J Mol Evol (1994) 39:418–430

Journal of Molecular Evolution

The Posterior Probability Distribution of Alignments and Its Application to Parameter Estimation of Evolutionary Trees and to Optimization of Multiple Alignments

L. Allison,* C.S. Wallace

Department of Computer Science, Monash University, Australia 3168

Received: 27 June 1993 / Revised: 25 March 1994 / Accepted: 8 May 1994

Abstract. How to sample alignments from their posterior probability distribution given two strings is shown. This is extended to sampling alignments of more than two strings. The result is first applied to the estimation of the edges of a given evolutionary tree over several strings. Second, when used in conjunction with simulated annealing, it gives a stochastic search method for an optimal multiple alignment.

Key words: Alignment — Estimation — Evolutionary tree — Gibbs sampling — Inductive inference — Monte-Carlo method — Multiple alignment — Sampling — Stochastic

Introduction

Estimating the evolutionary ''distances'' between some given strings, for example, DNA sequences, and finding an alignment for them, are examples of inductive inference problems. We rarely know the ''right'' answer to a question of evolution and must be satisfied with a ''good'' hypothesis. In previous papers (Allison et al. 1992a; Yee and Allison 1993), minimum message-length encoding was used to infer the relation, if any, between two strings and to compare models of evolution or relation. It is not possible to extend the algorithms given there directly to $K > 2$ strings in any simple way due to rapidly increasing algorithmic complexity. A stochastic process is one possible way around this difficulty. Such a process is described for two problems given a set of $K \geqslant 2$ strings and given an evolutionary tree over them: first estimating the evolutionary ''distances'' on the edges or arcs of the tree and second finding a good alignment of all of the strings. The method relies on sampling from the posterior probability distribution of alignments.

In an inductive inference problem we are given some data D and wish to form a ''good'' hypothesis H about the data. The language of hypotheses should be tailored to the problem we wish to solve. The following standard definitions are useful:

$$\begin{aligned} P(H\&D) &= P(H).P(D|H) \\ &= P(D).P(H|D) \text{—joint probability} \end{aligned}$$

Definitions:

$P(H|D)$ = posterior probability of H given D

$P(H)$ = prior probability of H

$P(D|H)$ = likelihood of H

$ML(E)$ = $-\log_2(P(E))$—message length of event E

$$\begin{aligned} ML(H\&D) &= ML(H) + ML(D|H) \\ &= ML(D) + ML(H|D)\text{—joint message length} \end{aligned}$$

$$\begin{aligned} ML(H|D) &- ML(H'|D) \\ &= ML(H) - ML(H') + ML(D|H) - ML(D|H') = -\log_2 \text{ posterior odds ratio of } H \,\&\, H' \end{aligned}$$

Correspondence to: L. Allison
* *email address:* lloyd@cs.monash.edu.au

In the present context, the data consist of some given strings. Alignments and estimates of evolutionary distances are different kinds of hypotheses about those strings. $P(D|H)$ is called the *likelihood* of the hypothesis H; it is a function of the data given the hypothesis and it is not the probability of the hypothesis. $P(H)$ is the prior probability of the hypothesis H. $P(H|D)$ is the posterior probability of H.

The message length (ML) of an event E is the minimal length, in bits, of a message to transmit E using an optimal code. A long message indicates a low probability and a short message indicates a high probability. It is convenient to work with message lengths, rather than probabilities, for a number of reasons: Typical probability values can be very small and the message lengths are of a more convenient magnitude for handling by computer and by person. The message paradigm reinforces the view that there should be no hidden parameter values associated with the hypothesis itself. All such values are costed explicitly in ML(H). All real-valued parameters must be stated to some optimum but finite accuracy, as described by Wallace and Boulton (1968). In sequence comparison there is a natural *null theory* which has a message length, being that required to state the given strings individually; this takes approximately two bits per character in the case of DNA. It provides a method of hypothesis testing. The null theory assumes that there is no pattern or structure in the data. It includes the assumption that an individual string is random. Wallace and Freeman (1987) gave the statistical foundations of minimum message-length encoding.

We often wish to find the "best" hypothesis, from a specified class of hypotheses. (Depending on the application, this might be the best evolutionary tree, a good estimate of the evolutionary "distance" between strings, or the best alignment). It is generally possible to calculate, or at least to give a good approximation of, ML(H) and ML($D|H$) under reasonable assumptions. ML(H) can even be ignored if it is a constant for all members of a class of hypotheses. It is not often possible to calculate ML(D), which is unfortunate, for it would yield ML($H|D$). However, by subtracting ML($H'\&D$) from ML($H\&D$) it is possible to get a posterior $-\log_2$ odds-ratio for two competing hypotheses H and H'; the shorter message indicates the more likely hypothesis. If one hypothesis is the null theory this also gives the hypothesis test.

An evolutionary tree which is a good hypothesis makes good predictions about the sorts of tuples of characters that occur in good alignments. (A *tuple* consists of the characters that appear in a column of a conventional alignment.) For example, given four related strings $s1$ to $s4$ and assuming that all edges are similar, the tree (($s1$ $s2$)($s3$ $s4$)) predicts that tuples of the form *xxyy* appear more often than *xyxy*, whereas (($s1$ $s3$)($s2$ $s4$)) favors *xyxy* over *xxyy*. Similarly, a good alignment contains many highly probable tuples, which leads to a short message length for the data. In the extreme case of the strings being identical they can all be transmitted for little more than the cost of transmitting one.

An optimal alignment of a set of strings can be used to infer an estimate of evolutionary distances and is sometimes used primarily for that purpose. However, a single alignment is a much more detailed sort of hypothesis that an estimate of distances because it also states which characters of the descendant strings are related. In statistical terms, the optimal alignment problem has many *nuisance parameters*—if the question is one of evolutionary distances. The message for the data given the estimate of distances should not be based on one alignment. It is not the case that distance estimates are better than alignments or vice versa; they are answers to different questions. Alignments are useful in their own right, for some purposes.

Yee and Allison (1993) showed that in order to obtain an unbiased estimate of the evolutionary "distance" between two strings it is necessary to use a weighted average of estimates from *all* alignments, whereas the use of a single optimal alignment gives a biased estimate. The average can be computed in a few steps, each step taking time proportional to the product of the string lengths. This is feasible for two strings but not more unless the strings are short. In this paper, a stochastic process is used to average over not all but many alignments of K strings so as to get the estimates of the distances on the edges of a tree over the strings in an acceptable time. This is an example of Gibbs sampling or a Monte-Carlo method (Hastings 1970). Random alignments are generated from the posterior probability distribution of alignments. When used in conjunction with simulated annealing this also gives a stochastic search process for a good alignment. We use *tree costs* because these correspond to explicit evolutionary hypotheses; the edges of a tree are modeled individually. We note that Lawrence et al. (1993) describe a Gibbs sampling strategy for finding *ungapped* signals in a set of protein sequences. That work relates each protein to a central model which implicitly represents the constraints of the typical member of the set. It is using a form of *star costs* which is probably more suitable for proteins, particularly if they are only distantly related.

There is important prior work in the treatment of alignment and evolutionary trees as inductive inference problems. Bishop and Thompson (1986) first cast pairwise alignment as a maximum-likelihood problem, summing the probabilities of all alignments. Thorne et al. (1991, 1992) extended the maximum-likelihood method to more general models of evolution, including conserved and variable regions, and related probabilities of mutation to time. Allison et al. (1992a) included the model or hypothesis cost in pairwise alignment and compared evolutionary models of different complexities on an equal footing. Felsenstein (1981, 1983) treated the inference of an evolutionary tree from a given multiple

```
string A = TAATACTCGGC
string B = TATAACTGCCG

mutation instructions:
   copy
   change(ch) NB. ch differs from the corr' char in string A.
   del
   ins(ch)

a mutation sequence:
   copy; copy; delete; copy; copy; ins(A); copy; copy;
   del; copy; change(C); copy; ins(G)

generation instructions:
   match(ch)
   mismatch(chA,chB) NB. chA≠chB
   insA(chA)
   insB(chB)

a generation sequence:
   match(T); match(A); insA(A); match(T); match(A); insB(A); match(C);
   match(T); insA(C); match(G); mismatch(G,C); match(C); insBG

equivalent alignment:
   TAATA-CTCGGC-
   || || || | |
   TA-TAACT-GCCG
```

Fig. 1. Basic models.

alignment as a maximum likelihood problem. All these are part of a large and important trend to make models of evolution explicit for better scrutiny and to place subsequent inferences on a sound statistical footing.

Models of Evolution

We model the evolution of a parent string A into a child string B by a finite-state *mutation machine.* Such a machine can *copy* a character, *change* a character, *insert* a character, or *delete* a character. Inserts and deletes are collectively called *indels.* The machine reads A and a sequence of mutation instructions and produces B.

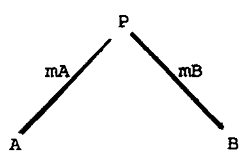
If we wish to consider A and B to be of equal standing, we model their relation under a finite-state *generation machine.* Such a machine can generate the same character in A and in B (match), generate different characters in A and in B (mismatch), generate a character in A only (ins$_A$), or generate a character in B only (ins$_B$). The machine reads a sequence of generation instructions and produces A and B.

Traditionally, a (pairwise) *alignment* is used to show a hypothetical relationship between two strings. Each string is padded out with zero or more *null* characters "—" until they have the same lengths. The padded strings are then written out one above the other. The two characters in a column are called a *pair,* a two-tuple, or just a *tuple.* The all-null tuple is not allowed. It is convenient to write a pair as $\langle x,y \rangle$ or just "xy" in text.

There is an obvious correspondence between alignments of two strings, sequences of generation instructions, and sequences of mutation instructions, as illustrated in Fig. 1, and they will be used interchangeably as convenient. Given a finite-state generation machine, say, the probability of a particular sequence of instructions of a given length is the product of the instructions' individual probabilities and the message length of the sequence is the sum of their individual message lengths. This depends on P(match), P(mismatch), P(ins$_A$), and P(ins$_B$), which are called the parameters of the machine. The parameters might be given a priori, but in general they must be estimated from the strings. They correspond to the evolutionary "distance" of the strings. Note that an estimate can be got by counting instructions in one, or more, instruction sequences. The parameters consist of several values; a machine has more than one degree of freedom. If it is necessary to reduce them to a single number representing, say, the number of years since divergence, then a *rates model* in the style of Thorne et al. (1991, 1992) could be fitted to them in some way. Our use of machines differs from the use of hidden Markov models by Haussler et al. (1993) in that a machine models an evolutionary process rather than a family of sequences or the typical member of a family. We believe that the former use best captures evolutionary relations and the latter best captures functional constraints.

The advantage of using finite-state machines is that there are standard techniques for manipulating them. The composition of two or more such machines is yet another one. Figure 2 illustrates the evolution of string A from string P as modeled by a mutation machine mA and of B from P as modeled by a mutation machine mB. If we know the parameters of mA and mB then we can calculate the parameters of an equivalent-generation machine $m = mA.mB$ which relates A and B directly. There are a number of possible explanation for each generation instruction of m in terms of instructions for mA and mB. We expect mA and mB to execute *more* instructions than there are characters in P because each copy, change, and delete acts on one character of P but an insert does not. The dashed quantities, pA'(c) etc., are the *rates* of instructions per character of P. Summing the rates of all explanations for all of m's instructions, we get a rate of $(1 + pA'(i) + pB'(i))$ instructions per character of P for m, the excess over 1 being due to the insertion operations of mA and mB. A further normalization step gives the parameters of m. Similar calculations are used when several machines are combined in the alignment of more than two strings. (See section $K \geqslant 2$ strings.) Note that we do not allow "$x \leftarrow \_ \rightarrow y$" to be an explanation of xy, for example, because the x and the y are not related in the explanation.

The dynamic programming algorithm (DPA) can be used to compare two strings under a variety of *costs* calculating, for example, the edit distance as in Fig. 3a. We take an *optimal* alignment to correspond to a *most probable* instruction sequence. An instruction sequence having the highest probability, and the minimum message length, can be found by a DPA using costs based on the message lengths of instructions, as in Fig. 3b. (Strictly, a term should be included for the number of instructions in the sequence but all plausible sequences have similar numbers of instructions and the $-\log_2$ of those numbers are very similar.) The alignment can be recovered by a traceback through the D[,] matrix or by Hirschberg's (1975) technique. If, on the other hand, the object is to infer the machine parameters, rather than an optimal alignment, something slightly different should be done. There are many optimal, near-optimal, and not-so-optimal instruction sequences or alignments in general. An estimate, E, of the machine parameters is



```
P                -- hypothetical, unknown parent
mA, mB           -- mutation machines
A            B   -- given descendants
m = mA.mB        -- equivalent generation machine
```

m's instns pair (a,b)	Explanation & rates per character of P.		
match(x) x x	x<---x--->x pA'(c).pB'(c)	x<---y--->x pA'(ch).pB'(ch)/3	
mismatch(x,y) x y	x<---x--->y pA'(c).pB'(ch)	x<---y--->y pA'(ch).pB'(c)	x<---z--->y pA'(ch).pB'(ch)2/3
$ins_A(x)$ x _	x<---x--->_ pA'(c).pB'(d)	x<---y--->_ pA'(ch).pB'(d)	x<---_--->_ pA'(i)
$ins_B(x)$ _ x	_<---x--->x pA'(d).pB'(c)	_<---y--->x pA'(d).pB'(ch)	_<---_--->x pB'(i)
_ _	_<---x--->_ pA'(d).pB'(d)	"invisible", not seen	

```
c-copy, ch-change, i-insert, d-delete

Define instn rates per character of P:
   pA'(instn) = pA(instn)/(1-pA(i)); pB'(instn) = pB(instn)/(1-pB(i))

Note pA'(c)+pA'(ch)+pA'(d) = pB'(c)+pB'(ch)+pB'(d) = 1

Sum rates = (1+pA'(i)+pB'(i)) = (1-pA(i).pB(i))/((1-pA(i)).(1-pB(i)))
```

Fig. 2. Composition of two machines, for DNA.

```
Boundary Conditions, i=1..|A|, j=1..|B|:
   D[0,0] = 0
   D[i,0] = D[i-1,0]+f(A[i], '-')
   D[0,j] = D[0,j-1]+f('-', B[j])

General Step, i=1..|A|, j=1..|B|:
   D[i,j] = g( D[i-1,j-1]+f(A[i],B[j]),    -- match or mismatch
               D[i-1,j  ]+f(A[i], '-'),    -- ins_A A[i]
               D[i,  j-1]+f('-', B[j]) )   -- ins_B B[j]

(a) For Edit Distance:
    f(x,x) = 0; f(x,y) = f(x,'-') = f('-',y) = 1
    g(a,b,c) = min(a,b,c)

(b) For most probable alignment:
    f(x,x)   =-log_2(P(match(x))); f(x,y)   =-log_2(P(mismatch(x,y)));
    f(x,'-')=-log_2(P(ins_A(x)));  f('-',y)=-log_2(P(ins_B(y)))
    g(a,b,c)= min(a,b,c)

(c) For r-theory:
    f( , ) as for (b)
    g(a,b,c)= logplus(a,logplus(b,c))
       where logplus(-log_2(p),-log_2(q)) = -log_2(p+q)
```

Fig. 3. Dynamic programming algorithm (DPA) variations.

itself a hypothesis. The best estimate minimizes ML(E) + ML(A&B|E), the parameters being stated to some optimal and finite accuracy. ML(A&B|E) is not the message length of any one instruction sequence. It is the $-\log_2$ probability that the machine-generated strings A and B, i.e., the $-\log_2$ of the sum of the probabilities over *all* possible instruction sequences or alignments that yield A and B. The hypothesis is that A and B are related by the inferred machine in some unspecified way and it is called the *r-theory*, the "*r*" standing for related. ML(A&B|E) can be calculated by a modified DPA which uses "logplus" instead of "min" in its general step, as in Fig. 3c, effectively adding the probabilities of alignments. For example, logplus(m,m) = $m - 1$. We now have $D[ij]$ = ML(A[1 ... i] & B[1 ... j]|E). Note that two different alignments represent exclusive explanations and their probabilities can therefore be added. Full details are given in Allison et al. (1992a). A code for efficiently transmitting A and B given E, using ML(A&B|E) bits and not based on a single alignment can be devised, but it is sufficient just to know what the message length would be in such a code.

The machine parameters must be estimated if they are not known a priori. An iterative approach is used; it may give only a local optimum. Assume we require an optimal alignment. Initial parameter estimates are "guessed." An alignment is found that is optimal for the current estimates using the DPA of Fig. 3b. New estimates are derived from the alignment. The process is iterated until it converges, which it must, invariably in a few steps. If, on the other hand, the objective is to estimate the machine parameters, rather than to find an optimal alignment, a similar approach is used but with the DPA of Fig. 3c, and this is further modified to accumulate weighted averages of instruction frequencies when alignments are combined in the general step. These values are used in the next iteration.

In what follows we consider only the simplest model of evolution, the one-state model, where the probabilities of instructions are independent of context. In particular, linear indel or gap costs require at least a three-state model. It is also assumed that $P(\text{insert}) = P(\text{delete})$ and that all characters are equally probable, all inserts are equally probable, and all changes are equally probable. In that case, we have for DNA and the generation machine:

$$\begin{aligned}
\text{ML}\,(\text{match}\,(x)) &= -\log_2(P\,(\text{match})) + \log_2(4) \\
&= -\log_2(P\,(\text{match})) + 2 \\
\text{ML}\,(\text{mismatch}\,(x,y)) &= -\log_2(P\,(\text{mismatch})) + \log_2(12) \\
\text{ML}\,(\text{ins}_A\,(x)) &= \text{ML}\,(\text{ins}_B\,(x)) \\
&= -\log_2(P\,(\text{ins}_A)) + 2 \\
&= -\log_2\,(P\,(\text{ins}_B)) + 2
\end{aligned}$$

These assumptions are made in the interests of simplicity as we are primarily interested in the underlying algorithms and methods. The model is simple but is not unacceptable for DNA. A more general model of changes would have to be incorporated if the methods were to be used on protein sequences. It is in any case the indels that are the main source of algorithmic complexity in string comparison.

The Posterior Distribution of Alignments

The stochastic processes to be described are based on a method of sampling alignments (or machine instruction sequences), given two strings A and B, with the correct frequency as determined by the posterior probability distribution of alignments. A sampling procedure that achieves this aim is sketched below with more details being given in the appendix. It is extended to multiple alignments of more than two strings in later sections. It strongly resembles Hirschberg's (1975) linear-space version of the DPA in its use of a recursive divide-and-conquer technique.

Assume that A is of length m and B is a length n. Informally, the procedure divides A at $i = m$ div 2 into $A1 = A[1 \ldots i]$ and $A2 = A[i + 1 \ldots m]$. It then chooses a point, j, and divides B into $B1 = B[1 \ldots j]$ and $B2 = B[j + 1 \ldots n]$; j might or might not equal n div 2. The procedure is then called recursively on $A[1 \ldots i]$ and $B[1 \ldots j]$ and on $A[i + 1 \ldots m]$ and $B[j + 1 \ldots n]$ until the base case is reached. The base case is that A contains just one character. It is straightforward to enumerate all the allowable ways of generating A and B in the base case and to sample from them according to probability. The choice of j in the general case is made according to the probability distribution of values taken over *all* alignments or instruction sequences. The DPA of Fig. 3c calculates the probability of A and B being generated—by summing over all possible alignments of A and B. (It actually works with the $-\log_2$ of probabilities.) It can also be used to calculate the probabilities of $A[1 \ldots i]$ and each *prefix* of B being generated; these values are contained in one row of $D[,]$. Running the DPA "in reverse" gives the probabilities of $A[i + 1 \ldots m]$ and each *suffix* of B being generated. These results are combined and normalized to give the probability distribution of j given i, and j is sampled accordingly. This is closely related to the notion of alignment density plots in Allison et al. (1992a, Figs. 6, 11).

The complete sampling procedure is just twice as slow as the DPA that it uses, i.e., it takes $O(|A| \times |B|)$ time. This is because it solves one full-size ($|A| \times |B|$) DPA problem, two problems whose total size ($|A| \times |B|/2$) is half that of the original problem, four problems whose total size is one-quarter that of the original problem, etc. In Hirschberg's algorithm the divide-and-conquer technique was used to reduce the use of space to $O(|B|)$—as it still does here—because only two rows of a matrix of length $|B|$ are required to calculate the values needed.

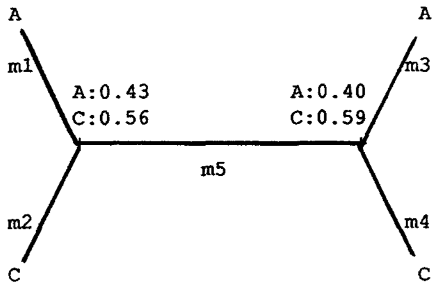
$K \geqslant 2$ Strings

A (multiple) alignment of K strings is formed by padding out each string with zero or more null characters so that they all have the same lengths. The padded strings are then written out one above another. Each column of characters is called a K-tuple or just a tuple. The all-null tuple is not allowed. There are two parts to the optimal alignment of several strings. The first is the search algorithm for finding a good alignment. The second is the cost function to be applied to alignments. As before, we take an optimal alignment to be a most probable alignment of the strings.

The naive extension of the DPA to K strings, each of approximately n characters, would require $O(2^K n^K)$ running time, which is infeasible for even modest values of K and n. However, the DPA can be extended so as to align two alignments. A string can be considered to a trivial case of an alignment and its characters to be one-tuples. Given an alignment AS over a set of L strings S and an alignment AT over a set of M different strings T, AS and AT can be aligned to give an alignment of $K = L + M$ strings $S \cup T$. The algorithm aligns AS, a string of L-tuples, with AT, a string of M-tuples. All that is necessary is that a cost be given to each K-tuple within the DPA; this is described below. The final alignment may not be optimal for the $L + M$ strings, but this algorithm can be used as an iterative step to improve a multiple alignment to at least a local optimum. This kind of *deterministic heuristic* is quite common, and an example has been described by Allison et al. (1992b): Given $K > 2$ strings and an evolutionary tree over them, an initial K-way alignment is found by some suboptimal process. The tree is then "broken" on some edge which partitions the strings into two disjoint sets, S of size L and T of size $M = K - L$. The K way alignment is projected onto these two sets of strings to give two subalignments, AS over S and AT over T, which are realigned with the DPA to give a new overall K-way alignment. The process is iterated and terminates when there is no further improvement in the full K-way alignment. The results are usually good although not guaranteed to give an optimal K-way alignment. The process may get stuck in local optima and results may depend on the initial alignment and the order in which edges are chosen during improvement. The alignment sampling process in the following section provides a way around these and other difficulties.

The other part of the optimal alignment problem is the assignment of a cost to an alignment of K strings. We use *tree costs* and model each edge of the evolutionary tree by a separate mutation machine. The machines can be combined so as to calculate the probability, and hence the message length, of each K-tuple examined by the DPA. The DPA can then find a good K-way alignment and calculate its total message length by summing K-tuple message lengths. A particular K-tuple of descendant characters can have several evolutionary explanations. If we knew the hypothetical ancestral characters at the internal nodes of the tree it would be a simple matter to combine the probabilities of the implied machine instructions on the edges. Since we do not, it is necessary to sum over all possible assignments to internal nodes in the style of Fig. 2. Fortunately, the combinatorial possibilities can be dealt with efficiently; Felsenstein (1981, 1983) describes the necessary algorithm. It involves calculating probabilities for each possible character value at each internal node of the tree; an example is given in Fig. 4. The probabilities of the various instructions for a given tuple can also be calculated and can be used to estimate instruction probabilities for each machine from a given K-way alignment. All this can be done by traversing the tree, in $O(K)$ time, for each tuple, but it is a significant computation and should not be repeated unnecessarily. Therefore results are stored in a lookup-table for fast access if ever the tuple is met again, as it probably will be. The lookup-table speeds up the algorithm considerably. Two alignments of four strings of length 500 can be aligned, to given an eight-way alignment, in 15 s on a SPARC station.

```
Actual machine parameters:
     copy change indel
m1:  0.9  0.05   0.05
m2:  0.9  0.08   0.02
m3:  0.7  0.2    0.1
m4:  0.8  0.1    0.1
m5:  0.75 0.1    0.15
```



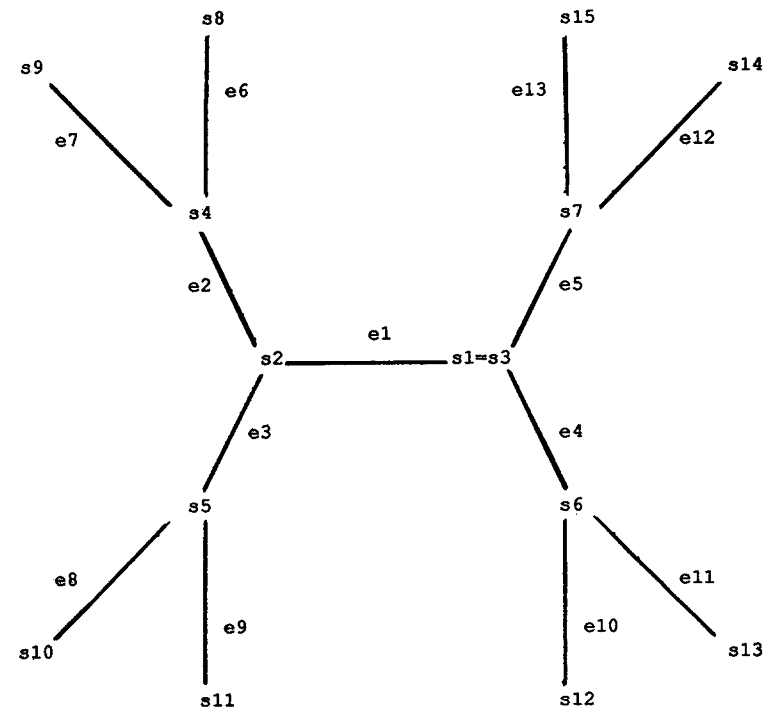
```
Probable hypothetical characters.
Tuple ACAC's probability = 0.00027

Estimated operations carried out:
     P(copy) P(change) P(indel)
m1:  0.43    0.57      <0.01
m2:  0.57    0.43      <0.01
m3:  0.40    0.60      <0.01
m4:  0.59    0.41      <0.01
m5:  0.94    0.06      <0.01
```

Fig. 4. Example, explanations for tuple ACAC.

Sampling Alignments of $K > 2$ Strings

It is infeasible to average over all K-way alignments of K strings for the purpose of estimating the edges of a given evolutionary tree but it can be sufficient to average over a large sample of alignments. Unfortunately it is also infeasible to extend the alignment sampling procedure of two sections previous directly to $K > 2$ strings for the same algorithmic reasons that the DPA cannot be directly extended. However, we can think of an alignment as the *state* of a complicated system having many *degrees of freedom*. It is sufficient to hold many of those degrees of freedom fixed while sampling the remainder from the conditional posterior distribution, so that is what is done: Given a multiple alignment, the tree is ''broken'' on a random edge which partitions the strings into two disjoint sets, as in the previous section. The multiple alignment is projected onto the two sets of strings to give two subalignments. A *random* realignment is sampled from the posterior distribution of alignments (of the subalignments) as described for just two strings (in section before last) and using the costs for K-tuples (as described in the last section). The realignment is sampled conditional on the subalignments. Each subalignment and the degrees of freedom that it specifies remain fixed during the realignment. Only the relation between the two subalignments is sampled in this step but the process is iterated many times, choosing random edges.

The machine parameters are estimated for each multiple alignment sampled. Results from all samples are averaged to give the final estimate of the machine parameters. Standard deviations are also calculated and give an indication of the accuracy of estimates. A working estimate of the machine parameters is needed to calculate the distributions in the algorithm and a weighted average from ''recent'' alignments is used for this purpose; the algorithm seems to be insensitive to the details of how this is done. To begin the sampling, an initial multiple alignment is found by the deterministic heuristic described previously.



Ancestor, s1=s3, length(s1)=500
1st generation: s2 and s3
2nd generation: s4 to s7
3rd generation: s8 to s15
Probabilities for each edge during evolution:
P(copy)=.9; P(change)=.05; P(insert)=P(delete)=.025

Fig. 5. Full tree.

Simulated Annealing

The alignment sampling procedure of the previous section is trivially modified for use in a simulated annealing approach to optimal multiple alignment. At the heart of the sampling procedure a point j of string or alignment B is chosen to correspond to the midpoint i of string or alignment A. The point j is sampled from the probability distribution of possible values. If the values of the probability distribution are raised to some power $p \geqslant 0$, and the distribution is renormalized, the sampling of j, and hence of alignments, is modified in a useful way. When $p = 0$, j is chosen from a uniform distribution. When $p = 1$, j is chosen from the posterior distribution implied by alignments. When $p > 1$, j is biased toward the more probable values. When p is very large, j is chosen so as to lead to an optimal (pairwise) alignment; the procedure becomes Hirschberg's algorithm in effect. Increasing p implements lowering the temperature in simulated annealing. Since the algorithm actually uses message lengths, the message lengths are multiplied by p, which is equivalent to raising the probabilities to the power p.

This strategy is very different in action from the simulated annealing strategy of Ishikawa et al. (1992), which makes small perturbations to a multiple alignment. The present strategy can make large perturbations to an alignment, especially when p is small.

Results

A program was written containing the alignment sampling procedure and the simulated annealing method described above. It was first tested on artificially generated DNA data. Figure 5 shows the artificial evolution of three generations of strings. The original ancestor, $s1$, is a random string of 500 characters. The first-generation descendants are $s2$ and $s3$; $s3$ is identical with $s1$ to make all edges similar because we are dealing with unrooted trees. Every edge in the tree corresponds to a mutation machine with the following parameters: $P(\text{copy}) = 0.9$, $P(\text{change}) = 0.05$; $P(\text{insert}) = P(\text{delete}) = 0.025$. One expects something like 10% mutation from parent to child but it could be more or less as the evolutionary process is random. One expects something like 20% mutation between $s8$ and $s9$ say, 30% mutation between $s4$ and $s6$, 40% mutation between $s8$ and $s10$, and 50% mutation between $s8$ and $s12$. Note that alignments can be found for $s8$ and $s12$ with more than 50% matches (their edit distance is 189, not 250) and that an alignment with 60–70% matches can be found even for two random, unrelated DNA sequences of similar length.

In three separate trials, first-, second-, and third-generation strings were used as data—each with the correct evolutionary tree. The tree for trial 3 includes that for trial 2 which includes the trivial tree for trial 1. Three

Table 1. Estimated edges from evolution at 10% mutation/edge[a]

(i) data = 1st generation, *s*2 and *s*3:										
Edge	Actual			det			Gibbs			
*e*1:	0.895	0.037	0.068	0.900	0.040	0.061	0.893	0.038	0.070	(0.005)
(ii) data = 2nd generation, *s*4 to *s*7:										
Edge	Actual			det			Gibbs			
*e*1:	0.895	0.037	0.068	0.894	0.044	0.063	0.878	0.038	0.085	(0.009)
*e*2:	0.902	0.053	0.045	0.910	0.054	0.036	0.899	0.052	0.049	(0.007)
*e*3:	0.919	0.041	0.039	0.916	0.057	0.026	0.917	0.048	0.035	(0.005)
*e*4:	0.898	0.055	0.047	0.899	0.059	0.043	0.891	0.054	0.056	(0.007)
*e*5:	0.908	0.041	0.051	0.915	0.055	0.030	0.908	0.052	0.041	(0.006)
Means over subsets of edges:										
*e*2–5:	0.907	0.048	0.046	0.910	0.056	0.034	0.904	0.052	0.045	
*e*1–5:	0.904	0.045	0.050	0.907	0.054	0.040	0.899	0.049	0.053	
(iii) data = 3rd generation *s*8 to *s*15:										
Edge	Actual			det			Gibbs			
*e*1:	0.895	0.037	0.068	0.878	0.060	0.062	0.863	0.043	0.094	(0.016)
*e*2:	0.902	0.053	0.045	0.928	0.048	0.025	0.900	0.054	0.046	(0.007)
*e*3:	0.919	0.041	0.039	0.915	0.061	0.025	0.920	0.048	0.033	(0.008)
*e*4:	0.898	0.055	0.047	0.920	0.038	0.043	0.906	0.036	0.058	(0.009)
*e*5:	0.908	0.041	0.051	0.913	0.055	0.032	0.904	0.061	0.035	(0.008)
*e*6:	0.891	0.058	0.051	0.896	0.066	0.038	0.893	0.062	0.045	(0.006)
*e*7:	0.926	0.035	0.039	0.924	0.035	0.042	0.918	0.031	0.052	(0.007)
*e*8:	0.922	0.029	0.049	0.937	0.037	0.026	0.932	0.026	0.041	(0.006)
*e*9:	0.898	0.035	0.067	0.909	0.042	0.050	0.895	0.032	0.073	(0.009)
*e*10:	0.893	0.060	0.048	0.898	0.048	0.055	0.879	0.052	0.069	(0.009)
*e*11:	0.902	0.049	0.049	0.907	0.057	0.036	0.906	0.053	0.042	(0.007)
*e*12:	0.898	0.049	0.053	0.909	0.045	0.046	0.902	0.040	0.059	(0.008)
*e*13:	0.916	0.049	0.035	0.912	0.064	0.024	0.906	0.064	0.030	(0.007)
Means over subsets of edges:										
*e*2–5:	0.907	0.048	0.046	0.919	0.051	0.031	0.908	0.050	0.043	
*e*6–13:	0.906	0.046	0.049	0.912	0.048	0.040	0.904	0.045	0.051	
*e*1–13:	0.905	0.045	0.049	0.911	0.050	0.039	0.902	0.046	0.052	

[a] Evolution: Ancestor length = 500. *P* (copy) = 0.9; *P* (change) = 0.05; *P* (insert) = *P* (delete) = 0.025 for each edge. Key: *P* (copy) *P* (change) *P* (indel); actual = frequencies as measured during evolution, det = inferred from a good alignment by deterministic heuristic, Gibbs = averaged from 1,000 × sampled alignments

analyses were performed in each trial. First, the deterministic heuristic was used to find a good alignment and parameters were estimated from this alone. The alignment was used as a starting point for the next two analyses. Second, parameters were estimated from 1,000 stochastically sampled alignments. Third, simulated annealing was used over 1,000 trials with the message-length multiplier increasing linearly from 1.0 to 4.0.

The results of these trials are summarized in Table 1. The figures marked *actual* give information from the evolution of the strings which is unknown to the analysis program. During evolution the machine on each edge had parameters *P*(copy) = 0.9, *P*(change) = 0.05, *P*(insert) = 0.025, *P*(delete) = 0.025 but there is variation and so the actual figures are given.

The figures marked *det* give the parameter estimates from the putative optimal alignment found by the deterministic heuristic in each trial. There is the beginning of a trend to overestimate *P*(copy) and, with the exception of *e*7 and *e*10, to underestimate *P*(indel). This is consistent with previous results on two strings (Yee and Allison 1993). In order to avoid repeated qualification in what follows, we often refer to an alignment found by the deterministic heuristic or by simulated annealing as an "optimal alignment" even though it may only be a near-optimal alignment.

The figures marked *Gibbs* give the estimates from the 1,000 sampled alignments in each trial. (The standard deviation of the estimate of *P*(indel) is the largest and is the only one reproduced.) The actual proportion of indels on each edge lies within about two standard deviations of the estimate of *P*(indel). Note that the standard deviation of the estimates for *e*1, which is common to each trial, increases as the data gets farther from *e*1, roughly doubling with each extra generation, and that the estimates for this edge are the worst.

Table 2. Estimated edges from evolution at 15% mutation/edge[a]

Edge	Actual			det			SA			Gibbs			(SD)
*e*1	0.858	0.056	0.086	0.854	0.089	0.057	0.876	0.073	0.051	0.850	0.073	0.078	(0.015)
*e*2	0.849	0.073	0.079	0.862	0.093	0.046	0.869	0.080	0.051	0.838	0.067	0.095	(0.017)
*e*3	0.881	0.069	0.051	0.920	0.050	0.030	0.908	0.069	0.024	0.887	0.062	0.052	(0.012)
*e*4	0.839	0.085	0.076	0.863	0.103	0.034	0.864	0.119	0.018	0.836	0.107	0.058	(0.014)
*e*5	0.833	0.079	0.088	0.845	0.133	0.022	0.833	0.160	0.008	0.836	0.089	0.075	(0.019)
*e*6	0.862	0.063	0.076	0.877	0.059	0.064	0.875	0.041	0.085	0.850	0.047	0.103	(0.011)
*e*7	0.871	0.069	0.061	0.880	0.079	0.041	0.876	0.093	0.031	0.877	0.084	0.040	(0.009)
*e*8	0.885	0.049	0.066	0.890	0.062	0.049	0.896	0.053	0.051	0.890	0.053	0.057	(0.010)
*e*9	0.831	0.083	0.086	0.841	0.098	0.062	0.838	0.091	0.072	0.813	0.079	0.108	(0.012)
*e*10	0.873	0.066	0.060	0.895	0.045	0.060	0.880	0.047	0.074	0.877	0.052	0.071	(0.009)
*e*11	0.837	0.088	0.074	0.845	0.090	0.059	0.860	0.074	0.065	0.836	0.075	0.089	(0.013)
*e*12	0.856	0.078	0.066	0.876	0.083	0.041	0.889	0.078	0.033	0.861	0.090	0.049	(0.008)
*e*13	0.844	0.084	0.072	0.866	0.083	0.051	0.854	0.079	0.066	0.846	0.075	0.079	(0.011)
Means over subsets of edges:													
*e*2–5:	0.851	0.077	0.073	0.873	0.095	0.033	0.869	0.107	0.025	0.849	0.081	0.070	
*e*6–13	0.857	0.073	0.070	0.872	0.075	0.053	0.871	0.070	0.060	0.856	0.069	0.075	
*e*1–13:	0.855	0.072	0.072	0.871	0.082	0.047	0.871	0.081	0.048	0.854	0.073	0.073	
Message lengths (bits):				6,735			6,701			7,556 (±109) Mean (SD)			
Null: 8,045													

[a] Evolution: Ancestor length = 500; *P* (copy) = 0.85; *P* (change) = 0.075; *P* (insert) = *P* (delete) = 0.0375 for each edge. Key: frequencies or estimated probabilities of copy, change, and indel from actual—as counted during evolution, det—estimated from a "good" alignment by deterministic heuristic, SA—from an optimal (?) alignment by simulated annealing, Gibbs—from 1,000 × Gibbs sampling

The deterministic heuristic proved hard to beat in the search for an optimal alignment at this moderate level of mutation. It was not beaten by simulated annealing in any of the above trials although simulated annealing found many alignments with a message length just two bits more than that found by the heuristic. From other trials it also seems that four-way alignment might be rather easy, in that the heuristic was not beaten in several trials. On eight-way alignment with 10% mutation per edge, the heuristic was sometimes beaten, but never by more than a few bits on artificial data. Possibly the simulated annealing was cooled insufficiently or was cooled too quickly. It seems that there is a very large number indeed of near-optimal alignments and that the search-space is hardly informative close to them. It would take extremely time-consuming tests to map out the alignment landscape thoroughly. The search for the marginally "best" alignment may be rather pointless in any case.

Simulated annealing beat the deterministic heuristic by a significant 34 bits when the level of mutation was increased to 15% per edge. Table 2 gives results for a tree with the topology of Fig. 5 where the machine on each edge had parameters *P*(copy) = 0.085, *P*(change) = 0.075, *P*(insert) = *P*(delete) = 0.0375. Strings such as *s*8 and *s*12 are only tenuously related here. Estimates are given from single alignments by the deterministic heuristic (det) and simulated annealing (SA) and from stochastic sampling of 1,000 alignments (Gibbs). There is an increased tendency for (near) optimal alignments to underestimate *P*(indel). This effect is most marked on "internal" edges of the tree, as illustrated by the means over different sets of edges. For example, simulated annealing gives an average of 0.025 against a real figure of 0.073 over *e*2 to *e*5. Sampling gives an average of 0.070. Note that much of the improvement in message length in going from the heuristic to simulated annealing seems to be due to the latter "explaining away" more indels as changes in the inner edges. (A similar effect has been noted with algorithms under development for the most parsimonious alignment problem.) The standard deviations in sampling's parameter estimates increase with the level of mutation as is to be expected.

Some tests were also done on unbalanced trees with edges of different lengths. Sampling continued to yield better estimates of actual edges although accuracy decreased and standard deviations increased on the longer edges.

Various tests were done to study the asymptotic behavior of the algorithms and some results are given in Table 3. In order to reduce computer time, only the deterministic heuristic was used to find (near) optimal alignments to compare with sampling. First, ten data sets were generated for the tree of Fig. 5 at 20% mutation per edge. This is quite a high level of mutation; across the 10 data sets the message lengths range from 7,700 to 7,900 bits for an optimal alignment, from 8,000 to 8,150 bits for the null theory, and from 9,100 to 9,900 bits for an average alignment. (Note that the message length of the *r*-theory, if it could be calculated, would be less than that of an optimal alignment.) Averages over all edges and all ten data sets of actual and estimated parameters are

Table 3. Averages over all edges & 10 data sets each for four different settings[a]

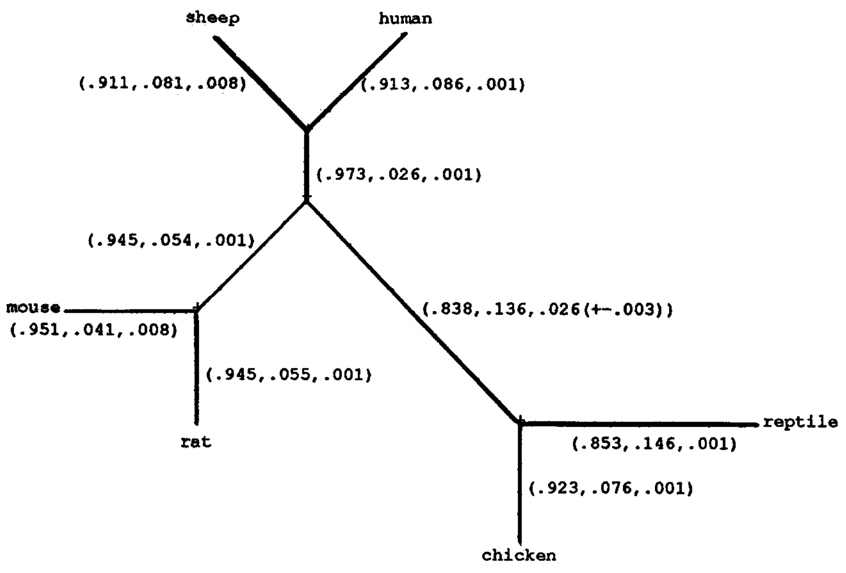
	Generation			Actual			det			Gibbs			SD
a)	0.8	0.1	0.1	0.801	0.101	0.099	0.828	0.125	0.048	0.790	0.107	0.103	(0.008 to 0.068)
b)	0.9	0.05	0.05	0.896	0.052	0.053	0.901	0.054	0.044	0.893	0.051	0.056	(0.003 to 0.008)
c)	0.8	0.1	0.1	0.795	0.105	0.101	0.815	0.124	0.061	0.785	0.104	0.111	(0.007 to 0.022)
d)	0.7	0.15	0.15	0.702	0.150	0.148	0.745	0.194	0.061	0.683	0.141	0.176	(0.012 to 0.051)

[a] Ancestor length = 500. Key: *P* (copy); *P* (change); *P* (indel). a): tree = (((*s*1 *s*2) (*s*3 *s*4)) ((*s*5 *s*6) (*s*7 *s*8))) unrooted, b), c) & d): tree = (*s*1 (*s*2 *s*3)) unrooted, generation: settings for mutation machine on each edge, actual: as counted during evolution, det: estimated from a "good" or optimal alignment, Gibbs: estimated by Gibbs sampling 1,000× per data set, std dev's: range of std dev's for sampling, across ten data sets

shown in Table 3, line (a). There is an increased tendency for optimal alignments to underestimate *P*(indel) at the 20% mutation level, particularly on inner edges of the tree. In going from Table 2 to Table 3 line (a), the average frequency of indels per edge has risen from 0.072 to 0.099 but the estimate from optimal alignment has remained at 0.048. Stochastic sampling gives good average estimates, within 0.01 of the actual figures. However for some data sets the estimates of some edges by sampling have standard deviations of over 0.06, implying that little more than one decimal place of such an estimate may be usable. Subsequently, ten data sets were generated for the tree of three leaves, (*s*1, (*s*2 *s*3)), at 10%, 20%, and 30% mutations per edge. Averages over all edges and all ten data sets of actual and estimated parameters are shown for each level of mutation in Table 3 lines (b)–(d). Sampling gives good estimates although standard deviations of estimates rise with the mutation level. At the same level of mutation per edge, the estimates from optimal alignment are better for the three-leaf than for the eight-leaf tree, presumably because the former has fewer degrees of freedom available for the maximization of alignment probability. The mutation level of 30% per edge is high and an optimal three-way alignment typically *fails* to be an acceptable hypothesis by a margin of 30–60 bits. The results also suggest that there may be a small bias in the sampling program to overestimate *P*(indel) at high levels of mutation. This possibility is being further investigated.

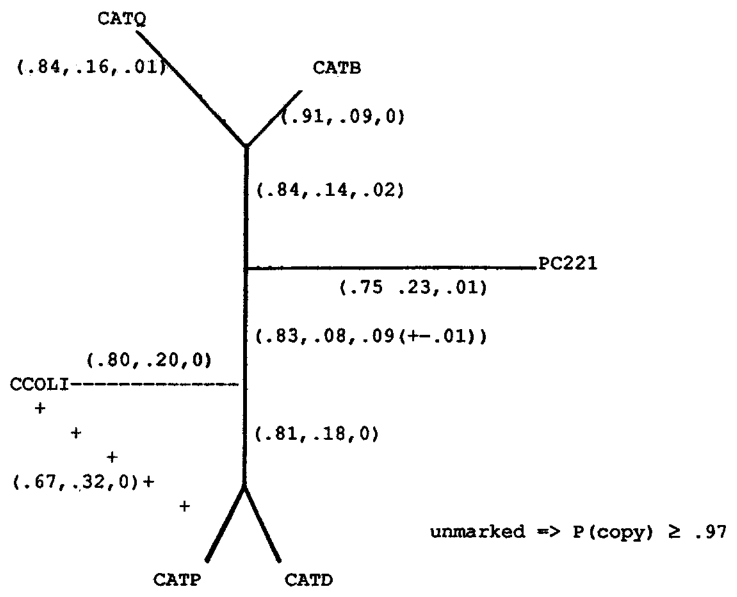
Transthyretin is a protein expressed in the brain. It is also expressed in the liver of some animals. Amongst other things, it is relevant to the evolution of birds, reptiles, and mammals (Schreiber et al. 1992). The cds sequences for transthyretin (Duan et al. 1991) from human, sheep, mouse, rat, chicken, and reptile (T. rugosa) were obtained from Genback. There is little doubt about the correct topology of the evolutionary tree, which is shown in Fig. 6 annotated with the estimates of the parameters for each edge as averaged from sampling 1,000 alignments. There must be considerable pressure of selection on these sequences, some relationships being close, and only the edge joining birds and reptiles to mammals shows significant numbers of indels. The standard deviations of estimates are low as the alignment is constrained. An optimal alignment gives similar estimates.

Huggins et al. (1992) constructed a tree of chloramphenicol acetyltransferase (CAT) monomers from various bacteria. Tests on five of the corresponding DNA sequences (Allison and Wallace 1994) revealed an interesting possibility. The tree (((CATQ CATB) (CATP CATD)) CCOLI) is weakly preferred on the basis of the message length of an optimal alignment. However, the tree ((CATQ CATB) ((CATP CCOLI) CATD)) is preferred on the basis of the average message length of sampled alignments, although the difference between the trees is less than the sum of the standard deviations in message lengths. This implies some uncertainty in the placing of CCOLI. Subsequently a sixth DNA sequence, PC221, was added to the data set. All test results are clear on how PC221 should be linked to each of the two trees above. For the six strings, the tree (((CATQ CATB) PC221) ((CATP CCOLI) CATD)) is supported over (((CATQ CATB) PC221) ((CATP CATD) CCOLI)) by both the optimal alignment and the average alignment criteria, although only weakly by the former. These trees are also closer, in terms of the average message length of alignments, than the trees on five sequences. As it happens the correct tree is almost certainly (((CATQ CATB) PC221) ((CATP CATD) CCOLI)) which is a subtree of the tree that Huggins et al. inferred using protein sequences and sequences from more organisms. The situation is illustrated in Fig. 7. The annotations on the edges come from Gibbs sampling 1,000 alignments of the six sequences under either tree as convenient; the two trees had good agreement on common edges. The significance of the above to the particular case of CAT genes is probably not great as the analysis is based on a simple model of evolution that is not the best for coding sequences and takes no account of expert biological knowledge. However, the results do illustrate the important general point that optimal alignment message length and average alignment message length may support different trees. An interesting question is, which one should be believed in such cases? The best answer is neither. The best criterion for choosing a tree would be based on the message length of the *r*-theory: the $-\log_2$ probability of getting the strings given the tree, i.e., the $-\log_2$ of the sum of the probabilities of all alignments given the tree. This could support a different tree entirely. Unfortunately it is not feasible to calculate its message length. The message length of an optimal alignment provides an upper bound on that of the *r*-theory, and a good alignment contributes



```
key: (P(match),P(change),P(indel))
Genbank ids: HUMPALA(27-470), OATTHYRE(12-452), MMALBR(27-467),
             RATPALTA(10-453), GDTRTHY(26-478), TRTRANST(16-468).
```

Fig. 6. Edge estimates by sampling for transthyretin cds.



```
Message Lengths (bits):
  Tree over 5 sequences                        opt'     Gibbs (SD)
  -----------------------------------          ----     ----------
  (((CATQ CATB)(CATP CATD))CCOLI) ---          4369     4555(+-40)
  ((CATQ CATB)((CATP CCOLI)CATD)) +++          4376     4502(+-26)
  Null: 6449 bits

  Tree over 6 sequences                        opt'     Gibbs (SD)
  -----------------------------------------    ----     ----------
  (((CATQ CATB)PC221)((CATP CATD)CCOLI)) ---   5395     5564(+-36)
  (((CATQ CATB)PC221)((CATP CCOLI)CATD)) +++   5391     5519(+-32)
  Null: 7752 bits
```

Data: chloramphenicol resistance gene, cds or orfs from

Genbank-id	organism		designation
m55620	Clostridium perfringens	459-1118	CATQ
m28717	Clostridium perfringens	145-768	CATP
x15100	Clostridium difficile	91-726	CATD
m35190	Campylobacter coli	309-932	CCOLI
m93113	Clostridium butyricum	145-800	CATB
spc221	Staphylococus aureus	2267-2914	PC221

Fig. 7. Alternative trees for bacterial sequences.

more to the probability of the r-theory than an average one contributes, but there are many more of the latter. It is only sensible to exercise caution in such cases.

Conclusions

An alignment of a set of strings can give an estimate of the edges of an evolutionary tree over the strings. However, the use of a good or optimal alignment gives a biased estimate of the true values, particularly on the inner edges of the tree. Forming the weighted average of estimates from all alignments would give an unbiased estimate of the edges. This can be done efficiently for two strings but is not feasible for more than two. However, averaging the estimates from many alignments sampled from their posterior probability distribution gives a good approximation and is feasible. In addition, sampling from the probability distribution raised to an increasing power (or from message lengths with an increasing multiplier) effects a simulated annealing search for an optimal alignment. A computer program embodying these ideas has been implemented for the simplest, one-state model of evolution. We intend to extend the model although this is not trivial for more than two strings. The current implementation is practical for ten strings of several hundred characters when used on a good work-station. With some optimization each limit could be increased somewhat. It is tempting to reduce the time complexity of the underlying DPA from quadratic to near linear by a windowing technique under the assumption that most of the probability in alignment space is concentrated near the current alignment. However, this may be a trap because the assumption may be invalid, particularly if the strings contain repeated and transposed sections. The sampling method is certainly a good candidate for implementation on a parallel computer. The tuning of simulated annealing algorithms is a difficult area and more work needs to be done on tuning the one described here.

It would be useful to be able to handle many more than ten strings. To do this it will probably be necessary to use a method related to the one described here, but one which samples definite strings for the internal, hypothetical nodes of the tree in a stochastic manner. (The current method makes only implicit, probabilistic assignments to internal nodes.) The problem can then be treated as an iterated three-descendant problem. Each step will be relatively simple but more of them will probably be required to explore the search space adequately. The resulting program would be a stochastic version of a method proposed by Sankoff et al. (1976). A program of this type is under development.

Our sampling and simulated annealing techniques could be used with other cost functions, such as star and all-pairs costs, in multiple alignment provided that the costs can be interpreted as something like the $-\log_2$ of a probability. It would be sufficient for the rank-ordering of the better alignments to be correct at low temperature. Simulated annealing could help with the problem of local optima that also affects alignment under such costs. However, it is not clear what the results of stochastic sampling would mean for these costs as they do not seem to have an obvious evolutionary interpretation.

Appendix: Sampling Alignment of Two Strings

The alignment sampling procedure given in the section The Posterior Distribution of Alignments is recursive and is in general called to act on a substring of string A and a substring of string B. If either substring is empty there is only one possible alignment. Otherwise, suppose we are given substrings $A[k \dots m]$ of string A where $m > k$ and $B[h \dots n]$ of string B where $n \geqslant h$. Let i be the middle position $(k + m)$ div 2 of $A[k \dots m]$. Consider an arbitrary alignment of $A[k \dots m]$ and $B[h \dots n]$. $A[i]$ appears somewhere in the alignment. It either occurs alone as $<A[i], ->$ which is equivalent to delete $(A[i])$ or as $\langle A[i], B[j]\rangle$ for some j which is equivalent to copy $(A[i])$ or to change $(A[i], B[j])$. In either case let $B[j]$ be the last character of $B[h \dots n]$, if any, that occurs with or before $A[i]$; $h - 1 \leqslant j \leqslant n$. There is a certain probability distribution over the possible values of j given i. The probability of a particular value "j" is proportional to the sum of the probabilities of all alignments that involve its choice, i.e., events (i) and (ii) below.

Divide-and-Conquer Cases of $A[k \dots m]\&B[l \dots n]$, $m > k$, $n \geqslant h$

(i) $A[k \dots i-1]$ & $B[h..j]$;	$<A[i], \rightarrow$;	$A[i+1 \dots m]$ & $B[j+1 \dots n]$
	... del $A[i]$ ...	
for some j, $h-1 <= j <= n$		
(ii) $A[k \dots i-1]$ & $B[h \dots j-1]$;	$<A[i], B[j]>$;	$A[i+1 \dots m]$ & $B[j+1 \dots n]$
	... copy $A[i]$ or	...
	... change $A[i]$ to $B[j]$	...
for some j, $h <= j <= n$		

Now $P(A[p \dots q]\&B[r \dots s])$ can be calculated for all "s" by the modified (logplus) DPA of Fig. 3c. Therefore the probabilities of each possible value of j can be calculated with two calls to the DPA: one on the forward strings $A[k \ldots i]\&B[h \ldots n]$ and one on the reversed strings $A[i+1 \ldots m]\&B[h \ldots n]$. The forward run is used to calculate the probabilities of $A[k \ldots i]$ and each prefix of $B[h \ldots n]$ being generated together, by *any* instruction sequence whose final instruction includes $A[i]$. The reverse run calculates the probabilities of $A[i+1 \ldots m]$ and each suffix of $B[h \ldots n]$ being generated together in any way. Combining the results, using logplus, gives the $-\log_2$ odds ratios of all possible ways of partitioning an instruction sequence for A and B at the instruction that includes $A[i]$. This allows j to be sampled from its correct probability distribution. The sampling procedure is then called recursively on $A[k \ldots i]\&B[h \ldots j]$ and on $A[i+1 \ldots m]\&B[j+1 \ldots n]$.

The coordinates (ij) are called an *internal terminal.* The coordinates $(|A|,|B|)$ are called the *external terminal.* Since the first half of the overall alignment was forced to end with either $\langle A[i], - \rangle$ or $\langle A[i],B[j]\rangle$ and not $\langle -,B[j]\rangle$, any subsequent reversed DPAs are forced to begin from an internal terminal with one of these alternatives by modification of the boundary conditions.

The lengths of the sections of A and B shrink with each recursive call to the sampling procedure. Eventually a single character of A remains and this leads to the *base cases* of the procedure. If the call is for an internal terminal, $A[i]$ must occur at the end of the mini-alignment. There are two possibilities:

Internal-Terminal Base-Cases of $A[i]\&B[j \ldots k],\ k \geqslant j$

```
(i)   -    ...  -        A[i]
      B[j] ...  B[k]     -
      ins  ...  ins      del
(ii)  -    ...  -        A[i]
      B[j] ...  B[k-1]   B[k]
           ...  ins      copy or
      ins  ...  ins      change
```

The probability of each possibility is easily calculated and they are sampled accordingly. If the base case is for the external terminal there are more possibilities but each is no more complex than before and they are easily sampled:

External-Terminal Base-Cases of $A[i]\&B[j \ldots k]$

```
(i) A[i] -    ...  -      or  -    A[i] -       ...  -     etc.
    -    B[j] ...  B[k]       B[j] -    B[j+1]  ...  B[k]
    del  ins  ...  ins        ins  del  ins     ...  ins

(ii) A[i]  -       ...  -      or -    A[i]    -       ...  -     etc.
     B[j]  B[j+1]  ...  B[k]      B[j] B[j+1]  B[j+2]  ...  B[k]
     copy/                             copy/
     chng  ins     ...  ins       ins  chng    ins     ...  ins
```

Essentially the same procedure is used to sample K-way alignments of K-strings as described in section 5.

Acknowledgments. This work is partly supported by an Australian Research Council grant.

References

Allison L, Wallace CS (1994) An information measure for the string to string correction problem with applications. 17th Australian Comp. Sci. Conf., Christchurch, New Zealand, pp 659–668

Allison L, Wallace CS, Yee CN (1992a) Finite-state models in the alignment of macro-molecules. J Mol Evol 35:77–89

Allison L, Wallace CS, Yee CN (1992b) Minimum message length encoding, evolutionary trees and multiple alignment. 25th Hawaii Int. Conf. Sys. Sci. 1:663–674

Bishop MJ, Thompson EA (1986) Maximum likelihood alignment of DNA sequences. J Mol Biol 190:159–165

Duan W, Achen MG, Richardson SJ, Lawrence MC, Wettenhall REH, Jaworowski A, Schreiber G (1991) Isolation, characterisation, cDNA cloning and gene expression of an avian transthyretin: implications for the evolution of structure and function of transthyretin in vertebrates Eur J Biochem 200:679–687

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

Felsenstein J (1983) Inferring evolutionary trees from DNA sequences. In: Weir BS (ed) Statistical analysis of DNA sequence data. Marcel Dekker, pp 133–150

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109

Haussler D, Krogh A, Mian S, Sjolander K (1993) Protein modelling using hidden Markov Models: Analysis of globins. 26th Hawaii Int. Conf. Sys. Sci. 1:792–802

Hirschberg DS (1975) A linear space algorithm for computing maximal common subsequences. Comm ACM 18(6):341–343

Huggins AS, Bannam TL, Rood JI (1992) Comparative sequence analysis of the catB gene from Clostridium butyricum. Antimicro agents Chemother 36(11):2548–2551

Ishikawa M, Toya T, Hoshida M, Nitta K, Ogiwara A, Kanehisa M (1992) Multiple sequence alignment by parallel simulated annealing. Institute for New Generation Computing (ICOT) TR-730

Lawrence CE, Altschul SF, Bogushki MS, Liu JS, Neuwald AF, Wooton JC (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262:208–214

Sankoff D, Cedergren RJ, Lapalme G (1976) Frequency of insertion-deletion, transversion, and transition in evolution of 5S ribosomal RNA. J Mol Evol 7:133–149

Schreiber G, Aldred AR, Duan W (1992) Choroid plexus, brain protein-homeostasis and evoluation. Today's Life Science Sept: 22–28

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

Thorne JL, Kishino H, Felsenstein J (1992) Inching towards reality: an improved likelihood model of sequence evolution. J Mol Evol 34: 3–16

Wallace CS, Boulton DM (1968) An information measure for classification. Comp J 11(2):185–194

Wallace CS, Freeman PR (1987) Estimation and inference by compact encoding. J R Stat Soc B 49:240–265

Yee CN, Allison L (1993) Reconstruction of strings past. Comp Appl Biosci 9(1):1–7