This trace-back algorithm uses the same three backward steps ↖, ↑ and ←, including back steps restriction, described above for the sampling algorithm. The root of the tree has as the last step, (I,J) as the current position, $t = 0$, and a score of zero. Iteratively, we build the trace-back tree from position (i,j) as follows:

(i) If the last step of the node is ←, there are three alternatives with the next step as ↖, ↑, ←. With ↖ as the next step, the score becomes $S^{t+1}(i-1,j-1) = S^t(i,j) + \boldsymbol{\Psi}_{R^{(1)}_i,R^{(2)}_j}$. With ↑ as the next step, $S^t(i-1,j) = S^t(i,j)$ and no new blocks are started. With ← as the next step, $S^t(i,j-1) = S^t(i,j)$ and no new blocks are started.

(ii) If the last step is ↑, there are two alternatives ↑ and ↖ as the next step, and nodes are updated as above (recall ← is not permitted to avoid double counting).

(iii) If the last step of the node is ↖, there are three alternatives ←, ↑, ↖ as the next step. For ← or ↑ as the next step, the score is updated as above. For ↖ as the next step, there are no next blocks, and the score is updated as $S^t(i-1,j-1) = S^t(i,j) + \boldsymbol{\Psi}_{R^{(1)}_i,R^{(2)}_j}$.

A search path is terminated if either of the following two conditions are met. (i) The score drops below the cutoff, i.e. if ↖ is the last step and $S^t(i,j) + W^{(k-t+1)}_{i,j} < S_{cutoff}$ or if ← and ↑ is the last step and $S^t(i,j) + W^{(k-t+1)}_{i,j} < S_{cutoff}$. (ii) Any of following three boundaries is encountered: $i = 0$; $j = 0$; or the last step is ← or ↑ and $t = k$.

If a leaf meets the following conditions, (i) $S^t(i,j) \geq S_{cutoff}$; (ii) has exactly k blocks, i.e. $t = k$ and the last step is ← or $W^{(k)}_{I,J}$ or (added to avoid double counting), then it corresponds to an alignment in the near-optimal set. While in principle this branch and bound algorithm must be applied for each combination of k and ψ, in practice we find that $W^{(k)}_{I,J} < S_{cutoff}$ for many values of k and ψ, and thus a search for these values is not required.

4. Results

In this section, we compare the performance of our algorithm with that of the popular Smith–Waterman algorithm in the identification of structurally similar proteins. We also illustrate distinctive features of the Bayes block alignment method using the alignment of two specific pairs of proteins: a pair of distantly related kinases, guanylate kinase and adenylate kinase; a pair of GTPases, elongation factor Tu (IET) and G (EF-G).

4.1. Database search of structural related proteins

The two most extensive comparisons of sequence alignment and database searching methods have been reported by Brenner *et al.* (1997) and Pearson (1995). Using a standard, SCOP (Murzin *et al.*, 1995), which is largely based on protein structure, Brenner *et al.* (1997) compared BLAST (Altschul *et al.*, 1990), FASTA (Pearson and Lipman, 1988) and Smith–Waterman's procedure (Smith and Waterman, 1981) implemented as SSEARCH. Pearson (1995) made a similar comparison based on superfamilies in PIR. Both reported that SSEARCH uniformly outperformed the other procedures. For our purpose, the approach of Brenner *et al.*, which relies extensively on a structural standard, is of more interest. The approach we take here is similar to that of Brenner *et al.*, but uses VAST (Madej *et al.*, 1995; Gibrat *et al.*, 1996), which uses only structural data, as standard.

Three databases of proteins whose pairwise identities were less than 25, 35 and 45% (pdb select 25, pdb select 35 and pdb select 45) were obtained from EMBL (Hobohm *et al.*, 1992). There are 553, 842 and 972 sequences in pdb25, 35 and 45, respectively. The structural neighbors of each of the proteins in the three databases were obtained using VAST (Madej *et al.*, 1995; Gibrat *et al.*, 1996). There are 2737, 4726 and 6101 pairs of similar structures among each database for pdb select:25, 35 and 45, respectively, indicating that each of the PDB select sequences has about six structure neighbors. In this analysis, all structural neighbors reported by VAST ($P < 0.1$) are taken as the exhaustive set of true positives. For each entry in pdb25, 35 and 45 which has structural neighbors in the same data set, we randomly sample 30 controls from the proteins not reported by VAST to be structural neighbors.

Table 1. Database search result comparison between Bayes aligner and SSEARCH with recommended parameter settings.

(a) Results for database pdb25 at 1% false-positive rate

Methods	Tuned parameters (criterion)	Coverage (%)
Bayes aligner	N/A	14.4
SSEARCH 3.0	BLOSUM45 –12/–1 (raw score)	13.7
SSEARCH 3.0	BLOSUM45 –12/–1 (*P* value)	13.7
SSEARCH 3.0	BLOSUM50 –12/–2 (raw score)	13.3
SSEARCH 3.0	BLOSUM50 –12/–2 (*P* value)	12.9
SSEARCH 3.0	BLOSUM55 –14/–1 (raw score)	13.9
SSEARCH 3.0	BLOSUM55 –14/–1 (*P* value)	13.2
SSEARCH 3.0	BLOSUM62 –8/–2 (raw score)	13.5
SSEARCH 3.0	BLOSUM62 –8/–2 (*P* value)	12.7

(b) Results for database pdb35 at 1% false positive rate

Methods	Tuned parameters (criterion)	Coverage (%)
Bayes aligner	N/A	20.2
SSEARCH 3.0	BLOSUM45 –12/–1 (raw score)	18.9
SSEARCH 3.0	BLOSUM45 –12/–1 (*P* value)	19.2
SSEARCH 3.0	BLOSUM50 –12/–2 (raw score)	19.3
SSEARCH 3.0	BLOSUM50 –12/–2 (*P* value)	18.2
SSEARCH 3.0	BLOSUM55 –14/–1 (raw score)	19.0
SSEARCH 3.0	BLOSUM55 –14/–1 (*P* value)	18.6
SSEARCH 3.0	BLOSUM62 –8/–2 (raw score)	18.9
SSEARCH 3.0	BLOSUM62 –8/–2 (*P* value)	17.8

(c) Resuts for database pdb45 at 1% false positive rate

Methods	Tuned parameters (criterion)	Coverage (%)
Bayes aligner	N/A	25.7
SSEARCH 3.0	BLOSUM45 –12/–1 (raw score)	24.3
SSEARCH 3.0	BLOSUM45 –12/–1 (*P* value)	25.3
SSEARCH 3.0	BLOSUM50 –12/–2 (raw score)	24.9
SSEARCH 3.0	BLOSUM50 –12/–2 (*P* value)	24.3
SSEARCH 3.0	BLOSUM55 –14/–1 (raw score)	24.7
SSEARCH 3.0	BLOSUM55 –14/–1 (*P* value)	24.6
SSEARCH 3.0	BLOSUM62 –8/–2 (raw score)	23.6
SSEARCH 3.0	BLOSUM62 –8/–2 (*P* value)	23.6

For the SW algorithm, we chose four parameter settings reported in the literature to be optimal: BLOSUM55 with –14/–1 gap penalty as the best setting for the identification of structural neighbors found by Brenner (1997), BLOSUM45 with –12/–1 gap penalty which was used by Brenner *et al.* (1997), BLOSUM50 with –12/–2 gap penalty as the best for database search values reported by Pearson (1995), and BLOSUM62 with –8/–2 gap penalty recommended by Pearson (1995) for the most popular scoring matrix. Results for SSEARCH using both raw scores and *P* values as a cutoff criterion were examined.

For pdb25, VAST identified 2737 pairs of related sequences, and 9900 pairs (30 for each sequence which has structural neighbors) of unrelated sequences were randomly sampled. There were 4728 pairs of related sequences, 15 180 pairs of unrelated sequences sampled for pdb35. For pdb45, their were 6101 pairs of related sequences, 17 910 pairs of unrelated sequences were sampled. Table 1 reports the coverage, i.e. the percentage of the VAST structural neighbors correctly identified when the score cutoff is set to yield 1% false positives. As shown, the Bayes aligner obtains higher coverage than SSEARCH with any of the parameter settings for all of the data sets. Furthermore, we found that for all three databases, the Bayes aligner uniformly achieved greater coverage over a wide range of cutoff values. The receiver operation characteristic (ROC) curve is given in Figure 1, to illustrate this effect for pdb35.

4.2. Specific applications

In this section, we illustrate distinctive features of the Bayes aligner by examining the alignment of two specific pairs of sequences.

4.2.1. Alignment of a pair of kinases. Guanylate kinase and adenylate kinase catalyze similar reactions: X-triphosphate + Y-monophosphate $\leftrightarrows$ X-diphosphate + Y-diphosphate, where X and Y are either adenine or guanine. Guanylate kinase from yeast (1GKY in PDB) and adenylate kinase from beef heart mitochondrial matrix (2AK3 chain A in PDB) are a pair of sequences in pdb35, which are VAST structural neighbors. We report this pair of alignments because it illustrates well some distinctive characteristics of a Bayes block alignment.

A distinctive characteristic of the Bayes block aligner stems from the fact that all alignments for a given number of blocks are equally likely. As a result, there is nothing equivalent to a gap extension penalty. It thus has advantages for alignment of sequences with unusual length gaps. The alignment of 1GKY and 2AK3 illustrates this feature. The similarity of these kinase sequences is not detected by SSEARCH with any parameter settings. For example, using BLOSUM50 with –12/–2 gap penalty the SSEARCH alignment score is 50, which is well below the cutoff value of 71 for 1% false positives. This low score stems from the fact that SSEARCH misaligns segments after a large insertion in 2AK3-A. The absence of a gap penalty in the Bayes aligner allows it to avoid this misalignment and thus correctly identify this VAST structural neighbor. The last peak in Figure 2 follows this large insertion.

The ability of these methods to characterize the uncertainty in all of the unknowns is their most important distinctive characteristic. The characterization has two features: (i) the marginal characterization of the uncertainty of each unknown; (ii) the characterization of the effect of the uncer-

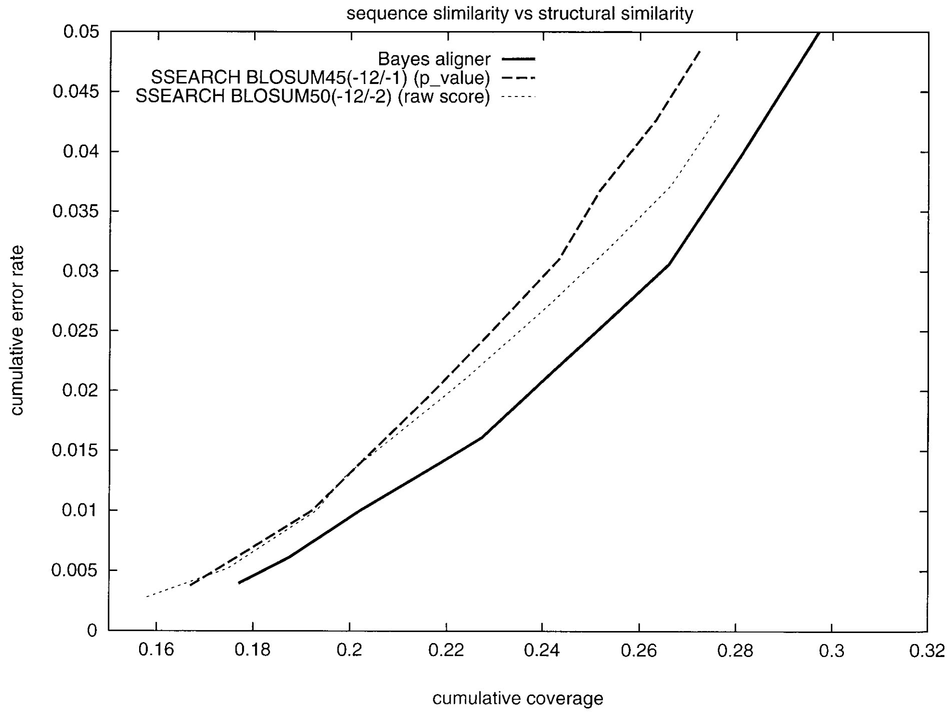

Fig. 1. Database search results for pdb35 using Bayes aligner, SSEARCH with two recomended parameter settings at different false-positive error rates.

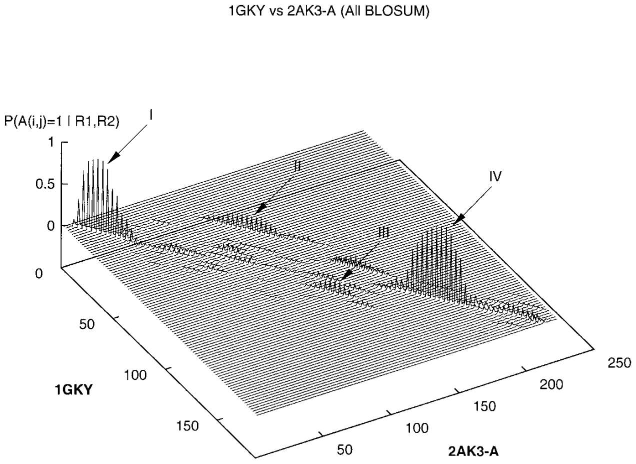

Fig. 2. Marginal posterior alignment distribution for 1GKY and 2AK3-A.

tainty in one unknown on another. The first of these is well illustrated in the alignment of 1GKY and 2AK3, while the latter is illustrated in the GTPase example described below. As shown in Table 2, the posterior distribution ψ for the BLOSUM series is quite flat with four matrices, BLOSUM45 to BLOSUM80, for which $P(\psi \mid R^{(1)},R^{(2)}) > 0.10$. The mode of this distribution is at BLOSUM62. As shown in Table 3, the posterior distribution ψ for the PAM series is not only flat, but also multimodal with modes at PAM 110, 140 and 200. These flat distributions indicate that no single matrix characterizes well the conservation between these two sequences, and suggests that there may be considerable variability in the degree of conservation over the lengths of the two sequences. We explore this suggestion more thoroughly in the next example. As shown in Table 4, the posterior distribution of the number of gaps is also flat. This characteristic is shared by the majority of the pairs of structural neighbors identified in the previous section. These flat distributions suggest that a gap opening model with a set value for the penalty parameter may not characterize alignments of this type well.

Table 2. Posterior probability distribution of BLOSUM score matrices for alignment of 1GKY and 2AK3-A

BLOSUM matrix index	Posterior probability
30	0.0257247
35	0.0449377
40	0.0825217
45	0.111502
50	0.175465
62	0.286682
80	0.234981
100	0.0381859

Table 3. Posterior probability distribution of PAM score matrices for alignment of 1GKY and 2AK3-A

PAM matrix index	Posterior distribution
40	0.000364128
50	0.000744213
60	0.00294976
70	0.028715
80	0.0205859
90	0.0129509
100	0.0495863
110	0.121702
120	0.0454115
130	0.0726501
140	0.163735
150	0.0215645
160	0.0357004
170	0.0656963
180	0.0209931
190	0.0630343
200	0.130371
210	0.0322463
220	0.0226838
230	0.0274658
240	0.0137122
250	0.0132439
260	0.0116769
270	0.0114211
280	0.00397427
290	0.00328901
300	0.00353304

Another important feature of this method is that it characterizes the space of all alignments rather than focusing on the single best alignment and/or near-optimal alignments. Using equation (7) and methods given in Sections 3.4 and 3.5, we can compare these two approaches. Applying equation (7) to the optimal alignment provides a measure of the value or quality of the optimal alignment. The probability of optimal alignment for these kinase sequences is 2.94×10^{-5}. Because this low value indicates that the optimal alignment does not cover much of the posterior alignment space, it suggests that it is not a good representation of the alignment of these sequences. Using the algorithm for the identification of near-optimal alignments described in Section 3.5 and equation (7), we find that the probability of all alignments whose probability is $>1.0 \times 10^{-5}$ is 1.0×10^{-3}. This suggests that even the set of near-optimal alignments does not represent well the alignment of these sequences.

To select a specific number of blocks to include in an averaged alignment, our experience, as guided by agreement with structural alignments, indicates that it is important not to select too many blocks, but also to be confident that the minimum number selected covers the major features of the alignment distribution. We meet these ends by requiring that there is at least a 90% chance of k blocks or more, i.e. $P(K \geq k \mid R^{(1)},R^{(2)}) > 0.9$. These alignments correspond reasonably with sequence alignments based on VAST. Figure 3 shows this correspondence for the kinase pair. Structures of 1GKY and 2AK3-A are superimposed according to this alignment, and shown as Figure 4 (segment I: red; II: cyan; III: orange; IV: green). Segments I, III and IV are involved in triphosphate binding. They correspond well to each other in the structures. As neither of the structures is co-crystallized with triphosphate, we chose a protein (2AKY) closely related to 2AK3-A to demonstrate triphosphate binding, shown as Figure 4C. Segment II is remote from mono- and triphosphate binding sites in both structures. While these segments share some similarities in secondary structure, they do not superimpose. Errors of this type, which have local structure similarity and are not near a ligand, are the most common structure prediction errors of the Bayes block aligner.

4.2.2. Alignment of GTPases. Perhaps the most interesting feature of the Bayes aligner is the fact that the posterior distribution of each of the unknowns reflects the uncertainty in the others. For example, as shown in equation (5), the posterior distribution of ψ is obtained by averaging over all alignments in proportion to their posterior probability. One of the examples previously presented by Zhu *et al.* (1997) briefly described the alignment of several GTPases, but does not describe important characteristics of this method. Specifically, the posterior PAM distribution for the pair, elongation factor Tu (IETU) and elongation factor G (EF-G), is shown to be bimodal with peaks at PAM 80 and PAM 140, but only a brief explanation of this important effect is given.

The multimodal nature of this posterior PAM distribution reflects variation in the alignment stemming from variation in conservation over the lengths of the two sequences. This effect can be seen by examining the posterior alignment distribution. The marginal posterior alignment distribution, $P(\boldsymbol{A} \mid R^{(1)},R^{(2)})$, of the two sequences, illustrated in Figure 5a, contains six conserved segments, which correspond to five well-known GTPase motifs G1, G2, G3, G4, G5 (Bourne *et al.*, 1991), and a motif specific to elongation factors (EF). The EF motif is a helix–loop–sheet structure. Figure 5b shows the posterior distribution of the alignment for PAM 80, $P(A \mid R^1,R^2, \psi = 80)$, and Figure 5c for PAM 140 $P(A \mid R^1,R^2, \psi = 140)$. As these two figures illustrate, the bimodality of the PAM distribution stems from variation in the conservation of these six motifs and in the intervening residues. There are several aspects of this feature. Motifs G2, G3 and EF are well conserved between these two sequences, but the intervening sequences are not so well conserved, as is reflected by the three distinct peaks associated with these three motifs at PAM 80 (see Figure 5b). As shown in Figure 5c, at a lower level of conservation (PAM 140) these three motifs and the interval sequences can be reasonably represented as a single conserved block. Furthermore, motif G4 is less well conserved than the other motifs. Accordingly, it contributes little to the alignment at PAM 80, but becomes an important component at PAM 140. Thus, even though each alignment use only one scoring matrix, the posterior alignment distribution illustrates well the variations in the degree of conservation.

5. Discussion

Dispensing with the need to set gap penalty and score matrix parameters is one of the chief advantages of the Bayesian block aligner. The other is the new products it produces. These products stem from the fact that the marginal distribution is obtained for all unknowns. As a result, the uncertainty of all unknowns is fully accounted for in these posterior distributions. For example, the posterior distribution of the distance between a pair of sequences in PAM units is obtained. This distance distribution is not dependent on any specific alignment, and the uncertainty in the distribution of alignment is fully accounted for in this distribution. The flat and multimodal nature of this distribution in the kinase example is typical of many we have examined. These distributions should be of interest in studies of molecular evolution because they provide a means in conjunction with distance-based methods for constructing an evolutionary tree without the requirement of obtaining a good alignment. It thus offers promise of extending the range over which the distant model may be applied. It so provides the means to incorporate alignment uncertainty in tree construction. The algorithm achieves these ends with a time complexity of $O(N^2)$, comparable to other pairwise alignment algorithms, but with a large constant.

Our applications illustrate another important feature of the algorithm. The algorithm will align only those subsequences of the two biopolymers which the data indicate are conserved. Not surprisingly, these subsequences often correspond to motifs that form ligand binding pockets. This feature has two facets. First, it improves the alignment by permitting the algorithm to ignore and thus not be confused by unrelated portions of the sequences. This characteristic can be seen as an extension of the concept of local alignment. Second, it points to the conserved subsequences which are likely to play important functional roles.

Table 4. Posterior probability distribution of number of blocks for alignment of 1GKY and 2AK3-A using BLOSUM matrices

Number of blocks	Posterior probability distribution $P(k \mid R^{(1)},R^{(2)})$	Cumulative posterior probability $P(K \geq k \mid R^{(1)},R^{(2)})$
0	0.0621149	1.0
1	0.00285821	0.937885
2	0.00653246	0.935027
3	0.0137866	0.928494
4	0.0227437	0.914708
5	0.0326225	0.891964
6	0.0430195	0.859342
7	0.0527596	0.816322
8	0.0606999	0.763563
9	0.0664768	0.702863
10	0.0702314	0.636386
11	0.0723108	0.566154
12	0.0731055	0.493844
13	0.0729739	0.420738
14	0.0722126	0.347764
15	0.0710524	0.275552
16	0.0696648	0.204499
17	0.0681728	0.134834
18	0.0666615	0.0666615

The Bayesian evidence that two sequences are related = 0.955.

```
A
  1 USRPIVISGPSGTGKSTLLKKLFAEYPDSFG  31
               ....
           .:::::::::.
          ::::::::::::..
         :::::::::::::::::.
       .::::::::::::::::::::::.
  5 RLLRAAIMGAPGSGKGTVSSRITKHFELKHL  35

 54 VSVDEFKSMIKNNEFIEWAQF  74      126 VEDLKKRLE 134
     ....:...........                  ....:::::
 73 LVLHELKNLTQYNWLLDGFPR  93      117 FEVIKQRLT 125

135 GRGTETEESINKRLSAAQAELAYAE 159
           ....:::...
         ..::::::::::::.
     ...::::::::::::::::.
    ::::::::::::::::::::::...
159 QREDDRPETVVKRLKAYEAQTEPVL 183

B
123  PPS---VEDLKKR-LEGRGTETEESINKRLSAAQAE  154
143  PPKTMGIDDLTGEPLVQREDDRPETVVKRLKAYEAQ  178

C
2   SRPIVISGPSGTGKSTLLKKLFAEYP 27   81 STVASVKQV 89
6   LLRAAIMGAPGSGKGTVSSRITKHFE 31   73 LVLHELKNL 81

92 SGKTCILD 99  106 VKSVKAIP 112  114 LNARFLFIAPPSVEDLKKRLEGR 136
82 TQYNWLLD 89   96 PQAEALDR 102  105 YQIDTVINLNVPFEVIKQRLTAR 127

137 GTETEESINKRLSAAQAELAYAETGA 162
161 EDDRPETVVKRLKAYEAQTEPVLEYY 186
```

Fig. 3. Comparison of alignments of 1GKY and 2AK3-A obtained by Bayes aligner, SSEARCH and structural alignment VAST. (**A**) The alignment of 1GKY and 2AK3-A obtained by Bayes aligner using BLOSUM matrices. The dots indicate the confidence of the importance of the aligned pairs. (**B**) The alignment produced by SSEARCH with the BLOSUM50 score matrix and –12/–2 gap penalty. (**c**) The alignment based on structure alignment produced by VAST.

There are specific proteins for which the Smith–Waterman algorithm correctly finds more structural neighbors than the Bayes aligner, and vice versa, but our results indicate that the Bayes aligner finds more structural neighbors on average than the Smith–Waterman algorithm. While relaxing the requirement for the specification of parameters greatly increased the flexibility of the alignment, it does so at the price of a substantial increase in the number of free parameters, and a corresponding potential increase in Type I error, i.e., more false positives. Thus, we were somewhat surprised to find that the Bayes aligner with uninformed priors obtained better coverage than the Smith–Waterman algorithm which uses a parameter setting that had been pre-tuned by previous authors based on structural comparison. With the caveat that the algorithms differ in other ways, the fact that the improved flexibility outweighed the adverse effects of more free parameters indicates the importance of accounting for the uncertainties in pairwise sequence alignment.

The two methods compared are at the extremes of the specification of parameter settings: completely free/completely fixed. These extremes are not the only alternatives. Informed priors provide a means to combine these extremes and thus potentially improve performance. Incorporation of informed priors is straightforward and our preliminary analysis shows this step to be promising.

We have compared the Bayesian alignment algorithm with only one optimal alignment procedure, and only in the limited context of a set of recommended parameter values. Expert selection of parameter settings may improve alignment performance. Also, a method which somehow selects the best scoring matrix and the best gap penalties for each alignment may perform better. On the other hand, one could argue that since the database used by Brenner *et al.* (1997) to find good parameter settings is likely to overlap substantially the database we examined, our comparison was already biased in favor of the optimal alignment procedure—SSEARCH. Nevertheless, the comparison presented here may not have given full credit to traditional alignment methods because of our selection of the optimal alignment algorithm, our use of less than optimal parameter settings, or our use of existing methods which require the pre-selection of parameter settings.

Our approach is most similar to that presented by Thorne *et al.* (1991, 1992) and Allison *et al.* (1992) in that these methods simultaneously address gapping and mismatch scoring while summing over all possible alignments. Besides the obvious difference that we employ Bayesian rather than classical statistics, our approach differs from theirs in a number of other ways. While in principle their approaches for DNA alignments can be extended to proteins, such extensions appear difficult because of the large number of extra parameters associated with the large scoring matrices. Another noteworthy point is that their methods find point estimates of penalty parameters by using an EM-type algorithm, which only guarantees local optimality. The approach described here gives the complete posterior distribution, including its global mode. Perhaps the largest difference comes from the fact that their use of point estimates means that uncertainties associated with these parameters are not incorporated into the resulting alignment distribution. As we have seen, posterior distributions for penalty parameters are often flat and sometimes multimodal. Thus, this uncertainty appears to be a very important factor. Lastly, their algorithms are gap penalty-based methods, and thus they do not yield the extended local alignments that appear to be important for distantly related protein sequences.

The methods for the selection of scoring matrices by Altschul (1993) and Agarwal and States (1996) also bear similarity to our method, but they do not simultaneously address gapping parameters and scoring matrices. Furthermore, since both of these procedures consider only an optimal alignment, they fail to incorporate alignment uncertainty, which becomes increasingly important as the distance between the sequences increases.

Some extensions to this algorithm are worth mentioning. First, while the applications shown here are for protein sequences, the extension to nucleotide sequence alignment is straightforward and has been incorporated into our software.

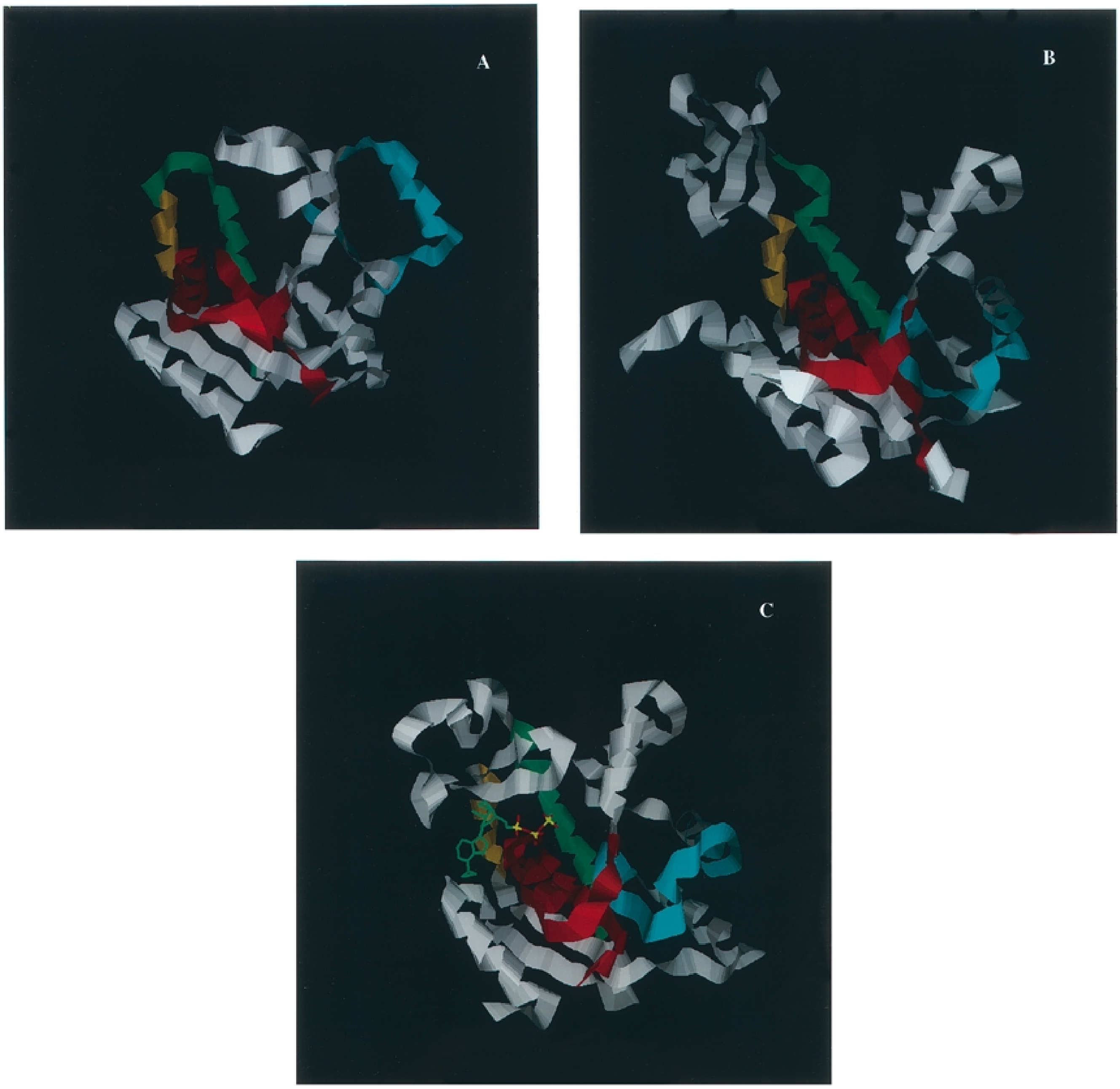

Fig. 4. Structural correspondence of aligned blocks identified by Bayes aligner. Segement I: red; II: cyan; III: orange; IV: green. (**A**) 1GKY. (**B**) 2AK3-A. (**C**) 2AKY, which is very similar to 2AK3-A, but with the ATP analogue co-crystallized (Stohle and Schulz, 1992).

We can further relax our alignment model to reflect different degrees of conservation at different sites such that each alignment pair has its own score matrix. This can be accomplished by increasing one order of computational complexity. As we showed in Section 4.2.2, we can achieve a similar goal without increasing computational complexity by examining the marginal posterior alignment distribution at various distances. Here we describe a Bayesian alignment algorithm which employs a recursion based on the algorithm of Sankoff. As shown by Liu and Lawrence (1998), there is a general framework for the conversion of a broad spectrum of dynamic programming algorithms used in bioinformatics to Baysian inference algorithms. Included in this framework are all popular alignment algorithms. Furthermore, this method can be extended to multiple sequence alignment through Gibbs sampling, using the approach described by

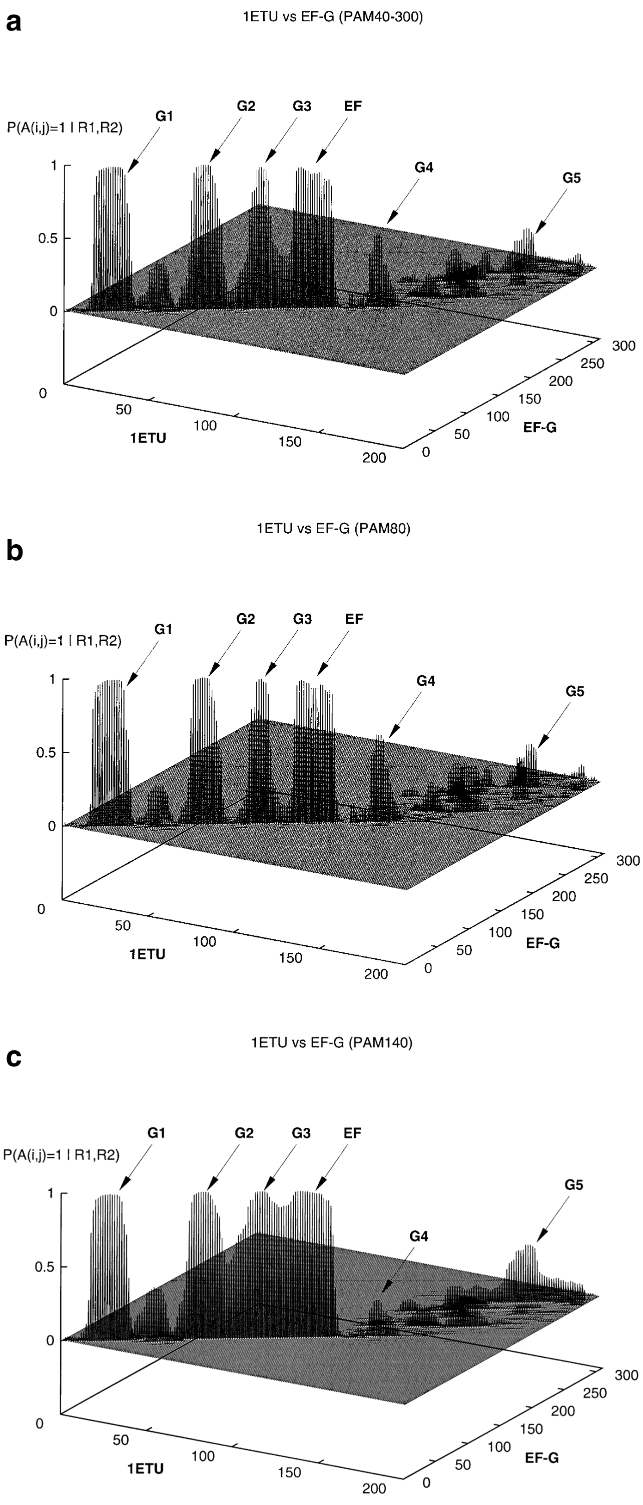

Fig. 5. Marginal posterior alignment distribution for 1ETU and EF-G using different scoring matrices. The peaks marked G1–G5 correspond to conserved segments G1–G5 in the GTPase superfamily. The segment corresponding to peak EF is conserved only in some elongation factors. (**a**) PAM40-300 matrices with a step interval of 10; (**b**) PAM80; (**c**) PAM140.

Liu and Lawrence (1995). Since the required summations cannot be practically completed for multiple sequence alignment problems, exact Bayesian inferences for these problems are not available. However, approximations involving presumed optimal solutions have shown their utility in these settings (Neuwald *et al.*, 1997).

Statistical algorithms have become increasingly popular in computational molecular biology. For example, in the last few years, HMMs (Baldi *et al.*, 1994; Krogh *et al.*, 1994), EM algorithms (Lawrence and Reilly, 1990; Bailey and Elkan, 1994) and Gibbs sampling algorithms (Lawrence *et al.*, 1993) have become important methods for multiple sequence alignment and other purposes. While the algorithms have received much attention, the associated statistical inferences which concern the description of unobserved population parameters and the making of predictions about variables that are yet to be observed have received far less. The results presented here suggest that Bayesian inference statistics promises to become an important tool for computational molecular biology.

Acknowledgements

This work was supported by NIH grant 5R01HG01257 and by DOE grant DEFG0296ER72266 to C.E.L., by NSF grant NSF-9501570 and by the Terman fellowship from Stanford University to J.S.L. We are thankful to Dr William Thompson for proof-reading the manuscript.

References

Agarwal,P. and States,D.J. (1996) A Bayesian evolutionary distance for parameterically aligned sequences. *J. Comput. Biol.*, **3**, 1–17.

Allison,L., Wallace,C.S. and Yee,C.N. (1992) Finite-state models in the alignment of macromolecules. *J. Mol. Evol.*, **35**, 77–90.

Altschul,S.F. (1991) Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, **219**, 555–565.

Altschul,S.F. (1993) A protein alignment scoring system sensitive at all evolutionary distances. *J. Mol. Evol.*, **36**, 290–300.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *ISMB*, **2**, 28–36.

Baldi,P., Chauvin,Y., McClure,M. and Hunkapiller,T. (1994) Hidden Markov models of biological primary sequence information. *Proc. Natl Acad. Sci. USA*, **91**, 1059–1063.

Berger,J. and Sellke,T. (1987) Testing a point null hypothesis: The irreconcilability of P values and evidence. *J. Am. Stat. Assoc.*, **82**, 112–122.

Bourne,H.R., Sanders,D.A. and McCormick,F. (1991) The GTPase superfamily: conserved structure and molecular mechanism. *Nature*, **349**, 117–127.

Box,G.E.P. (1980) Sampling and Bayesian inference in scientific modelling and robustness. *J. R. Stat. Soc.*, **143**, 383–430.

Brenner,S.E. (1997) Assessing sequence comparison methods. Poster at *Pacific Symposium on Biocomputing, 1997.*

Brenner,S.E., Chothia,C. and Hubbard,T. (1997) Assessing sequence comparison methods. Submitted.

Bronner,C.E. *et al*. (1994) Mutation in the DNA mismatch repair gene homologue h*MLH*1 is associated with hereditary non-polyposis colon cancer. *Nature*, **368**, 258–261.

Collins,J.F., Coulson,A.F.W. and Lyall,A. (1988) The significance of protein sequence similarities. *Comput. Applic. Biosci.*, **4**, 67–71.

Dayhoff,M.E., Eck,R.V. and Park,C.M. (1972) A model of evolutionary change in protein. In Dayhoff,M.E. (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC, Vol. 5, pp. 89–99.

Gelman,A., Carlin,J.B., Stern,H.S., Lyall,A. and Rubin,D.B. (1995) *Bayesian Data Analysis*. Chapman Hall, New York.

Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.

Goad,W.B. and Kanehisa,M.I. (1982) Pattern recognition in nucleic acid sequences. A general method for finding local homologies and symmetries. *Nucleic Acids Res.*, **10**, 247–263.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) Selection of a representative set of structures from the Brookhaven protein data bank. *Protein Sci.*, **1**, 409–417.

Krogh,A., Brown,M., Mian,I.S., Sjoelander,K. and Haussler,D. (1994) Hidden Markov models in computational biology: Application to protein modelling. *J. Mol. Biol.*, **235**, 1501–1531.

Lawrence,C.E. (1997) Bayesian bioinformatics: tutorial in ISMB97. http://www.wadsworth.org/res&res/bioinfo

Lawrence,E.C. and Reilly,A.A (1990) An Expectation Maximization (EM) alogrithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, **7**, 41–51.

Lawrence,E.C., 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.

Lipman,D.J. and Pearson,W.R. (1985) Rapid and sensitive protein similarity searches. *Science*, **227**, 1435–1441.

Liu,J.S. and Lawrence,C.E. (1998) Bayesian analysis of biopolymer sequences. In press.

Liu,J.S. and Lawrence,C.E. (1995) Statistical models for multiple sequence alignment: unification and generalization. In *Proceedings of the American Statistical Association.* ASA Press, Orlando, FL, Statistical Computation Section 21:1–8.

Madej,T., Gibrat,J.-F. and Bryant,S.H. (1995) Threading a database of protein cores. *Proteins*, **23**, 356–369.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Neuwald,A.F., Liu,J.S., Lipman,D.J. and Lawrence,C.E. (1997) Extracting protein alignment models from the sequence database. *Nucleic Acids Res.*, **25**, 1665–1677.

Papadopoulos,N. *et al.* (1994) Mutation of a *mut*L homolog in hereditary colon cancer. *Science*, **263**, 1625–1629.

Pearson,W.R. (1991) Searching protein sequence libraries: comparison of the Smith-Waterman and FASTA algorithms. *Genomics*, **11**, 635–650.

Pearson,W.R. (1995) Comparison of methods for searching protein sequence databases. *Protein Sci.*, **4**, 1145–1160.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Sankoff,D. (1972) Matching sequences under deletion/insertion constraints. *Proc. Natl Acad. Sci. USA*, **69**, 4–6.

Schwartz,R.M. and Dayhoff,M.O. (1978) Matrices for detecting distant relationships. In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC, Vol. 5, Suppl. 3, pp. 353–358.

Sellers,P.H. (1984) On the theory and computation of evolutionary distances. *J. Appl. Math.*, **26**, 787–793.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

States,D.J., Harris,N.L. and Hunter,L. (1993) Computationally efficient cluster representation in molecular sequence megaclassification. In *Proceedings of ISMB93*. AAAI Press, Bethesda, MD, pp. 387–394.

Stehle, T. and Schulz,G.E. (1992) Refined structure of the complex between guanylate kinase and its substrate GMP at 2.0 A resolution. *J. Mol. Biol.*, **224**, 1127–1141.

Thorne,J., Kishino,H. and Felsenstein,J. (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.*, **33**, 114–124.

Thorne,J., Kishino,H. and Felsenstein,J. (1992) Inching toward reality: An improved likelihood model of sequence evolution. *J. Mol. Evol.*, **34**, 3–16.

Waterman,M.S. (1994) Parametric and ensemble sequence alignment algorithms. *Bull. Math. Biol.*, **56**, 743–767.

Waterman,M.S., Eggert,M. and Lander,E. (1992) Parametric sequence comparisons. *Proc. Natl Acad. Sci. USA*, **89**, 6098–6093.

Zhu,J., Liu,J.S and Lawrence,C.E. (1997) Bayesian adaptive alignment and inference. In: *Proceedings of ISMB97*. pp. 358–368.