Bayesian Adaptive Alignment and Inference

Jun Zhu [1], Jun Liu [2] and Charles Lawrence [1]

(1) Wadsworth Center for Laboratories and Research, Albany, NY, 12201: junzhu, lawrence@wadsworth.org, 518-473-3382, FAX: 518-474-7992
(2) Dept. of Statistics, Stanford University, Stanford, CA, 94305-4065 jliu@paolu.stanford.edu, 415-723-2623, FAX 415-725-8977

Abstract

Sequence alignment without the specification of gap penalties or a scoring matrix is attained by using Bayesian inference and a recursive algorithm. This procedure's recursive algorithm sums over all possible alignments on the forward step to obtain normalizing constants essential to Bayesian inferences, and samples from the exact posterior distribution on the backward step. Since both terminal and intervening unrelated subsequences will often be excluded from an alignment, the resulting alignments may be seen as extensions of local alignments. An alignment's significance is assessed using the Bayesian evidence. A shuffling simulation shows that Bayesian evidence against the null hypothesis tends to be a conservative measure of significance compared to classical p-values. An application to proteins from the GTPase superfamily shows that the posterior distribution of the number of gaps is often flat and that the posterior distribution of the evolutionary distance is often flat and sometimes bimodal. An alignment of 1GIA with 1ETU shows good correspondence with a structural alignment.

Introduction

The alignment of biopolymer sequences has played a critical role in the identification of related genes and proteins and in the prediction of protein structure and function. Many efficient algorithms for the alignment of pairs of sequences have been developed. 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 only share a common substructure. For example when a domain has been added to the end of one protein and is not present in the other one it should be ignored in the alignment. Accordingly, the development of local alignment algorithms which identify and align the best common subsequences was an important advance (Smith and Waterman 1981). In distantly related proteins the only traces of sequence conservation may be in disjoint segments of proteins that are involved in the binding of ligands. Subsequences between these conserved segments may be completely unrelated. Thus, algorithms which extend the concept of local alignment to simultaneously identify multiple disjoint locally conserved segments are needed.

A serious limitation of most current alignment algorithms is their need for the specification of gap penalty and scoring matrix parameters. This input process is neither intuitive nor rigorous, but often strongly influences the alignment. The appropriate scoring matrix depends on the distance between a pair of sequences, which is often unknown. Several earlier works have addressed the issue of using multiple scoring matrices (Schwartz and Dayhoff 1978) (Collins, Coulson, and Lyall 1988). More recently, 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, Harris, and Hunter 1993). A Bayesian model for measuring evolutionary distance using optimal un-gaped DNA alignments have been presented by Agarwal and States (1996). Similarly, several works have addressed the choice of gap penalty parameters (Waterman, Eggert, and Lander 1992) (Pearson 1995). For DNA sequence comparisons, systematic search procedures for finding the best gap penalty parameters have been developed (Waterman, Eggert, and Lander 1992) (Waterman 1994). Statistical approaches to sequence alignment which find the optimal gap and mismatch penalties for DNA alignments using iterative algorithms have been developed (Thorne, Kishino, and

Felsenstein 1991) (Thorne, Kishino, and Felsenstein 1992) (Allison, Wallace, and Yee 1992). However, the large size of the scoring matrices makes these approaches difficult for protein sequence alignment. Practical and statistically rigorous methods which simultaneously address gaping and the selection of scoring matrices for protein sequences require further investigation.

Among the approaches taken to address gaping in pairwise alignments, the algorithm by Sankoff (1972) is of particular interest here. This algorithm by-passes the need to specify gap penalties through the use of constrained optimization. Specifically, it finds the optimal alignment subject to the constraint that there are no more than k aligned blocks (or, equivalently, k-1 internal gaps). It has the added feature that those portions of the sequences that are not included in the aligned blocks are completely ignored, thus extending the concept of local alignment discussed above. However, it does so at the price of an even more vexing problem: the requirement for the specification of k. Furthermore, like other alignment algorithms it requires the specification of a scoring matrix. As we show below, Bayesian inference methods provide the means to overcome both of these limitations.

Bayesian statistics start with the specification of a joint distribution of all the quantities, both observed and unobserved, involved in the problem to be analyzed. Then basic probability rules are used to derive posterior distributions of the unknown quantities of interest. Final inferential statements are made rigorously based on these posterior distributions i.e., the conditional distribution of the variable(s) of interest given the observed data (Box 1980) (Gelman et al. 1995). However, a major difficulty for virtually all Bayesian analyses is computational complexity. While it is often easy to write down expressions for posterior distributions up to a normalizing constant, it is frequently infeasible to compute useful numerical summaries for the quantities of interest or normalizing constants, since they typically involve complicated high-dimensional integrations or summations.

Here we show how the pairwise alignment problem can be formulated as a Bayesian inference problem. In the Method section, we developed a Bayesian model to overcome the two major limitations of the Sankoff algorithm: the determination of the number of gaps and selection of a scoring matrix. When the series of scoring matrices are indexed by a distance measure, eg. the number of point accepted mutations, the complete posterior distribution of the distance is produced. Furthermore, the "evidence" against the null, the Bayesian analog of the classical p-value is calculated exactly. We present a modification of the optimal alignment algorithm of Sankoff (1972), to compute the large sum over all possible alignments with a time complexity of $O(kN^2)$.

In the Results section, we examine the behavior of Bayesian evidence in randomly shuffled sequences. We also illustrate this method with applications to sequences in the GTPases superfamily, giving posterior distributions of distances gaping, and the alignment. We show that this alignment corresponds well to a structural alignment.

Method

We begin by posing the alignment problem in terms of a joint distribution of the sequences of residues in a pair of biopolymers. Consider a pair of sequences $R^{(1)}=\{R_1^{(1)}\cdots R_I^{(1)}\}$ and $R^{(2)}=\{R_1^{(2)}\cdots R_J^{(2)}\}$, the alignment problem may be characterized by the following joint distribution,

$$logP(R_i^{(1)}, R_j^{(2)} \mid \Theta, \Psi, I) = \theta_{R_i^{(1)}} + \theta_{R_j^{(2)}} + I_{i,j}\Psi_{R_i^{(1)}R_j^{(2)}} \qquad (1),$$

where θ_{R_j} is the log marginal probability of observing each residue type R_j; Ψ_{R_j,R_i} is a matrix of the logarithms of residue interactions, ie. an alignment score matrix, such as PAM or BLOSUM matrices; $I_{i,j}$ is equal to one or zero and $\sum_I I_{i,j}\leq 1$ and $\sum_J I_{i,j}\leq 1$. In this formulation the alignment is characterized by the matrix I of missing values which specify which residue pairs align. We will assume in the remainder that the composition of both sequences are such that the first two terms on the right hand side of equation (1) can be dropped.

The assumption of collinearity requires that if $I_{i,j} = 1$, then

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

where for all $\Delta, \delta > 0$. With this assumption, an alignment is composed of aligned segments (or blocks) in both sequences interspersed with gaps. An aligned segment of length m, as determined by three indices i, j, and m, satisfies $I_{i,j} = 0$, $I_{i+m+1,j+m+1} = 0$, and

$$I_{i+l,j+l} = 1 \quad for \ \ l = 1,2,\ldots,m \qquad (3).$$

If no additional constraints are given, optimal alignments yield biologically unrealistic solutions which contain many short aligned 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 the alignments with at most (k-1) internal gaps.

Inferences on the number of gaps and on the scoring matrices can be made by examining the posterior distributions:

$$P(k \mid R^{(1)}, R^{(2)}) =$$

$$\frac{\sum_{\psi}\sum_{I} P(R^{(1)}, R^{(2)} \mid I, k, \psi)\, P(I \mid k)\, P(k)\, P(\psi)}{\sum_{k}\sum_{\psi}\sum_{I} P(R^{(1)}, R^{(2)} \mid I, k, \psi)\, P(I \mid k)\, P(k)\, P(\psi)} \quad (4),$$

and

$$P(\psi \mid R^{(1)}, R^{(2)}) =$$

$$\frac{\sum_{k}\sum_{I} P(R^{(1)}, R^{(2)} \mid I, k, \psi)\, P(I \mid k)\, P(k)\, P(\psi)}{\sum_{k}\sum_{\psi}\sum_{I} P(R^{(1)}, R^{(2)} \mid I, k, \psi)\, P(I \mid k)\, P(k)\, P(\psi)} \quad (5),$$

where we assume ψ and I are independent *a priori*. Here, ψ takes values on a finite set of scoring matrices, eg. the PAM or BLOSUM series of matrices.

Without any *a priori* information, we assume all the possible score matrices are equally likely, i.e., $P(\psi) = 1/N_{\psi}$ where N_{ψ} is the number of scoring matrices in the series. We further assume that all alignments with k segments are equally likely, i.e., $P(I \mid k) = \frac{1}{N_{I,J}^{(k)}}$ where $N_{I,J}^{(k)}$ is the number of collinear alignments with k or fewer blocks. If there are no blocks then,

$$I_{i,j} = 0, \quad i=1,2,\ldots I; j=1,2,\ldots J.$$

We assume that *a priori* the probability that two sequences are related (there are some matching blocks) or not related (null model: zero matching segments) are equal, formally $P(k=0)=0.5$ and $P(k>0)=0.5$. Furthermore, if the two sequences are related, all possible numbers of matching blocks are equally likely, $P(K=k \mid K>0) = \frac{1}{\kappa}$, 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

$$\kappa = \min\{\frac{L_s}{15}, 20\},$$

where L_s is the length of the shorter sequence.

In Bayesian statistics, the "evidence" for alternative hypotheses is obtained by examining the posterior probabilities of the alternatives. Here the evidence that the two sequences are related is obtained as follows:

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

where the supremum is taken over all prior distributions on K such that half the mass is on the null, ie. $P(K=0) = \pi_0 = 0.5$. This test can be ex-tended to the database search framework by setting π_0 to a large enough value to reflect the *a priori* chance that the query sequence is similar to a sequence taken at random from the database.

Algorithms

An algorithm for completing the sums in equations (4-6) is outlined in the first subsection. The second subsection briefly describes an algorithm for obtaining samples from the exact posterior distribution of I. Zhu, Liu and Lawrence (1997) give a more complete description of these and related algorithms.

Completing the Sums

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. Here we describe a recursive algorithm for completing sums over the large number of alignments. To complete this sum the algorithm below recursively builds a series of partial sums considering one residue or matched pair of residues at a time. Accordingly, at each iteration of this recursion the partial sum, say up to residue i in sequence 1 and residue j in sequence 2 with no more than t blocks, contains three components. These components correspond to the following steps: 1) a match of residue i, j denoted by ↘, yields

the partial sum $PC_{i,j}^{(t)}$; 2) an insertion in sequence 1 (deletion in sequence 2), denoted by $\downarrow$, yield the partial sum $PD_{i,j}^{(t)}$; and 3) an insertion in sequence 2 (deletions in sequence 1), denoted by $\rightarrow$, yields the partial sum $PR_{i,j}^{(t)}$. To avoid multiple counting of gaps, we impose the rule that a $\rightarrow$ move can not be followed by $\downarrow$. We complete the desired sums recursively, with initial values $PC_{i,j}^{(0)}=0$, $PD_{i,j}^{(0)}=0$ and $PR_{i,j}^{(0)}=1$ for $i\neq 0$ and $j\neq 0$; $PC_{i,0}^{(t)}=0$, $PD_{i,0}^{(t)}=1$ and $PR_{i,0}^{(t)}=0$; $PC_{0,j}^{(t)}=0$, $PD_{0,j}^{(t)}=0$ and $PR_{0,j}^{(t)}=1$ as follows:

1) A match extends a block, or starts a new block after an insertion in either sequences,

$$PC_{i,j}^{(t)} = [PC_{i-1,j-1}^{(t)} + PD_{i-1,j-1}^{(t-1)} + PR_{i-1,j-1}^{(t-1)}] * \exp(\Psi_{R_i^{(1)},R_j^{(2)}}) \quad (7);$$

2) An insertion in sequence 2 can follow any previous move and starts no new blocks,

$$PR_{i,j}^{(t)} = PC_{i,j-1}^{(t)} + PD_{i,j-1}^{(t)} + PR_{i,j-1}^{(t)} \quad (8);$$

3) An insertion in sequence 1 can extend an insertion in sequence 1 or follow a matching block, but because of the restriction to avoid multiple counting it can not follow an insertion in sequence 2,

$$PD_{i,j}^{(t)} = PC_{i-1,j}^{(t)} + PD_{i-1,j}^{(t)} \quad (9).$$

The marginal likelihood for exactly k blocks with Ψ given is

$$P(R^{(1)},R^{(2)} \mid k,\Psi) = PC_{I,J}^{(k)} + PD_{I,J}^{(k)} + PR_{I,J}^{(k)} - (PC_{I,J}^{(k-1)} + PD_{I,J}^{(k-1)} + PR_{I,J}^{(k-1)}) \quad (10).$$

The time complexity for completing this sum is $O(kIJ)$. Counts of the number of alignments with k or less blocks can be obtained by an analogous algorithm (Zhu, Liu and Lawrence, 1997).

Back Sampling Alignments

In the Bayesian view the posterior distribution of the alignment is required to characterizes the space of alignments. It is traditional in computational molecular biology to focus on the single best alignment. However, this mode of the posterior distribution will reasonably characterize the space only when it dominates all other alignments. When the sequences are subtly related or when favorable alignments contain many gaps, the optimal alignment describes only a very small fraction, in terms of probability, of the complete alignment space. An algorithm for exactly computing the posterior marginal alignment probability is given by Zhu, Liu, and Lawrence (1997). The joint posterior distribution can be approximated to any desired degree of accuracy with a sufficient sample from its posterior distribution. Samples from the posterior alignment distribution $P(I \mid R^{(1)},R^{(2)})$ are drawn through "backward" sampling in the following manner. Firstly, we draw a scoring matrix from $P(\Psi \mid R^{(1)},R^{(2)})$. Then we sample the total number of allowed gaps, k, from $P(k \mid R^{(1)},R^{(2)},\Psi)$. Finally, the alignment is drawn recursively as follows: starting from (I,J), there are three choices $\leftarrow$, $\uparrow$ and $\nwarrow$, which relate to forward step choices $\rightarrow$, $\downarrow$ and $\searrow$, respectively..Now draw a sample of these from the following posterior probabilities $\frac{PR_{I,J}^{(k)}}{PR_{I,J}^{(k)}+PD_{I,J}^{(k)}+PC_{I,J}^{(k)}}$, $\frac{PD_{I,J}^{(k)}}{PR_{I,J}^{(k)}+PD_{I,J}^{(k)}+PC_{I,J}^{(k)}}$, and $\frac{PC_{I,J}^{(k)}}{PC_{I,J}^{(k)}+PD_{I,J}^{(k)}+PC_{I,J}^{(k)}}$, respectively. If $\leftarrow$ is chosen, then we move to position (I,J-1); if $\uparrow$ is chosen, we move to (I-1,J); if $\nwarrow$ is chosen, we moved to (I-1,J-1). As described in greater detail by Zhu, Liu and Lawrence (1997), in a similar manner we sample recursively from any position (i,j), with at most t matching blocks to reach (1,1). The sampling probabilities depend on the previously sampled alignment step in a manner that is analogous to that taken on the forward summation stage of the algorithm.

Results

It is well known that Bayesian evidence tends to be conservative compared to a traditional p-value (Casella and Berger 1987). To examine how conservative it tends to be in this setting we aligned human α Hemoglobin with 10,000 random shuffles of Flagellar switch protein. As Figure 1 shows, Bayesian evidence is also conservative in this setting. For example a traditional p-value of .05 corresponds to a $P(K=0 \mid R^{(1)},R^{(2)}) = 0.170$.

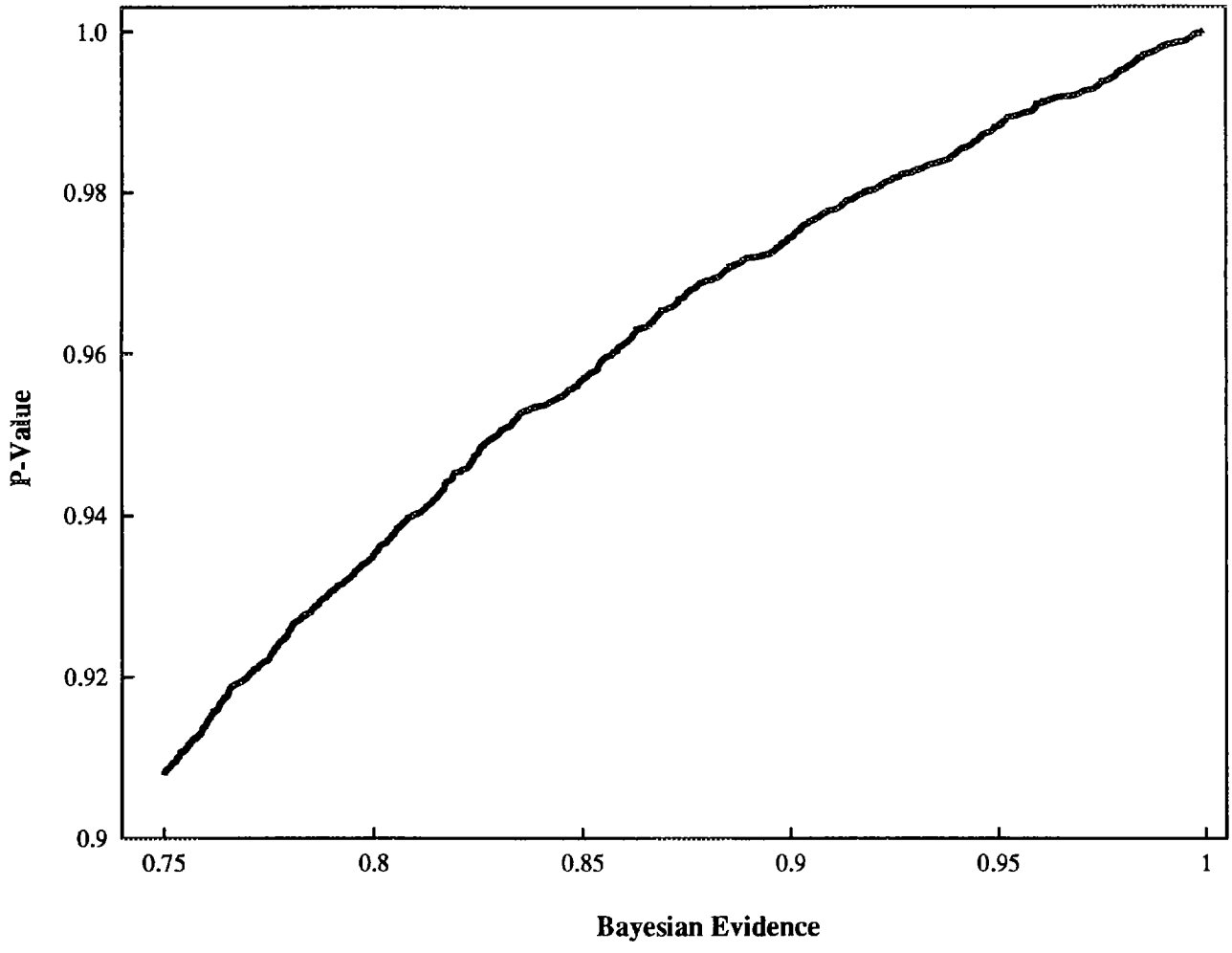


Figure 1. One minus the P-values vs. Bayesian evidence against null model. P-value is the fraction of the random shuffle at or above the given level of Bayesian evidence.

PAM	P32132	EF-G	1GIA	P32889
40	0.67	1.2e-3	2.8e-3	3.5e-3
60	0.25	0.03	3.3e-3	0.01
80	0.08	0.19	0.02	0.06
100	1.2e-3	0.10	0.07	0.09
120	2.8e-4	0.04	0.06	0.06
140	9.8e-6	0.63	0.22	0.14
160	1.4e-6	7.0e-3	0.11	0.15
180	1.8e-8	2.3e-3	0.04	0.04
200	1.6e-8	6.9e-3	0.16	0.15
220	3.2e-10	2.8e-4	0.07	0.05
240	3.2e-10	2.9e-4	0.08	0.08
260	6.0e-11	1.7e-4	0.05	0.08

Table 1. Posterior distribution of PAM distances $P(\psi \mid R^{(1)}, R^{(2)})$. $R^{(1)}$ is 1ETU, $R^{(2)}$ are P32132, EF-G, 1GIA, and P32889, which are members of GTPase superfamily.

Next we present results from the alignment of *E-coli* elongation factor τ (1ETU) with 4 other GTPases which were chosen to span a wide range of distance from 1ETU. Table 1 shows the posterior distributions of PAM distances. For three of the sequences ψ has a bimodal distribution. For example, in the alignment with EF-G there are modes at PAM80 and PAM140. This bimodality arises because of the variation of the level of conservation over these sequences. There are 4 well conserved motifs in the GTPase family, G1, G2, G3, and G4. The sequences corresponding to these motifs are well conserved, but the intervening sequences are not. Specifically using PAM80 there are strong peaks in the posterior alignment distribution corresponding to the G2 and G3 motifs, but not for the intervening sequences. While in the PAM140 the alignment extends across both peaks and includes the intervening sequences. Both of these provide good descriptions of the alignment, but compromises represented by PAM100 and PAM120 do not.

We also aligned the sequences using the BLOSUM series of matrices. Table 2 shows the distribution of BLOSUM distances. While there is no bimodality the flatness of several of these distribution reflects the degree of uncertainty about the distances between these sequences. Table 3 shows the Bayesian evidence against the null that the sequences are unrelated and the posterior distribution of number of aligned blocks, K.

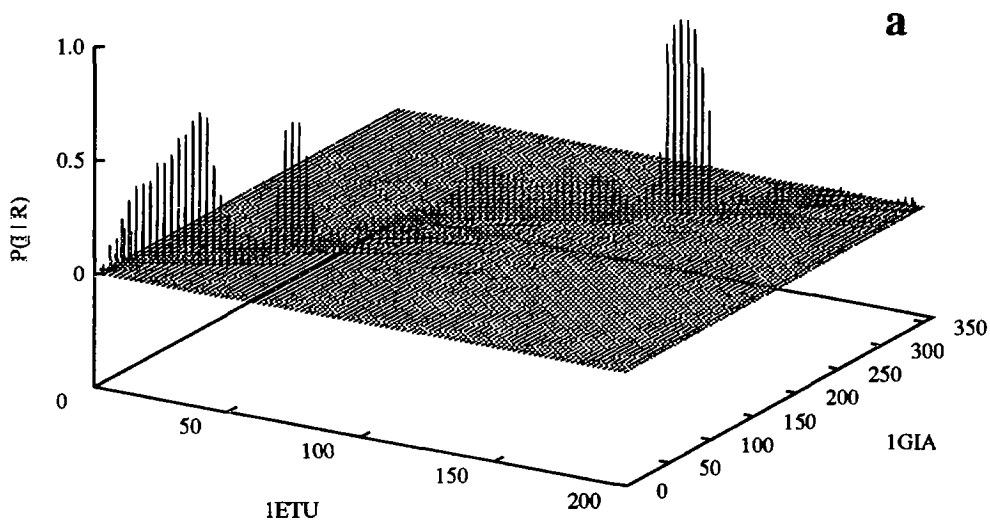
As Table 3 shows there is considerable uncertainty about the gaping in these sequences. As shown in Figure 2a, posterior distribution of the alignment of 1GIA to 1ETU, $P(I \mid R^{(1)}, R^{(2)})$, has four peaks. Overlapping of peaks, such as at the amino termini, and dips in the modes such as that shown in the third peak reflects the uncertainty in the gaping for this alignment. A more conventional representation of the alignment is shown in Table 4. As shown by the structural superposition in Figure 2b, this alignment corresponds well to a structural alignment. Gap penalty based alignment procedures tend not to align these sequences well (Zhu, Liu, and Lawrence 1997).

BLOSUM index	P32132	EF-G	1GIA	P32899
30	1.0e-8	1.3e-6	0.02	0.05
35	4.6e-6	1.9e-4	0.04	0.07
40	5.1e-5	1.5e-3	0.09	0.08
45	5.7e-4	0.02	0.18	0.10
50	3.7e-3	0.13	0.21	0.14
62	0.01	0.25	0.21	0.21
80	0.04	0.52	0.20	0.26
100	0.94	0.08	0.05	0.09

Table 2. Posterior distribution of BLOSUM matrices $P(\psi \mid R^{(1)}, R^{(2)})$. $R^{(1)}$ is 1ETU, $R^{(2)}$ are P32132, EF-G, 1GIA, and P32889, which are members of GTPase superfamily.

K	P32132	EF-G	1GIA	P32889
Evidence	1.0	1.0	0.95	0.91
1	1.1e-3	6.1e-5	6.2e-3	6.6e-3
2	0.01	7.1e-3	0.01	8.5e-3
3	0.01	0.07	0.02	0.01
4	0.05	0.10	0.04	0.02
5	0.10	0.12	0.06	0.02
6	0.15	0.12	0.07	0.03
7	0.15	0.12	0.08	0.04
8	0.14	0.12	0.08	0.05
9	0.12	0.10	0.08	0.06
10	0.09	0.08	0.08	0.08
11	0.07	0.06	0.08	0.09
12	0.05	0.04	0.08	0.11
13	0.03	0.03	0.08	0.11
14	0.02	0.02	0.08	0.11
15	0.01	0.01	0.08	0.11

Table 3. Posterior distribution of number of blocks $P(K \mid R^{(1)}, R^{(2)})$. $R^{(1)}$ is 1ETU, $R^{(2)}$ are P32132, EF-G, 1GIA, and P32889, which are members of GTPase superfamily.



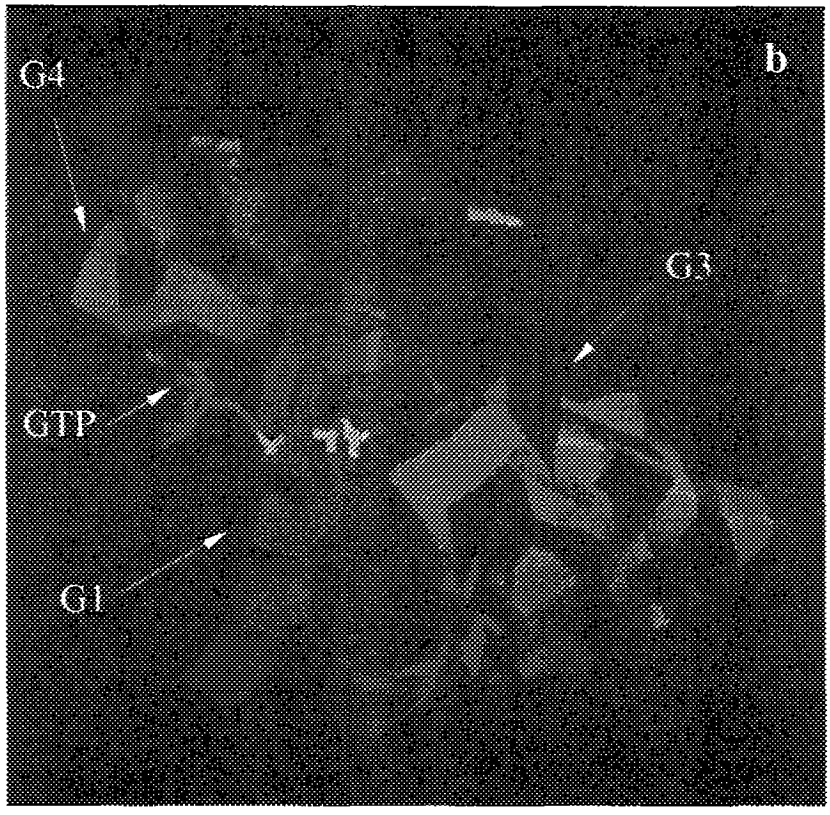


Figure 2. Alignment of 1ETU vs 1GIA obtained by Bayesian method. (a) Two dimensional histogram of $P(I \mid R^{(1)}, R^{(2)})$, which indicates that four motifs are conserved; (b) Structural superposition of motifs 1,3,4 found by Bayesian method. These motifs correspond to the well characterized motifs G1, G3, and G4 of the GTPase superfamily, respectively. As it is well known the G2 motif is not conserved over the GTPase superfamily, motif 2 does not correspond to the G2 motif. Motif 2 is not shown because almost none of the coordinates for its residues in 1ETU are available. However, from the small portion that is available it appears unlikely that motif 2 will correspond to the structural alignment.

```
3   KEKFERTKPHVNVGTIGHVDHGKTTLTAAITTV  35   36 LAKTYGGAARAFDQ  49
                          ....
                 ...............                  ..::::.
     ...:::::::::::::::::::::::..              .:::::::::.
   ::::::::::::::::::::::::::::::::::          ..::::::::::::
23 REDGEKAAREVKLLLLGAGESGKSTIVKQMKII  55   90 LKIDFGDAARADDA 103

 65 TSHVEYDTPTRHYAHVDCPGHADYVKNMITGAAQMDGAILVVAATDGPMP 114
                       .                          .
    ..........:::::::::::::::::::::::::::::::::::::....
183 IVETHFTFKDLHFKMFDVGGQRSERKKWIHCFEGVTAIIFCVALSDYDLV 232

129 PYIIVFLNKCDMVDDE 144
           ...
        ::::::::::.
       ::::::::::::.
     .:::::::::::::::
     ::::::::::::::::
261 TSIILFLNKKDLFEEK 276
```

Table 4. The alignment of 1ETU (top) and 1GIA (bottom) obtained by Bayesian method. The number of dots between the aligned residues are proportional to $P(I \mid R^{(1)}, R^{(2)})$. The first, third and fourth aligned segment corresponds to structural/functional motif G1, G3 and G4 as labeled in Fig. 2b, respectively.

Conclusion

In the applications presented here we have used uninformed *a priori* distributions for all unobserved variables. Accordingly the results were obtained without *a priori* knowledge of these variables. When appropriate *a priori* information is available it can be used to improve the performance of these procedures. Further investigation of useful prior distributions appears to be warranted. Furthermore, while only protein alignments were given, application to DNA alignments is straight forward.

The approach is most similar to that presented by Thorne et al. (1991 and 1992) and Allison et al. (1992) in that these methods simultaneously ad-dress gaping and mismatch scoring while summing over all possible alignments. Besides the obvious difference that we employ Bayesian rather classical statistics, our approach differs from theirs in a number of ways. While in principle their approaches for DNA alignments can be extended to that for 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 the EM algorithm, which only guarantees local optimality. The approach described here gives the complete posterior distribution including its global mode. Furthermore, their use of point estimates implies that uncertainty associated with these parameters is not incorporated into resulting alignment distribution. As shown by our results that posterior distributions for penalty parameters are often flat and sometimes bimodal, 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 Agrawal and States (1996) also bear similarity to our method but they do not simultaneously address gaping 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 features of this algorithm are worth summarizing. The algorithm has a time complexity of $O(kN^2)$ which is comparable with other pairwise alignment algorithms. The method can be extended to multiple sequence alignment through Gibbs sampling, using the approach described by Liu and Lawrence (1995). The posterior distributions contain substantially more informa-

tion than a point estimate as illustrated by the flatness and bimodality of some posteriors of PAM distances between a pair of sequences. These distributions should provide a useful means to reflect uncertainty in molecular evolution.

The GTPase example illustrates another important feature of the algorithm. The algorithm will align only those subsequences of the two biopolymers which the data indicate to be conserved. Not surprisingly here and in other examples these subsequences often correspond to motifs that form ligand binding pockets (Zhu, Liu and Lawrence, 1997). This feature has two facets. First, it improves the alignment by permitting the algorithm to ignore and thus not be confused by unrelated proportions of the sequences. Second, it points to the conserved subsequences which are likely to play important functional or structural roles.

Over the last few years HMMs and Gibbs sampling algorithm have shown the potential of statistically based algorithms to contribute to the challenging problems in computational molecular biology. The benefits of these earlier efforts have been primarily algorithmic. Here we show that the statistical inference features of Bayesian statistics also promise to make important contributions to computational molecular biology.

Acknowledgment

This work was partially supported by NIH Grant # 5R01-1HG0125702, Department of Energy Grant # DEFG0296ER2266, and the Computational Molecular Biology and Statistics core of the Wadsworth Center.

References

Agrawal, R. and States, D. J. 1996. A Bayesian evolutionary distance for parametrically alignment sequences. *J. Comp. Biol.* 3(1):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.

Box G.E.P. 1980. Sampling and Bayesian inference in scientific modelling and robustness. *J. Royal Stat. Soc.* 143:383-430.

Casella, G., Berger, R.L. 1987 Reconciling Bayesian and Frequentist evidence in the one-sided testing problem. *J. Amer. Stat. Assoc.* Vol. 82, 106-111.

Collins, J.F., Coulson, A.F.W., and Lyall, A. 1988. The significance of protein sequence similarities. *Comput. Appl. Biosci.* 4:67-71.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. eds. 1995. *Bayesian data analysis.* New York: Chapman Hall.

Liu, J.S. and Lawrence, C.E. 1995. Statistical models for multiple sequence alignment: unifications and generalizations. In Proc. Amer. Statist. Assoc., Statistical Computing Section, 21:1-8. Orlando, FL: ASA Press.

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

Pearson, W.R. 1995. Comparison of methods for searching protein sequence databases. *Protein Science* 4:1145-1160.

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, vol. 5, suppl. 3. Natl. Biomed. Res. Found, Washington, pp.353-358.

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

States, D. J., Harris, E. L., and Hunter, L. 1993. Computationally efficient cluster representation in molec-

ular sequence megaclassification. In Proc. ISMB93. 1: 387-394. Bethesda, MD: AAAI Press.

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., Eggert, M., and Lander, E. 1992. Parametric sequence comparisons. *Proc. Natl. Acad. Sci. USA* 89:6098-6093.

Waterman, M.S. 1994. Parametric and ensemble sequence alignment algorithms. *Bull. Math. Biol.* 56:743-767.

Zhu, J., Liu, J.S. and Lawrence, C.E. 1997 Bayesian adaptive alignment algorithms. Forthcoming.