J. Mol. Biol. (1973) **77**, 159–164

A Test for Nucleotide Sequence Homology

DAVID SANKOFF

Centre de recherches mathématiques, Université de Montréal
c.p. 6128 Montréal 101, Québec, Canada

AND

R. J. CEDERGREN

Département de biochimie, Université de Montréal
c.p. 6128 Montréal 101, Québec, Canada

(*Received 4 December 1972*)

Two macromolecular sequences which have evolved from a common ancestor sequence will tend to include a large number of elements unaffected by replacement mutations in both sequences, as long as the evolutionary rate is not too high or the divergence time is not too great. The positions of corresponding elements may have changed in either daughter sequence due to deletion/insertion mutations involving other sequence elements, but their order can be expected to be the same in both sequences. These sets of correspondences, called matches, may be computed by a recursive algorithm which incorporates constraints on the number of deletion/insertion mutations hypothesized to have occurred. A test is developed which computes the significance of each deletion/insertion hypothesized, based on Monte-Carlo sampling of random sequences with the same base composition as the experimental sequences being tested. Applying the test to 5 S RNAs confirms the relation of *Escherichia coli* and KB carcinoma 5 S RNAs and establishes the previously undetected homology between *Pseudomonas fluorescens* and KB 5 S RNAs.

1. Introduction

When the nucleotide or amino acid sequences of functionally related macromolecules are found to be very similar, this can be adduced as evidence for the existence of a common historical antecedent†. If the sequences involved are almost identical, there can be little doubt about such an inference. On the other hand, randomly generated sequences, especially random models of nucleotide sequences based on only four symbols, will often bear a surprising degree of accidental resemblance; so that inferences of relationship when there is only a moderate degree of resemblance are dubious without some test of significance (Needleman & Wunsch, 1970; Morazain & Cedergren, 1973). In the case of nucleotide sequences, moreover, it is often difficult to choose between two or more plausible patterns of base-by-base correspondences between sequences.

† Such similarities between *different* organisms may indicate phylogenetic relationship; similarities between two macromolecules in a *single* organism may justify the hypothesis of their functional and structural differentiation from a single, more general, historical precursor.

Barker *et al.* (1969), using a scoring procedure for comparing the 120-base nucleotide sequences of the 5 S RNAs of *Escherichia coli* (Brownlee *et al.*, 1968) and KB carcinoma (Forget & Weissman, 1967), hypothesized that they could have evolved through a total of only six deletions or insertions of short subsequences and 46 base replacements, with 70 bases remaining unchanged in both organisms. They then generated eight pairs of random sequences of 120 terms, using the base proportions of *E. coli* and KB carcinoma 5 S RNAs, and found that the highest degree of resemblance in the random pairs, measured by their criterion, was considerably less than in the experimental pair.

Our purpose in this paper is to formulate the general problem of this type in statistical terms, to give a procedure for producing and applying tests of significance, and to carry out this procedure on the sequences of 5 S RNAs (Brownlee *et al.*, 1968; Forget & Weissman, 1967; DuBuy & Weissman, 1971). The basis of our method is an algorithm (Sankoff, 1972) for constructing "best matches" between two sequences under constraints on the number of "deletions/insertions" allowed. The probability distributions for the tests of significance are calculated by a Monte-Carlo method. Throughout, our discussion will be in terms of nucleotide sequences. The methods, however, are general and could be applied to other types of sequences.

2. Matches with Deletion/Insertion Constraints

In our model of mutation, we assume that there are three basic ways a nucleotide sequence can change: through base replacement, insertion of a number of consecutive bases, or deletion of a number of consecutive bases. The result of a number of such steps can change the sequence drastically, but always subject to the following constraint. Suppose base X precedes base Y (not necessarily immediately) in the original sequence, and they both remain unreplaced and undeleted in the final sequence. Then X must also occur before Y in the final sequence. This motivates a definition of a "match" between two sequences.

DEFINITION: Let $a_1, \cdots, a_m$ and $b_1, \cdots, b_n$ be two sequences of letters chosen from A, C, G and U. Consider pairs of numbers (i,j) where i can range from 1 to m and j can range from 1 to n. A subset M of these pairs is a *match* if for all (i,j) in M, we have $a_i = b_j$; and if both (i,j) and (h,k) are in M, then $i < h$ if and only if $j < k$. A *best match* is one where $P(M)$, the number of pairs in M, is as large as possible.

EXAMPLE: Let (a_1,a_2,a_3) = (A,G,C) and (b_1,b_2,b_3,b_4) = (C,A,C,U). Then the pairs (1,2) and (3,3) constitute a best match M, where $P(M) = 2$. The pairs (1,2) and (3,1), on the other hand, do not satisfy the definition of a match.

In general, the construction of a match M is motivated by the hope that the ordered pairs in M might reflect the ancestral sequence common to the two experimental sequences; so that we might, to some extent, infer the history of replacement, insertion and deletion from the nature of the gaps between successive ordered pairs in the match (see Sankoff *et al.*, 1973). In practice, *best* matches tend to imply a history of very high rates of insertion and deletion compared to replacement†, and this is not justified by what is known about the processes of macromolecular evolution. For this reason, we have developed the following approach for controlling the number of gaps in a match and assessing their significance.

† To achieve the best match size of 81 between *E. coli* and KB 5 S RNAs, we must infer at least 23 deletions or insertions, compared to the 6 realistically hypothesized by Barker *et al.* (1969).

Suppose (i,j) and (h,k) are two consecutive pairs in a match, and these pairs each reflect a base in the ancestral sequence. If $h - i = k - j$, then the same number of bases intervene between i and h in the first sequence as between j and k in the second. The non-correspondence of these intervening bases could well have arisen through base replacement in one or both evolutionary lines. If, on the other hand, $h - i > k - j$, then there must have been either an insertion in the first sequence between a_i and a_h or a deletion of some of the bases in the second sequence between b_j and b_k. This observation motivates a definition of the deletion/insertion index of a match.

DEFINITION: Let M be a match between two sequences. The deletion/insertion (DI) index of M is the number of successive pairs of pairs (i,j), (h,k) in M such that $h - i \neq k - j$.

EXAMPLE: Suppose $a_1, \cdots, a_{12}$ is depicted above $b_1, \cdots, b_{12}$ as follows:

A A A A G G G C C C A A
A A A A U U U G G G A A.

Then three different matches are:

M_1: (1,1), (2,2), (3,3), (4,4), (11,11), (12,12)
M_2: (1,1), (2,2), (3,3), (4,4), (5,8), (6,9), (7,10)
M_3: (1,1), (2,2), (3,3), (4,4), (5,8), (6,9), (7,10), (11,11), (12,12)

and

$$P(M_1) = 6, \qquad DI(M_1) = 0,$$
$$P(M_2) = 7, \qquad DI(M_2) = 1,$$
$$P(M_3) = 9, \qquad DI(M_3) = 2.$$

From now on we shall be interested in matches, such as those in the example, which contain the *largest number of pairs possible without exceeding a given DI value.* A construction of such matches (Sankoff, 1972) is based on the matrices V_q defined as follows:

For $i = 0,1, \cdots, m$; $j = 0,1, \cdots, n$; and $q = 0,1, \cdots$,

$$V_q(0,j) = V_q(i,0) = 0.$$

For $i = 1, \cdots, m$; and $j = 1, \cdots, n$,

$$V_0(i,j) = V_0(i-1,j-1) + 1 \qquad \text{if} \quad a_i = b_j,$$
$$V_0(i,j) = V_0(i-1,j-1) \qquad \text{if} \quad a_i \neq b_j,$$

and, for $q = 1,2, \cdots$,

$$V_q(i,j) = \max_{\substack{0 \leq h < i \\ 0 \leq k < j}} \{V_{q-1}(i-1,k),\ V_q(i-1,j-1),\ V_{q-1}(h,j-1)\} + 1 \qquad \text{if} \quad a_i = b_j,$$
$$V_q(i,j) = \max_{\substack{0 \leq h < i \\ 0 \leq j < k}} \{V_{q-1}(i-1,k),\ V_q(i-1,j-1),\ V_{q-1}(h,j-1)\} \qquad \text{if} \quad a_i \neq b_j.$$

It can be shown that if M is the largest match with $DI(M) \leq q$, then $P(M) = V_q(m,n)$. Once V_q is calculated, an appropriate M can be constructed by working backwards on the variables i,j and q, noting which term on the right-hand side in the above equations produces a maximum for each matrix element. The entire procedure is easily programmed in an efficient manner, though there may be heavy storage requirement if mn is large. The number of calculations required to construct best

matches satisfying $DI(M_0) = 0$, $DI(M_1) \leq 1$, $\cdots$, $DI(M_q) \leq q$, is proportional to mnq, and the largest value of q necessary (i.e. to produce an unconstrained best match) is usually much less than m or n.

3. A Test

Using our constrained best matches procedure, we can construct a test of homology which examines whether $V_q - V_{q-1}$ is statistically significantly greater for the experimental pair of sequences being considered than for randomly generated pairs†, for certain values of q. Significance in this case justifies, to some extent, the inference that the two sequences are genetically related and that their history of divergence has involved at least q deletions and/or insertions.

To understand the test in more detail, consider two hypothetical sequences of length m and n, historically related, whose evolution has included a total of q distinct deletions and insertions, and r replacements, the replacements being distributed more or less randomly along the two sequences. The best match allowing a zero DI index will (usually) include a number of successive pairs representing the largest segment uninterrupted by deletions or insertions in *both* sequences. The best match with index 1 will tend to accommodate the second largest such segment, and so on. Because of the historical relationship, each of these segments should produce more pairs of terms than we would expect to find in a similar analysis of a randomly generated pair of sequences. Thus we would expect that for each increment of one in the DI index, the historically related sequences should produce a larger increment in best match size than a pair of randomly generated sequences. This would hold up to a DI index of q. That is, V_0, $V_1 - V_0$, $\cdots$, $V_q - V_{q-1}$, should all be statistically significantly greater in the case of the genuinely related sequences than in the random case.

As time progresses, both q and r increase, introducing further dissimilarity between the sequences, adding "noise" to any analysis of homology, and reducing the significance to be expected from any test of homology. The increase in q weakens one aspect of the test, and the increase in r weakens another aspect. When q is large with respect to m and n, the qth largest segment uninterrupted by deletions or insertions in both sequences will tend to be a very short segment. This will make it very unlikely that $V_q - V_{q-1}$ is significantly different for the related pair and a random pair of sequences. This may be true for $V_{q-1} - V_{q-2}$ and the preceding few comparisons as well. When r is large, many terms in the largest segment uninterrupted by deletions or insertions in both sequences will have been replaced, reducing the contribution of this segment to V_0. When combined with the more or less random correspondences occurring elsewhere in the sequences and contributing to V_0, the effect due to historical relationship will be masked and no statistically significant difference will be found for the comparison of V_0 in the related pair from V_0 in a random pair. This may hold true for $V_1 - V_0$, and the next few comparisons as well. Thus, as time and evolution proceed, we can hope for relatively little information from either the first few or the last few terms of the sequence V_0, $V_1 - V_0$, $\cdots$, $V_{q-1} - V_{q-2}$, $V_q - V_{q-1}$. What we *can* expect, is that for some s and t, where $0 \leq s < t \leq q$, all or most of $V_{s+1} - V_s$, $V_{s+2} - V_{s+1}$, $\cdots$, $V_{t-1} - V_{t-2}$, $V_t - V_{t-1}$, will be significantly greater for the historically related pair of sequences than for random pairs of sequences.

† We use the abbreviation V_q for $V_q(m,n)$ when this is not ambiguous.

Accordingly, to test for homology between *E. coli* and KB carcinoma 5 S RNAs, and between *Pseudomonas fluorescens* and KB 5 S RNAs, the following procedure was adopted.

First, V_0, V_1, $\cdots$ was calculated for the pair of sequences being tested. Then the same was done for each of 100 pairs of sequences, with the same base composition as the experimental pair but with randomly permuted terms. This produced an estimated probability distribution for each of V_0, $V_1 - V_0$, $\cdots$. From this we obtained significance levels for the experimental values of V_0, $V_1 - V_0$, $\cdots$ and these are plotted in Figure 1. As predicted by our consideration of the effects of "noise" on

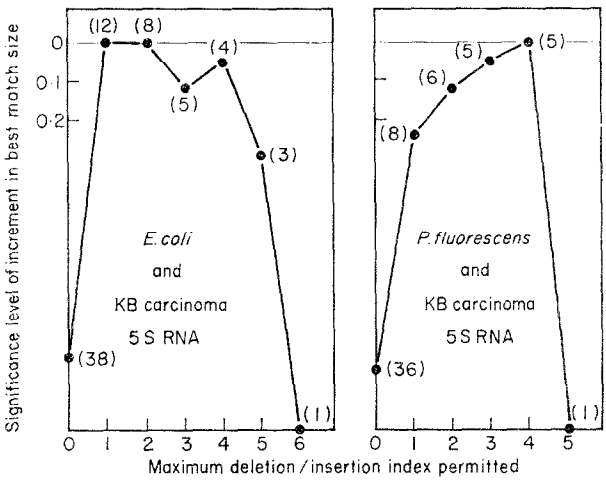

FIG. 1. Test of homology using significance level of best match size when *DI* index is allowed to increase by 1. *E. coli* and KB significant for 1st through 4th deletion/insertion. *P. fluorescens* and KB significant for 2nd through 4th deletion/insertion. Numbers in parentheses beside points indicate values of V_0 or $V_q - V_{q-1}$, as the case may be.

the test, there is little information from the first or the later comparisons. Nevertheless, there is statistical significance for a range of values of the *DI* index, clearly indicating genetic homology. As a bonus, this test indicates the number of deletions or insertions which can legitimately be inferred to have occurred historically—there may well have been more, but their presence cannot be discerned above the noise level represented by the pairs of random sequences. This property enables this test to be incorporated into procedures for the partial reconstruction of proto-sequences, work which is presently under way on 5 S RNA and other types of sequences.

An indication of the power† of our method is that although the homology of *E. coli* and KB 5 S RNAs can be inferred by the methods of Fitch (1966) or Sackin (1971), as shown by Morazain & Cedergren (1973), that of *P. fluorescens* and KB 5 S RNAs is not detectable by these methods. Another advantage of our test is that it is independent of any arbitrarily assigned scoring criteria, such as those imposed by Needleman & Wunsch (1970) or Barker *et al.* (1969).

† Even this approach is limited, of course; an attempt to directly demonstrate the remote genetic homology between human and *P. fluorescens* cytochrome *c* using this method failed to show any significance.

REFERENCES

Barker, W. C., Wallace, D. C. & Dayhoff, M. O. (1969). In *Atlas of Protein Sequence and Structure*, (Dayhoff, M. O., ed.) vol. 4, pp. 95–98. National Biomedical Research Foundation, Silver Spring, Md.

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

DuBuy, B. & Weissman, S. M. (1971). *J. Biol. Chem.* **246**, 747–761.

Fitch, W. M. (1966). *J. Mol. Biol.* **16**, 9.

Forget, B. G. & Weissman, S. M. (1967). *Science*, **158**, 1695.

Morazain, R. & Cedergren, R. J. (1973). *J. Mol. Evol.* In the press.

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

Sackin, M. J. (1971). *Biochem. Genetics*, **5**, 287–313.

Sankoff, D. (1972). *Proc. Nat. Acad. Sci., U.S.A.* **69**, 4–6.

Sankoff, D., Morel, C. & Cedergren, R. J. (1973). *Centre de recherches mathématiques*, University of Montreal, Canada. Technical report no. 269.