1974

# On the Theory of Computation and Evolutionary Distances

Peter H. Sellers
*The Rockefeller University*

SIAM J. APPL. MATH.
Vol. 26, No. 4, June 1974

# ON THE THEORY AND COMPUTATION OF EVOLUTIONARY DISTANCES*

PETER H. SELLERS†

**Abstract.** This paper gives a formal definition of the biological concept of evolutionary distance and an algorithm to compute it. For any set $S$ of finite sequences of varying lengths this distance is a real-valued function on $S \times S$, and it is shown to be a metric under conditions which are wide enough to include the biological application. The algorithm, introduced here, lends itself to computer programming and provides a method to compute evolutionary distance which is shorter than the other methods currently in use.

**1. Introduction.** The method explained in this paper for finding the distance or degree of unlikeness between any two finite sequences is particularly suited to the biological problem of finding the evolutionary distance between two DNA sequences. The simplest way to define this distance is as the smallest number of mutations and deletions by which the two sequences can be made alike. In this paper we assume that each mutation and deletion is given a *weight*, expressed by a positive real number, and that the distance between two sequences is the total of the weights of the mutations and deletions, which are chosen not only so as to make the two sequences alike, but also to have the smallest possible total weight. The first and simpler distance is equivalent to the situation in which all the weights are equal, and an algorithm for this case was first given in [1], a paper which was more concerned with the enumeration of mutations and deletions than with the distance itself.

Here we consider a new and more efficient algorithm which computes the weighted distance and, of course, the simpler distance of [1] as well. If two sequences have lengths $m$ and $n$, the algorithm calculates the distance between them in essentially $mn$ steps, where each step consists of choosing the smallest of 3 numbers.

Before coming to the algorithm, we define evolutionary sequences and distances formally. We prove that evolutionary distance satisfies the axioms of a metric, a fact which is indispensable in some applications, such as, for instance, in the construction of an evolutionary tree by the method of Margoliash and Fitch [2].

**2. A metric on evolutionary sequences.** The mathematical treatment of evolutionary sequences is simplified by writing them in the product form, $a_1a_2a_3\cdots$, where $a_1, a_2, a_3, \cdots$ are the *terms*, and by introducing a neutral element 1, which can be placed in a position of the sequence from which a term has been deleted. For example, if $a_2$ is deleted from $a_1a_2a_3$, we may write it as $a_1 1 a_3$ instead of $a_1a_3$, but both expressions represent the same evolutionary sequence. Evolutionary sequences are finite, but the representation $a_1a_2a_3\cdots$ is used on the assumption that, from a certain position on, every term is equal to 1.

DEFINITION 1. Let $M$ be an arbitrary set which contains a unique element 1, called the neutral element:

* Received by the editors April 9, 1973.

† The Rockefeller University, New York, New York 10021.

(i) $a_1a_2a_3 \cdots$ is an *M-sequence* if its terms belong to $M$ and only finitely many of them differ from 1.

(ii) Two $M$-sequences are *equivalent* if the subsequence of nonneutral terms is the same in both.

(iii) An *evolutionary sequence* $\overline{a_1a_2a_3 \cdots}$ consists of the class of all $M$-sequences equivalent to a single one, $a_1a_2a_3 \cdots$.

It follows from this definition that equivalent $M$-sequences $a_1a_2 \cdots$ and $b_1b_2 \cdots$ represent identical evolutionary sequences:

$$\overline{a_1a_2 \cdots} = \overline{b_1b_2 \cdots}.$$

Suppose $M$ is a metric space; i.e., there is a distance $d(a, b)$ between any $a$ and $b$ in $M$, which is zero if $a = b$ and a positive real number if $a \neq b$, such that $d(a, b) = d(b, a)$, and for any $a$, $b$, and $c$ in $M$

$$d(a, b) + d(b, c) \geqq d(a, c).$$

This metric can be extended to $M$-sequences by the formula

$$d(a_1a_2 \cdots, b_1b_2 \cdots) = \sum_{i=1}^{\infty} d(a_i, b_i),$$

which is well-defined, because there can only be finitely many nonzero summands on the right. This metric leads, in turn, to a metric $\bar{d}$ on evolutionary sequences if we let the distance between two equivalence classes $\overline{a_1a_2 \cdots}$ and $\overline{b_1b_2 \cdots}$ equal the distance between their nearest elements, as in the following definition.

DEFINITION 2. Let $(M, d)$ be a metric space, and let the distance between $M$-sequences be given by

$$d(a_1a_2 \cdots, b_1b_2 \cdots) = \sum_{i=1}^{\infty} d(a_i, b_i).$$

Then the *evolutionary distance* $\bar{d}(\overline{a_1a_2 \cdots}, \overline{b_1b_2 \cdots})$ between two evolutionary sequences equals

$$\min d(a_1a_2 \cdots, b_1b_2 \cdots),$$

where the minimum is taken over all $M$-sequences $a_1a_2 \cdots$ and $b_1b_2 \cdots$ in their respective equivalence classes.

It can be seen that this definition gives the weighted distance between evolutionary sequences, which was defined informally in the Introduction: suppose that $a_1a_2 \cdots$ and $b_1b_2 \cdots$ are representatives of $\overline{a_1a_2 \cdots}$ and $\overline{b_1b_2 \cdots}$, respectively, such that $d(a_1a_2 \cdots, b_1b_2 \cdots)$ is minimal; then

$$\bar{d}(\overline{a_1a_2 \cdots}, \overline{b_1b_2 \cdots}) = \sum_{i=1}^{\infty} d(a_i, b_i).$$

Each summand on the right is a positive weight except in the trivial case, when $a_i = b_i$. Assume $a_i \neq b_i$; if neither equals 1, then $d(a_i, b_i)$ is the weight of the mutation which makes $a_i$ and $b_i$ alike, whereas if $a_i = 1$, then $d(a_i, b_i)$ is the weight associated with the deletion of $b_i$, and, likewise if $b_i = 1$, then $d(a_i, b_i)$ is the weight associated with the deletion of $a_i$. The total of these weights over all $i$ is the *weighted*

*distance* mentioned in the Introduction, because the matching between $a_i$ and $b_i$ has been set up to make the total weight a minimum.

The simple unweighted distance, mentioned in the Introduction and treated in references [1] and [3], corresponds to the situation where the metric on $M$ is defined merely by

$$d(a, b) = \begin{cases} 0 & \text{for } a = b. \\ 1 & \text{for } a \neq b. \end{cases}$$

The most obvious example of an evolutionary sequence is a DNA sequence. Then,

$$M = \{a, g, c, t, 1\},$$

where $a$, $g$, $c$, $t$ are the 4 bases of which DNA is made, and 1 is the neutral symbol for a deleted base. The metric on $M$ can be completely specified by 10 positive integers $d(a, g), d(a, c), \cdots, d(t, 1)$, one for each combination of two elements of $M$.

THEOREM 1. *$\bar{d}$ is a metric on evolutionary sequences.*

*Proof.* It is easily seen that $d$ is a metric on $M$-sequences, and for evolutionary sequences, the only property of a metric which is not obvious is the triangle inequality. Therefore, let us prove that

$$\bar{d}(\overline{a_1a_2\cdots}, \overline{b_1b_2\cdots}) + \bar{d}(\overline{b_1b_2\cdots}, \overline{c_1c_2\cdots}) \geqq \bar{d}(\overline{a_1a_2\cdots}, \overline{b_1b_2\cdots}).$$

The left side equals

$$d(a_1a_2\cdots, b_1b_2\cdots) + d(b'_1b'_2\cdots, c_1c_2\cdots),$$

where the evolutionary sequences have been replaced by suitably chosen representative $M$-sequences. This choice can be modified, so that $b_1b_2\cdots$ and $b'_1b'_2\cdots$ will become alike. We proceed as follows: since $b_1b_2\cdots$ and $b'_1b'_2\cdots$ differ only by neutral terms, we can insert 1's in both sequences, so as to make them alike. Furthermore, wherever a 1 is put into $b_1b_2\cdots$, we put 1 in the same position of $a_1a_2\cdots$, and wherever a 1 is put into $b'_1b'_2\cdots$, we put 1 in the same position in $c_1c_2\cdots$. This will prevent $d(a_1a_2\cdots, b_1b_2\cdots)$ and $d(b'_1b'_2\cdots, c_1c_2\cdots)$ from changing value as $b_1b_2\cdots$ and $b'_1b'_2\cdots$ are being made alike. Therefore, assuming this procedure has been carried out, we can omit the primes from $b'_1b'_2\cdots$. Then,

$$\begin{aligned} \bar{d}(\overline{a_1\cdots}, \overline{b_1\cdots}) + \bar{d}(\overline{b_1\cdots}, \overline{c_1\cdots}) &= d(a_1\cdots, b_1\cdots) + d(b_1\cdots, c_1\cdots) \\ &\geqq d(a_1\cdots, c_1\cdots) \\ &\geqq \bar{d}(\overline{a_1\cdots}, \overline{c_1\cdots}). \end{aligned}$$

The first inequality holds because $d$ is a metric on $M$-sequences, and the second holds because $\bar{d}(\overline{a_1\cdots}, \overline{c_1\cdots})$ is the smallest distance between any two members of $\overline{a_1\cdots}$ and $\overline{c_1\cdots}$, respectively.

**3. The algorithm.** Let $a_1a_2\cdots a_m$ and $b_1b_2\cdots b_n$ denote $M$-sequences in which no term is equal to the neutral element 1, except for the unwritten ones which follow $a_m$ and $b_n$, and those unwritten ones are all neutral. Every evolutionary sequence except the trivial $\bar{1}$ has a unique representative of this form. Therefore, the evolutionary distance $\bar{d}$ is a well-defined metric on such expressions.

Formally,

$$\overline{d}(a_1 \cdots a_m, b_1 \cdots b_n) = \overline{d}(\overline{a_1 \cdots a_m}, \overline{b_1 \cdots b_n}).$$

The algorithm in the next theorem computes this number. It is based on a technique introduced by Sankoff [4] and Needleman and Wunsch [5].

(Notice that $\bar{d}$ is well-defined on all $M$-sequences, but the distance between equivalent $M$-sequences is 0. $\bar{d}$ is a *pseudometric* on $M$-sequences and a *metric* on $M$-sequences selected, as above, with no two from the same equivalence class.)

THEOREM 2. *The evolutionary distance*

$$\bar{d}(a_1 a_2 \cdots a_m, b_1 b_a \cdots b_n)$$

*is determined by mathematical induction as follows. Let* $i = 0, 1, \cdots, m$ *and* $j = 0, 1, \cdots, n$, *and interpret* $a_1 a_2 \cdots a_i$ *and* $b_1 b_2 \cdots b_j$ *as* 1 *when* $i = 0$ *and* $j = 0$, *respectively. The induction is initiated by the formulas*

$$\bar{d}(a_1 a_2 \cdots a_i, 1) = \sum_{h=1}^{i} d(a_h, 1)$$

*and*

$$\bar{d}(1, b_1 b_2 \cdots b_j) = \sum_{h=1}^{j} d(1, b_h),$$

*and the inductive step is made by giving*

$$\bar{d}(a_1 a_2 \cdots a_i, b_1 b_2 \cdots b_j)$$

*the minimum of the following three values*:

(i) $\bar{d}(a_1 a_2 \cdots a_{i-1}, b_1 b_2 \cdots b_j) + d(a_i, 1)$,

(ii) $\bar{d}(a_1 a_2 \cdots a_{i-1}, b_1 b_2 \cdots b_{j-1}) + d(a_i, b_j)$,

(iii) $\bar{d}(a_1 a_2 \cdots a_i, b_1 b_2 \cdots b_{j-1}) + d(1, b_j)$.

*Proof.* The formulas for $\bar{d}(a_1 a_2 \cdots a_i, 1)$ and $\bar{d}(1, b_1 b_2 \cdots b_j)$ conform with Definition 2 except for minimizing, which is not necessary because in these cases there is only one value in the set to be minimized.

The sequences $a_1 a_2 \cdots a_i$ and $b_1 b_2 \cdots b_j$ can be elongated by the insertion of the neutral element 1 in certain positions to give new sequences $a'_1 a'_2 \cdots a'_k$ and $b'_1 b'_2 \cdots b'_k$, respectively, such that

$$\begin{aligned}\bar{d}(a_1 a_2 \cdots a_i, b_1 b_2 \cdots b_j) &= d(a'_1 a'_2 \cdots a'_k, b'_1 b'_2 \cdots b'_k) \\ &= d(a'_1 a'_2 \cdots a'_{k-1}, b'_1 b'_2 \cdots b'_{k-1}) + d(a'_k, b'_k) \\ &= \bar{d}(a'_1 a'_2 \cdots a'_{k-1}, b'_1 b'_2 \cdots b'_{k-1}) + d(a'_k, b'_k).\end{aligned}$$

The last equality holds because if $d(\cdots a'_{k-1}, \cdots b'_{k-1})$ were not minimal, then $d(\cdots a'_k, \cdots b'_k)$ would not be.

It is clear that the elongations can be performed so that $a'_k$ and $b'_k$ are not both equal to 1. Therefore, the last expression above equals

(i) $\bar{d}(a_1 \cdots a_{i-1}, b_1 \cdots b_j) + d(a_i, 1)$ if $a'_k \neq 1$ and $b'_k = 1$,

or

(ii) $\bar{d}(a_1 \cdots a_{i-1}, b_1 \cdots b_{j-1}) + d(a_i, b_j)$ if $a'_k \neq 1$ and $b'_k \neq 1$,

(iii) $\bar{d}(a_1 \cdots a_i, b_1 \cdots b_{j-1}) + d(1, b_j)$ if $a'_k = 1$ and $b'_k \neq 1$.

One of these must give the correct value, and it must be less than or equal to each of the other two values. Therefore, if we make the inductive assumption that the value of $\bar{d}$ is known whenever one or both of the sequences $a_1 \cdots a_i$ and $b_1 \cdots b_j$ is replaced by a partial sequence, then the above three values are known. Their minimum must equal $\bar{d}(a_1 \cdots a_i, b_1 \cdots b_j)$, and this proves the theorem.

The above proof does more than give us the value of the evolutionary distance between two known evolutionary sequences. It gives us an algorithm by which we can insert neutral elements into the two known sequences, $a_1 \cdots a_m$ and $b_1 \cdots b_n$, so that they become $a'_1 \cdots a'_k$ and $b'_1 \cdots b'_k$, and

$$\bar{d}(a_1 \cdots a_m, b_1 \cdots b_n) = \sum_{h=1}^{k} d(a'_h, b'_h).$$

The process may be described as finding a *best matching* between $a_1 \cdots a_m$ and $b_1 \cdots b_n$. It is accomplished in $(m + 1)(n + 1)$ steps, but is not unique, because each step may require an arbitrary choice: suppose best matches are already known for the following three pairs:

(i) $a_1 \cdots a_{i-1}$, $b_1 \cdots b_j$,

(ii) $a_1 \cdots a_{i-1}$, $b_1 \cdots b_{j-1}$,

(iii) $a_1 \cdots a_i$, $b_1 \cdots b_{i-1}$.

Then a best match for the pair $a_1 \cdots a_i$, $b_1 \cdots b_j$ is derivable from one of the three, and possible alternative best matches are derivable from the other two. This depends on whether 1, 2, or 3 of the expressions

(i) $\bar{d}(a_1 \cdots a_{i-1}, b_1 \cdots b_j) + d(a_i, 1)$,

(ii) $\bar{d}(a_1 \cdots a_{i-1}, b_1 \cdots b_{j-1}) + d(a_i, b_j)$,

(iii) $\bar{d}(a_1 \cdots a_i, b_1 \cdots b_{j-1}) + d(1, b_j)$,

assume the minimum value—which suggests that if we wished to list all the best matches between $a_1 \cdots a_m$ and $b_1 \cdots b_n$, we would have to repeat our algorithm $3^{mn}$ times, or less. In practice, we use the algorithm once to find the $(m + 1)(n + 1)$ numerical values of $\bar{d}(a_1 \cdots a_i, b_1 \cdots b_j)$ for $i = 0, 1, \cdots, m$ and $j = 0, 1, \cdots, n$, and then by inspection of these values we may find precise limits for the range of all best matches. In collaboration with W. Einar Gall of the Rockefeller University, I have written a computer program which carries out this process.

*Example.* Find $\bar{d}(acbba, abca)$ on the assumption that $d$ is given by

$$1 = d(1, a) = d(1, b) = d(1, c) = d(a, b) = d(a, c) = d(b, c).$$

First, let us solve the problem by inspection. It has been shown that, when $d$ has the same value for all distinct pairs, the value of $\bar{d}$ is the number of mutations and deletions necessary to make the sequences alike. If we delete $c$ from the first sequence and change $c$ to $b$ in the second sequence, then both become *abba*. Hence, the distance is 2; it is obviously not 0 or 1.

Now let us apply Theorem 2: our two sequences are of the form $a_1 \cdots a_5$ and $b_1 \cdots b_4$, and we shall find $\bar{d}(a_1 \cdots a_i, b_1 \cdots b_j)$ for all of its 30 possible cases. Construct a $6 \times 5$ matrix whose first column contains the values

$$\bar{d}(1, 1), \bar{d}(a, 1), \bar{d}(ac, 1), \bar{d}(acb, 1), \bar{d}(acbb, 1), \bar{d}(acbba, 1)$$

of the distances between partial sequences of *acbba* and 1. The values are given by the initial step of the induction.

| | 1 | *a* | *b* | *c* | *a* |
|---|---|---|---|---|---|
| 1 | 0 | 1 | 2 | 3 | 4 |
| *a* | 1 | 0 | 1 | 2 | 3 |
| *c* | 2 | 1 | 1 | 1 | 2 |
| *b* | 3 | 2 | 1 | 2 | 2 |
| *b* | 4 | 3 | 2 | 2 | 3 |
| *a* | 5 | 4 | 3 | 3 | 2 |

Likewise, the values in the first row of the matrix are given by the initial step of the induction. The inductive step fills the remaining $5 \times 4$ submatrix, in which entry $i, j$ is the distance between the $i$th partial sequence of *acbba* and the $j$th partial sequence of *abca*. The value of $\bar{d}(acbba, abca)$ appears in the lower right corner.

Now let us use this matrix to find a best matching between the two sequences. We start in the lower right corner, which says $\bar{d}(acbba, abca) = 2$. This was derived by taking the minimum of the values

$$\bar{d}(acbba, abc) + d(1, a) = 3 + 1,$$

$$\bar{d}(acbb, abc) + d(a, a) = 2 + 0,$$

$$\bar{d}(acbb, abca) + d(a, 1) = 3 + 1,$$

where the $\bar{d}$ values come from the matrix positions which adjoin the lower right position. The middle expression is the smallest, and therefore,

$$\bar{d}(acbba, abca) = \bar{d}(acbb, abc) + d(a, a).$$

Now we expand $\bar{d}(acbb, abc)$ by the same argument, and the right side becomes

$$d(acb, ab) + d(b, c) + d(a, a).$$

Continuing to move back through the matrix this way, we eventually get

$$d(a, a) + d(c, 1) + d(b, b) + d(b, c) + d(a, a).$$

Therefore,

$$\bar{d}(acbba, abca) = d(acbba, a1bca).$$

This describes what we have called a best matching, and in this example it happens to be the only one, because at each step in the construction we had only one choice.

## REFERENCES

[1] P. H. SELLERS, *An algorithm for the distance between two sequences*, J. Combinatorial Theory, to appear.

[2] W. M. FITCH AND E. MARGOLIASH, *Construction of philogenetic trees*, Science, 155 (1967), pp. 279–284.

[3] S. M. ULAM, *Some combinatorial problems studied experimentally on computing machines*, Applications of Number Theory to Numerical Analysis, S. K. Zaremba, ed., Academic Press, New York, 1972.

[4] D. SANKOFF, *Matching sequences under deletion insertion constraints*, Proc. Nat. Acad. Sci. U.S.A., 69 (1972), pp. 4–6.

[5] S. B. NEEDLEMAN AND C. D. WUNSCH, *A general method applicable to the search for similarities in the amino acid sequence of two proteins*, J. Molecular Biology, 48 (1969), pp. 443–453.