Monte Carlo Sequence Alignment

Gary A. Churchill
Biometrics Unit
Cornell University
Ithaca, NY 14853

1 Introduction

1.1 Motivation

The alignment of molecular sequences is a problem central to many important questions in molecular biology and evolution. The main theme in the development of sequence alignment methods has been to obtain an optimal alignment between two (or more) sequences. The optimality criterion is typically based on the sum of scores assigned to substitutions, insertions and deletions required to transform one sequence into another. The scores themselves are fixed and often arbitrary.

The problem studied here lies at the intersection of two lines of research. The first is concerned with the sensitivity of alignments to choice of scores. Fitch and Smith (1983) introduced the idea of dividing the parameter space of possible scoring systems into regions within which the (set of) optimal alignment(s) is invariant. Distinct optimal alignments are obtained when scores are chosen from different regions of the space. A recent review of this problem is provided by Vingron and Waterman (1994). A second line of research is the study of suboptimal alignments. This work has been driven by the observation that an optimal alignment is not neccessarily a biologically correct alignment. However, biologically correct alignments are often nearly optimal when the scoring system is well chosen. A recent review of suboptimal alignment methods is provided by Vingron (1996).

We consider sequence alignment in the context of an explicit stochastic model of sequence evolution. The model parameters have a direct interpretation as rates or probabilities of sequence transformations. These parameters are generally not known and cannot be estimated to arbitrary precision. In fact, given the small amounts of data available in most problems, accurate estimation of evolutionary model parameters may be impossible. We believe that it is possible to use available data to estimate alignment parameters but that uncertainty in these estimates should be explicitly accounted for in the assessment of the reliability of an alignment.

Ideally, alignment inference would consider all sets of scores and allow biologically realistic sets to have a greater impact on the inference. Furthermore it will often be useful to consider several alternative alignments and to associate some measure of confidence with each. These considerations motivate our Bayesian approach to the study of probability distributions associated with sequence alignments. We have developed sampling algorithms because these distributions are not available in any simple form. We demonstrate that sampling algorithms are useful for assessing the reliablity of a multiple alignment. They also have the potential to provide tools for studying the reliability of inferences, such as phylogenetic tree construction, that are based on sequence alignments.

1.2 Approach

Our goal in this work is to develop algorithms to sample from the marginal posterior distribution of an alignment. Let $\mathbf{Y}$, θ, $\vec{\alpha}$ denote the sequence data, the model parameters and a sequence alignment, respectively. The desired marginal distribution can be obtained by integration with respect to θ of the joint distribution on alignments and model parameters,

$$\Pr(\vec{\alpha} \mid \mathbf{Y}) = \int \Pr(\vec{\alpha}, \theta \mid \mathbf{Y}) \partial\theta. \qquad (1)$$

Unfortunately there is no simple representation for this distribution. Instead we resort to a Markov chain Monte Carlo (MCMC) algorithm (Gelfand and Smith, 1990) to generate samples. The algorithm works by iteratively

sampling from conditional distributions

$$\theta^{(s)} \sim \Pr\left(\theta \,|\, \mathbf{Y}, \vec{\alpha}^{(s-1)}\right) \quad (2)$$

$$\vec{\alpha}^{(s)} \sim \Pr\left(\vec{\alpha} \,|\, \mathbf{Y}\theta^{(s)}\right). \quad (3)$$

In the limit as $s \to \infty$, the sampled alignments will have distribution approaching (1).

The conditional distribution of the model parameters $\Pr(\theta \,|\, A, B, \vec{\alpha})$ is defined on a (subset of) Euclidean space and is generally straightforward to sample from. The conditional probability distribution on alignment paths $\Pr(\vec{\alpha} \,|\, \mathbf{Y}, \theta)$ is defined on the space of sequence alignments and presents a more challenging task.

We will first outline the sampling algorithm for pairwise alignments and then extend this to a special case of multiple sequence alignment. Details for specific implementations can be found in Churchill (1995), Thorne and Churchill (1995) and Churchill and Lazareva (1996). An example is provided to demonstrate the utility of alignment sampling.

2 Pairwise Alignment

2.1 The Path Graph

The observable data are two sequences of characters $\mathbf{Y} = \{A, B\}$ where $A = a_1 a_2 \cdots a_{n_A}$ and $B = b_1 b_2 \cdots b_{n_B}$ are assumed to be related by descent from a common ancestor. If the model of evolution is time reversible, we can ignore the common ancestor (Felsenstein,1981) and assume that B is a descendant of A. Thorne et al. (1991) describe a time reversible model of sequence evolution with insertion rate λ, deletion rate μ and substitution rate s. There are only two free parameters in this model $\theta = \{\lambda, s\}$ due to the reversibility constraint. More elaborate models can also be considered, e.g., Thorne et al. (1992).

Pairwise alignments can be represented (Figure 1) as a directed graph on a two dimensional grid of vertices indexed by $i = 0, \ldots, n_A$ and $j = 0, \ldots, n_B$. The sequence A is shown along the top margin of the grid such that the base a_i falls between the columns indexed by $i-1$ and i. Similarly, the sequence B is shown down the left margin of the grid. An alignment is shown as a path, a connected sequence of arcs, traversing the matrix from the upper left vertex to the lower right vertex by a series of east ($\rightarrow$), southeast ($\searrow$) and south ($\downarrow$) moves. Thus, an alignment can be summarized as a sequence

$$\vec{\alpha} = \alpha_1 \alpha_2 \cdots \alpha_n \quad (4)$$

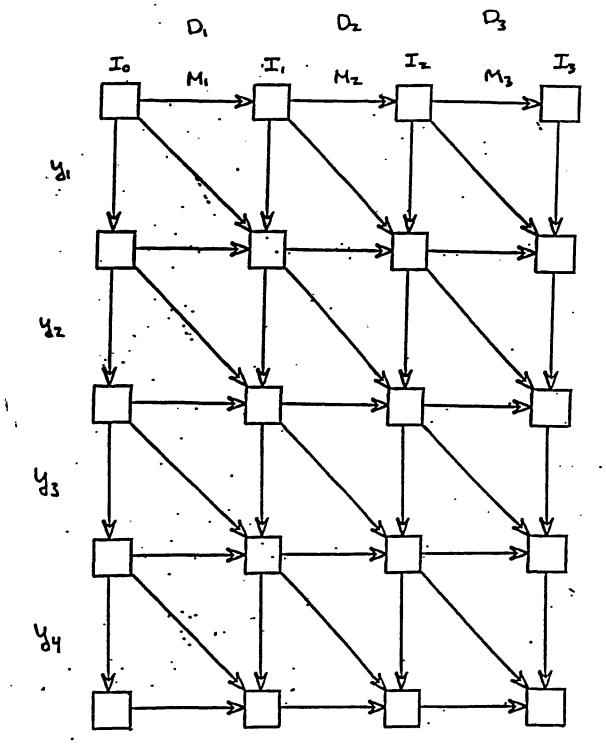

Figure 1: A Pairwise Alignment Path Graph

where n is the number of arcs in the path and

$$\alpha_h = \begin{cases} \rightarrow & \text{deletion of a base from } A \\ \searrow & \text{substitution of a base from } A \text{ into } B \\ \downarrow & \text{insertion of a base into } B. \end{cases}$$

2.2 The Alignment Sampling Algorithm

The algorithm employed to sample from the distribution (3) is similar in style to standard dynamic programming. A forward pass through the matrix is used to compute conditional probabilities for partial alignments that end with each node in the path graph. However, instead of choosing the optimal score at each step, our algorithm sums over the three arcs entering the node. Thus we compute an integrated likelihood over all possible paths. The result of the forward pass algorithm is a set of conditional probabilities

$$q_{kl}(i,j) \equiv \Pr\left(\mathcal{A}_k(i,j) \,|\, \mathcal{A}_l^r(i,j), A_i, B_j, \theta\right) \quad (5)$$

where $\mathcal{A}_k(i,j)$ is the set of all paths entering node (i,j) on a k–arc, $\mathcal{A}_l^r(i,j)$ is the set of paths leaving node (i,j) on an l–arc, $A_i = a_1, \ldots, a_i$ and $B_j = b_1, \ldots, b_j$.

Given the probabilities as computed in the forward pass, we sample an alignment by tracing back from the

lower right corner of the pathgraph to the upper left. In contrast to the usual dynamic programming traceback that chooses a fixed (set of) optimal path(s), the sampling algorithm selects a sequence of arcs at random to generate a probable paths. Given that the traceback has reached a node (i,j) and the last arc sampled was an l–arc, $\alpha_{h+1} = l$, the next arc sampled will be a k–arc, $\alpha_h = k$, with probability $q_{kl}(i,j)$. The probability of sampling an alignment path $\vec{\alpha}$ by this algorithm is the product of conditional probabilities $q_{kl}(i,j)$ at each step taken in the traceback. Thus the algorithm generates a sample from the desired probability distribution (3).

3 Multiple Alignment

3.1 The Setting

We consider a set of sequences

$$\mathbf{Y} \equiv \left\{ \begin{array}{lcl} \mathbf{y}_1 & = & y_{1,1}\, y_{1,2}\, \cdots, y_{1,n_1} \\ \vdots & & \vdots \\ \mathbf{y}_N & = & y_{N,1}\, y_{N,2}\, \cdots, y_{N,n_N} \end{array} \right\}.$$

and assume that they have evolved independently from a common *prototype* sequence, $\mathbf{r} = r_1, \ldots, r_L$ by a process that introduces substitutions, deletions and insertions. This process of independent evolution can be represented as a hidden Markov model (HMM) (Krogh et al. 1994). A schematic is shown in Figure 2. The backbone of the model consists of states $\{M_1, M_2, \ldots, M_L\}$ such that each M–state is associated with an element of the prototype sequence, *i.e.*, M_i is associated with r_i. As the hidden Markov chain is traversed, the states output characters and generate one of the observed sequences, $\mathbf{y}_i$. A substitution occurs when the output letter of state M_i differs from the prototype r_i. A deletion error occurs when the state D_i is visited, thus bypassing M_i. D–states generate no output. Insertions are generated by the states I_i. There are two sets of parameters associated with the hidden Markov model. The parameters Π determine the output distributions of individual states. The parameters Λ govern the rate of transitions between states in the hidden Markov chain.

The sequences of paths through the hidden Markov chain that produced $\mathbf{Y}$ will be denoted by

$$\mathbf{S} \equiv \left\{ \begin{array}{l} \mathbf{s}_1 = s_{1,1}\, s_{1,2}\, \cdots, s_{1,n_1} \\ \vdots \\ \mathbf{s}_N = s_{N,1}\, s_{N,2}\, \cdots, s_{N,n_N} \end{array} \right\}.$$

We note that there is a one-to-one correspondence between the possible sequence of states $\mathbf{s}_i$ and the paths from $(0,0)$ to (L,n) on an alignment path graph (Figure 1). To see this, let $\vec{\alpha}_i = \{\alpha_{i,1}, \ldots, \alpha_{i,n_i}\}$ where

$$\alpha_{i,j} = \left\{ \begin{array}{ll} \rightarrow & \text{if } s_{i,j} \text{ is an } I\text{-state,} \\ \searrow & \text{if } s_{i,j} \text{ is an } M\text{-state,} \\ \downarrow & \text{if } s_{i,j} \text{ is a } D\text{-state.} \end{array} \right.$$

Thus the problem of sampling $\mathbf{s}$ can be substituted by the problem of sampling $\vec{\alpha}$.

The alignment sampling algorithm is now essentially the same as described above for pairwise alignments. One additional step is required to sample the prototype sequence $\mathbf{r}$ and each sequence $\mathbf{y}_i$ is aligned independently to the sampled prototype. We note that forward pass algorithm is similar to a Baum–Welch algorithm (Rabiner, 1989). The differences between the HMM setting and the pairwise alignment setting are 1) that the protoype sequence $\mathbf{r}$ plays the role of the ancestor and 2) that the rates of substitution, insertion and deletion are no longer constrained to be constant.

3.2 An Example

Table 1 shows an example of six DNA sequences $(\mathbf{y}_1, \ldots, \mathbf{y}_6)$. These sequences are from a shotgun sequencing experiment and include the ambiguous base character N. Because they are distinct copies of the same DNA region, the independent evolution from a prototype model is plausible.

A maximum *a posteriori* alignment of these sequences is shown in Figure 3. In a run of 100,000 MCMC steps, using these sequences, 17,488 distinct multiple alignments were explored. The most frequent variants of the multiple alignment are summarized in Table 2. These variants identify three regions where the multiple alignment is least reliable. The 20 most probable alignments include variant 1a with all combinations of 3a,b,c,d,e and 5a,b,c,d. The next 20 include variant 1b with all combinations of 3a,b,c,d,e and 5a,b,c,d. The next ten include all combinations of variants 1a,b and 3a,b,c,d,e together with 5e.

The rates of insertion and deletion were held constant across all sites in our HMM but substitution rates were allowed to vary from site to site. The posterior mean for the insertion and deletion rates were, respectively, $\hat{\lambda} = 0.0186$ ($\widehat{\text{sd}} = 0.00813$) and $\hat{\mu} = 0.022$ ($\widehat{\text{sd}} = 0.00871$). The sequence labeled "consensus" in Figure 3 is the estimated posterior mode of the prototype sequence $\mathbf{r}$. The character n in position 10 reflects uncertainty in the assignment of r_{10}.

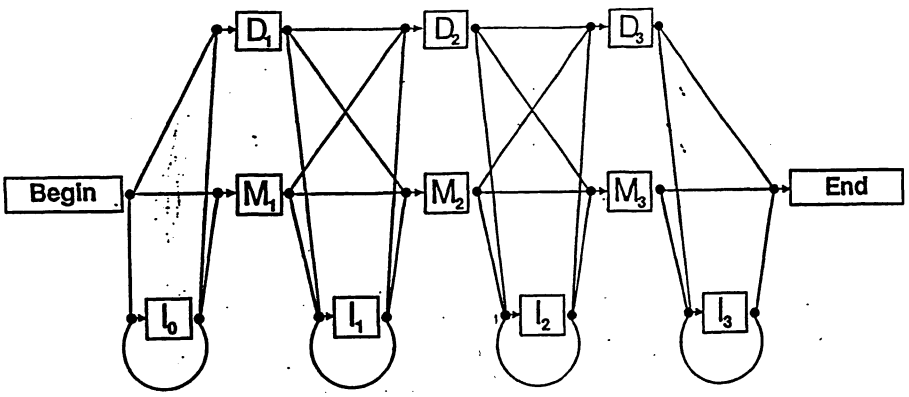

Figure 2: A Hidden Markov Model

TAGACAGGNGCCCCACTGGAGGAATGAGGTCACCAACCAACCTTCAAAAACTT
TAGACAGGGNCCCACTGGAGGAATGAGGTCACCAACCAACCTTCAAAAACTT
TAGANAGGGCCTCCACTGGGGAAATGAAGGTACCNACCAACCTTCAAAACTT
TAGACCAGGNGCTCCACTGGAGGAATGAGGTCACCAACCAACCTTCAAAAACTT
TAGACAGGGCCTCCACTGGAGATNTGAGGTCACCAACCAACCTTCAAAAACTT
TAGACAGGGGCTCCACTGGAGGAATGAGGTCACCAACCAACCTTCAAAAACTT

Table 1: An unaligned set of DNA sequences

4 Directions for Future Work

One advantage of the Bayesian approach to inference is that, through the posterior distribution, we can quantify the uncertainty in the inference of a complex discrete structure such as a sequence alignment. Our ability to summarize and visualize these distributions is limited, but with careful attention to particular examples, effective summaries of uncertainty can be developed. Further efforts to characterize the uncertainties associated with pairwise and multiple sequence alignments are needed. For pairwise alignments, simple graphical summaries based on the path graph are feasible. The problem of summarizing uncertainty in multiple alignments appears to be more challenging.

Sequence alignment and phylogeny inference are interconnected. The usual practice of basing a phylogenetic inference on a specific sequence alignment is inherently circular because the sequence alignment itself implicitly or explicitly assumes a specific evolutionary tree (Thorne and Kishino 1992). This circularity is troublesome and could, in principle, be addressed with a Markov chain Monte Carlo approach. The idea would be to alternately sample a tree given an alignment and an alignment given a tree. Progress toward this goal could be made by developing algorithms to sample alignments on a given tree. One approach would involve sampling ancestors at each interior node.

Development of more realistic stochastic models that can allow for rate heterogenity and events such as multiple insertions and deletions is needed. However, as models grow in complexity two problems arise. The first is computational and it is hoped that the ever increasing speed and efficiency of computing hardware will help us to keep abreast of this problem. The second problem is more fundamental and is related the fact that a model is never true. How much faith can we place in the answers provided by model based inferences? Only serious and hard analysis of robustness issues, perhaps with the support of extensive simulation studies will help us address this question.

References

[1] Churchill GA (1995) Accurate restoration of DNA sequences, *Case Studies in Bayesian Statistics* vol. II, eds. C. Gatsaris, J.S. Hodges,

Figure 3: A Multiple Sequence Alignment

sequence		variant
1	a	GG(N)GCC-
	b	GGNGCC
3	a	-AAAA
	b	A-AAA
	c	AA-AA
	d	AAA-A
	e	AAAA-
5	a	-GAT(N)TG
	b	G-AT(N)TG
	c	-GA(T)NTG
	d	G-A(T)NTG
	e	GATNTG

Table 2: Most frequent alignment variations. Brackets () indicate insertions.

R.E. Kass, N.D. Singpurwalla, Springer-Verlag, New-York, pp. 90-148.

[2] Churchill GA, Lazareva B (1996) Bayesian restoration of a hidden Markov chain with appliactions to DNA sequencing. Submitted: J. American Statistical Assoc.

[3] Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

[4] Fitch WM, Smith TS (1983) Optimal Sequence Alignments. Proc. Natl. Acad. Sci. USA 80:1382–1386.

[5] Gelfand AE, Smith AFM (1990) Sampling Based Approaches to Calculating Marginal Densities. J. Amer. Statist. Assoc. **85**:398–409.

[6] Krogh A, Brown M, Mian, IS, Sjölander K, Haussler D (1994) Hidden Markov models in computational biology: Applications to protein modeling. J. Mol. Biol., **235**: 1501-1531.

[7] Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE, 77, 257-286.

[8] Thorne JL, Churchill GA (1995) Estimation and Reliability of molecular sequence alignments. Biometrics, 51: 100-113.

[9] Thorne JL, Kishino H, Felsenstein J (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol. 33:114–124.

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

[11] Thorne JL, Kishino H (1993) Freeing phylogenies from artifacts of alignment. Mol. Biol. Evol. 9(6):1148–1162.

[12] Vingron M (1996) Near–optimal sequence alignment. Current Opinion in structural biology 6: 346-352.

[13] Vingron M and Waterman MS (1994) Sequence alignment and penalty choice: Review of concepts, case studies and implications. J. Mol. Biol. 235: 1-12.