J Mol Evol (1991) 33:114–124

Journal of
Molecular Evolution
© Springer-Verlag New York Inc 1991

# An Evolutionary Model for Maximum Likelihood Alignment of DNA Sequences

Jeffrey L. Thorne, Hirohisa Kishino,* and Joseph Felsenstein

Department of Genetics SK-50, University of Washington, Seattle, WA 98195, USA

**Summary.** Most algorithms for the alignment of biological sequences are not derived from an evolutionary model. Consequently, these alignment algorithms lack a strong statistical basis. A maximum likelihood method for the alignment of two DNA sequences is presented. This method is based upon a statistical model of DNA sequence evolution for which we have obtained explicit transition probabilities. The evolutionary model can also be used as the basis of procedures that estimate the evolutionary parameters relevant to a pair of unaligned DNA sequences. A parameter-estimation approach which takes into account all possible alignments between two sequences is introduced; the danger of estimating evolutionary parameters from a single alignment is discussed.

**Key words:** DNA sequence alignment — Maximum likelihood procedure — Dynamic programming — Evolutionary model — Insertion–deletion model

## Introduction

With the advent of modern molecular biology, the ability to collect biological sequence data has outpaced the ability to adequately analyze this data. One tool for reducing this surfeit of inadequately treated data is sequence alignment. A sequence alignment is designed to exhibit the evolutionary correspondence between different sequences. It is possible and, among some researchers, popular to align sequences by eyeball. The eyeball technique is time-consuming, tedious, and irreproducible. In 1970, Needleman and Wunsch presented a dynamic programming algorithm for the alignment of two biological sequences by computer. Computer-aided sequence alignment does not possess these disadvantages of the eyeball technique. The basic dynamic programming algorithm chooses the best alignment by finding the alignment with the minimum associated weight. This is assumed to be the best of all alignments between the two sequences in question. The evolutionary weight associated with an alignment is simply the sum of the weights of the evolutionary events implied by the alignment. In the case of an alignment between two sequences, insertions cannot be distinguished from deletions. Therefore, the term indel is used to describe an evolutionary event that may be either an insertion or a deletion. Because a single-base indel leads to a single-base gap in the alignment and because a nucleotide mismatch in the alignment is caused by one or more nucleotide substitutions, the following alignment implies that at least three substitutions and two single-base indels took place:

```
A T A G A G - T T T G T A C G
- T A G C G G T T C G T T C G
```

The dynamic programming algorithm has subsequently been improved (e.g., Gotoh 1982) but, in its most basic form, there is a weight for each single gap and a weight for each mismatch. If the weight of a mismatch is 1 and the weight of a single-base gap is 5, then the weight associated with the above alignment is 13 (= 1 + 1 + 1 + 5 + 5). A complete

*Offprint requests to:* J.L. Thorne

[*] *Present address:* Ocean Research Institute, University of Tokyo, 1-15-1, Minami-dai, Nakano-ku, Tokyo 164, Japan

explanation of the dynamic programming algorithm can be found in Sankoff and Kruskal (1983).

The weakness of the basic dynamic programming method and its subsequent modifications is the lack of an objective procedure to choose the relative weights of gaps and mismatches. The result of this weakness is that researchers are forced to use either of two flawed approaches to obtain an alignment between two sequences. One approach is to arbitrarily choose these weights and then obtain an alignment. If this alignment is aesthetically pleasing to the researcher, the process stops. Otherwise, the researcher continues to adjust the weights until an aesthetically pleasing alignment is obtained. Obviously, the subjective nature of this approach is not ideal. Another approach is to use the same set of weights for every pairwise alignment. This approach is less subjective than the former approach—only the initial choice of weights is subjective.

A few objective alignment techniques have been proposed (e.g., Reichert et al. 1973; Fitch and Smith 1983; Allison and Yee 1990) but only Bishop and Thompson (1986) have described an objective technique that is based upon an evolutionary model. Because evolution is the force that promotes divergence between biological sequences, it is desirable to view biological sequence alignment algorithms in the context of evolution. The weights of evolutionary events should be a function of evolutionary rates and divergence times. Under this interpretation, the basic dynamic programming procedure assumes that the types of evolutionary events that can change a biological sequence fall into three categories. For a DNA sequence, these three possible types of events are insertion of exactly one base, deletion of exactly one base, and substitution of one base for another. The basic dynamic programming procedure assigns an evolutionary weight to each type of evolutionary event. The evolutionary weight should be proportional to the negative logarithm of the probability of the evolutionary event (Felsenstein 1981a). Thus, the most basic alignment algorithm requires one evolutionary weight for a substitution and another evolutionary weight for a single-base indel. It is incorrect to use the same set of weights for every pairwise alignment because the probabilities of evolutionary events depend on the particular pair of sequences to be aligned.

In this paper, we present a maximum likelihood approach to the alignment of a pair of DNA sequences. This maximum likelihood approach is an extension and modification of the pioneering approach of Bishop and Thompson (1986). The Bishop and Thompson approach is completely objective but is approximate and is most effective for short divergence times. Our more general approach yields explicit calculations of likelihood and a method for estimating evolutionary parameters. This procedure can adjust the evolutionary weights to the sequences to be aligned. We also examine the bias that is generated when only a single alignment is used for the estimation of evolutionary parameters. Our method for estimating evolutionary parameters is accurate and avoids this bias because it maximizes the likelihood of two sequences. In other words, our method maximizes the sum—taken over all possible alignments between two sequences—of the likelihood of individual alignments.

## Statistical Model of DNA Sequence Evolution

Our maximum likelihood approach is based upon an evolutionary model that allows only substitutions, single-base insertions, and single-base deletions. It is our hope to eventually replace this evolutionary model with a more realistic version that can allow other evolutionary events such as inversions, large insertions, and large deletions. This evolutionary model is a Markov process; the probability of a transition from the current state of a sequence is independent of previous states of the sequence. The likelihood of a pair of modern sequences, *A* and *B*, separated from a common ancestral sequence *C* by divergence time *t* is

$$P_t(A, B) = \sum_{C} P_\infty(C)P_t(A \mid C)P_t(B \mid C) \qquad (1)$$

Here $P_t(A \mid C)$ is the transition probability from sequence *C* to sequence *A*, and $P_\infty(C)$ is the equilibrium probability of sequence *C*. It should be understood that the values of these probabilities all depend on the particular values of the parameters that are pertinent to the evolutionary process. The evolutionary process described in this paper is reversible. The reversibility property implies that the joint probability of sequence *A* and sequence *C* is not influenced by the fact that sequence *A* is a descendant of sequence *C*: the joint probability of these two sequences would be the same if *C* were a descendant of *A* or if both were descendants of a third sequence. For a reversible process [i.e., $P_\infty(C)P_t(A \mid C) = P_\infty(A)P_t(C \mid A)$ for every *A*, *C*, and $t > 0$], Eq. (1) reduces to

$$P_t(A, B) = P_\infty(A)P_{2t}(B \mid A) \qquad (2)$$

When the evolutionary process is reversible, it is therefore not necessary to sum over all possible ancestral sequences to compute the probability of two modern sequences arising from a common ancestral sequence. Instead, it is sufficient to treat one modern sequence as if it were the ancestor and the other modern sequence as if it were the descendant for the computation of $P_t(A, B)$.

### *Specification of Evolutionary Process*

immortal link (λ). A newborn link is always a normal where $0 < \lambda < \pi$. The mean and variance are easily calculated:

$$E(n) = \frac{\frac{\lambda}{\mu}}{1 - \frac{\lambda}{\mu}} \qquad (4)$$

$$\mathrm{Var}(n) = \frac{\frac{\lambda}{\mu}}{\left(1 - \frac{\lambda}{\mu}\right)^2}$$

*Likelihood Expression of a Pair of DNA Sequences*

Consider two DNA sequences. The first, sequence *A*, is TGTC. The second, sequence *B*, is GCACA. Various paths are possible for a transition from the first sequence to the second sequence. For example, one possible path consists of the first three bases of the former sequence (TGT) undergoing substitution to the first three bases of the latter sequences (GCA) and the rightmost base of the latter sequence arising via insertion. The transition probability from one sequence to another is the sum of the probabilities of all possible paths connecting the two sequences. The particular path of a transition from one sequence to another can be expressed well by alignment. As an example of an alignment or transition path from sequence *A* to sequence *B*, consider the following improbable alignment which will be denoted as $\alpha$:

```
- T G T - C -
G - C - A C A
```

The information on presence and absence of bases in alignment $\alpha$ will be termed $\alpha'$ and, when $\alpha'$ is represented in terms of links, $\alpha'$ can be represented as:

```
• -   ★   ★   ★ -   ★ -
• ★   -   ★   - ★   ★ ★
```

The links have been clustered in the above representation of alignment $\alpha'$ for the purpose elucidating the form of $P(\alpha' \mid \theta)$. The probability of the specific transition path represented by alignment $\alpha$ [i.e., $P(\alpha \mid \theta)$ where $\theta$ is the collection of parameters $\mu t$, $\lambda t$, $st$, $\pi_A$, $\pi_G$, $\pi_C$, and $\pi_T$] can be decomposed into two components, $P(\alpha' \mid \theta)$ (the transition probability of insertion–deletion) and $P(\alpha \mid \alpha', \theta)$. This decomposition is possible because $\alpha$ contains all of the information of $\alpha'$. In other words

$$P(\alpha \mid \theta) = P(\alpha, \alpha' \mid \theta) = P(\alpha \mid \alpha', \theta)P(\alpha' \mid \theta) \quad (5)$$

$P(\alpha' \mid \theta)$ will be a product of $n + 2$ terms. The first term is the equilibrium probability of an ancestral sequence with $n$ bases (or $n + 1$ links) and the second term is a transition probability for the immortal link. The remainder of the terms are transition probabilities for normal links. The specific transition probability for each link depends on the type of link (normal or immortal), whether the link survived, and the number of descendant links. The number of descendant links for a particular ancestral link is easily determined by depicting the information on presence and absence of bases in terms of links. The number of descendant links of a particular ancestral link is one (if the particular ancestral link survives) plus the number of descendant links to the right of the particular ancestral link and to the left of the particular ancestral link's neighbor on the right.

Concerning the fate of an individual link over time, three types of transition probabilities are considered: $p_n(t)$ is the probability after a timespan of length $t$ that $n$ links are descended from a normal link and one of them is the original, $p'_n(t)$ is the probability that $n$ links are descended from a normal link and the original dies, and $p''_n(t)$ is the probability that the immortal link has $n$ descendants including itself. In the above example

$$P(\alpha' \mid \theta) = \gamma_4 p''_2(t)p'_0(t)p_1(t)p'_1(t)p_2(t)$$
$$P(\alpha \mid \theta, \alpha') = \pi_G f_{GC}(t)\pi_A f_{CC}(t)\pi_A \qquad (6)$$

It can be proven by induction with respect to sequence length that our model is reversible. In fact it is reversible with respect to each particular history $\alpha$.

By their definitions, $p_0(t) = p''_0(t) = 0$. The remainder of the transition probabilities can be obtained by solving the differential equations governing this birth–death process. These differential equations can be formerly expressed:

$$\frac{dp_n(t)}{dt} = \lambda(n-1)p_{n-1}(t) - (\lambda + \mu)np_n(t) + \mu np_{n+1}(t) \quad n > 0$$

$$\frac{dp'_n(t)}{dt} = \lambda(n-1)p'_{n-1}(t) - (\lambda + \mu)np'_n(t) + \mu(n+1)p'_{n+1}(t) + \mu p_{n+1}(t) \quad n > 0 \quad (7)$$

$$\frac{dp'_0(t)}{dt} = \mu p'_1(t) + \mu p_1(t)$$

$$\frac{dp''_n(t)}{dt} = \lambda(n-1)p''_{n-1}(t) - [\lambda n + \mu(n-1)] \cdot p''_n(t) + \mu np''_{n+1}(t) \quad n > 0,$$

$$
\begin{aligned}
&p_1(0) = p''_1(0) = 1 \\
&p_n(0) = p''_n(0) = 0 \qquad n = 2, 3, \ldots \qquad (8)\\
&p'_n(0) = 0 \qquad n = 0, 1, \ldots
\end{aligned}
$$

Equations (7) can be solved. The explicit forms of the transition probabilities corresponding to the above differential equations are

$$
\begin{aligned}
&p_n(t) = e^{-\mu t}[1 - \lambda\beta(t)][\lambda\beta(t)]^{n-1} \qquad n > 0 \\
&p'_n(t) = [1 - e^{-\mu t} - \mu\beta(t)][1 - \lambda\beta(t)] \\
&\qquad \cdot[\lambda\beta(t)]^{n-1} \qquad n > 0 \qquad (9)\\
&p'_0(t) = \mu\beta(t) \\
&p'_n(t) = [1 - \lambda\beta(t)][\lambda\beta(t)]^{n-1} \qquad n > 0
\end{aligned}
$$

where

$$\beta(t) = \frac{1 - e^{(\lambda-\mu)t}}{\mu - \lambda e^{(\lambda-\mu)t}} \qquad (10)$$

It is important to note that there is a slight discrepancy between the conventional form of alignment and our model. Previously, the following alignment denoted by $\alpha$ was presented

- T G T - C -
G - C - A C A

If the fourth and fifth positions of alignment $\alpha$ are switched, the result is

- T G - T C -
G - C A - C A

It is not clear how the meaning of this modified alignment and the meaning of alignment $\alpha$ differ when viewed conventionally, but the two alignments clearly differ in meaning when viewed with reference to our model. This difference is easier to understand if the top sequence in each alignment is viewed as the ancestor and the bottom sequence in each alignment is viewed as the descendant. According to the likelihood model, the link associated with the T in the top sequence at the fourth position of alignment $\alpha$ was deleted but not before it gave rise, via insertion, to a descendant link associated with the A that can be found in the lower sequence at the fifth position of alignment $\alpha$. In the modified alignment, the same A—now in the fourth alignment position—is a descendant of the link associated with the G in the top sequence at the third position. This difference stems from the fact that, in our model, a newborn link is always inserted directly to the right of its parental link.

$B$ be $s_B$. Evolutionary parameters can be estimated by maximizing the likelihood

$$L_\theta(A, B) = \pi_A{}^{r_A}\pi_G{}^{r_G}\,\pi_C{}^{r_C}\pi_T{}^{r_T}\,\gamma_{s_A}P_t(B \mid A, \theta) \quad (11)$$

where $r_A$, $r_G$, $r_T$, and $r_C$ are the number of occurrences of each type of nucleotide in sequence $A$. To simplify notation, we write $L_\theta(A, B)$ instead of $P(A, B \mid \theta)$. Because for $k \geq 1$,

$$
\begin{aligned}
p_k(t) &= p_1(t)[\lambda\beta(t)]^{k-1} \\
p'_k(t) &= p'_1(t)[\lambda\beta(t)]^{k-1} \qquad (12)\\
p''_k(t) &= p''_1(t)[\lambda\beta(t)]^{k-1},
\end{aligned}
$$

all insertion–deletion transition probabilities can be written as a function of $p_1(t)$, $p'_1(t)$, $p''_1(t)$, $p'_0(t)$, and $\lambda\beta(t)$. This fact enables development of two recursive algorithms, the alignment algorithm, and the parameter estimation algorithm, which are very similar to the conventional dynamic programming algorithm.

Denote the subsequent consisting of the first $m$ bases of sequence $A$ by $A_m$ and denote the subsequence consisting of the first $n$ bases of sequence $B$ by $B_n$. Because our model is reversible, we can without loss of generality consider sequence $A$ to be an ancestor of sequence $B$. This implies that all links in sequence $B$ are descendants of links in sequence $A$. Define $S(A_m, B_n)$ as the set of all possible alignments between $A_m$ and $B_n$. Each possible alignment $\alpha(A_m, B_n)$ between $A_m$ and $B_n$ is a member of exactly one of three subsets of $S(A_m, B_n)$:

$S^0(A_m, B_n) = \{\alpha(A_m, B_n)$ where rightmost link of $A_m$ has no descendant links in $B_n\}$

$S^1(A_m, B_n) = \{\alpha(A_m, B_n)$ where rightmost link of $A_m$ has exactly one descendant link in $B_n\}$

$S^2(A_m, B_n) = \{\alpha(A_m, B_n)$ where rightmost link of $A_m$ has at least two descendant links in $B_n\}$

To refer to a particular alignment between $A_m$ and $B_n$ which happens to be a member of the subset $S^i(A_m, B_n)$, the notation $\alpha^i(A_m, B_n)$ will be used.

### *Alignment Algorithm*

First, we introduce the alignment algorithm. This recursive algorithm can produce the maximum likelihood alignment between sequence $A$ and sequence $B$ and its likelihood for a given value of $\theta$. The procedure consists of gradually filling in the entries in the matrices constructed by our procedures are not weights but are alignment likelihoods. The likelihood of a specific subsequence alignment $\alpha(A_m, B_n)$ for a certain value of $\theta$ will be written as $l_\theta[\alpha^i(A_m, B_n)]$ where $i = 0, 1, 2$. The value of $i$ depends on the subset to which $\alpha(A_m, B_n)$ belongs. Let us denote the alignment of highest likelihood in $S^i(A_m, B_n)$ for a certain value of $\theta$ by $\alpha^i_{\max}(A_m, B_n)$, i.e.,

$$l_\theta[\alpha^i_{\max}(A_m, B_n)] = \max_{\alpha^i(A_m, B_n)} l_\theta[\alpha^i(A_m, B_n)]$$

In addition, let

$$\begin{aligned} l_\theta[\alpha_{\max}(A_m, B_n)] = \max\{ & l_\theta[\alpha^0_{\max}(A_m, B_n)], \\ & l_\theta[\alpha^1_{\max}(A_m, B_n)], \\ & l_\theta[\alpha^2_{\max}(A_m, B_n)]\} \end{aligned}$$

The maximum likelihood alignment between sequence $A$ and sequence $B$ for a particular value of $\theta$ can be determined by a recursive procedure that updates $l_\theta[\alpha^0_{\max}(A_m, B_n)]$, $l_\theta[\alpha^1_{\max}(A_m, B_n)]$, and $l_\theta[\alpha^2_{\max}(A_m, B_n)]$.

Let $a_m$ denote the type of nucleotide at the $m$th position of sequence $A$ and let $b_n$ denote the type of

So the maximum likelihood alignment between sequence $A$ and sequence $B$ has likelihood

$$l_\theta[\alpha_{\max}(A, B)] = \max\{l_\theta[\alpha^0_{\max}(A, B)], l_\theta[\alpha^1_{\max}(A, B)], l_\theta[\alpha^2_{\max}(A, B)]\}$$

Similar to be conventional dynamic programming procedure, recovery of the actual maximum likelihood alignment is obtained by tracing back through the likelihood matrix on the path that led to the maximum likelihood value. Although it is often true that there are many different $\alpha_{\max}(A, B)$ that attain $\max_\alpha l_\alpha[\alpha(A, B)]$, and although high likelihood alignments could be recovered by employing the algorithm of Waterman (1983), our current computer implementation only returns a single one of these equally good maximum likelihood alignments.

### *Evolutionary Parameter Estimation Algorithm*

The second recursive procedure is designed to calculate the likelihood of two sequences for a given

Then, the likelihood of two sequences is obtained by

$$L_\theta(A, B) = L_\theta^0(A, B) + L_\theta^1(A, B) + L_\theta^2(A, B) \qquad (15)$$

To find the maximum likelihood estimate of $\theta$, this procedure can be used in conjunction with a numerical maximization routine. This strategy for the estimation of $\theta$ [i.e., the estimation of $\theta$ by the value of $\theta$ that satisfies $\max_\theta L_\theta(\mathrm{A, B})$] will be referred to as the sum approach.

There may be applications where the posterior probability of a specific alignment is of interest. If there is a specific alignment $\alpha(A, B)$ between sequence $A$ and sequence $B$ that is of interest, the posterior probability of $\alpha(A, B)$—the fraction of the total likelihood contributed by $\alpha(A, B)$—can be calculated

$$P[\alpha(A, B) \mid \theta, A, B] = \frac{l_\theta[\alpha(A_m, B_n)]}{L_\theta(A_m, B_n)} \qquad (16)$$

The numerical maximization routine used to produce the results in this paper is adapted from the simplex method. The computer code for this maximization routine was published in Press et al. (1988). Press et al. used the algorithm of Nelder and Mead

a descendant sequence $B$. The evolutionary process in the simulation was consistent with our evolutionary model except that the length of the ancestral sequence was fixed instead of being drawn from a geometric distribution. The purpose of this intentional violation was to eliminate the effect of variable initial sequence length on the estimation of evolutionary parameters. For the simulated evolutionary process, $\lambda$ was fixed at $\frac{\mu s_A}{s_A + 1}$ where $s_A$ is the length of ancestral sequence $A$. This is the maximum likelihood estimate of $\lambda$ for a given value of $\mu$ and $s_A$ under our evolutionary model. The base composition was set to $\pi_A = \pi_G = \pi_C = \pi_T = 0.25$ [the Jukes–Cantor model (1969)] and the divergence time was $t = 1.0$.

Conceivably, $\lambda t$, $\mu t$, $st$, $\pi_A$, $\pi_G$, $\pi_C$, and $\pi_T$ could all be estimated with regard to each pair of sequences. This would be the ideal situation. Our parameter estimation process was not this complete because a complete analysis would be prohibitively slow. To simplify the parameter estimation process, equilibrium base frequencies ($\pi_A$, $\pi_G$, $\pi_C$, and $\pi_T$) were estimated by the frequency with which each type of estimation algorithm. As with the sum approach, the direct alignment approach requires an amount

pairs of sequences under the same value of $\theta$. From this sample of sequence pairs, the sample standard

**Table 1.** A comparison of the sum approach and the direct alignment approach

| | $\widehat{\mu t}$ | $\widehat{st}$ | $(A)$ | $(B)$ |
|---|---|---|---|---|
| $\mu t = 0.01$, $st = 0.01$ | | | | |
| I | 0.0108 ± 0.0034<br>± 0.0034 | 0.0105 ± 0.0036<br>± 0.0055 | 790.8 | 786.7 |
| II | 0.0105 ± 0.0031<br>± 0.0032 | 0.0107 ± 0.0035<br>± 0.0053 | 790.8 | 786.7 |
| $\mu t = 0.01$, $st = 0.1$ | | | | |
| I | 0.0105 ± 0.0049<br>± 0.0034 | 0.0974 ± 0.0205<br>± 0.0171 | 920.8 | 915.2 |
| II | 0.0098 ± 0.0043<br>± 0.0031 | 0.0974 ± 0.0205<br>± 0.0168 | 920.8 | 915.3 |
| $\mu t = 0.01$, $st = 0.5$ | | | | |
| I | 0.0103 ± 0.0036<br>± 0.0038 | 0.5140 ± 0.0381<br>± 0.0477 | 1218.1 | 1207.9 |
| II | 0.0080 ± 0.0027<br>± 0.0028 | 0.5141 ± 0.0405<br>± 0.0460 | 1217.7 | 1208.2 |
| $\mu t = 0.01$, $st = 1.0$ | | | | |
| I | 0.0101 ± 0.0038<br>± 0.0044 | 1.0456 ± 0.1033<br>± 0.0925 | 1348.0 | 1334.1 |
| II | 0.0060 ± 0.0021<br>± 0.0024 | 1.0540 ± 0.1151<br>± 0.0863 | 1346.8 | 1334.8 |
| $\mu t = 0.1$, $st = 0.1$ | | | | |
| I | 0.1081 ± 0.0127<br>± 0.0159 | 0.1007 ± 0.0292<br>± 0.0258 | 1197.9 | 1143.7 |
| II | 0.0775 ± 0.0139<br>± 0.0091 | 0.1211 ± 0.0385<br>± 0.0194 | 1194.4 | 1146.7 |
| $\mu t = 0.1$, $st = 0.5$ | | | | |
| I | 0.1023 ± 0.0205<br>± 0.0220 | 0.4920 ± 0.0572<br>± 0.0674 | 1390.7 | 1311.8 |
| II | 0.0409 ± 0.0072<br>± 0.0065 | 0.5882 ± 0.1007<br>± 0.0518 | 1372.8 | 1320.2 |
| $\mu t = 0.1$, $st = 1.0$ | | | | |
| I | 0.1110 ± 0.0345<br>± 0.0382 | 0.9758 ± 0.1289<br>± 0.1470 | 1476.4 | 1373.3 |
| II | 0.0083 ± 0.0067<br>± 0.0027 | 1.9009 ± 0.5301<br>± 0.2164 | 1410.4 | 1387.2 |
| $\mu t = 0.5$, $st = 0.5$ | | | | |
| I | 0.5176 ± 0.2086<br>± 0.2647 | 0.5292 ± 0.3722<br>± 0.4114 | 1660.5 | 1383.9 |
| II | 0.0147 ± 0.0106<br>± 0.0036 | 1.7892 ± 0.4286<br>± 0.1886 | 1434.8 | 1396.4 |

The average results of the sum approach and the direct alignment approach are presented for various values of $\mu t$ and $st$. The average results were obtained from 20 pairs of sequences. To produce a pair of sequences separated by a particular value of $\mu t$ and $st$, a descendant sequence was evolved as described in the

native inferences are both relatively probable becomes more common. The sum approach is not forced to make the same kind of choice between alternatives. It can estimate parameters by considering each alternative in proportion to its likelihood. In other words, the best alignment can often be a poor source of information about the actual values of evolutionary parameters.

We believe that the detected bias is not particular to our model. Instead, this bias is likely to arise any time evolutionary parameters are being estimated from a single alignment; it does not matter whether this alignment is a maximum likelihood alignment or a subjective alignment. Because phylogeny inference techniques tend to be based on the analysis of single multiple-sequence alignments, the estimates of evolutionary parameters obtained by phylogeny inference techniques will be biased, especially when distantly related sequences are being considered. The significance, if any, of this bias on the inference of phylogenetic tree topology is unknown.

The estimates of standard error derived from the inverse of the information matrix were quite similar to the sample standard errors (Table 1). This similarity is fortunate because sample standard errors cannot be calculated for actual data whereas the inverse of the information matrix can be calculated. This similarity implies that the inverse of the information matrix yields a reliable predictor of parameter estimate precision.

The quality of the performance of all three approaches deteriorates as the evolutionary distance separating a pair of sequences increases because it becomes more difficult to correctly infer which events are responsible for the differences between two sequences as the number of differences accumulates. In addition, it was found that the precision of parameter estimation increases with increasing sequence length (Fig. 1). This result is expected because long sequences can be viewed as large data sets and short sequences can be viewed as small data sets.

The different parameter estimates obtained by the sum approach and the direct alignment approach can have a pronounced effect on the appearance of the maximum likelihood alignments

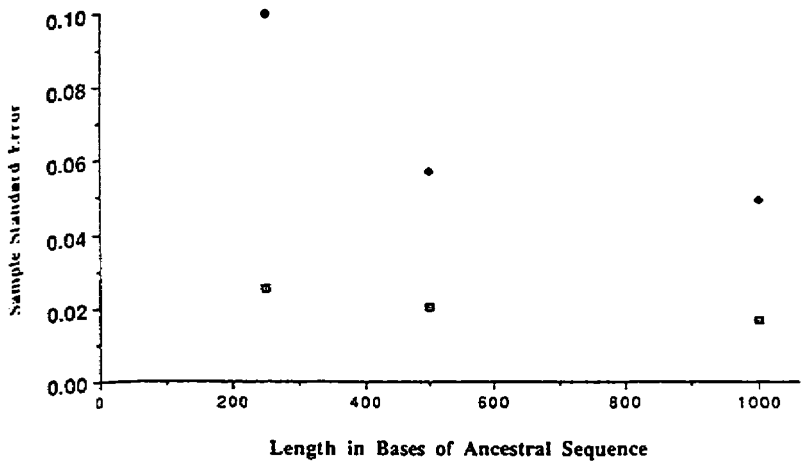

Fig. 1. The effect of sequence length on the standard error of *μt* and *st*. Pairs of sequences with *μt* = 0.1 and *st* = 0.5 were simulated as described in the text. Parameter estimates were obtained from the sum approach and sample standard errors as calculated from the analysis of 20 pairs of sequences are shown. Data points represent the standard errors associated with ancestral sequence lengths of 250, 500, or 1000 bases. The square symbols represent standard errors of *μt* and the filled diamond symbols represent standard errors of *st*.

A:

```
GACAAATCC-C-TGAGACCCC-TTCAGTAGTTAACACGTA-ATC-ATTGTT-TGTC-CGTAGCGGTAAGA
G-CTAATCCGCCCGTGACCCCCTTC-CAAGGAAAAACCCACATCCACTGTGCTACCGCGTAGT-TCACGA

CAGATACGAACCTACTCCTCGCAC-AGCGAAGGTGCGAAACAA-TAATTGCGAAGTGAGTAACTTGATTG
AGGGAACGTA-CTACGGAT-GCAGGAAGGAGGGTGC-AAAGAATTAATGGAGCACTTAGTAA-ATGATTG
```

B:

```
GACAAATC-CCTGAGACCCCTTCAGTAGTTAACACGTA-ATC-ATTGTTTGTCCGTAGCGGTAA-GACAG
```

approach and the direct alignment approach. We believe that the contrast between this alignment and the alignment of Bishop and Thompson (Fig. 3) is

principal investigator Joseph Felsenstein and by NSF grant number BSR 8918333. Jeff Thorne was also supported by a National Science Foundation Graduate Fellowship.

J Mol Evol (1992) 34:91–92

Journal of
Molecular Evolution

## Erratum

An Evolutionary Model for Maximum Likelihood Alignment of DNA Sequences
J. L. Thorne, H. Kishino, J. Felsenstein
J Mol Evol (1991) 33:114–124

Equations (6) should read:

$$P(\alpha' \mid \theta) = \gamma_4 p_2''(t) p_0'(t) p_1(t) p_1'(t) p_2(t)$$
$$P(\alpha \mid \theta, \alpha') = \pi_G \pi_T \pi_G f_{GC}(t) \pi_T \pi_A \pi_C f_{CC}(t) \pi_A .$$

The last formula of Equations (9) should begin with $p_n''(t)$ instead of $p_n'(t)$. Thus, Equations (9) should read:

$$p_n(t) = e^{-\mu t}(1 - \lambda\beta(t))(\lambda\beta(t))^{n-1} \quad n > 0$$

$$p_n'(t) = (1 - e^{-\mu t} - \mu\beta(t))(1 - \lambda\beta(t))(\lambda\beta(t))^{n-1} \quad n > 0$$

$$p_0'(t) = \mu\beta(t)$$

$$p_n''(t) = (1 - \lambda\beta(t))(\lambda\beta(t))^{n-1} \quad n > 0$$

## Announcements

The **Fifth International Conference on the Cell and Molecular Biology of** ***Chlamydomonas*** will be held, May 26–31, 1992 at the Asilomar Conference Center in Pacific Grove, CA. The meeting will consist of platform and poster sessions devoted to all aspects of the molecular biology and genetics of *Chlamydomonas*. Platform sessions will include:

| | Session | Chair |
|---|---|---|
| I. | Cell Differentiation and Life Cycle | Ursula Goodenough |
| II. | Photosynthesis | Richard Sayre |
| III. | Molecular Biology of Dynein | David Mitchell |
| IV. | Biochemistry and Metabolism | Emilio Fernandez |
| V. | Mating, Signal Transduction, and Behavioral Response | Herman van den Ende |
| VI. | Innovations in Genetics and Molecular Biology of *Chlamydomonas* | Paul Lefebvre |
| VII. | The Flagellar Apparatus: Basal Bodies and Assembly | Joel Rosenbaum |
| VIII. | Organelle Genetics and Molecular Biology | Elizabeth Harris |

There will also be one or two other platform sessions to be announced. For further information, please contact Dr. George Witman, Organizer, The Worcester Foundation for Ex-