J.Mol.Evol.7,133-149 (1976)

Journal of Molecular Evolution

Frequency of Insertion-Deletion, Transversion, and Transition in the Evolution of 5S Ribosomal RNA

DAVID SANKOFF, R. J. CEDERGREN, GUY LAPALME
Centre de recherches mathématiques et Département de biochimie, Université de Montréal

Received December 4, 1974; April 4, 1975

Summary. The problem of choosing an alignment of two or more nucleotide sequences is particularly difficult for nucleic acids, such as 5S ribosomal RNA, which do not code for protein and for which secondary structure is unknown. Given a set of 'costs' for the various types of replacement mutations and for base insertion or deletion, we present a dynamic programming algorithm which finds the optimal (least costly) alignment for a set of N sequences simultaneously, where each sequence is associated with one of the N tips of a given evolutionary tree. Concurrently, protosequences are constructed corresponding to the ancestral nodes of the tree. A version of this algorithm, modified to be computationally feasible, is implemented to align the sequences of 5S RNA from nine organisms. Complete sets of alignments and protosequence reconstructions are done for a large number of different configurations of mutation costs. Examination of the family of curves of total replacements inferred versus the ratio of transitions/transversions inferred, each curve corresponding to a given number of insertions-deletions inferred, provides a method for estimating relative costs and relative frequencies for these different types of mutation.

Key words: 5S rRNA - Nucleotide Sequence Homology - Evolution - Mutation Frequencies

1. EVOLUTIONARY INFERENCE AND SEQUENCE ALIGNMENT

The conservatism of certain RNAs, like tRNA and 5S ribosomal RNA, makes them attractive possibilities for the study of very early molecular evolution. This is in contrast to DNA or RNA coding for proteins, most of which tend to evolve more rapidly, and hence to lose archaic structure more rapidly. In addition, sequences of tRNA and rRNA are known

Table 1
Organisms for which 5S RNA sequences are known. *B.subtilis* has not yet been incorporated into the analysis described in the present paper

Organism	Reference
E.coli	Brownlee et al. (1968)
P.fluorescens	DuBuy & Weissman (1971)
B.stearothermophilus	Marotta et al. (1973)
B.subtilis	Rosenberg et al. (1974)
S.carlsbergensis	Hindley & Page (1972)
T.utilis	Nishikawa & Takemura (1974)
Human	Forget & Weissman (1967)
Chicken	Pace et al. (1974)
Xenopus L.	Brownlee et al. (1972)
Chlorella P.	Jordan et al. (1973)

explicitly, whereas coding sequences must generally be inferred, somewhat ambiguously, from amino acid sequences, only a few messages having been directly sequenced. Counterbalancing these advantages of conservatism and unambiguity, there are two difficulties which hinder the use of non-coding RNAs for evolutionary inference. First, though some sixty tRNAs have been sequenced, this set is partitioned among almost all of the twenty amino acids, so that no single amino acid is represented by sequences from more than a few diverse organisms. With 5S RNA, the picture is somewhat brighter, at least ten sequences now being available from the bacteria, yeast, algae, and animal species listed in Table 1, with work almost complete on sequences from other phylogenetically diverse lines. With this molecule, however, the alignment problem, which is a minor nuisance in studying protein or tRNA homology between species, becomes a major stumbling block.

To illustrate this problem, Fig.1 depicts two alignments of human and *E.coli* 5S RNA. In the first, 81 pairs of identical bases are aligned, of the 120 total in each sequence. In the second, only 70 identical pairs are aligned. Nevertheless the second alignment is the more plausible of the two because of the higher incidence of parallelism between successive aligned identical pairs. Where there is parallelism, i.e. equal numbers of unmatched bases intervening in the two sequences between successive aligned identical pairs, we may infer that one of each aligned pair of these inter-

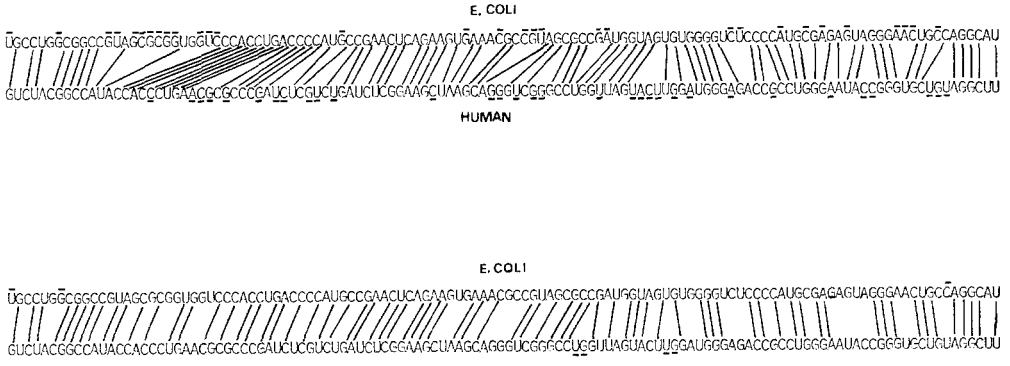

Fig.1
Two alignments of *E.coli* and human 5S RNA sequences. Lines connect identical aligned bases. Bars indicate bases which are inferred to have been deleted or inserted. Upper alignment has more matched pairs, but this is outweighed by the large number of insertions/deletions it implies, so that the lower alignment becomes the more plausible of the two

vening bases results from a base replacement mutation. Where unequal numbers of bases intervene between successive aligned identical pairs, as is frequent in the first alignment in Fig.1, we are forced to postulate either that one of the sequences has had one or more bases inserted into it, or else that the other sequence has had the same number deleted from it. Alignments implying relatively few insertions and deletions generally seem more plausible than those implying many. On the other hand, without any insertion or deletions it is usually impossible to capture all or even most of the obvious homology which may be present between the two sequences, and consequently such alignments imply very large numbers of base replacement mutations. Thus the alignment problem leads to the question of how many insertion-deletion mutations one is willing to infer, or of the relative explanatory 'cost' of insertion-deletions versus replacement mutations. In the present paper we develop a method for assessing these costs.

As we have noted, the alignment problem is more serious in 5S RNA than in nucleic acids which code for protein, since in the latter case the fact that twenty different amino acids occur, together with frame-shift restrictions on insertion-deletion possibilities, means that given any reasonably large number of homologous protein sequences, alignments can be produced rapidly just by inspection. Similarly with tRNA, the existence of modified bases occupying corresponding positions in several sequences, together with the structural constraints summarized by the cloverleaf

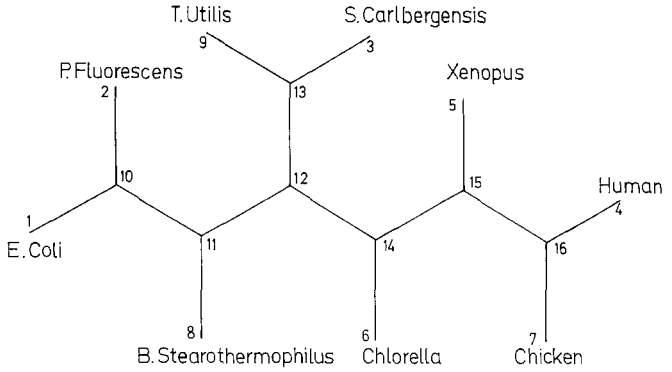

Fig.2

Phylogenetic relationships of nine organisms. Root of tree could be placed between vertices 11 and 12

model of secondary structure, makes for easy and rapid alignment. With 5S RNA, there are no amino acid correspondences to fall back on, and what structural constraints are known do not seem to carry across major phylogenetic divisions. Even among bacteria, 5S RNA sequence homology seems rather independant of secondary structural homology (Bellemare, 1974).

2. ALIGNMENT ON TREES

We have previously shown how to solve the mathematical problem of constructing optimal alignments between a pair of sequences with a given number of insertion-deletions (Sankoff, 1972; Sankoff & Sellers, 1973) and how to assess the statistical significance of such an alignment (Sankoff & Cedergren, 1973). Once there are more than two sequences to be considered, however, a new dimension is added to the alignment problem. We are no longer directly interested in all possible pairwise alignments, only those alignments of pairs of sequences from organisms which are relatively closely related phylogenetically, i.e. which are close neighbours on a phylogenetic tree. Consider, for example, the sequences which come from the organisms listed in Table 1. Perhaps the most reasonable hypothesis as to their phylogenetic relationships is as presented in Fig.2. Here, the construction of an alignment just between, say, *E.coli* and *Xenopus* sequences is of little interest, since much more information about the evolution of their 5S RNAs can be obtained by first aligning each known sequence with its immediate neighbors on the tree, constructing protosequences for the non-terminal nodes of the tree, and similarly

aligning neighboring protosequences until a complete chain of alignments on the tree is constructed between *E.coli* and *Xenopus*. Then the mutations which intervene between these two species are more likely to be the sum of those inferred from each of the intervening alignments, than those inferred from a direct alignment between their two sequences.

Thus the reconstruction of the evolutionary history of the molecule in question necessarily involves the reconstruction of protosequences. This is a much more difficult problem than the pairwise alignment of sequences. For example, in constructing the protosequence for yeast, represented by node 13, Fig.2, we should take into account not only the sequences of *T.utilis* and *S.carlsbergensis* , but also suitable information about the bacterial sequences on one hand, and the remaining eukaryotes on the other, to ensure a reasonable alignment on the third edge incident to the protoyeast node. By *edge*, we mean any line in the tree diagram directly joining two numbered nodes.

3. DYNAMIC PROGRAMMING AND THE RECONSTRUCTION OF PROTOSEQUENCES

From our discussion of the examples in Fig.1, it is clear that in constructing an alignment of two sequences, one tries to maximize the number of aligned identical pairs, and to minimize the number of insertion-deletions inferred; so that an identical alignment pair is preferable to an aligned pair which are non-identical, which is, in turn, preferable to a base which is not aligned with any base in the other sequence. This may be quantified by assigning a zero cost to a pair of identical aligned bases, various costs $c(X,Y)$ greater than zero to the alignments of different bases X and Y, and the greatest cost $c(X,O) = c(O,X)$ to unaligned bases. The idea is then to try to find an alignment which minimizes total cost. It should be clear that *cost* here means the explanatory cost of the evolutionary inference corresponding to the alignment. It has nothing to do with the population genetic cost (which is an expense in terms of "genetic" deaths required to effect the substitution of a variant allele).

Let $S_1(1),S_1(2),\ldots,S_1(n_1)$ and $S_2(1),S_2(2),\ldots,S_2(n_2)$ be the two nucleotide sequences of lengths n_1 and n_2, and let $C(i,j)$ be the minimum cost possible for alignments between the partial sequences $S_1(1),\ldots S_1(i)$ and $S_2(1),\ldots S_2(j)$. Then the dynamic programming solution for finding the optimal alignment is contained in the recursion

$$(1) \qquad C(i,j) = \min_{\substack{\delta_1 \in \{0,1\} \\ \delta_2 \in \{0,1\}}} \{C(i-\delta_1, j-\delta_2) + c(\delta_1 S_1(i), \delta_2 S_2(j))\}$$

where if $S_1(i)$, say, is the nucleotide X, then $\delta_1 S_1(i)$ is O or X depending on whether δ_1 is O or 1, respectively. The minimal cost $C(n_1,n_2)$ is found by applying this formula $n_1 n_2$ times, starting with the initial conditions $C(i,O)=c(S_1(1),O) + \ldots + c(S_1(i),O)$ and $C(O,j)=c(O,S_2(1)) + \ldots + c(O,S_2(j))$. An optimal alignment is then produced by a procedure called backtracking. The last (L-th) pair (or unpaired base) in the alignment must be $(\delta_1^{(L)} S_1(n_1), \delta_2^{(L)} S_2(n_2))$ using values of $\delta_1^{(L)}$ and $\delta_2^{(L)}$ which give the minimum in calculating $C(n_1,n_2)$ in the final application of recursion (1). The next-to-last pair (or unpaired base) in the alignment must be $(\delta_1^{(L-1)} S_1(n_1 - \delta_1^{(L)}), \delta_2^{(L-1)} S_2(n_2 - \delta_2^{(L)}))$ for some minimizing choice of $\delta_1^{(L-1)}$ and $\delta_2^{(L-1)}$ used in calculating $C(n_1 - \delta_1^{(L)}, n_2 - \delta_2^{(L)})$ by (1), and so on.

The dynamic programming approach to aligning pairs of sequences according to various criteria has been followed by Needleman & Wunsch (1970), Sankoff (1972), Sankoff & Sellers (1973), Sankoff & Cedergren (1973), Sellers (1974a,b), Wagner & Fischer (1974), Cohen et al. (1974), Haton (1974).

What of the case where not two, but N sequences must aligned, and where this alignment must be optimal in terms of a given tree? It can be proved, following Sankoff (1975), that formula (1) may be generalized to

$$(2) \qquad C(i,j,\ldots,z) = \min_{\substack{\delta_1 \in \{0,1\} \\ \delta_2 \in \{0,1\} \\ \vdots \\ \delta_N \in \{0,1\}}} \{C(i-\delta_1, j-\delta_2, \ldots, z-\delta_N) + f(\delta_1 S_1(i), \delta_2 S_2(j), \ldots, \delta_N S_N(z))\}$$

where $C(i,j,\ldots,z)$ represents the cost of the optimal alignment of the appropriate partial sequences terminating in $S_1(i), S_2(j), \ldots, S_N(z)$, respectively, and which includes only the pairwise alignment costs of just those pairs of sequences which correspond to two nodes joined by an edge of the tree. The function f depends on the cost function c and the topological shape of the tree. Starting from suitable initial conditions, it takes $n_1 n_2 \ldots n_N$ applications of formula (2) to find $C(n_1,n_2,\ldots,n_N)$. In this case, the

backtracking to find the N-tuples contributing to the minimal cost alignment produces as well the reconstructed protosequences. In other words, though in examples such as Fig.2 where N=9, the only sequence pairs whose alignments contribute to the total cost $C(n_1,n_2,\ldots,n_N)$ consist of two reconstructed sequences or one reconstructed plus one data sequence, this total cost is computed, through (2), *before* any sequences have actually been reconstructed. The algorithm aligns not only the N given data sequences, but all the protosequences to be reconstructed, as well.

The function $f(X_1,\ldots,X_N)$ can be rapidly calculated by an algorithm of Sankoff & Rousseau (1975) which generalizes the method variously described by Fitch (1971), Hartigan (1973), and Moore et al. (1973). Let A be the set of nucleotide bases A, C, G, U together with the symbol O. Let ρ denote any non-terminal node of the tree and direct all edges of the tree away from ρ. Then for any non-terminal node γ (including $\gamma=\rho$), let $\gamma\beta_1,\gamma\beta_2,\ldots,\gamma\beta_p$ be the edges incident to γ and directed away from γ. For the tree in Fig.2, p = 3 when $\gamma = \rho$, p = 2 elsewhere. Consider the recursion

$$(3) \qquad g_\gamma(Y) = \min_{\substack{Y_1 \in A \\ Y_2 \in A \\ \vdots \\ Y_p \in A}} \sum_{i=1}^{p} \{g_{\beta_i}(Y_i) + c(Y,Y_i)\}$$

with initial condition $g_\beta(Z) = 0$ or $g_\beta(Z) = \infty$ depending on whether $Z = X_r$ or $Z \neq X_r$ where β is the terminal node representing sequence S_r. Then $\min_{Y\in A} g_\rho(Y)$ is $f(X_1,\ldots,X_N)$. The backtracking part of this algorithm, itself also of the dynamic programming type, is what produces the terms of the protosequences as well as the optimal alignment of both the given sequences and the protosequences.

Unfortunately this straightforward and precise method for reconstructing protosequences and optimal alignments is computationally impractical. For example, if all N sequences to be aligned are of length n, then computer storage requirements are of the order of n^N to carry out recursion (2) for all values of the N-tuple (i,j,...,z) and execution time requirements are proportional to $(2n)^N$. For 5S RNA, since n is about 120, even N = 4 becomes unfeasible. The case N = 3, however, is feasible, and this forms the basis of a restricted version of the method which is quite practical although it may conceivably result in protosequences which are slightly less than optimal.

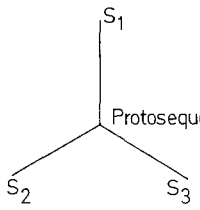

Fig.3
Simplest generalization of alignment problem to the case of three given sequences S_1, S_2 and S_3. Required to find protosequence S such that the sum of best alignment costs SS_1, SS_2, SS_3 is minimized

4. AN ITERATIVE METHOD FOR LOCAL OPTIMIZATION

Consider the N = 3 case involving one protosequence as in Fig.3. Recursion (2) becomes

(4)

$$C(i,j,k) = \min_{\substack{\delta_1 \in \{0,1\} \\ \delta_2 \in \{0,1\} \\ \delta_3 \in \{0,1\}}} \{C(i-\delta_1, j-\delta_2, k-\delta_3) + f(\delta_1 S_1(i), \delta_2 S_2(j), \delta_3 S_3(k))\}$$

We can calculate the minimum possible sum of costs for the three alignments represented by the three edges of the tree, and then find the optimal protosequence. In this case, Eq. (3) reduces to

$$(5) \qquad f(X_1,X_2,X_3) = \min_{Y \in A} \{c(X_1,Y) + c(X_2,Y) + c(X_3,Y)\}$$

and we will distinguish only four values of c; $c(X,X) = 0$ for all $X \in A$, $c(X,Y) = t$ if $X \neq Y$ but they are both purines or both pyrimidines, $c(X,Y) = v$ if one of X or Y is a purine and the other is a pyrimidine, and $c(X,0) = c(0,X) = d$ for any base X. Thus we are assigning cost d to insertion-deletions, v to transversion mutations, and t to transitions.

It is to be noted that the tree in Fig.2 may be considered to be made up of seven overlapping versions of the tree in Fig.3, each of which contains exactly one protosequence node. This forms the basis for our iterative method as illustrated in Fig.4. As a first approximation for each protosequence, we simply set it equal to one of the known sequences close to it on the tree. Then we begin the second approximation by recalculating the more peripheral of the protosequences as indicated in Fig.4a. Using the second approximations for these protosequences, we recalculate second approximations for the less peripheral protosequences indicated in Fig.4b, and so on until the most interior protosequence has been recalculated as in Fig.4d. The procedure is then reversed, i.e. the calculation in 4c is repeated, then 4b, and the third

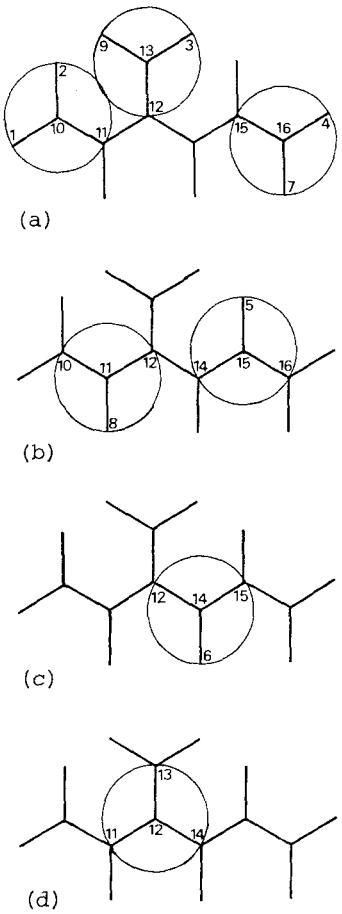

Fig.4
Decomposition of 9-sequence alignment problem to seven three-sequence cases. Algorithm performed first for the peripheral cases in (a), then (b), (c), (d), then (c), (b), (a), ... until convergence

cycle starts again with 4a. Whenever one complete cycle has been performed without any change in the individual protosequence costs, the algorithm stops. In our experience, this always occurs before the fifth cycle.

Using the protosequences thus calculated, we then align, according to recursion (1), all pairs of sequences or protosequences which correspond to an edge of the tree. This automatically produces an overall alignment of all the known and protosequences. Since the algorithm is only locally optimal, this alignment can sometimes be somewhat improved, such as through the use of recursion (3) of Section 3 applied to each set of aligned bases (or O's), one from each sequence.

5. COSTS OF MUTATION TYPES, AND THEIR FREQUENCIES

To apply our method to the 5S RNA data, it remains only to specify the costs t, v and d of the various mutation types. This, however, turns out to be a major problem. There does

Table 2
Estimated number of each mutation type in the evolution of five 5S sequences. Bottom row represents insertions; extreme right column lists deletions. Transition mutation types underlined. Input costs v=t=d=1

Original base	New base A	C	G	U	None
A		5.0	8.0	6.3	3.3
C	3.7		6.2	11.3	4.3
G	7.3	4.0		6.3	4.0
U	3.0	14.0	4.2		3.0
None	6.3	4.3	6.0	6.3	

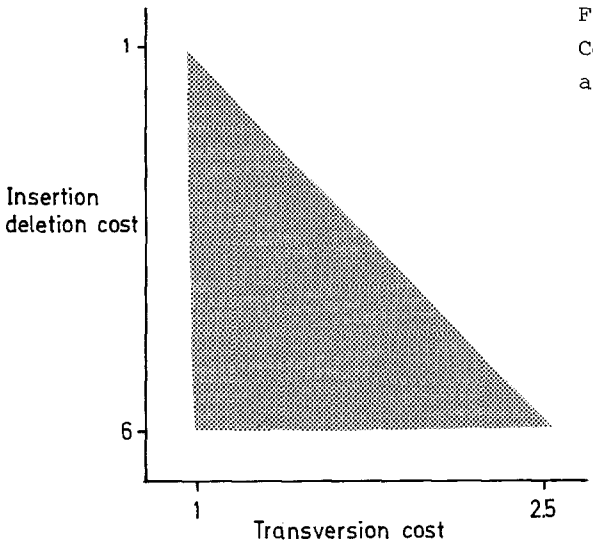

Fig.5
Cost range searched for best alignments

not seem to be any a priori way of assessing these costs which is applicable to different types of sequences, and this has been a drawback of much work in molecular evolution. For example, in our earlier alignment of only five of the sequences in Table 1 (Sankoff et al., 1973), we set $t = v = d$, but once the alignments were constructed we found as in Table 2 that more of each of the four transition types of mutation was inferred than each of the eight transversion types, or each of the eight types of base deletion or insertion. This suggests that transitions occur, or are fixed, much more easily and should 'cost' less than the others. Setting $t = 1$, the shaded area in Fig.5 shows what might be the conceivable range in which v and d covary. We carried out the computations in the previous section for more than 70 points within this area, plus a few others somewhat be-

yond it. Needless to say, different cost configurations usually resulted in different configurations of mutation type frequency, some 50 different configurations in all.

How shall we evaluate these different alignments? Consider first what we might expect if the true evolutionary history of our sequences involved D* insertion-deletions, V* transversions, and T* transitions. Then with the correct choice of costs d* and v*, we should hope to recover an alignment which implies a mutational pattern very similar to the true pattern. This follows from the general principle of most parsimonious evolutionary explanation, a principle which though not uncontroversial (Felsenstein, 1973), is equivalent to maximum likelihood in many contexts (Farris, 1973) and underlies the dynamic programming approach to molecular homology.

What should happen if we use incorrect values of d and v? Let us narrow our consideration to those incorrect values which still produce D* insertion-deletions, but $T \neq T^*$ and $V \neq V^*$ transitions and transversions. In 'spoiling' some of the correct replacement mutations we must add some incorrect ones, and the parsimony principle leads us to expect that it should in general take more incorrect ones to explain the data than it took with the correct mutations. Thus we expect that $T + V > T^* + V^*$. This is illustrated in Fig.6a where the curve of T + V versus T/V takes on a minimum of $T^* + V^*$ at T^*/V^*, for alignments having D* insertion-deletions.

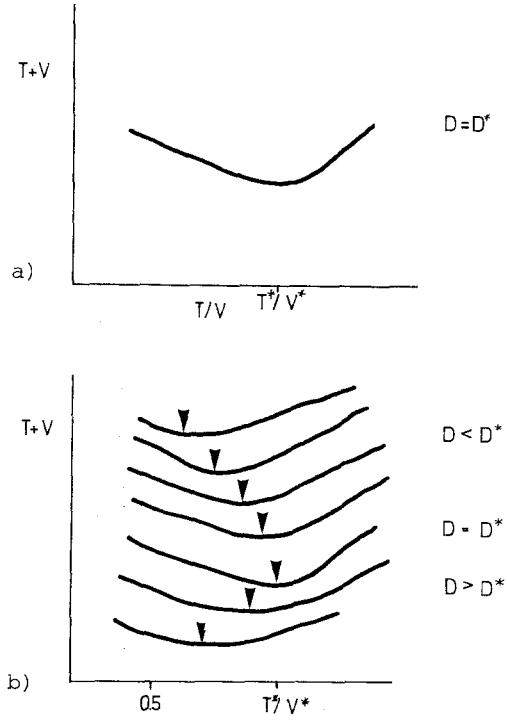

Fig.6
a) Total replacements inferred as a function of transitions-transversions ratio, when the correct number D of insertions and deletions is inferred
b) Same as a) but varying the number of insertions-deletions

Now consider those d and v producing $D \neq D^*$ insertion-deletions. Such alignments contain two classes of replacement mutations, correct ones with a transitions/transversions ratio of about T^*/V^*, and incorrect ones with a random pattern of replacements, i.e. a transitions/transversions ratio of about 1/2, since there are two transversion mutations possible for each base, and only one transition. Thus while we should still expect the same general shape for the T + V versus T/V curve for those D other than D*, we should expect the minimum to occur not at T^*/V^*, but somewhere between T^*/V^* and 1/2, depending on how close D is to D*. This expected pattern is illustrated by Fig.6b.

We cannot compare directly the expected pattern of Fig.6b with the actual alignments, since with only 50 different alignments, for no D could we place more than four or five points on its curve, and for many D we have only two, one or no corresponding value of T and V. Nevertheless, for each of v = 1, 1.02, 1.1, 1.25, 1.5, 1.75, 2, and 2.5 we had tried at least seven values of d, producing a range of values of D, T, and V. For each fixed value of v, we ordered the observed alignments according to D, and for each D not observed between the maximum D and minimum D for the group, we interpolated linearly to find corresponding values of T and V. For example, for v = 1.5, we observed eight alignments including two with D = 29, T = 88, V = 87 and D = 40, T = 83, and V = 73, and no alignments with D between 30 and 39 inclusive. To interpolate T and V corresponding to these values of D, we used the formulae

$$(6) \qquad T = 88 - \frac{D-29}{40-29} \times (88-83)$$

$$(7) \qquad V = 87 - \frac{D-29}{40-29} \times (87-73)$$

Based on these interpretations we drew Fig.7 to compare with Fig.6b.

The theoretical pattern in Fig.6b shows up clearly in the curves of Fig.7, though the minima are not always easy to locate precisely. Nevertheless, these results suggest that the best estimates for D*, T* and V* are 41, 82 and 71 respectively. Three of our assignments for v and d produced alignments with this configuration, namely v = 1.75, d = 2.25; v = 1.60, d = 2.15; v = 1.67, d = 2.25.

We present in Fig.8 an alignment of all the known sequences and protosequences with D = 41, T = 82 and V = 71, and in Table 3 a detailed breakdown of each type of mutation, analogous to the earlier results of Table 2.

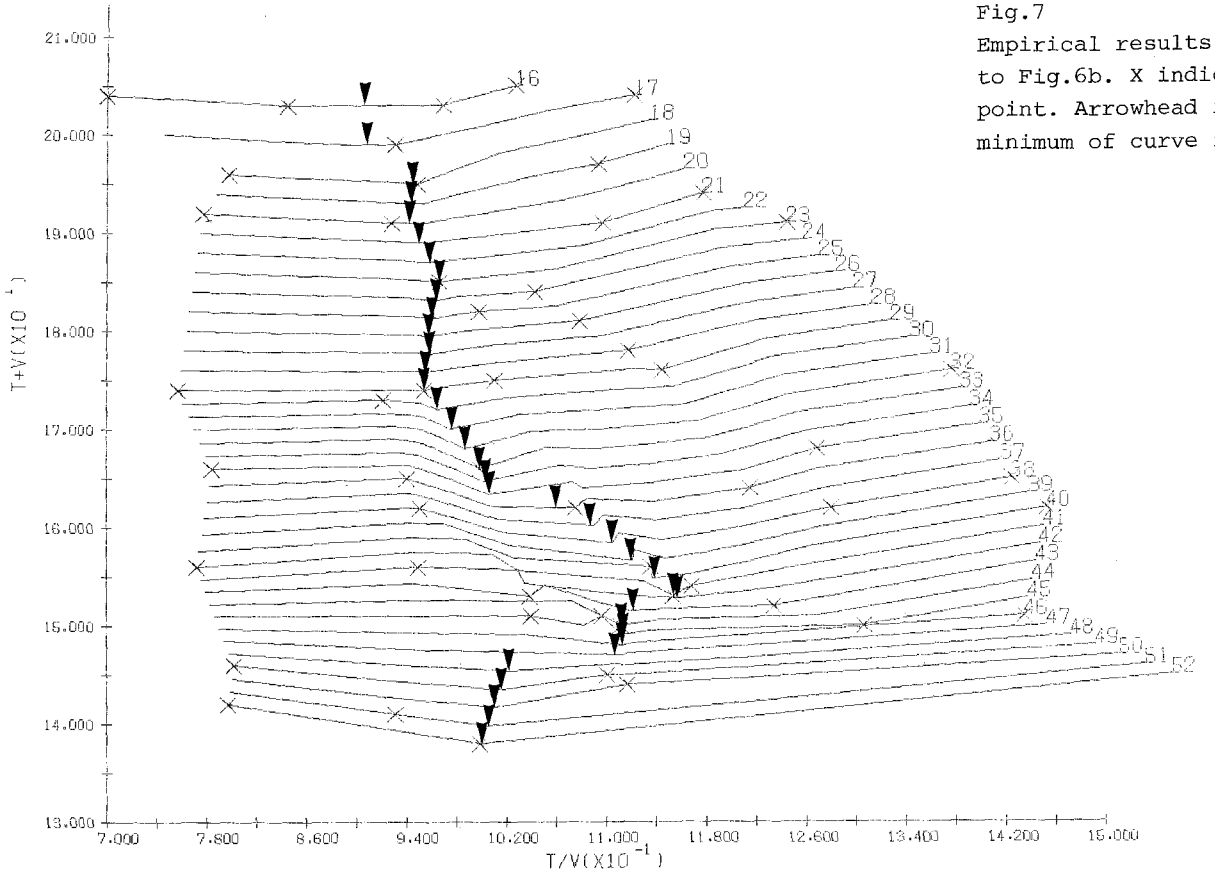

Fig.7
Empirical results corresponding to Fig.6b. X indicates a data point. Arrowhead indicates minimum of curve for each D

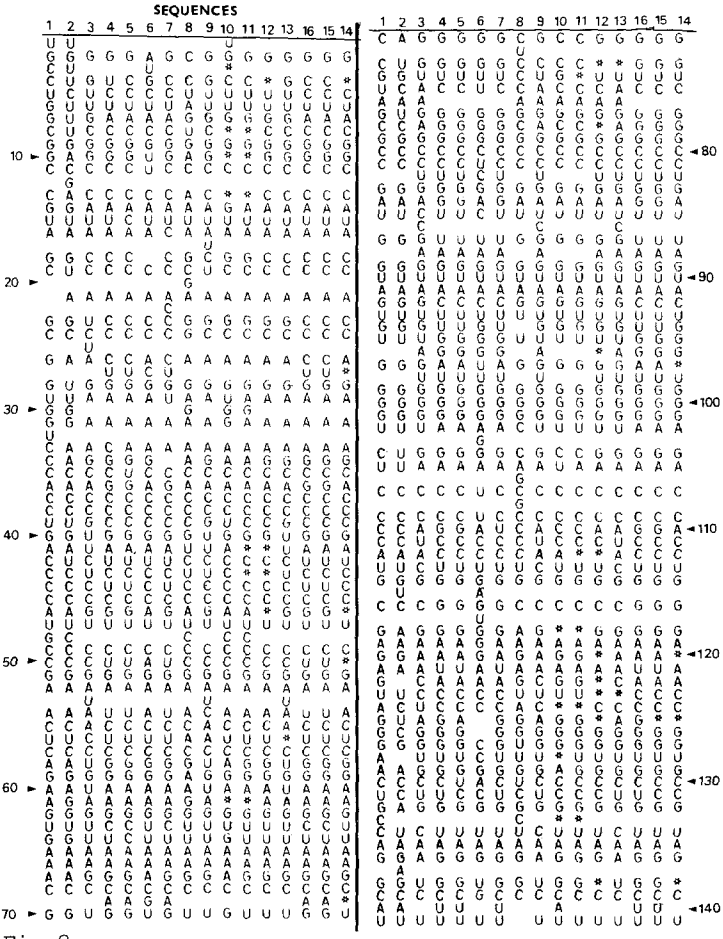

Fig.8

Alignment based on t = 1, v = 1.75, d = 2.25. * indicates uncertainty as to base in reconstructed sequence. T = 82, V = 71, D = 41

Table 3

Estimated number of each mutation type in the evolution of nine 5S sequences. Bottom row represents insertions; extreme right column lists deletions. Input costs t = 1, v = 1.75, d = 2.25. Inference of the directionality of mutation occurrences, e.g. C→A rather than A→C, requires that a root or ancestor node be designated on the tree and all edges assigned the direction which leads away from the root. Here root is placed between nodes 11 and 12 in Fig.2

Original base	New base A	C	G	U	None
A		9.8	15.0	8.3	4.3
C	9.3		8.8	29.8	1.8
G	19.0	9.8		13.5	5.3
U	5.3	18.3	6.5		3.8
None	3.8	5.3	6.3	10.8	

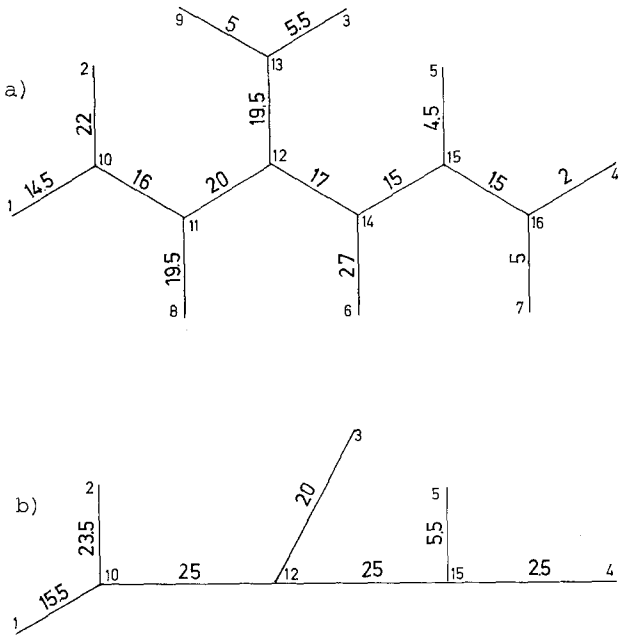

Fig.9
Evolutionary distances between nodes in phylogenetic tree
a) as estimated from nine 5S RNA sequences
b) as estimated from five sequences

6. EVOLUTIONARY DISTANCES AND THE PROTOSEQUENCES

Given the alignment in Fig.8, it becomes possible to infer the number of mutations which intervene between any two adjacent nodes on the tree in Fig.2. These estimates are summarized in Fig.9. In our earlier study (Sankoff et al., 1973), we estimated some of the same distances, also summarized in Fig.9, in an alignment based on $t = v = d = 1$, and it is of interest to compare the two sets of estimates. The estimates for those edges in the old tree (9b) which have not been interrupted in the more complete tree (9a) by any new nodes, remain relatively unchanged. This involves edges (10,1), (10,2) and (15,5), and illustrates the relative stability of the mutational distance estimates with respect to the parameters t, v, and d. Those edges which have been interrupted by a new node, i.e. (10,12), (12,3), (12,15) and (15,4) have 23% - 40% larger estimates in the new tree. This is a well known property of minimum distance or minimum cost methods (cf. Moore et al., 1973). Given that there is a 'true' underlying phylogenetic tree representing evolutionary history, such methods tend to underestimate evolutionary distances when based on only a few data points. As the amount of starting data becomes larger, the estimates tend to increase towards the true value.

What of the protosequences themselves? Let us examine the most critical case, sequence 12. Of the 120 possible bases, including 7 ambiguous or non-determined positions in the earlier construction, a total of 97 are identical in the construction presented here. There are only 5 cases where corresponding bases are different in the two sequences and only five positions in one or the other sequence which have no counterpart in the other. The remaining non-correspondences involve ambiguous cases in one sequence where the ambiguity does not appear in the other. The agreement between the two sequences is thus of the order of 80% - 90%. For protosequence 10, the agreement between the two constructions is slightly better, and for protosequence 15 it is about 99%.

7. CONCLUSIONS

It is now well established that fixed natural mutations are not random with respect to the bases involved or with respect to the locus in the gene (see Fitch & Markowitz, 1970, for the latter type of non-randomness). Much of the current evidence indicates that transitions are favoured over transversions no matter what type of data is studied - DNA, RNA, or protein (Fitch, 1967; Dayhoff, 1972; Vogel, 1972; Robertson & Jeppesen, 1972; Sankoff et al., 1973; as well as the present paper).

In the case of 5S ribosomal RNA, our new method of estimating fixation rates indicates a ratio of 1.15:1.0 for transitions over transversions. This is equivalent to 2.3 times as many of each type of transition as each type of transversion, on the average, there being twice as many transversion mutation types (8) as transition types (4). Our theory suggests that a minimal cost solution is a good estimate of the true phylogenetic tree when transversions cost 1.75 times as much as transitions, and insertions and deletions 2.25 times as much. In future work, these values may well be taken as known, sparing the extensive and costly computation necessary to estimate them.

Hopefully, the results of our quantitative approach to mutation frequency can be correlated with data on stereochemical conformation and theories of enzymatic DNA repair mechanisms.

The procedures used in this work for reconstructing protosequences and estimating evolutionary distances are seen to be internally self-consistent. Results of an earlier study of five data sequences and the present study seem to be converging, hopefully to an accurate inference of evolutionary history.

REFERENCES

Bellemare, G. (1974). Unpublished manuscript

Brownlee, G.G., Cartwright, E., McShane, T., Williamson, R. (1972). FEBS Lett. 25, 8

Brownlee, G.G., Sanger, F., Barrell, B.G.(1968). J.Mol.Biol. 34, 379

Cohen, D.N., Reichert, T.A., Wong, A.K.C. (1974). Unpublished manuscript

Dayhoff, M.O. (1972). In: Atlas of protein sequence and structure 5, p. 44. Washington, D.C. : Nat.Biomed.Res.Found.

DuBuy, B., Weissman, S.M. (1971). J.Biol.Chem. 246, 747

Farris, J.S. (1973). Syst.Zool. 22, 250

Felsenstein, J. (1973). Syst.Zool. 22, 240

Fitch, W.M. (1967). J.Mol.Biol. 26, 499

Fitch, W.M. (1971). Syst.Zool. 20, 406

Fitch, W.M., Markowitz, E. (1970). Biochem.Genet. 4, 579

Forget, B.G., Weissman, S.M. (1967). Sci. 158, 1695

Hartigan, J.A. (1973). Biometrics 29, 53

Haton, J.-P. (1974). C.R.Acad.Sci.Paris 278-A, 1527

Hindley, J., Page, S.M. (1972). FEBS Lett. 26, 157

Jordan, B.R., Galling, G., Jourdan, R. (1973). FEBS Lett. 37, 333

Marotta, C.A., Levy, C.C., Weissman, S.M., Varricchio, F. (1973). Biochem. 12, 2901

Moore, G.W., Barnabas, J., Goodman, M. (1973). J.Theoret.Biol. 38, 459

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

Nishikawa, K., Takemura, S. (1974). FEBS Lett. 40, 106

Pace, N.R., Walker, T.A., Pace, B., Erikson, R.L. (1974). J.Mol.Evol. 3, 151

Robertson, H.D., Jeppesen, P.G.N. (1972). J.Mol.Biol. 68, 419

Rosenberg, M., Pace, N.R., Sogin, M. (1974). Personal communication

Sankoff, D. (1972). Proc.Nat.Acad.Sci., U.S.A. 69, 4

Sankoff, D. (1975). SIAM J.Appl.Math. 28, 35

Sankoff, D., Cedergren, R.J. (1973). J.Mol.Biol. 77, 159

Sankoff, D., Morel, C., Cedergren, R.J. (1973). Nature New Biol. 245, 232

Sankoff, D., Rousseau, P. (1975). Math.Programming 9, 240

Sankoff, D., Sellers, P.H. (1973). Discrete Math. 4, 287

Sellers, P.H. (1974a). J.Comb.Theory 16, 253

Sellers, P.H. (1974b). SIAM J.Appl.Math. 26, 787

Vogel, F. (1972). J.Mol.Evol. 1, 334

Wagner, R.A., Fischer, M.J. (1974). J.Assn.Comp.Mach. 21, 168

David Sankoff and Guy Lapalme
Centre de recherches mathématiques
Université de Montréal
c.p. 6128 Montréal H3C 3J7
Québec, Canada

R.J. Cedergren
Département de biochimie
Université de Montréal
c.p. 6128 Montréal H3C 3J7
Québec, Canada