J Mol Evol (1992) 34:3–16

Journal of Molecular Evolution

Inching toward Reality: An Improved Likelihood Model of Sequence Evolution

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

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

Summary. Our previous evolutionary model is generalized to permit approximate treatment of multiple-base insertions and deletions as well as regional heterogeneity of substitution rates. Parameter estimation and alignment procedures that incorporate these generalizations are developed. Simulations are used to assess the accuracy of the parameter estimation procedure and an example of an inferred alignment is included.

Key words: Alignment — Maximum likelihood procedure — Dynamic programming — Evolutionary model — Insertion–deletion model

1 Introduction

Accompanying the recent explosion in the amount of available DNA and protein sequence data has been a smaller burst of methods designed to analyze these data. Statistically rigorous methods exist to treat data sets consisting of aligned sequences, but it is often the case that the appropriate alignment between sequences is far from obvious. In these cases, the method of alignment becomes important. Many elegant dynamic programming algorithms for the alignment of sequences have been developed but these algorithms lack a rigorous statistical basis. One outcome of this lack of a statistical foundation is the incorporation of subjectivity into the most widely used alignment methods.

A brief description of the most widely used alignment algorithms will suffice to reveal their subjective nature. These algorithms can be categorized as similarity-based or distance-based. Because these two categories are closely related (Smith et al. 1981) and because their subjective properties are shared, this description will focus on distance-based algorithms. Distance-based algorithms search for the least-penalized alignment. An alignment is penalized for each evolutionary event that it postulates. Assume, for example, that substitutions, insertions, and deletions are the only evolutionary events being considered; more exotic evolutionary events such as inversions will be ignored. The following alignment exhibits two mismatches, a gap of length one, and a gap of length three:

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

The penalty associated with this alignment would be $2m + G_1 + G_3$ where m is the penalty for a mismatch and G_i is the penalty for a gap of length i. If $G_i = a + bi$, then the penalty of the above alignment would be $2m + 2a + 4b$. The appropriate values of m, a, and b are not obvious. As a consequence, subjective decisions are often made to choose these values. Noteworthy attempts to avoid this subjectivity include the Monte Carlo method of Fitch and Smith (1983) and the minimum message length method of Allison and Yee (1990).

A more subtle type of subjectivity of widely used alignment methods also should not be ignored. This type of subjectivity involves the form of the function that relates the evolutionary events postulated by an alignment to the penalty associated with the alignment. For instance, it is not clear that the form of the gap penalty should be $G_i = a + bi$. Certain

* *Current address:* Department of Plant Breeding and Biometry, Cornell University, Ithaca NY 14853, USA
† *Current address:* Ocean Research Institute, University of Tokyo, 1-15-1, Minami-dai, Nakano-ku, Tokyo 164, Japan
Offprint requests to: J.L. Thorne

implicit assumptions about the evolutionary process and particularly about the insertion–deletion process must underly the assignment of the function $a + bi$ to the gap penalty. Without a statement of the assumptions that govern the evolutionary process, it is not even clear that an arbitrarily chosen function such as $G_i = \sqrt{a + bi}$ is less plausible than $G_i = a + bi$. The truth is that G_i is commonly set to $a + bi$ because Gotoh (1982) invented a clever computationally feasible alignment algorithm for this form of G_i. Without a model of sequence evolution and the accompanying theoretical justification, the criteria used to select good alignments are highly subjective.

Our approach is to develop a likelihood model of the evolutionary process. The advantages of this approach include explicit assumptions, a model of sequence change based upon actual biological phenomena instead of arbitrary criteria for sequence comparison, and the vast statistical theory concerned with likelihood methods. Bishop and Thompson (1986) were the pioneers of this approach, and we (Thorne et al. 1991) modified and improved it. Our earlier evolutionary model allowed substitutions, single-base insertions, and single-base deletions. By solving the set of differential equations that govern this evolutionary model, explicit forms for the transition probabilities of sequence evolution were obtained. We showed that these transition probabilities can serve as the basis for recursive algorithms that estimate evolutionary parameters and infer sequence alignments. In this paper, we generalize our earlier likelihood model to allow approximate treatment of multiple-base insertions and deletions as well as regional heterogeneity of substitution rates, we develop recursive algorithms for parameter estimation and alignment inference that incorporate these generalizations, and we present simulations that assess the performance of our algorithms. The focus of this paper is DNA sequence analysis, but the ideas are easily extended to the analysis of protein sequences.

2 The Evolutionary Model

2.1 The Insertion–Deletion Process: The Fragment Model

The two components of our evolutionary model are the insertion–deletion process and the substitution process. The multiple-base insertion–deletion model presented here treats each biological sequence as a sequence of fragments. It should be noted immediately that the fragment is merely a convenient theoretical unit and not a biological entity. Each fragment of a DNA sequence consists of one or more adjacent nucleotides. The insertion–deletion process inserts and deletes fragments. The insertion or deletion of a fragment is independent of the insertion or deletion of all other fragments within the sequence. This means that the probability of more than one fragment being inserted or deleted at a specific instant is negligible. Fragment boundaries do not change over evolutionary time. Therefore, if several adjacent nucleotides are inserted at a specific time then these belong to the same fragment and will continue to belong to the same fragment as evolution progresses. Because our evolutionary model operates at the level of fragments, a nucleotide in a fragment can only be deleted at a specific time if all other nucleotides in this fragment are deleted at the same time.

In Thorne et al. (1991), we presented the insertion–deletion process as a birth–death process of imaginary links that separate the DNA bases of a sequence. There are two types of links: normal and immortal. Under the single-base insertion–deletion model, each normal link is associated with exactly one nucleotide. Specifically, there is a normal link to the right of each base. In addition, there is an immortal link to the left of the leftmost base in the sequence. Both types of links can be associated with births. The birth rate per normal link (λ) is equal to the birth rate per immortal link (λ). A newborn link and its associated nucleotide are always inserted directly to the right of the parent link. Only normal links can die; the deletion rate per normal link is μ. The death of a normal link is accompanied by deletion of its associated nucleotide. Therefore, a sequence of length n nucleotides experiences single base deletions at rate $n\mu$ and single-base insertions at rate $(n + 1)\lambda$.

To adapt the earlier single-base birth–death model to a fragment birth–death model, we now allow each normal link to be associated with one or more nucleotides. Each normal link and its associated nucleotide(s) can be viewed as a fragment. To the right of each fragment is a normal link. To the left of the leftmost fragment in the sequence is the immortal link.

For example, under the fragment model it is possible for the DNA sequence GCA to be in any of four possible states (Fig. 1). These states are defined by how the DNA sequence GCA can be depicted in terms of fragments. State A of Fig. 1 portrays the G a being in one fragment and the CA as being in another. Alternatively, state B of Fig. 1 portrays the GC as being in one fragment and the A as being in another. State C and state D, the other two possible depictions of GCA, respectively correspond to the possibility that GCA contains exactly three normal links and the possibility that GCA contains exactly one normal link.

In terms of links, the birth–death process of the fragment model is identical to the birth–death process of the earlier single-base insertion–deletion

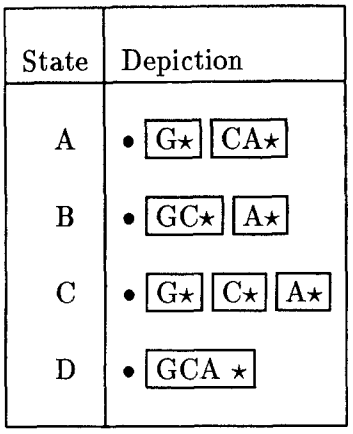

Fig. 1. The four possible states of the DNA sequence GCA under the fragment model. The symbol ● represents the immortal link. The symbol ★ represents a normal link. Nucleotides that belong to the same fragment are encased by a rectangle.

model. Concerning the fate of an individual link over time, three types of transition probabilities are considered: $p_n(t)$ is the probability that n links are descended from a normal link and one of them is the original after a timespan of length t, $p'_n(t)$ is the probability that n links are descended from a normal link and the original dies during a timespan of length t, and $p''_n(t)$ is the probability that the immortal link has n descendants including itself during a timespan of length t. By their definitions, $p_0(t) = p''_0(t) = 0$. The remainder of the transition probabilities (Thorne et al. 1991) are

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

where

$$\beta(t) = \frac{1 - e^{(\lambda-\mu)t}}{\mu - \lambda e^{(\lambda-\mu)t}}. \tag{2}$$

Under the fragment model, the distribution of the number of nucleotides per fragment is as important as these transition probabilities. Let $h(n)$ be the probability that a normal link is associated with exactly n nucleotides where $n = 1, 2, \ldots$. It is not known which distribution of $h(n)$ most accurately reflects the actual biological insertion–deletion process. We choose to put $h(n)$ in the form of a geometric distribution,

$$h(n) = (1 - r)r^{n-1} \quad 0 \le r < 1, \; n = 1, 2, \ldots \tag{3}$$

It can be seen that our earlier single-base insertion–deletion model was the special case of the geometric fragment model where $r = 0$. It is possible to develop a sequence alignment algorithm and an evolutionary parameter estimation algorithm for any arbitrary form of $h(n)$. If N is the number of nucleotides in the longer sequence and M is the number of nucleotides in the shorter sequence, then the amount of computation required by these algorithms should be, at most, proportional to N^2M. As will be shown later, the amount of computation required by the sequence alignment algorithm and the parameter estimation algorithm is proportional to NM when $h(n)$ is distributed geometrically. The geometric form of $h(n)$ yields an alignment algorithm resembling that of Gotoh (1982).

Let ρ_n be the equilibrium probability of sequences with n normal links. In Thorne et al. (1991), it was shown that

$$\rho_n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n \quad 0 < \lambda < \mu, \quad n = 0, 1, \ldots \tag{4}$$

Let γ_n be the equilibrium probability under the geometric fragment model of sequences n nucleotides in length. Because a sequence of length zero nucleotides must contain exactly zero normal links, $\gamma_0 = \rho_0$. Because a sequence of length one nucleotide must contain exactly one normal link that is associated with a fragment of size one nucleotide, $\gamma_1 = \rho_1 h(1)$. Examination of the probability-generating function (e.g., Feller 1968) for sequence length reveals that $\gamma_n = \gamma_{n-1}[\frac{\lambda}{\mu}(1 - r) + r]$ when $n \ge 2$. Therefore, the equilibrium sequence length distribution is nearly, but not quite, a geometric distribution:

$$\gamma_0 = 1 - \frac{\lambda}{\mu}$$

$$\gamma_n = \left(1 - \frac{\lambda}{\mu}\right)\frac{\lambda}{\mu}(1 - r)\left[\frac{\lambda}{\mu}(1 - r) + r\right]^{n-1} \quad n \ge 1$$

Although more realistic than our earlier single-base insertion–deletion model, the fragment model is obviously not an ideal mathematical description of the actual biological process of insertion and deletion. An inherent flaw in the fragment model is more easily understood by considering two modern sequences that are descended from the same ancestral sequence. Assume that a deletion occurs in the evolutionary lineage to one descendant sequence, and another overlapping but nonidentical deletion occurs in the evolutionary lineage to the other descendant sequence. Under the fragment model, it is not possible to explain these two overlapping deletions with a scenario of just two events; the fragment model requires at least three events (Fig. 2). The severity of this flaw rises as the evolutionary distance between sequences increases because the probability of independent overlapping deletions in related lineages rises as evolutionary distance in-

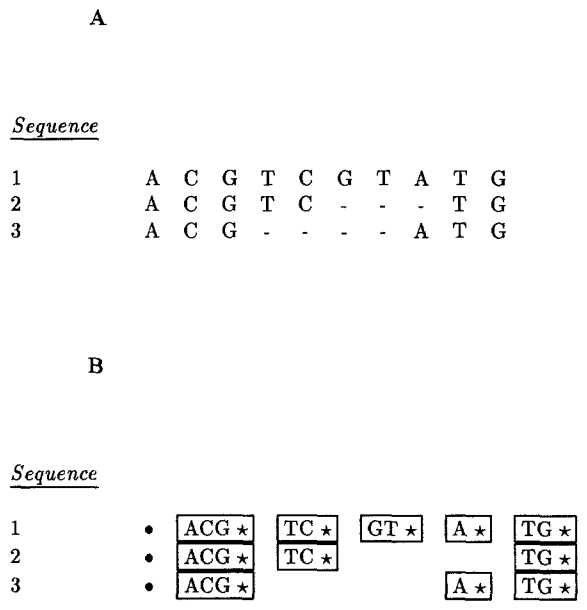

Fig. 2. A flaw of the fragment model. **A** An alignment between an ancestral sequence and two descendant sequences. The gaps shown in this alignment can be explained by two overlapping deletion events. **B** A depiction of the alignment in Fig. 2A which is consistent with the fragment model. Possible fragment boundaries are shown. The fragment model cannot explain the gaps in the alignment of Fig. 2A by two deletion events; it requires at least three events to explain these overlapping deletions.

creases. Although flawed, the fragment model is still superior to a single-base insertion–deletion model. The alignment in Fig. 2A can be explained by a minimum of two deletion events in reality, by a minimum of three deletion events under the fragment model, and by a minimum of five deletion events under a single-base insertion–deletion model.

2.2 *The Substitution Process*

Many substitution models could be used in the present context. We have selected a substitution model that was developed by Felsenstein and is used in the PHYLIP computer package, versions 2.6 and later (Felsenstein 1989). It is almost identical to one independently developed by Hasegawa et al. (1985); both of these models were first described by Hasegawa et al. (1985). This substitution model combines attributes of Kimura's (1980) two-parameter model and Felsenstein's (1981) model. This is a reversible substitution model that allows transition rates to exceed transversion rates. The model requires the use of equilibrium probabilities of the four types of nucleotides. The equilibrium probability of a type of nucleotide is equal to its expected frequency. The equilibrium probabilities of A, G, C, and T will be denoted by π_A, π_G, π_C, and π_T. The equilibrium probability of purines, π_R, is $\pi_A + \pi_G$. The equilibrium probability of pyrimidines, π_Y, is $\pi_T + \pi_C$.

This substitution model permits two types of substitution events. The first type can replace a purine only with a purine and can replace a pyrimidine only with a pyrimidine. We will refer to this type as a within-group substitution event. If a within-group substitution event occurs at a position occupied by a T, for example, the probability that the T is replaced by a C is π_C/π_Y and the probability that the T is replaced by another T is π_T/π_Y. We will refer to the second type of substitution event as a general substitution event. For a general substitution event, the type of nucleotide being replaced is not important; the base will be replaced by A, G, C, or T with respective probabilities π_A, π_G, π_C, and π_T. Let $f_{ij}(t)$ be the transition probability that a nucleotide which begins as type i is of type j at time t. To improve notation, let $H(i)$ indicate whether type i is purine or pyrimidine. In other words, $H(A) = H(G) = R$ and $H(C) = H(T) = Y$. If w is the within-group substitution rate and g is the general substitution rate, then

$$f_{ij}(t) = \begin{cases} e^{-(g+w)t} + e^{-gt}(1 - e^{-wt})\dfrac{\pi_j}{\pi_{H(j)}} & \\ \quad + (1 - e^{-gt})\pi_j & i = j \\ e^{-gt}(1 - e^{-wt})\dfrac{\pi_j}{\pi_{H(j)}} & i \neq j, \\ \quad + (1 - e^{-gt})\pi_j & H(i) = H(j) \\ (1 - e^{-gt})\pi_j & H(i) \neq H(j) \end{cases} \tag{5}$$

2.3 *Regional Heterogeneity of Substitution Rates*

It has been well established that different portions of a sequence can evolve at different rates. To allow heterogeneity of substitution rates from region to region within a DNA sequence, we simply postulate the existence of several varieties of fragments. Each fragment variety possesses a specific general substitution rate and a specific within-group substitution rate. When a fragment is inserted into a sequence, the probability that the fragment is of a specific variety is independent of the fragment varieties in the neighborhood of the insertion site. This type of insertion process does not produce clustering (or overdispersion) of fragment varieties within a sequence. The probability that a newly inserted fragment is of a specific variety will be termed the equilibrium frequency of the fragment variety. It is possible to postulate as many varieties of fragments as desired. It is even possible to allow variation in fragment size distribution between fragment varieties, although allowing this variation is likely to decrease the speed of the algorithms used to analyze the sequences. It may not be practical to postulate

the existence of many varieties of fragments because the limited length of DNA sequences can make the resolution of the characteristics (e.g., substitution rate, size distribution) of each type of fragment difficult.

Our practice has been to use a very simple model of substitution rate heterogeneity. This model assumes that there are only two varieties of fragments. One variety of fragments experiences substitutions relatively rarely (slow fragments) and the other variety experiences substitutions relatively often (fast fragments). This simple model of substitution rate heterogeneity also assumes that the fragment size distribution of these two varieties is identical and that the ratio of the within-group substitution rate to the general substitution rate is identical between the slowly evolving fragments and the quickly evolving fragments. Under this scenario, slow fragments experience a within-group substitution rate of w and a general substitution rate of g. Furthermore, it is necessary to estimate two more parameters (p_f and k_f) relevant to the substitution process. The parameter, p_f, represents the equilibrium frequency of fast fragments. From the value of p_f, the equilibrium frequency of slow fragments (p_s) is immediate because $p_f + p_s = 1$. The parameter, k_f, relates the substitution rates in slow fragments and fast fragments. The within-group substitution rate of fast fragments is $k_f w$, and the general substitution rate of fast fragments is $k_f g$ where $k_f > 1$.

The fragment size distribution is intricately related to both the insertion–deletion process and the pattern of regional heterogeneity of substitution rates. This is a flaw; the relationship may not be realistic and would not be necessary in a more advanced evolutionary model. A more advanced evolutionary model also might permit increased rates of insertion and deletion in regions that experience high rates of substitution.

3 Procedures

3.1 Estimation of Evolutionary Parameters

The procedure for estimating evolutionary parameters under the model of regional homogeneity of substitution rates will be referred to as the homogeneity procedure. The procedure for estimating evolutionary parameters under the aforementioned model of regional heterogeneity of substitution rates will be referred to as the heterogeneity procedure. Both procedures utilize a function that returns the likelihood value for an input set of evolutionary parameter values and both procedures require a numerical maximization routine that calls this function. Because the two procedures are conceptually similar and because explanation of the homogeneity procedure requires less notation, only the details of the homogeneity procedure are introduced in this section. The heterogeneity procedure is briefly described in the Appendix.

The homogeneity procedure is a generalization of the parameter estimation procedure given in Thorne et al. (1991). Assume that two DNA sequences, A and B, are to be analyzed. The unknown evolutionary parameters to be estimated are represented by a vector $\theta = (\pi_A, \pi_G, \pi_C, \pi_T, r, \lambda t, \mu t, gt, wt)$. Without supplementary information, the evolutionary rates and the divergence time cannot be separately estimated; measures of evolutionary distance (the product of evolutionary rates and divergence time) can be estimated. Consider the calculation of the likelihood of A and B,

$$L_\theta(A, B) = P(A, B \mid \theta) \tag{6}$$

The likelihood of two sequences is the sum of the likelihood of all possible alignments between the two sequences. As we demonstrated in the earlier paper, the precision of parameter estimation is vastly enhanced and the amount of bias in parameter estimates is greatly reduced by considering all alignments instead of a single alignment.

A recursive algorithm for the calculation of $L_\theta(A, B)$ can be formulated, but several definitions are necessary. Assume that the length of sequence A is s_A and the length of sequence B is s_B. Denote the subsequence consisting of the first m bases of sequence A by A_m and denote the first n bases of sequence B by B_n. Let a_m be the mth base of sequence A and b_n be the nth base of sequence B. Because the evolutionary model is reversible, sequence A can be considered an ancestor of sequence B without loss of generality. This implies that all links in sequence B are descendants of links in sequence A. Define $S(A_m, B_n)$ to be the set of all possible alignments between A_m and B_n. $S(A_m, B_n)$ can be partitioned into six subsets by considering the rightmost link of A_m. Each possible alignment of $\alpha(A_m, B_n)$ between A_m and B_n is a member of exactly one of these six subsets of $S(A_m, B_n)$:

$S^1(A_m, B_n)$ = [$\alpha(A_m, B_n)$ where the rightmost link of A_m survives to become the rightmost link of B_n]

$S^2(A_m, B_n)$ = [$\alpha(A_m, B_n)$ where the rightmost link of A_m has no descendant links in B_n]

$S^3(A_m, B_n)$ = [$\alpha(A_m, B_n)$ where the rightmost link of A_m has at least two descendant links in B_n]

$S^4(A_m, B_n)$ = [$\alpha(A_m, B_n)$ where the rightmost link of A_m does not survive but has the rightmost link of B_n as its sole descendant. In addition, the fragment

associated with the rightmost link of A_m is exactly the same length as the fragment associated with the rightmost link of B_n]

$S^5(A_m, B_n)$ = [$\alpha(A_m, B_n)$ where the rightmost link of A_m does not survive but has the rightmost link of B_n as its sole descendant. In addition, the fragment associated with the rightmost link of A_m is longer than the fragment associated with the rightmost link of B_n]

$S^6(A_m, B_n)$ = [$\alpha(A_m, B_n)$ where the rightmost link of A_m does not survive but has the rightmost link of B_n as its sole descendant. In addition, the fragment associated with the rightmost link of A_m is shorter than the fragment associated with the rightmost link of B_n]

Each of the six alignments depicted in Fig. 3 belongs to a different one of these six subsets. To keep track of the likelihood of each of these six subsets of $S(A_m, B_n)$ define

$$L_\theta^i(m, n) = P[\alpha(\mathrm{A}_m, \mathrm{B}_n) \in S^i(A_m, B_n) \mid \theta] \qquad i = 1, 2, \ldots, 6 \quad (7)$$

With these six likelihood terms, a recursive algorithm to calculate $L_\theta(A, B)$ can be developed.

The first base in each sequence has to be specially treated. If a sequence only consists of a single base then it must be the case that the sequence consists of exactly one single-base fragment; the single base of this sequence cannot be part of a larger fragment. For this reason, the indicator κ_i is introduced:

$$\kappa_i = \begin{cases} 0 & i = 1 \\ 1 & i \neq 1 \end{cases} \quad (8)$$

The boundary conditions of the recursive algorithm are

$$L_\theta^1(0, 0) = \gamma_0 p''_1(t)$$

$$L_\theta^i(0, 0) = 0 \qquad i = 2, 3, 4, 5, 6,$$

$$L_\theta^2(1, 0) = \gamma_1 p''_1(t)\pi_{a_1} p'_0(t)$$

$$L_\theta^2(m, 0) = \gamma_1 p''_1(t)\pi_{a_1} p'_0(t) \cdot \prod_{j=2}^{m} \pi_{a_j}[r + p'_0(t)(1 - r)\frac{\lambda}{\mu}] \quad m \geq 2$$

$$L_\theta^i(m, 0) = 0 \qquad m \geq 1, \quad i = 1, 3, 4, 5, 6$$

$$L_\theta^3(0, 1) = \gamma_0 p''_2(t)(1 - r)\pi_{b_1}$$

$$L_\theta^3(0, n) = \gamma_0 p''_2(t)(1 - r)\pi_{b_1} \cdot \prod_{j=2}^{n} \pi_{b_j}[r + \lambda\beta(t)(1 - r)] \quad n \geq 2$$

$$L_\theta^i(0, n) = 0 \qquad n \geq 1, \quad i = 1, 2, 4, 5, 6$$

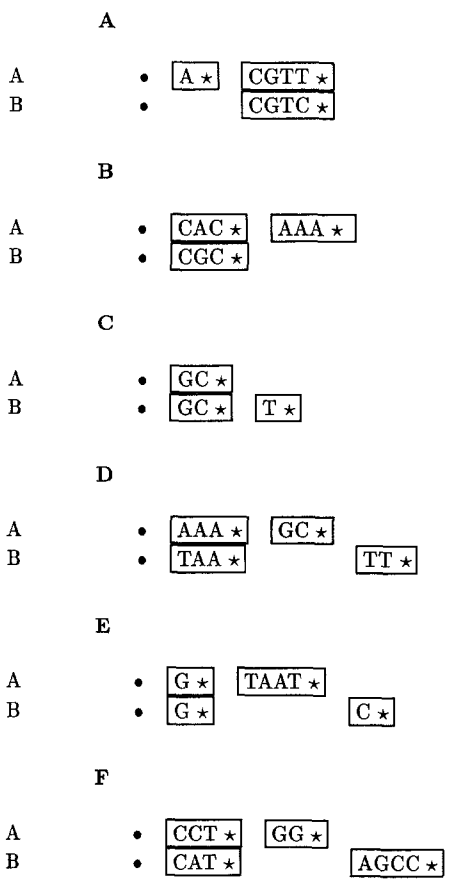

Fig. 3. The rightmost fragment of an alignment can be used to place a specific alignment between two sequences into a particular subset of all possible alignments between the sequences. Examples of pairwise alignments along with specification of the subset to which they belong are shown. **A** An alignment that belongs to $S^1(A_5, B_4)$. **B** An alignment that belongs to $S^2(A_6, B_3)$. **C** An alignment that belongs to $S^3(A_2, B_3)$. **D** An alignment that belongs to $S^4(A_5, B_5)$. **E** An alignment that belongs to $S^5(A_5, B_2)$. **F** An alignment that belongs to $S^6(A_5, B_7)$.

For $1 \leq m \leq s_A$ and $1 \leq n \leq s_B$, the recursive algorithm follows these rules:

$$L_\theta^1(m, n) = \pi_{a_m} f_{a_m b_n}(t)[r\kappa_m\kappa_n L_\theta^1(m - 1, n - 1) + p_1(t)(1 - r)\frac{\lambda}{\mu} \cdot \sum_{i=1}^{6} L_\theta^i(m - 1, n - 1)]$$

$$L_\theta^2(m, n) = \pi_{a_m}[r\kappa_m L_\theta^2(m - 1, n) + p'_0(t)(1 - r)\frac{\lambda}{\mu}\sum_{i=1}^{6} L_\theta^i(m - 1, n)]$$

$$L_\theta^3(m, n) = \pi_{b_n}[r\kappa_n L_\theta^3(m, n - 1)$$

$$+ \lambda\beta(t)(1-r) \sum_{i=1,3,4,5,6} L_\theta^i(m, n-1)]$$

$$L_\theta^4(m, n) = \pi_{a_m}\pi_{b_n}[r^2\kappa_m\kappa_n L_\theta^4(m-1, n-1)$$

$$+ p'_1(t)(1-r)^2\frac{\lambda}{\mu}\sum_{i=1}^{6} L_\theta^i(m-1, n-1)]$$

$$L_\theta^5(m, n) = \pi_{a_m}\kappa_m\kappa_n r \sum_{i=4}^{5} L_\theta^i(m-1, n)$$

$$L_\theta^6(m, n) = \pi_{b_n}\kappa_m\kappa_n r \sum_{i=4,6} L_\theta^i(m, n-1)$$

Then, the likelihood of the two sequences is obtained by

$$L_\theta(A, B) = \sum_{i=1}^{6} L_\theta^i(s_A, s_B) \qquad (9)$$

A maximum likelihood estimate of the evolutionary parameters can be obtained by finding the value of θ that satisfies $\max_\theta L_\theta(A, B)$. To find this value of θ, a numerical maximization routine can be used in conjunction with the above method for calculating $L_\theta(A, B)$. The computer code of the numerical maximization routine, which produced the results presented in this paper, was written by Press et al. (1988) and is an implementation of the simplex maximization routine of Nelder and Mead (1965).

3.2 Alignment Inference

Two useful types of alignments will be considered with respect to the framework of the fragment model. The maximum likelihood algorithms for inferring both types of alignment are dynamic programming algorithms. In both cases, entries of an $(s_A + 1) \times (s_B + 1)$ matrix are iteratively calculated. Like the matrix used by the evolutionary parameter estimation procedure, the set of all alignments between A and B [i.e., $S(A, B)$] is partitioned into subsets. A site (m, n) of the matrix contains entries representing the likelihood of the most likely alignment in each specific subset of $S(A_m, B_n)$. A traceback procedure is used to find the best path through this matrix. This path specifies the inferred alignment. The first type of alignment is conventional; it exhibits only the relationships between the nucleotides of the two different DNA sequences and does not specify boundaries between fragments. This type of alignment would usually be preferred in situations where the possibility of regional heterogeneity of substitution rates will be ignored. The second type of alignment is useful when regional heterogeneity of substitution rates is suspected. This type of alignment is obtained by inferring the relationships between the nucleotides of the two sequences, the boundaries between fragments, and whether each fragment is fast or slow. Just as the evolutionary parameter estimation algorithm of section 3.1 is a generalization of the conceptually similar algorithm of Thorne et al. (1991), the algorithms for the inference of the two types of alignment discussed here are generalizations of the conceptually similar alignment algorithm of Thorne et al. Therefore, the specific description of the two new alignment algorithms is omitted.

4 Simulation Studies

4.1 Design

Parameter estimation properties were investigated by simulation study. Pairs of sequences were generated by evolving from an ancestral sequence A to a descendant sequence B. The evolutionary process in the simulation was consistent with the fragment model except that the number of fragments in the ancestral sequence was fixed so that the expected length of an ancestral sequence with this number of fragments would be 500 bases. Because the average fragment length is $1/(1 - r)$ bases, the number of fragments in the ancestral sequence was equal to $500/(1 - r)$. The purpose of this intentional violation was to reduce the effects of variable initial sequence length on the estimation of evolutionary parameters. For the simulated evolutionary process,

$$\lambda = \mu\frac{s_A(1 - r) - 2r + \sqrt{s_A^2(1-r)^2 + 4r}}{2(s_A + 1)(1 - r)} \qquad (10)$$

This is the maximum likelihood estimate of λ for given values of μ, s_A, and r under our evolutionary model. The divergence time was set to $t = 1.0$. The base composition was $\pi_A = \pi_G = \pi_C = \pi_T = 0.25$. For this base composition, the substitution model described in section 2.2 reduces to Kimura's two-parameter model (Kimura 1980).

Ideally, all evolutionary parameters—including λt and the equilibrium base frequencies—would be simultaneously estimated under the maximum likelihood framework. The finite amount of computer time available makes this ideal impractical. To make sequence analysis practical, each equilibrium base frequency (π_A, π_G, π_C, π_T) was estimated by the frequency of appearances of that type of base in the evolved sequences. The observed base frequencies may not be the maximum likelihood estimates of the equilibrium base frequencies. To further reduce the number of parameters to be estimated, λt was fixed at

$$\lambda t = \phi t \qquad (11)$$

where

$$\phi = \mu(s_A + s_B)(1 - r) - 4r + \sqrt{(s_A + s_B)^2(1 - r)^2 + 16r} /(2(s_A + s_B + 2)(1 - r)) \quad (12)$$

For two unrelated sequences evolving according to given values of μ, r, s_A, and s_B, the maximum likelihood value of λ is equal to ϕ. For related sequences, ϕt may not be the maximum likelihood estimate of λt. The possible effects of not obtaining maximum likelihood estimates for all parameters are unknown.

In the future, it may be determined that certain parameters (e.g., λ, μ, r, t, and the equilibrium base frequencies) can be predicted accurately prior to sequence analysis. If this proves to be the case, it may be unnecessary to estimate these parameters for each specific pair of sequences. Unfortunately, we believe that current knowledge of molecular evolution is generally insufficient for accurate a priori parameter estimation.

It is interesting to study the behavior of the heterogeneity procedure when there is no regional heterogeneity of substitution rates. Likewise, it is interesting to study the behavior of the homogeneity procedure when there actually is regional heterogeneity of substitution rates. To fulfill these objectives, each pair of simulated sequences was analyzed by both the homogeneity procedure and the heterogeneity procedure regardless of whether the pair of sequences was evolved with or without regional heterogeneity of substitution rates.

In addition to calculating the sample standard error of parameter estimates by analyzing many replicate pairs of sequences that were evolved under the same value of θ, a more approximate measure of standard error and the covariance structure in general was investigated. This approximation of the asymptotic error has the advantage that it can be calculated from analyzing a single pair of sequences. The approximation can be obtained in the standard way by evaluating the inverse of the Fisher information matrix (i.e., the Hessian matrix of the negative log likelihood: Kendall and Stuart 1973) for a pair of sequences.

4.2 Simulation Results and Discussion

Analysis of simulated pairs of sequences showed that evolutionary parameters can be reasonably accurately estimated by the homogeneity procedure when there is no regional heterogeneity of substitution rates (Table 1). The substitution process parameters, *gt* and *wt*, were especially well estimated. The parameter that determines the fragment size distribution (r) tends to be underestimated. Because maximum likelihood estimators can be biased, this is not surprising. Because long sequences contain more information about the evolutionary process than do short sequences, they yield relatively accurate estimates of r and the other evolutionary parameters (Table 2).

When μt is small and r is large, it is common for no deletion events to occur during the evolution from an ancestral sequence of moderate length (e.g., 500 bases) to a descendant sequence. When the situation does occur, the maximum likelihood estimate of μt should be near zero. Appropriately, this situation did produce estimates of μt near zero when either the homogeneity procedure or the heterogeneity procedure was employed. Zero is at the boundary of the parameter space of μt. Unfortunately, the estimates of the covariance structure that were derived from the Fisher information matrix were not satisfactory when parameter estimates were near the boundary of the parameter space. For example, variance estimates from the Fisher information matrix were often negative. For pairs of sequences that were evolved without regional heterogeneity of substitution rates and then analyzed by the homogeneity procedure, the estimate of the variance of μt from the Fisher information matrix was negative in 1 out of 100 cases when $\mu t = 0.01$ and $r = 0.5$, 40 out of 100 cases when $\mu t = 0.01$ and $r = 0.9$, and 0 out of 100 cases when $\mu t = 0.1$ and $r = 0.5$. As would be expected, the estimate of the variance of r from the Fisher information matrix was also occasionally negative. Problems with the Fisher information matrix became even more severe when pairs of sequences that were evolved without regional heterogeneity were analyzed by the heterogeneity procedure. Values of $p_f = 1$, $p_f = 0$, and $k_f = 1$ all indicate a lack of regional heterogeneity in substitution rates. All of these boundary values were capable of yielding Fisher information matrices that contained negative variance estimates.

This lack of success is evidently due to failure of the numerical maximization routine to converge at the maximal point of the likelihood surface. In this case, the Fisher information matrix would not be expected to yield the desired variance covariance structure. The parameter space searched by the homogeneity procedure is a specific portion of the parameter space searched by the heterogeneity procedure. Therefore, the maximum likelihood value [i.e., $L_\theta(A, B)$] returned by the heterogeneity procedure should always be greater than or equal to the maximum likelihood value returned by the homogeneity procedure. As further evidence of the failure of the numerical maximization routine near the boundaries of the parameter space, the maximum likelihood value returned by the homogeneity procedure was not always greater than the maximum likelihood value returned by the heterogeneity procedure in practice; when the heterogeneity proce-

Table 1. Performance of the homogeneity procedure for pairs of sequences with no regional heterogeneity of substitution rates

μt	r	gt	wt
A) Cases with low deletion probability per fragment ($\mu t = 0.01$) and small fragments ($r = 0.5$)			
0.01	0.5	0.1	0.1
0.0082 ± 0.0045	0.42 ± 0.18	0.098 ± 0.022	0.094 ± 0.032
± 0.0047	± 0.21	± 0.021	± 0.029
0.01	0.5	0.1	0.5
0.0107 ± 0.0056	0.38 ± 0.18	0.098 ± 0.017	0.493 ± 0.041
± 0.0052	± 0.19	± 0.021	± 0.066
0.01	0.5	0.1	1.0
0.0097 ± 0.0040	0.46 ± 0.20	0.097 ± 0.021	1.028 ± 0.164
± 0.0057	± 0.18	± 0.021	± 0.131
0.01	0.5	0.5	0.5
0.0112 ± 0.0042	0.47 ± 0.14	0.484 ± 0.037	0.519 ± 0.105
± 0.0061	± 0.17	± 0.059	± 0.108
0.01	0.5	0.5	1.0
0.0106 ± 0.0082	0.36 ± 0.27	0.527 ± 0.065	1.045 ± 0.228
± 0.0049	± 0.19	± 0.063	± 0.198
B) Cases with low deletion probability per fragment ($\mu t = 0.01$) and large fragments ($r = 0.9$)			
0.01	0.9	0.1	0.1
0.0088 ± 0.0183	0.81 ± 0.22	0.090 ± 0.022	0.107 ± 0.030
± 0.0151	± 0.35	± 0.020	± 0.029
0.01	0.9	0.1	0.5
0.0099 ± 0.0160	0.86 ± 0.15	0.100 ± 0.020	0.506 ± 0.070
± 0.0156	± 0.10	± 0.021	± 0.068
0.01	0.9	0.1	1.0
0.0121 ± 0.0147	0.89 ± 0.08	0.097 ± 0.018	1.023 ± 0.153
± 0.0222	± 0.15	± 0.021	± 0.128
0.01	0.9	0.5	0.5
0.0090 ± 0.0133	0.74 ± 0.33	0.512 ± 0.058	0.520 ± 0.104
± 0.0140	± 0.17	± 0.062	± 0.111
0.01	0.9	0.5	1.0
0.0097 ± 0.0199	0.80 ± 0.27	0.496 ± 0.059	1.054 ± 0.195
± 0.0272	± 0.09	± 0.058	± 0.189
C) Cases with high deletion probability per fragment ($\mu t = 0.1$)			
0.1	0.5	0.1	0.1
0.0881 ± 0.0196	0.45 ± 0.05	0.100 ± 0.027	0.099 ± 0.031
± 0.0190	± 0.07	± 0.027	± 0.035
0.1	0.5	0.1	0.5
0.0969 ± 0.0285	0.48 ± 0.05	0.095 ± 0.035	0.518 ± 0.085
± 0.0219	± 0.08	± 0.028	± 0.080
0.1	0.5	0.1	1.0
0.1047 ± 0.0315	0.47 ± 0.10	0.088 ± 0.031	0.975 ± 0.165
± 0.0238	± 0.07	± 0.029	± 0.143
0.1	0.5	0.5	0.5
0.1115 ± 0.0431	0.48 ± 0.09	0.504 ± 0.114	0.506 ± 0.173
± 0.0350	± 0.10	± 0.100	± 0.157
0.1	0.5	0.5	1.0
0.1027 ± 0.0334	0.44 ± 0.14	0.502 ± 0.080	1.058 ± 0.407
± 0.0361	± 0.13	± 0.108	± 0.304
0.1	0.9	0.1	1.0
0.0936 ± 0.0420	0.88 ± 0.04	0.097 ± 0.020	0.986 ± 0.147
± 0.0437	± 0.04	± 0.023	± 0.132

Each row of the table contains average parameter estimates for a specific set of true values of μt, r, gt, and wt. True parameter values are underlined. The average parameter estimate is below the true parameter value. To the right of the average parameter estimate is the sample standard error. Directly below each sample standard error is the average estimate of the standard error obtained from the information matrix. As noted in the text, failure of the numerical maximization routine occasionally resulted in the information matrix producing negative estimates of variance. The reported average is an average among cases where the information matrix yielded positive estimates of variance

Table 2. Effect of sequence length on parameter estimation by the homogeneity procedure

	μt	r	gt	wt
Truth	0.01	0.5	0.1	1.0
Expected length, 500	0.010 ± 0.004	0.46 ± 0.20	0.097 ± 0.021	1.028 ± 0.164
Expected length, 1500	0.010 ± 0.002	0.48 ± 0.09	0.102 ± 0.013	0.959 ± 0.075

Twenty simulated pairs of sequences with expected length of 500 nucleotides were produced by an evolutionary process with no regional heterogeneity of substitution rates. These pairs of sequences were analyzed by the homogeneity procedure. The same values of μt, r, gt, and wt were used to generate 20 pairs of sequences of expected length 1500 nucleotides and these pairs of sequences were also analyzed by the homogeneity procedure. This table shows the true evolutionary parameter values, the average parameter estimates from the sequences of expected length 500 nucleotides, and the average parameter estimates from the sequences of expected length 1500 nucleotides. Sample standard errors are to the right of the average parameter estimates

Table 3. Performance of the heterogeneity procedure for pairs of sequences with regional heterogeneity of substitution rates

μt	r	gt	wt	p_f	k_f
0.01	0.9	0.05	0.05	0.5	4.0
0.011 ± 0.012	0.87 ± 0.07	0.074 ± 0.049	0.066 ± 0.027	0.44 ± 0.25	4.7 ± 2.9
± 0.015	± 0.06	± 0.037	± 0.041	± 0.27	± 5.4
0.01	0.9	0.05	0.05	0.5	8.0
0.012 ± 0.015	0.88 ± 0.09	0.065 ± 0.044	0.062 ± 0.032	0.49 ± 0.14	8.1 ± 4.7
± 0.016	± 0.07	± 0.035	± 0.036	± 0.18	± 5.4
0.01	0.9	0.05	0.05	0.8	4.0
0.012 ± 0.017	0.79 ± 0.24	0.101 ± 0.060	0.097 ± 0.069	0.63 ± 0.28	5.7 ± 9.3
± 0.015	± 0.18	± 0.059	± 0.054	± 0.44	± 22.5
0.01	0.9	0.05	0.05	0.8	8.0
0.015 ± 0.025	0.84 ± 0.15	0.092 ± 0.060	0.099 ± 0.067	0.71 ± 0.21	8.6 ± 10.3
± 0.021	± 0.08	± 0.079	± 0.088	± 0.19	± 19.7
0.01	0.9	0.05	0.2	0.5	4.0
0.011 ± 0.012	0.87 ± 0.07	0.057 ± 0.029	0.206 ± 0.073	0.46 ± 0.22	4.5 ± 1.4
± 0.016	± 0.06	± 0.026	± 0.083	± 0.23	± 2.2
0.01	0.9	0.05	0.2	0.5	8.0
0.009 ± 0.010	0.87 ± 0.06	0.043 ± 0.023	0.173 ± 0.092	0.52 ± 0.12	42.9 ± 127.4
± 0.012	± 0.06	± 0.019	± 0.066	± 0.13	± 62.9
0.01	0.9	0.05	0.2	0.8	4.0
0.012 ± 0.011	0.83 ± 0.21	0.041 ± 0.019	0.173 ± 0.093	0.77 ± 0.16	9.1 ± 13.9
± 0.013	± 0.07	± 0.030	± 0.121	± 0.22	± 13.3
0.01	0.9	0.05	0.2	0.8	8.0
0.009 ± 0.010	0.86 ± 0.09	0.048 ± 0.030	0.217 ± 0.131	0.76 ± 0.13	20.5 ± 34.5
± 0.011	± 0.08	± 0.033	± 0.123	± 0.11	± 24.9

Each row of the table contains average parameter estimates for a specific set of true values of μt, r, gt, wt, p_f, and k_f. The average parameter estimates were obtained from the simulated evolution and analysis of 20 pairs of sequences by the heterogeneity procedure. To produce a pair of sequences from particular values of the evolutionary parameters, a descendant sequence was evolved as described in the text from an ancestral sequence of expected length 500 nucleotides. True parameter values are underlined. The average parameter estimate is below the true parameter value. To the right of the average parameter estimate is the sample standard error. Directly below each sample standard error is the average estimate of the standard error obtained from the information matrix. As noted in the text, failure of the numerical maximization routine occasionally resulted in the information matrix producing negative estimates of variance. The reported average is an average among cases where the information matrix yielded positive estimates of variance

dure returned a parameter estimate that was near the boundary of the parameter space, the maximum likelihood value was sometimes less than the value returned by the homogeneity procedure.

In contrast, when estimates of parameters were in the interior of the parameter space, failure of the maximization routine was uncommon and the covariance structure estimated by the Fisher information matrix tended to be more reasonable. Insight can be assisted by examination of the sample correlations and correlations from the Fisher information matrix. For example, the correlation between μt and r is positive. Because the value of r determines the expected number of fragments within a sequence, a high value of r indicates that a sequence will have a few large fragments. If a certain number of deletions has occurred since the divergence of two sequences, then the estimate of the probability of deletion per fragment will be higher if the sequence is thought to contain a few large

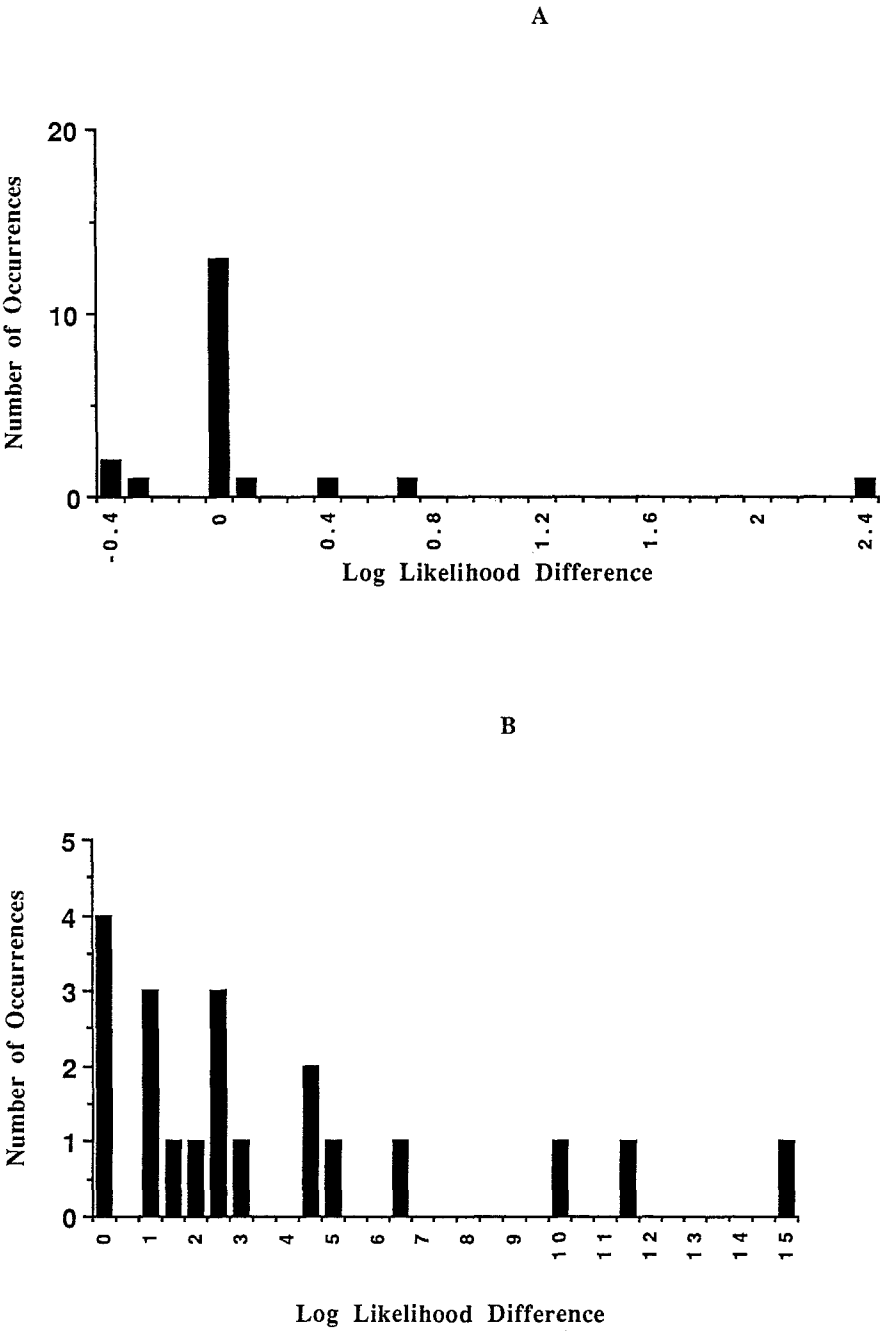

Fig. 4. The distribution of the log likelihood difference between the value of θ suggested by the heterogeneity procedure and the value of θ suggested by the homogeneity procedure is affected by whether regional heterogeneity of substitution rates actually occurs during the evolutionary process. The data shown represent the logarithm of the maximum likelihood value returned by the heterogeneity procedure minus the logarithm of the maximum likelihood value returned by analysis of the same pair of sequences by the homogeneity procedure. All negative values occurred when the heterogeneity procedure obtained parameter estimates near the boundary of the parameter space. Results were obtained from the simulated evolution and analysis of 20 pairs of sequences. To produce a pair of sequences from particular values of the evolutionary parameters, a descendant sequence was evolved from an ancestral sequence of expected length 500 nucleotides. **A** A histogram of the distribution of the log likelihood difference when there is no regional heterogeneity of substitution rates during the evolutionary process. Each pair of sequences was simulated with these evolutionary parameter values $r = 0.5$, $\mu t = 0.01$, $gt = 0.1$, and $wt = 1.0$. Each column of the histogram represents all log likelihood differences within a 0.1 log likelihood unit interval. **B** A histogram of the distribution of the log likelihood difference when there is regional heterogeneity of substitution rates during the evolutionary process. Each pair of sequences was simulated with these evolutionary parameter values $r = 0.9$, $\mu t = 0.01$, $gt = 0.05$, $wt = 0.2$, $p_f = 0.8$, and $k_f = 4.0$. Each column of the histogram represents all log likelihood differences within a 0.5 log likelihood unit interval.

fragments (i.e., a high value of r) than it would be if the sequence is thought to contain many small fragments (i.e., a low value of r).

Reassuringly, the evidence for regional heterogeneity of substitution rates was stronger when regional heterogeneity of substitution rates actually occurred than when it did not occur (Fig. 4). The ability to accurately estimate the substitution process parameters when regional heterogeneity in substitution rates occurs (p_f, k_f, g, w) is highly dependent on the value of r. This value determines the scale on which regional heterogeneity of substitution rates occurs. If the value of r is close to zero, fragments will tend to be very short. If fragments tend to be short, adjacent nucleotides are likely to belong to different fragments, and, under the fragment model, the substitution rates of adjacent fragments are independent. The power to distinguish between heterogeneity of substitution rates and no heterogeneity of substitution rates is tied to the amount of nonindependence in the substitution process between adjacent nucleotides. Therefore, small values of r yield little power to detect regional substitution rate heterogeneity. In fact, the value of r has to be quite high to detect heterogeneity according to our simulations. In sets of simulated sequence pairs where

A

```
TTTTCTGAGAATTTGATCTTGGTTCAGATTGAACGCTGGCGGCGTGGATGAGGCATGCAAGTCGAACGGA
-ATACGAAGAGTTTGATCCTGGCTCAGGATTAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAACGGA

A----------------TAATGACTTCGGTTGTTATTTAGTGGCGGAAGGGTTAGTAATACATAGATAAT
AGTTTAAGCAATTAAAC------------------TTTAGTGGCGAACGGGTGAGTAACGCGTAAGCAAT

CTGTCCTCAACTTGGGAATAACGGTTGGAAACGACCGCTAATACCGAATGTG------------------
CTGCCCCTAAGACGAGGATAACAGTTGGAAACGACTGCTAAGACTGGATAGGAGACAAGAAGGCATCTTC

-------------------------GTATGTTTAGGCATCTAAAACATATTAAAGAAGGGGATCTTCGGA
TTGTTTTTAAAAGACCTAGCAATAGGTATGCTTAGG---------------------------------

CCTTTCGGTTGAGGGAGAGTCTATGGGATATCAGCTTGTTGGTGGGGTAATGGCCTACCAAGGCTTTGAC
-------------GAGGAGCTTGCGTCACATTAGTTAGTTGGTGGGGTAAAGGCCTACCAAGACTATGAT

GTCTAGGCGGATTGAGAGATTGACCGCCAACACTGGGACTGAGACACTGCCCAGACTTCTACGGAAGGCT
GTGTAGCCGGGCTGAGAGGTTGAACGGCCACATTGGGACTGAGACACGGCCCAAACTCCTACGGGAGGCA
```

B

```
TTTTCTGAGAATTTGATCTTGGTTCAGATTGAACGCTGGCGGCGTGGATGAGGCATGCAAGTCGAACGGA
-ATACGAAGAGTTTGATCCTGGCTCAGGATTAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAACGGA
 fffffsssssssssssssssssssssssssssssssssssssssffffffsssssssssssssss

ATAATGACTTCGGTTGTTATTTAGTGGCGGAAGGGTTAGTAATACATAGATAATCTGTCCTCAACTTGGG
A--GTTTAAGCAATTAAACTTTAGTGGCGAACGGGTGAGTAACGCGTAAGCAATCTGCCCCTAAGACGAG
s  fffffffffffffffssssssssssssssssssssssssssssssssssssssssssfffffffff

AATAACGGTTGGAAACGACCGCTAATACCGAATGTGGTATGTTTAGGCATCTAAAACATATTAAAGAAGG
GATAACAGTTGGAAACGACTGCTAAGACTGGATAGGAGACAAGAAGGCATCTTCTTGTTTTTAAA----A
fsssssssssssssssssssssssssssssssssfffffffffffssssssssfffffffssssss    f

GGATCTTCGGACCTTTCGGTTGAGGGAGAGTCTATGGGATATCAGCTTGTTGGTGGGGTAATGGCCTACC
GACCTAGCAATAGGTATGCTTAGGGAGGAGCTTGCGTCACATTAGTTAGTTGGTGGGGTAAAGGCCTACC
fffffffffffffffffffffffffffffffffffffffffsssssssssssssssssssssssssssss

AAGGCTTTGACGTCTAGGCGGATTGAGAGATTGACCGCCAACACTGGGACTGAGACACTGCCCAGACTTC
AAGACTATGATGTGTAGCCGGGCTGAGAGGTTGAACGGCCACATTGGGACTGAGACACGGCCCAAACTCC
sssssssssssssssssssssssssssssssssssssssssssssssssssssssssssssssssssssss

TACGGAAGGCTGCAGTCGAGAATCTTTCGCAATGGACGAAAGTCTGACGAAGCGACGCCGCGTGTGTGAT
TACGGGAGGCAGCAGTAGGGAATTTTCGGCAATGGAGGAAACTCTGACCGAGCAACGCCGCGTGAACGAT
sssssssssssssssssssssssssssssssssssssssssssssssssssssssssssssssssssssss
```

Fig. 5. Comparison of an alignment produced under the assumption of regional homogeneity of substitution rates with an alignment produced under the assumption of regional heterogeneity of substitution rates. To produce these alignments, fragment boundaries were inferred but not shown. The specific sequences that were aligned were the 1552-bp sequence of the 16S rRNA sequence of *Chlamydia psittaci* (Weisburg et al. 1986) and the 1535-bp sequence of the 16S rRNA sequence of a mycoplasma-like organism (Lim and Sears 1989). In the case of each alignment, the top sequence is from *C. psittaci* and the bottom sequence is from the mycoplasma-like organism. Only the first 420 positions of each alignment are shown. **A** The two sequences were analyzed by the homogeneity procedure. The number of iterations required by the numerical maximization procedure was 155. The specific parameter values suggested by the homogeneity procedure and used to produce the alignment are $\mu t = 0.169 \pm 0.064$, $r = 0.947 \pm 0.016$, $gt = 0.205 \pm 0.022$, and $wt = 0.212 \pm 0.030$. The negative log likelihood of the two sequences for these parameter values is 3522.48. **B** The two sequences were analyzed by the heterogeneity procedure. The number of iterations required by the numerical maximization procedure was 340. The specific parameter values suggested by the heterogeneity procedure and used to produce the alignment are $\mu t = 0.022 \pm 0.009$, $r = 0.812 \pm 0.053$, $gt = 0.142 \pm 0.024$, $wt = 0.171 \pm 0.033$, $p_f = 0.212 \pm 0.053$, and $k_f = 10.9 \pm 4.1$. The negative log likelihood of the two sequences for these parameter values is 3499.31. To distinguish paired regions that were inferred to evolve quickly from regions that were inferred to evolve slowly, an "f" is placed under those nucleotide sites that were inferred to belong to a fast fragment, and an "s" is placed under those nucleotide sites that were inferred to belong to a slow fragment.

$r = 0.75$, results were disappointing (data not shown). A value of $r = 0.9$ led to better evolutionary parameter estimates (Table 3). If more than two sequences could be simultaneously analyzed under the fragment model, it would become easier to detect heterogeneity of substitution rates for small values of r. Given enough sequences, it is possible that individual quickly evolving nucleotide sites could be distinguished from individual slowly evolving nucleotide sites.

As demonstrated by comparing the 1552-bp sequence of the 16S rRNA sequence of *Chlamydia psittaci* (Weisburg et al. 1986) with the 1535-bp sequence of the 16S rRNA sequence of a mycoplasma-like organism (Lim and Sears 1989), the appearance of an alignment suggested by the parameter estimates of the homogeneity procedure can be quite different from the appearance of an alignment suggested by the parameter estimates of the heterogeneity procedure (Fig. 5). This striking difference arises because the two sequences are closely related throughout much of their length but they contain interior regions that are highly diverged. The homogeneity procedure yields evolutionary parameter estimates that make it more probable that the high degree of divergence of these interior regions is due to the insertion–deletion process than the substitution process; the heterogeneity procedure can more easily account for highly diverged interior regions.

5 Future Directions

An explicit model of biological sequence evolution can potentially provide a theoretical basis for the study of molecular evolution. We believe that the maximum likelihood methodology presented in this paper will prove to be a step toward such a theoretical basis even though the current model has severe limitations. A model incorporating other common evolutionary events (e.g., recombination, inversion) in addition to insertions, deletions, and substitutions would be an improvement. It would also be useful to allow local DNA context to affect the probability of evolutionary events. In addition, it must be recognized that sequence data are produced by human researchers and the decisions of these researchers affect which data are included in the data set. For this reason, terminal indels are often not a result of the evolutionary process but are, instead, an artifact of the data collection process. An objective treatment of terminal insertions and deletions based upon the data collection process would be desirable.

Another important problem is alignment of more than two sequences. Reliable information concerning the topology and the branch lengths of the phylogenetic tree can substantially improve the validity of an alignment. Hein (1990) considered concurrent inference of phylogenies and multiple-sequence alignments by the approach of maximum parsimony. The long-term goal of developing a computationally feasible method that uses a maximum likelihood framework to simultaneously infer phylogenies and multiple-sequence alignments is still distant, but it is not unrealistic.

Acknowledgments. We thank Mary K. Kuhner for her suggestions. This research and the computing facilities were supported by NSF grant BSR 8918333 and NIH grant 5R01 GM41716, to principal investigator Joseph Felsenstein. Jeff Thorne was also supported by a National Science Foundation Graduate Fellowship.

References

Allison L, Yee CN (1990) Minimum message length and the comparison of macromolecules. Bull Math Biol 52:431–453

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

Feller W (1968) An introduction to probability theory and its applications, vol 1, ed 3. McGraw-Hill, New York, pp 264–269

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

Felsenstein J (1989) PHYLIP—phylogeny inference package (version 3.2). Cladistics 5:164–166

Fitch WM, Smith TF (1983) Optimal sequence alignments. Proc Natl Acad Sci USA 80:1382–1386

Gotoh O (1982) An improved algorithm for matching biological sequences. J Mol Biol 162:705–708

Hasegawa M, Kishino H, Yano T (1985) Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

Hein J (1990) A unified approach to alignment and phylogenies. In: Doolittle RF (ed) Methods in enzymology, vol 183. Academic Press, San Diego, pp 626–645

Kendall M, Stuart A (1973) The advanced theory of statistics, vol 2, ed 3. Charles Griffen, London, pp 45–46

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Lim P-O, Sears BB (1989) 16s rRNA sequence indicates that plant-pathogenic mycoplasmalike organisms are evolutionarily distinct from animal mycoplasmas. J Bacteriol 171:5901–5906

Nelder JA, Mead R (1965) A simplex method for function minimization. Comput J 7:308–313

Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1988) Numerical recipes in C. Cambridge University Press, New York, pp 305–309

Smith TF, Waterman MS, Fitch WM (1981) Comparative biosequence metrics. J Mol Evol 18:38–46

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

Weisburg WG, Hatch TP, Woese CR (1986) Eubacterial origin of Chlamydiae. J Bacteriol 167:570–574

Received April 22, 1991/Revised and accepted August 20, 1991

Appendix

The Heterogeneity Procedure. Because the heterogeneity procedure closely resembles the homogeneity procedure, only the details unique to the heterogeneity procedure are described in this Appendix. The notation is consistent with the notation of section 3.1 unless otherwise noted.

The heterogeneity procedure classifies each alignment between A_m and B_n as a member of exactly one of seven subsets of $S(A_m, B_n)$. Five of these subsets (e.g., $S^i(A_m, B_n)$ for $i = 2, 3, \ldots, 6$) were defined in section 3.1. The remaining two subsets, $S^{1s}(A_m, B_n)$ and $S^{1f}(A_m, B_n)$, are obtained by partitioning $S^1(A_m, B_n)$. The definitions of these two subsets are:

$S^{1s}(A_m, B_n)$ = [$\alpha(A_m, B_n)$ where the rightmost link of A_m is associated with a slow fragment and survives to become the rightmost link of B_n]

$S^{1f}(A_m, B_n)$ = [$\alpha(A_m, B_n)$ where the rightmost link of A_m is associated with a fast fragment and survives to become the rightmost link of B_n]

Let $\alpha(A_m, B_n)$ be an alignment between A_m and B_n and define $L_\theta^{1s}(m, n)$ and $L_\theta^{1f}(m, n)$ as follows:

$$L_\theta^{1s}(m, n) = P[\alpha(A_m, B_n) \in S^{1s}(A_m, B_n) \mid \theta] \quad i = 1, 2, \ldots, 6$$
$$L_\theta^{1f}(m, n) = P[\alpha(A_m, B_n) \in S^{1f}(A_m, B_n) \mid \theta] \quad i = 1, 2, \ldots, 6$$

In this appendix, the use of $L_\theta^i(m, n)$ for $i = 2, 3, \ldots, 6$ is consistent with the definitions of section 3.1 but the definition of $L_\theta^1(m, n)$ will become

$$L_\theta^1(m, n) = L_\theta^{1s}(m, n) + L_\theta^{1f}(m, n)$$

The heterogeneity procedure allows the substitution rates of fast and slow fragments to differ. Let $f_{ij}^s(t)$ be the transition probability that a nucleotide associated with a slow fragment, which begins as type i, is of type j at time t. Let $f_{ij}^f(t)$ be the transition probability that a nucleotide associated with a fast fragment, which begins as type i, is of type j at time t.

From the above definitions, a recursive algorithm that calculates $L_\theta(A, B)$ can be developed. The algorithm is identical to that of section 3.1 except that the recursive equation for $L_\theta^1(m, n)$ is not used. Instead, recursive equations for $L_\theta^{1s}(m, n)$ and $L_\theta^{1f}(m, n)$ must be specified. These equations are

$$L_\theta^{1s}(m, n) = \pi_{a_m} f^s_{a_m b_n}(t)\Big[r\kappa_m\kappa_n L_\theta^{1s}(m-1, n-1) + p_s p_1(t)(1-r)\frac{\lambda}{\mu} \cdot \sum_{i=1}^{6} L_\theta^i(m-1, n-1)\Big]$$

$$L_\theta^{1f}(m, n) = \pi_{a_m} f^f_{a_m b_n}(t)\Big[r\kappa_m\kappa_n L_\theta^{1f}(m-1, n-1) + p_f p_1(t)(1-r)\frac{\lambda}{\mu} \cdot \sum_{i=1}^{6} L_\theta^i(m-1, n-1)\Big]$$

All other details, including the necessity of a numerical maximization routine, are shared by the homogeneity procedure and the heterogeneity procedure.