Maximum Likelihood Alignment of DNA Sequences

M. J. Bishop[1] **and E. A. Thompson**[2]†

[1] *University of Cambridge, Computer Laboratory*
Corn Exchange Street, Cambridge CB2 3QG, U.K.

[2] *University of Cambridge, Statistical Laboratory*
16 Mill Lane, Cambridge CB2 1SB, U.K.

(Received 16 September 1985, and in revised form 3 April 1986)

The optimal alignment problem for pairs of molecular sequences under a probabilistic model of evolutionary change is equivalent to the problem of estimating the maximum likelihood time required to transform one sequence to the other. When this time has been estimated, various alignments of high posterior probability may be written down. A simple model with two parameters is presented and a method is described by which the likelihood may be computed. Maximum likelihood estimates for some pairs of tRNA genes illustrate the method and allow us to obtain the best alignments under the model.

1. Introduction

When two molecular sequences are aligned, there exist sets of substitutions, insertions and deletions of the letters in one sequence which would transform it to the other (Fig. 1). If a weight is attached to each kind of substitution, insertion or deletion then every possible set has an associated score that is the sum of the weights of the component transformations.

```
Sequence 1   ● A A G G T A T T A G A A A A A C C A T
                 | | | |   | | | |   | | | | |     |
Sequence 2   A G A G G T G T T A G T A A A A C ● ● T
```

Figure 1. One possible alignment of 2 DNA sequences. Dots indicate gaps. Sequence 1 may be converted to sequence 2 by 2 substitutions of A → G, 1 substitution of A → T, deletion of CA and insertion of A.

Given two unaligned sequences, we may determine the set of alignments of minimum score, given the rules about assigning the weights. This problem (formulated in terms of maximum match) was first solved by Needleman & Wunsch (1970). In the simple version of the problem, the insertion and deletion events are considered to occur singly. It is also possible to consider the insertion and deletions of blocks of $(1, 2, \ldots, n)$ sequence symbols as single events (Fredman, 1984). A summary of the theory and practice of sequence comparison is provided by Sankoff & Kruskal (1983).

No stochastic model for an evolutionary process under which sequences are transformed is involved in the Needleman–Wunsch comparisons. Neither has there been any general agreement on the assessment of the significance of the alignments achieved, though one approach has been described by DeLisi & Kanehisa (1984). Zuckerkandl & Pauling (1965) suggested that comparisons of pairs of proteins could provide information about the time of divergence of the genes concerned. They did not consider insertions and deletions, but restricted their model to substitutions. The mathematical theory of molecular sequence divergence was developed by Kimura (1969), Neyman (1971), Holmquist (1972) and others. Felsenstein (1981) solved the general problem of joint estimation of divergence times from DNA sequences related by an evolutionary tree. Bishop & Friday (1985) have summarized the earlier work and have presented methods and results based on sequences from mitochondrial genomes.

We present here a maximum likelihood solution to the DNA sequence alignment problem under a stochastic model. This extends previous models under which estimation of the evolutionary process is possible, and thus provides a basis for quantitative statistical comparison of alternative alignments. Felsenstein (1982, 1983) has discussed the advantages of a likelihood approach to evolutionary inferences of this type.

2. Methods

(a) *The model*

Consider two sequences B and C, assumed to have diverged from a common ancestor A existing at some

† Present address: Department of Statistics GN22, University of Washington, Seattle, WA 98195, U.S.A.

0022-2836/86/140159–07 $03.00/0

time t in the past. Let $\pi(\mathrm{B})$ denote the overall probability of observing a sequence B, and $P_t(\mathrm{A} \rightarrow \mathrm{B})$ the transition probability from A to B over time t. Then the probability of observing sequences B and C is:

$$\Sigma_{\mathrm{A}} \pi(\mathrm{A}) P_t(\mathrm{A} \rightarrow \mathrm{B}) P_t(\mathrm{A} \rightarrow \mathrm{C}), \qquad (1)$$

assuming independent evolution in the 2 lines of descent. This probability of observed data is then the likelihood of parameters of the model providing the probabilities π, and P_t. Now, if this model assumes an underlying reversible stochastic process (Kelly, 1979, p. 15), then:

$$\pi(\mathrm{A}) P_t(\mathrm{A} \rightarrow \mathrm{B}) = \pi(\mathrm{B}) P_t(\mathrm{B} \rightarrow \mathrm{A}). \qquad (2)$$

The implication of this assumption is that, presented only with sequences A and B, there is no feature of them that can be used to distinguish which is the ancestor and which is the descendant. For sequences of genes in species that are not too divergent, this seems a reasonable assumption (see the Appendix). Then (1) becomes:

$$\pi(\mathrm{B}) \Sigma_{\mathrm{A}} P_t(\mathrm{B} \rightarrow \mathrm{A}) P_t(\mathrm{A} \rightarrow \mathrm{C}) = \pi(\mathrm{B}) P_{2t}(\mathrm{B} \rightarrow \mathrm{C}) = \pi(\mathrm{C}) P_{2t}(\mathrm{C} \rightarrow \mathrm{B}). \qquad (3)$$

So, rather than concerning ourselves with a hypothetical ancestor A, we need consider only transition from either one of our 2 current sequences (B and C) onto the other, over time period $2t$.

The model for evolution of the sequence should incorporate substitution, deletion and insertion. For stationarity the rates of insertion and deletion (per base) must be equal, and for simplicity we assume all base substitutions equiprobable. As a first approximation to a model, we consider each base separately. Suppose that substitutions occur at some rate r, and that deletion of a base and insertions of a base between any 2 existing bases are events occurring at some rate s. Then after time τ, a base will have become any one of the 3 other bases with probability (Bishop & Friday, 1985):

$$q = q(r, \tau) = (1 - \exp(-r\tau))/4,$$

and will have suffered deletion or an adjacent insertion with probability:

$$p = p(s, \tau).$$

The form of p does not matter, since for convenience we shall write $p = q/w$ and consider the estimation of q (or equivalently $r\tau$) and of w (which under our model will also depend upon τ) rather than of s or p. The model allows for the independent insertions of several bases at any location, with probabilities that are functions of powers of p.

These considerations imply a transition matrix from bases of B to bases of C (or *vice versa*) over the time period $\tau = 2t$. This matrix is given in Table 1: the term for insertions requires some explanation. Note that an insertion may be of any one of the 4 base types, and occurs between 2 bases. In order that the total probability of all possible transitions of a complete set of bases sums to 1, it is necessary to consider such an insertion as occurring with probability 1/2 before a given base and probability 1/2 after it (see the Appendix). (In addition insertions occur with the same probability after the preceding base and before the following.) Any observed insertion event, therefore, must be ascribed a probability $p/8$, so that the total between 2 bases is:

$$(\text{number of possible bases}) \times (\text{number of locations between bases}) \times (p/8) = 4 \times 2 \times (p/8) = p.$$

Table 1
Transition matrix from bases of sequence B to bases of sequence C over time period 2t

	A	C	G	T	Insert†	Delete
A	R	q	q	q	$p/8$	p
C	q	R	q	q	$p/8$	p
G	q	q	R	q	$p/8$	p
T	q	q	q	R	$p/8$	p

$R = 1 - 3q - 2p$.
† This probability applies to each of the 8 possible insert events described in the text.

Note that the total of each row in Table 1 is 1. An alternative representation of Table 1 is shown in Fig. 2; the node (n, m) denotes a location distance n bases along one sequence and m along the other. The diagonal transition (R or q) denotes the substitution of the next base in 1st sequence by the next base of the other. (If the bases are the same there has been no substitution.) The horizontal transition denotes insertion of a base into the 2nd sequence, and a vertical transition denotes loss of a base.

Let $\mathbf{X}_k$ denote the set of 4^k possible sequences of length k, and X be a given sequence of length k. Then any given base b of sequence B may:

(1) Disappear; probability p.
(2) Remain of length 1 *via* substitution or insertion + deletion; probability:

$$\Sigma_{\mathrm{X} \in \mathbf{X}_1} P(\mathrm{b} \rightarrow \mathrm{X}) = R + 3q + 2 \times 4 \times p \times p/8 = 1 - 2p + p^2.$$

(3) Become of length 2; probability:

$$\Sigma_{\mathrm{X} \in \mathbf{X}_2} P(\mathrm{b} \rightarrow \mathrm{X}) = (p/8)(8R + 24q + 16 \times 3 \times p \times p/8) = p - 2p^2 + 3p^3/4.$$

(4) Become of length 3; probability:

$$\Sigma_{\mathrm{X} \in \mathbf{X}_3} P(\mathrm{b} \rightarrow \mathrm{X}) = 3p^2/4 + \text{higher order terms}.$$

The probabilities for larger values of k may be similarly computed, but involve only powers of p of 3 or greater.

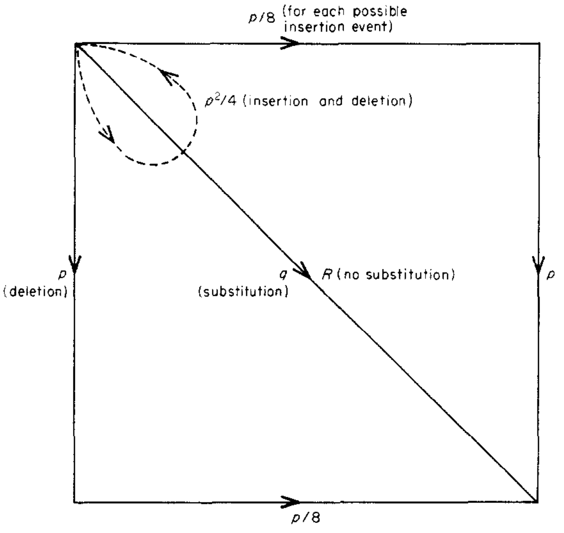
p/8 (for each possible insertion event)
p²/4 (insertion and deletion)
p (deletion)
q (substitution)
R (no substitution)
p
p/8

Figure 2. The basic cell showing the possible transformations of a single base, and the probabilities of each possible event.

The reader may note that the sum of the above probabilities is less than 1. The missing probability of order $p^2/4$ may be ascribed to the event of independent loss of a base in both sequences, or of gain + loss in transition from one to the other. This event may be represented by a loop from the node (n, m) to itself (Fig. 2), since there is no resultant base in either sequence, and this representation also allows the event to be included in the likelihood computations described below. However, since such events can never be substantiated from evidence within the 2 sequences alone, we omit it from the likelihood. Thus, with this proviso, the transition matrix defines a probability distribution for evolution of a single base. Since each base of sequence B is then treated independently in its contribution to the overall probability, the matrix does indeed define a probability distribution over all possible evolutions of a sequence (see the Appendix).

(b) *The likelihood*

In order to compare alternative values for the parameters of the model, we require the likelihood; that is, the probability of the data under the model. In effect, this is the transition probability $P(\mathrm{B} \rightarrow \mathrm{C})$, for the probabilities π do not involve features of the divergence of sequences but only their equilibrium distributions. Consider the 2 sequences laid out as shown in Fig. 3, along the sides of a lattice. Let $\mathrm{B}^{(n)}$ denote the B sequence up to the nth base, and similarly $\mathrm{C}^{(m)}$ the first m bases of sequence C. Then $P(\mathrm{B} \rightarrow \mathrm{C}) = P(\mathrm{B}^{(N)} \rightarrow \mathrm{C}^{(M)})$, the probability of transition of the total sequences of length N and M, respectively, can be computed inductively as follows.

Let $L_{n,m} = P(\mathrm{B}^{(n)} \rightarrow \mathrm{C}^{(m)})$, and consider the addition of an extra "cell" to the sequences under consideration (Fig. 2). The final base of each sequence may correspond to a substitution (denoted by a diagonal transition in the cell) or to a base being present in one sequence but not the other (corresponding to the vertical and horizontal transitions implied by insertion and deletion). Further, given that it is a substitution event that is under consideration, a substitution *has* occurred if the bases are different in the 2 sequences, and *has not* if they are the same. Thus:

$$L_{n,m} = R(1-d)L_{n-1,m-1} + qdL_{n-1,m-1} + pL_{n-1,m} + (p/8)L_{n,m-1}, \quad (4)$$

where $d = 1$ if the (n, m) bases are different and $d = 0$ if they are the same. From (4), $L_{n,m}$ may be computed for all points of the lattice, starting from $L_{0,0} = 1$, to obtain eventually $L_{N,M}$. This likelihood may be evaluated for different values of q and p (or of q and w), to obtain maximum likelihood estimates. Note that:

$$P(\mathrm{B} \rightarrow \mathrm{C}) \neq P(\mathrm{C} \rightarrow \mathrm{B})$$

since the sequences will normally have different equilibrium probabilities (eqn 2) particularly if of different length. So, if the sequences are interchanged, the likelihood values $L_{n,m}$ will be changed in a non-trivial way. However, the same maximum likelihood estimates of parameters are found.

In principle, in making these estimates no constraints on the range of parameter values are necessary. The parameters w and q are identifiable. If the 2 sequences are identical, the maximum likelihood estimate is $p = q = 0$. If they differ widely, the maximum likelihood estimate may be that they do not overlap at all, and the estimate of w becomes very small (insertions are frequent, substitutions are irrelevant). To avoid such aberrations, in practice it is convenient to limit the range of w in searching the likelihood surface. However, these constraints do not affect the estimates when sequences of similar genes in related species are considered.

Note that each "southeasterly" route through the lattice of Fig. 3, from (0, 0) to (N, M), defines a sequence of insertion, substitution and deletion events. The intermediate likelihoods $L_{n,m}$, evaluated at the maximum likelihood estimates of q and w, give the relative posterior probabilities for each particular sequence of transitions from (0, 0) to (n, m). By searching for a route of nodes (n, m) with high $L_{n,m}$, we have an interpretation of the precise sequence of events involved in the transition between the 2 sequences. In particular, backtracking from (N, M) to (0, 0) by moving to the previous node of highest $L_{n,m}$ enables us to write an alignment whose corresponding implied evolutionary events provide the best interpretation of the data.

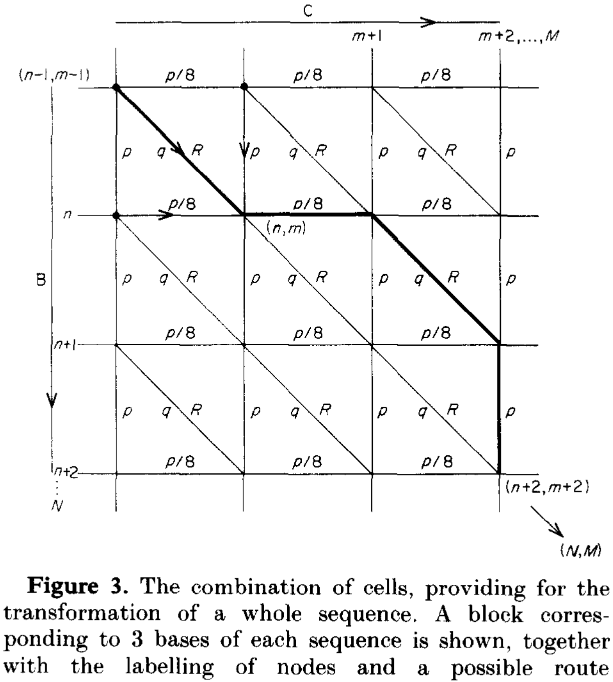

Figure 3. The combination of cells, providing for the transformation of a whole sequence. A block corresponding to 3 bases of each sequence is shown, together with the labelling of nodes and a possible route corresponding to a particular sequence of events.

(c) *Example*

As an example of the ideas of this section consider a comparison of the 2 base sequences ACT[†] and ACGC (Fig. 4). The most plausible explanation is that 1 of C or G has been substituted by T, but there are other possibilities. For example, the T could have suffered deletion, and the GC be an insertion. All the possible series of events contribute to the overall likelihood, which thus has many terms; every route through the lattice makes its contribution. When the likelihood is evaluated and maximum likelihood estimates made, we obtain $r\tau = 0{\cdot}56$ (hence $q = 0{\cdot}11$), $w = 0{\cdot}88$ (hence $p = 0{\cdot}12$), and hence $R = 0{\cdot}44$. For the parameter values, the resulting values of $L_{n,m}$ at all nodes are shown in Fig. 4. (The given values are in fact $-2 \log_e L$.) The largest values of L (smallest of $-2 \log_e L$) are approximately down the

[†] Sequence hyphens are omitted throughout for clarity.

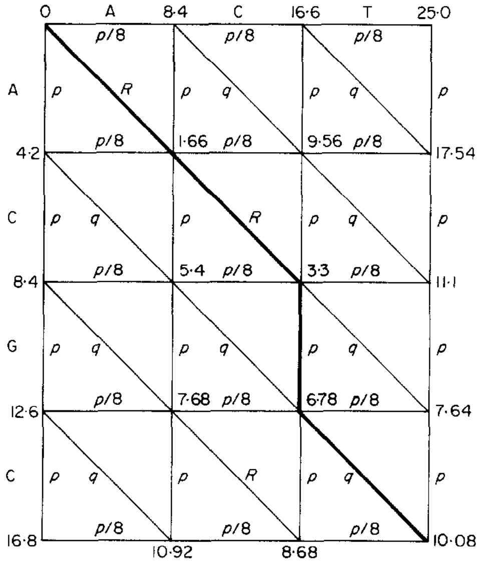

Figure 4. The example discussed in the text, of possible transformations from sequence ACGC to ACT. For this small example, the likelihood values $L_{n,m}$ can be evaluated by hand to check computer results.

Table 2

Maximum likelihood estimates of q *and* w *for tRNA genes for aspartic acid from the mitochondrial genomes of man, mouse and ox*

Sequences	$-2\log_e L$	$r\tau$	q	w	p
Man/man	0·0	0·0	0·0	$>10^3$	0·0
Man/permutation	173·0	0·91	0·15	6·2	0·024
Man/mouse	85·8	0·16	0·04	1·9	0·021
Man/ox	121·6	0·39	0·08	10·1	0·008
Mouse/ox	134·8	0·42	0·09	4·4	0·020

Permutation is a random permutation of the sequence for man with the same base composition.

diagonal, showing that indeed a single substitution and a single deletion provide the best estimate of the transition events between the 2 sequences.

3. Results

To illustrate the maximum likelihood alignment method, we chose to study the tRNA genes for aspartic acid from the mitochondrial genomes of *Homo sapiens* (man), *Mus musculus* (mouse) and

Table 3

Alignments corresponding to routes of highest likelihood values of the tRNA genes for aspartic acid from the mitochondrial genomes of man, mouse and ox

```
Man/permutation    AAGGTATTAGAAAAACCATTTCATAACTTTGTCAAAGTTAAATTATAGGCTAAAT.CCTATATATCTTA..
                   ::    : : :: ::: ::: : :   :   : ::: :  :::::      :: : : :  : ::: ::
                   AAATAGTCAAAACAACAATTACTTTGTT.G.TTAAATTCTAATTACTTTGAAA.TTCATGGACATCATAGA

Man/mouse          AAGGTATTAGAAAAACCA.TTTCATAACTTTGTCAAAGTTAAATTATAGGCTAA.A.TCCTATATATCTTA
                   ::: :::::: :::: :: :: :::::::::::::::::::::::::::   :: : ::  ::::::::::
                   AAGATATTAGTAAAATCAATTACATAACTTTGTCAAAGTTAAATTATAGATCAATAATCT.ATATATCTTA

Man/ox             AAGGTATTAGAAAAACCATTTCATAACTTTGTCAAAGTTAAATTATAGGCTAAAT.CCTATATATCTTA
                    :::: :::: :::::  ::  :::: :::::::::::::: ::: : :  :::  ::: :: : :: :
                   GAGGTGTTAGTAAAACA.TTATATAATTTTGTCAAAGTTAAGTTACAAGTGAAAGTCCTGTACACCTCA

Mouse/ox           AAGATATTAGTAAAATCAATTACATAACTTTGTCAAAGTTAAATTATAGATCAATAATCT.ATATATCTTA
                    :: : :::::::::  : : : :::: :::::::::::::: ::: :  : ::   ::   :: : :: :
                   GAGGTGTTAGTAAAAC.ATT.ATATAATTTTGTCAAAGTTAAGTTACAAGTGAAA.GTCCTGTACACCTCA
```

Permutation is a random permutation of the sequence for man with the same base composition.
Dots indicate gaps.

Table 4

Interpretation of the number of events for the alignments of Table 3 of the tRNA genes for aspartic acid from the mitochondrial genomes of man, mouse and ox

Sequences	Matches	Substitutions	Deletions	Insertions
Man/man	68	0	0	0
Man/permutation	36	29	3	3
Man/mouse	59	8	3	1
Man/ox	50	17	1	1
Mouse/ox	48	19	1	3

Permutation is a random permutation of the sequence for man with the same base composition.

Bos taurus (ox) (Anderson *et al.*, 1981, 1982*a*,*b*; Bibb *et al.*, 1981). We also generated a sequence that was a random permutation of the *H. sapiens* tRNA gene (with the same base composition).

The results of maximum likelihood estimation are shown in Table 2. The maximum likelihood is expressed in terms of $-2 \log_e L$, and estimates of q and w are given. Table 3 presents alignments corresponding to routes of highest likelihood values, and Table 4 gives the interpretation of these alignments in terms of numbers of substitution, insertion and deletion events. Note that when identical man/man sequences are compared there is zero estimated divergence time, and no evidence for insertion or deletion. All other comparisons give finite non-zero estimates, the least similar pair of sequences being man and a random permutation thereof. Also, in every case the estimate of w is greater than 1, indicating that deletion (probability q) is less frequent than any particular substitution (probability p). With the exception of the man/ox comparison, estimates of p are remarkably similar.

4. Discussion

Although the methodology of this paper parallels techniques that have been used in the parsimony comparisons of DNA sequences (compare the results of Anderson *et al.* (1982*a*) with those presented here), the basis of the analysis is very different. Here, a model for the evolution of sequences is developed, and inferences are made and parameters are estimated on the basis of the model.

The model is, of course, only a first approximation, and, on the basis only of pairs of sequences, inferences are necessarily limited. None the less, two major parameters have been estimated: the "time", $\tau/2$, of divergence in terms of substitution rate, r, of bases, and the relative rates, w, of substitutions to insertions and deletions. These are both identifiable parameters of our model and can be jointly and severally estimated. The model assumes that the same values of r and of w apply to both sequences; that is, *stochastically* the two sequences have evolved from a common ancestor under the same process. Note also that evolutionary stability requires that w is dependent on τ. Our estimate thus relates only to the two sequences concerned, and perhaps to similar sequences evolving over similar periods of time.

A major restriction of the model is that every base is assumed to evolve independently; this is clearly not the case. (Another restriction is that the substitution probability q is the same for all substitutions, but this is not essential to either the model or the method; clearly, different relevant probabilities q could be used.) Also, the restriction of considering only single bases is not essential, although it is more cumbersome to circumvent. Figure 3 shows the computation of the likelihood, each value depending on the values at the three preceding (northwesterly) nodes. In all there are 63 routes from node $(n-1, m-1)$ to node $(n+2, m+2)$. Under a more general model, each of the sets of events corresponding to each route would have a probability, no longer necessarily determined by the single-cell events. By computing the likelihood at each node, not just using the three routes from adjacent nodes but summing over 63 possible routes, one could take into account models in which three cells evolved jointly. Although cumbersome, this is feasible in principle. A first step might be to allow insertion and deletion of several bases. Rather than the deletion of three bases (the vertical route from (n, m) to $n+3, m)$) having probability of order p^3, it could be assigned a probability of order p.

APPENDIX

The Evolutionary Model and Resulting Likelihood

Various properties of the evolutionary model are used without proof in the text. For completeness, these are detailed more fully here. The results quoted for Markov chains and birth and death processes are given by Feller (1968).

(a) *Reversibility*

The property of reversibility allows simultaneous independent evolution from a common ancestor to be analyzed as evolution from one descendant sequence to the other (eqns (1), (2), and (3)). The theoretical requirement for reversibility is that:

$$\pi(\mathrm{A})P_t(\mathrm{A} \rightarrow \mathrm{B}) = \pi(\mathrm{B})P_t(\mathrm{B} \rightarrow \mathrm{A}).$$

The model for substitution, deletion and insertion proposed in the text is indeed reversible, provided an upper limit on length of sequence is imposed.

(i) *Proof*

For a given length of sequence, the substitution model provides for a positively recurrent Markov process with a single communicating class of states. A unique equilibrium distribution to the jump chain therefore exists, determined by the relative frequencies of alternative transitions but not dependent upon the overall rate. (In the case of all transitions equiprobable, this equilibrium distribution is uniform over all 4^k sequences of length k.)

The insertion/deletion process independently determines the length of sequence; the process is a birth and death process. Suppose that a sequence length n has infinitesimal rates of increase and decrease $\lambda(n)$ and $\mu(n)$ respectively. Then the probability $P_n(t)$ of sequence length n at time t satisfies:

$$\frac{\mathrm{d}P_n}{\mathrm{d}t} = -(\lambda(n)+\mu(n))P_n(t) + \lambda(n-1)P_{n-1}(t)+\mu(n+1)P_{n+1}(t), \quad (5)$$

for $n = 1, \ldots, N$ and:

$$\frac{\mathrm{d}P_0}{\mathrm{d}t} = -\lambda(0)P_0(t)+\mu P_1(t).$$

The equilibrium probabilities $\pi(n)$ thus satisfy:

$$(\lambda(n)+\mu(n))\pi(n) = \lambda(n-1)\pi(n-1) + \mu(n+1)\pi(n+1), \quad n = 1, \ldots, N-1,$$

and:

$$\lambda(0)\pi(0) = \mu(1)\pi(1)$$

which reduce to:

$$\lambda(n)\pi(n) = \mu(n+1)\pi(n+1), \quad n = 1, \ldots, N-1. \quad (6)$$

The probabilities $\pi(n)$ are unique and strictly positive, provided that for some large N, $\lambda(N) = 0$, while $\lambda(n) > 0$ for $n = 0, \ldots, N-1$, and $\mu(n) > 0$ for $n = 1, \ldots, N$. That is, no sequence can ever grow beyond N bases. This slightly artificial mathematical constraint makes sense biologically and statistically, in that there is presumably some point at which DNA sequences cannot sustain further growth and in that there is certainly a limit upon the lengths of sequence we shall choose to analyze.

Now, the infinitesimal transitions (and hence the process) are reversible if:

$$\pi(n)\lambda(n) = \pi(n+1)\mu(n+1), \quad \text{for } \textit{every } n, \quad (7)$$

but this is identical with equation (6) for the equilibrium probabilities. This concludes the proof.

(ii) *Note*

(1) The process differs from the simple birth and death process in that $\lambda(0) > 0$. That is, a sequence can be "created" by "insertion" in a sequence of length zero. Again, noting that a sequence is simply that portion of the genome chosen for analysis, this is a natural assumption.

(2) The model described in this paper implies a simple birth and death process with:

$$\lambda(n) = np,$$
$$\mu(n) = np \quad \text{for } n = 1, \ldots, N-1,$$
$$\mu(N) = Np,$$

where p is the insertion and deletion rate per base, but with:

$$\lambda(0) = p.$$

Then:

$$\pi(n) = \frac{n^{-1}}{\Sigma} \quad \text{on } n = 1, \ldots, N, \quad \pi(0) = \frac{1}{\Sigma},$$

where $\Sigma = 1 + \sum_{1}^{N} r^{-1}$.

(3) If we consider all insertions at the ends of the sequence to be a part of it, rather than one-half of them:

$$\lambda(n) = (n+1)p, \quad n = 0, \ldots, N-1$$
$$\text{and} \quad \mu(n) = np, \quad n = 0, 1, \ldots, N,$$

since there are now $(n+1)$ gaps in a sequence length n. Then:

$$\pi(n) = \frac{1}{(N+1)} \text{ on } n = 0, \ldots, N.$$

There is little difference between these distributions for a sequence of any substantial length. Other minor modifications to the model can be similarly included. Note that in all cases, $\pi(n)$ are determined by the relative values of the infinitesimal rates $\mu(n+1)$ and $\lambda(n)$. Thus provided the insertion and deletion rates per base are equal, the equilibrium distribution is uninformative about the evolutionary parameters and can be omitted from the likelihood, as asserted in the text.

(b) *The total probability of transitions*

A second set of results concerns the fact that the model covers (at least to first order) all eventualities; the total probability of all transitions from a given base (and hence any given sequence) sum to one. The length distribution at time rt is given by the solution of equation (5) with initial conditions:

$$P_1(0) = 1, \quad P_j(0) = 0, \quad \text{for } j \neq 1.$$

The sequential computation method of this paper provides an approximation to this solution. First, it is computationally necessary to assign events on a per-base basis. Thus an insertion between two bases of the sequence must be (artificially) ascribed to one or the other. This is accomplished by ascribing such events with probability 1/2 to either one of the two neighbouring bases. It is, of course, not necessary to distinguish the two possibilities. The point of the artefact is to ensure that the probability of insertion is not counted twice in summing over bases.

The computation method then gives the transition probability from a given base to any given sequence X of length k ($k > 0$) as:

$$(k+1)p(p/8)^k + (p/8)^{k-1}((k-d_{\mathrm{X}})q + d_{\mathrm{X}}R),$$

where d_{X} is the number of bases in X of the same type as the original. The first term here corresponds to the case where the original base is deleted and k insertions occur, the factor $(k+1)$ corresponding to the $(k+1)$ alternative locations of the deleted original base relative to the final sequence. In the second expression there are $(k-1)$ insertions, and a total of k alternative substitution (or non-substitution) events. Summing this over all the 4^k sequences of length k gives:

$$(k+1)p^{k+1}2^{-k}+kqp^{k-1}2^{-k+3}+ (R-q)\mathrm{p}^{k-1}8^{-k+1}\sum_{\mathrm{X}} d_{\mathrm{X}},$$

which reduces to:

$$kp^{k-1}2^{-k+1}-kp^{k}2^{-k+2}+(k+1)p^{k+1}2^{-k}. \tag{8}$$

The total over all k (including probability p for $k=0$) is:

$$\frac{(1-2p)}{(1-1/2p)^2}+2\frac{(2(1/2p)^2-(1/2p)^3)}{(1-(1/2p))^2}+p,$$

which reduces to:

$$1-\left[\frac{1/2p}{(1-1/2p)}\right]^2 \tag{9}$$

and, to first order in p (the insertion/deletion probability), the probabilities sum as required. The omitted probability arises because the sequential computation method takes no account of bases that may have been present in a hypothetical ancestor, but are no longer present in either current sequence or, more generally, of bases gained but subsequently lost in the evolution from one sequence to the other. The probability of such losses and gains is a second-order effect in the deletion probability p. While it would be desirable to incorporate this effect into the computation to reduce the level of approximation involved, there are undoubtedly other aspects of the procedure that result in greater departures from reality.

One such aspect is the assumption that all bases can be treated alike. It is the nature of a likelihood approach that evolutionary probabilities must be expressed as functions of a small number of parameters that are then estimated, by computing this probability for the observed data at alternative parameter values. It is necessary, therefore, to assume that p, q (and w) are the same at all points of the lattice. Although it would be possible to extend the parameter set, allowing substitution probabilities to be base-type dependent for example, the scope for such extensions is limited. The separate bases of the sequence provide the separate realizations of the evolutionary process upon which estimates can be based. Without an assumption of a common evolutionary process, likelihood estimation would be futile.

M.J.B. thanks the SERC for a Research Grant. A Fortran 77 program implementing the alignment method described here is available on request.

References

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H. L., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J. H., Staden, R. & Young, I. G. (1981). *Nature, (London)*, **290**, 457–465.

Anderson, S., de Bruijn, M. H. L., Coulson, A. R., Eperson, I. C., Sanger, F. & Young, I. G. (1982*a*). *J. Mol. Biol.* **156**, 683–717.

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H. L., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J. H., Staden, R. & Young, I. G. (1982*b*). In *Mitochondrial Genes* (Slonimski, P., Barst, P. & Attardi, G., eds), pp. 5–43, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Bibb, M. J., Van Etten, R. A., Wright, C. T., Walberg, M. W. & Claydon, D. A. (1981). *Cell*, **26**, 17–180.

Bishop, M. J. & Friday, A. E. (1985). *Proc. Roy. Soc. ser. B*, **226**, 271–302.

DeLisi, C. & Kanehisa, M. (1984). *Math Biosci.* **69**, 77–85.

Feller, W. (1968). *An Introduction to Probability Theory and its Applications*, vol. 1, 3rd edit., Wiley, New York.

Felsenstein, J. (1981). *J. Mol. Evol.* **17**, 368–376.

Felsenstein, J. (1982). *Quart. Rev. Biol.* **57**, 379–404.

Felsenstein, J. (1983). *J. Roy. Statist. Soc. ser. A*, **146**, 242–272.

Fredman, M. L. (1984). *Bull. Math. Biol.* **46**, 553–566.

Holmquist, R. (1972). *J. Mol. Evol.* **1**, 115–133.

Kelly, F. P. (1979). *Reversibility and Stochastic Networks*, Wiley, New York.

Kimura, M. (1969). *Proc. Nat. Acad. Sci., U.S.A.* **63**, 1181–1188.

Needleman, S. B. & Wunsch, C. D. (1970). *J. Mol. Biol.* **48**, 443–453.

Neyman, J. (1971). In *Statistical Decision Theory and Related Topics* (Gupta, S. S. & Yackel, J., eds), pp. 1–27, Academic Press, New York.

Sankoff, D. & Kruskal, J. B. (1983). Editors of *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, Addison-Wesley, Reading, MA.

Zuckerkandl, E. & Pauling, L. (1965). In *Evolving Genes and Proteins* (Bryson, V. & Vogel, H. J., eds), pp. 97–166, Academic Press, New York.

Edited by S. Brenner