JOURNAL OF ALGORITHMS 1, 359–373 (1980)

The Theory and Computation of Evolutionary Distances: Pattern Recognition

PETER H. SELLERS

Department of Mathematics, Rockefeller University, New York, New York 10021

Received November 25, 1979; revised April 18, 1980

A method of finding pattern similarities between two sequences is given. Two portions, one from each sequence, are similar if they are close in the metric space of evolutionary distances. The method allows a complete list to be made of all pairs of intervals, one from each of two given sequences, such that each pair displays a maximum local degree of similarity, and, if the lengths of the given sequences are m and n, then the procedure takes on the order of mn computational steps. This result lends itself to finding similarities by computer between pairs of biological sequences, such as proteins and nucleic acids.

1. INTRODUCTION

The principal object of this paper is to introduce an algorithm for finding each pair of regions, one in a given finite sequence A and the other in a given finite sequence B, such that the two regions most resemble each other. A region will consist of consecutive terms of a sequence, and the degree of resemblance between two such regions will be the smallness of the *evolutionary distance* between them. This distance is a metric, which was introduced by Ulam [4] and is discussed in detail in Refs. [1–3, 5].

An algorithm for computing evolutionary distance is well known, having been discovered independently in two contexts: one of them [2, 3] as a means of estimating the ancestral relationship between genetic sequences, and the other [1] as a means of estimating the relatedness between strings of symbols which differ by random errors of transcription. Both are cases of finite sequences, subject to mutations, insertions, and deletions of terms. The algorithm, when applied to sequences of lengths m and n, takes on the order of mn steps to compute the distance between them. It is reviewed in Section 2. Then, in the remainder of the paper it is shown that it takes essentially the same computation time mn to recognize all pairs of intervals, one from each sequence, which most resemble each other, in a sense still to be made precise.

0196-6774/80/040359-15$02.00/0

2. Evolutionary Distance

It is shown in Ref. [3] that a set of finite sequences can be made into a metric space in the following manner:

(i) Each finite sequence A corresponds to the class $\bar{A}$ of all infinite sequences, each of which contains A as a subsequence and has the neutral element 1 in every other position.

(ii) The terms of every sequence A, as well as the neutral element 1, belong to a given metric space, in which the distance between any two elements, a and b, is denoted by $d(a, b)$.

(iii) The *evolutionary distance* $d(A, B)$ between any two sequences is defined as

$$\min\left\{\sum_{i=1}^{\infty} d(a_i, b_i)\right\},$$

in which the minimum is taken over every pair of sequences

$$(a_1, a_2, \ldots), (b_1, b_2, \ldots)$$

in $\bar{A}$ and $\bar{B}$, respectively.

The idea underlying the above definition is to insert neutral elements in sequences A and B, so as to bring like terms into alignment and to have terms which are unmatched aligned with neutral elements. We achieve this by considering all alignments and choosing those in which the term-by-term distances add up to a minimum. Abstractly, an alignment is a sequence of pairs, each being a term of A or a neutral element followed by a term of B or a neutral element. If the terms of A and B are numbered consecutively, then an alignment is expressible as a sequence of number pairs, a neutral element being represented by a repetition of the number of the term before it. There is no need to align two neutral elements. Therefore, in going from one number pair to the next, one number or both always increase by 1. Accordingly, the known facts [1–3] about evolutionary distances can be expressed in convenient mathematical terms by regarding an alignment as a path of adjoining positions in a matrix whose row and column numbers increase monotonically as the path is followed.

For the purposes of this paper a *path* in a matrix is a monotone sequence of matrix positions, in which each consecutive pair is of one of the three forms

$$(i-1, j), (i, j),$$
$$(i-1, j-1), (i, j),$$
$$(i, j-1), (i, j).$$

When two sequences $a_1 \dots a_m$ and $b_1 \dots b_n$ are under consideration, there are three known values, $d(a_i, 1)$, $d(a_i, b_j)$, and $d(1, b_j)$, associated respectively with the three two-term paths listed above. Accordingly, we can define the *value associated with a path* as the sum of the values associated with all its successive pairs. Then the value of $d(a_1 \dots a_m, b_1 \dots b_n)$ can be characterized simply as the smallest value associated with any path from $(0,0)$ to (m,n). Also in this terminology it is easy to state and prove the known algorithm [1–3] for calculating the distance $d(a_1 \dots a_m, b_1 \dots b_n)$ in mn steps:

Construct the matrix $(d(a_1 \dots a_i, b_1 \dots b_j))$, with (i,j) ranging from $(0,0)$ to (m,n), as follows: If $i = 0$ or $j = 0$, which means $a_1 \dots a_i = 1$ or $b_1 \dots b_j = 1$, where 1 is the empty sequence, then the matrix values are

$$d(1,1) = 0.$$

$$d(a_1 \dots a_i, 1) = \sum_{h=1}^{i} d(a_h, 1).$$

$$d(1, b_1 \dots b_j) = \sum_{h=1}^{j} d(1, b_h).$$

To determine $d(a_1 \dots a_i, b_1 \dots b_j)$ in general, consider each path from $(0,0)$ to (i,j). It must pass through $(i-1,j)$, $(i-1,j-1)$, or $(i,j-1)$. Therefore, if we make the inductive assumption that we know the smallest values associated with any paths from $(0,0)$ to each of these three intermediate positions, then we know that the smallest value $d(a_1 \dots a_i, b_1 \dots b_j)$ associated with any path from $(0,0)$ to (i,j) is equal to the minimum of the three values

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

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

$$d(a_1 \dots a_i, b_1 \dots b_{j-1}) + d(1, b_j).$$

Therefore, to calculate $d(a_1 \dots a_m, b_1 \dots b_n)$, which is the value in the lower right corner of the matrix, we have to construct the whole matrix, which takes mn steps, each of which involves choosing the smallest of three values.

3. Finding Pattern A in Sequence B

Consider two finite sequences, A and B, expressed explicitly by $a_1 a_2 \dots a_m$ and $b_1 b_2 \dots b_n$, respectively, and think of A as being much shorter than B.

Let the notation $I \subset B$ mean that I is an interval in B of the form

$$b_p b_{p+1} \cdots b_q,$$

where $1 \leq p \leq q \leq n$. One way to find every interval $I \subset B$ for which the evolutionary distance $d(A, I)$ is an absolute minimum would be to calculate the distance between A and each one of the $n(n-1)/2$ intervals in B. Since a distance calculation takes on the order of mn steps, the above procedure would take on the order of mn^3 steps.

Two preliminary theorems are given in this section. Theorem 1 replaces the above procedure by an algorithm which takes mn steps, and Theorem 2 solves the same problem under the more general condition that $d(A, I)$ be a local minimum.

DEFINITION 1. Let d be the evolutionary distance; then sequence A *most resembles* $I \subset B$, if

$$d(A, I) \leq d(A, J)$$

for all $J \subset B$.

THEOREM 1. *The sequences* $a_1 a_2 \cdots a_m$ *and* $b_1 b_2 \cdots b_n$ *are given. The following procedure determines each interval* $b_p b_{p+1} \cdots b_q$ *such that* $a_1 a_2 \cdots a_m$ *most resembles* $b_p b_{p+1} \cdots b_q$:

(i) *Construct an* $(m+1) \times (n+1)$ *matrix* $(e(i,j))$ *by an induction, in which the initial values are*

$$e(0,j) = 0,$$

$$e(i,0) = \sum_{h=1}^{i} d(a_h, 1)$$

for $i = 1, 2, \ldots, m$ *and* $j = 0, 1, \ldots, n$, *and the remaining values are determined by making* $e(i,j)$ *equal to the smallest of the three values*

$$e(i-1,j) + d(a_i, 1).$$

$$e(i-1,j-1) + d(a_i, b_j),$$

$$e(i,j-1) + d(1, b_j).$$

(ii) *In the last row of the matrix find each position* q *at which there is a minimum entry, that is, where*

$$e(m,q) = \min_h e(m,h);$$

then b_q *is the last term of one of the desired intervals.*

(iii) *In the first row of the matrix find each position* p, *such that* $e(0,p-1)$ *is connected by a sequence of inductive steps to* $e(m,q)$. (*The inductive steps, referred to here, are those characterized in step* (i).) *Then* $b_p b_{p+1} \ldots b_q$ *is one of the desired intervals, all of which may be found in this manner.*

Proof. Let (m,h) be any position in the last row of the matrix constructed in part (i) of the theorem. Using the path terminology, we can state that $e(m,h)$ is the smallest value associated with any path starting at a position in the first row and ending at the position (m,h). This statement can be proved by induction. The fact that

$$e(0,h) = 0$$

for all h is explained by including within the definition of *path* any path which starts and ends at the same position and giving it the value zero. Next, assume that $e(i-1,j)$, $e(i-1,j-1)$, and $e(i,j-1)$ are the smallest values associated with any paths starting in the first row and ending at positions $(i-1,j)$, $(i-1,j-1)$, and $(i,j-1)$, respectively. Then, since every path to (i,j) passes through one of the three positions, just given, the smallest value associated with any path from the first row to (i,j) must be the minimum of the three values,

$$e(i-1,j) + d(a_i,1),$$
$$e(i-1,j-1) + d(a_i,b_j),$$
$$e(i,j-1) + d(1,b_j),$$

which agrees with the inductive step used in the construction of $e(i,j)$ in the theorem.

It follows that $e(m,q)$, as defined in part (ii) of the theorem, is the smallest value associated with any path from the first row to the last row of the matrix. If one such path is chosen, as was done in part (iii) of the theorem, its initial position being denoted by $(0,p-1)$, then the value $e(m,q)$, associated with it, is bound to be the smallest value associated with any path from $(0,p-1)$ to (m,q), which in turn is equal to

$$d(a_1 \ldots a_m, b_p \ldots b_q),$$

as observed in Section 2. In other words, if $b_p \ldots b_q$ is written as I, then $d(A,I)$ is the smallest value associated with any path from the first to the last row of the matrix. For any other interval $J \subset B$, the distance $d(A,J)$ is equal to the value associated with some path from the first to the last row

of the matrix. Therefore,

$$d(A,I) \leq d(A,J),$$

which means that A most resembles $I \subset B$. This proves the theorem.

DEFINITION 2. Let d be the evolutionary distance; then sequence A *most resembles* $I \subset B$ *locally*, if

$$d(A,I) \leq d(A,H)$$

and

$$d(A,I) \leq d(A,J)$$

for all H and J, where $H \subset I \subset J \subset B$.

In the above definition the use of the word *locally* is justified by the fact that I belongs to a family $\mathfrak{N}$ of neighboring intervals such that A most resembles I in the global sense of Definition 1 relative to $\mathfrak{N}$. More precisely,

$$d(A,I) \leq d(A,J)$$

for all $J \in \mathfrak{N}$. The exact set $\mathfrak{N}$ over which Definition 2 guarantees the above inequality can be characterized as follows: Recall that $d(A,I)$ is the smallest value associated with a family of paths. Hence, there will be one or more paths which actually take that value. If such a path is described (for the moment) as an I-path, then $J \in \mathfrak{N}$ if, and only if, there is a J-path which intersects an I-path. This fact will emerge during the proof of the next theorem.

THEOREM 2. *The sequences* $a_1a_2 \ldots a_m$ *and* $b_1b_2 \ldots b_n$ *are given. The following procedure determines each interval* $b_pb_{p+1} \ldots b_q$ *such that* $a_1a_2 \ldots a_m$ *most resembles* $b_pb_{p+1} \ldots b_q$ *locally*:

(i) *Construct the matrix* $(e(i,j))$ *as in Theorem* 1. *Here* (i,j) *ranges from* $(0,0)$ *to* (m,n).

(ii) *Construct the matrix* $(f(i,j))$, *where* (i,j) *ranges from* $(1,1)$ *to* $(m+1, n+1)$ *by an induction in which the initial values are*

$$f(i,n+1) = \sum_{h=0}^{m-i} d(a_{m-h}, 1),$$

$$f(m+1,j) = 0$$

for $i = 1,2,\ldots,m$ *and* $j = 1,2,\ldots,n+1$, *and the remaining values are*

determined by making $f(i,j)$ equal to the smallest of the three values

$$f(i+1,j) + d(a_i, 1),$$
$$f(i+1,j+1) + d(a_i, b_j),$$
$$f(i,j+1) + d(1, b_j).$$

(iii) *Select a monotone sequence*

$$(0,p-1),\ldots,(m,q)$$

of positions in the e-matrix and the corresponding sequence (all numbers 1 greater)

$$(1,p),\ldots,(m+1,q+1)$$

of positions in the f-matrix, such that each value except the first in

$$e(0,p-1),\ldots,e(m,q)$$

is determined by an inductive step from its immediate predecessor and each value except the last in

$$f(1,p),\ldots,f(m+1,q+1)$$

is determined by an inductive step from its immediate successor. Then $b_p b_{p+1} \ldots b_q$ is one of the desired intervals, all of which may be found in this manner.

Before proving the above theorem, let us consider the following example: Find all intervals in *abcacac*, which *bcab* most resembles, first in the absolute sense, then in the local sense. Let the set

$$\{1,a,b,c\}$$

be a metric space, determined by

$$d(x,y) = 0, \qquad \text{if } x = y,$$
$$= 1, \qquad \text{if } x \neq y.$$

The matrices $(e(i,j))$ and $(f(i,j))$ are shown in Fig. 1. The line segments in these matrices mark the pairs of matrix positions which are joined by inductive steps. In the first matrix every path from the first row to the last (made up of segments, going straight down or down at an angle or to the right) corresponds to a sequence of inductive steps by which the value at the end of the path is determined. In the last row there are two positions in which the minimum value 1 is taken, and there are two paths leading to

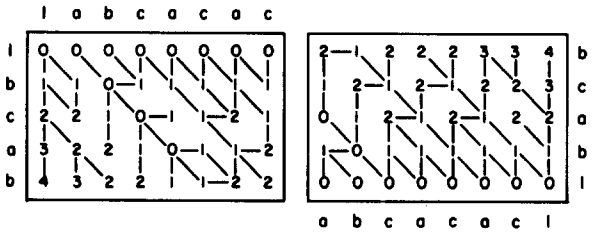

FIGURE 1

these positions. It follows from Theorem 1 that the two intervals *bca* and *bcac* in the sequence *abcacac* are those which the sequence *bcab* most resembles.

The intervals in *abcacac* which *bcac* most resembles locally are found, according to Theorem 2, by selecting those paths from the first to the last row which are common to both matrices. The intervals corresponding to such paths are *ab*, *bca*, *bcac*, *aca*, *acac*, *ca* (the second time *ca* appears), and *cac* (the second time *cac* appears). The values at the ends of the paths tell us that

$$d(bca, bcab) = d(bcac, bcab) = 1$$

and that

$$d(ab, bcab) = d(aca, bcab) = d(acac, bcab) = d(ca, bcab) = d(cac, bcab) = 2.$$

Even though the last five distances are not absolute minima, they are local minima in the sense that no widening or narrowing of the intervals in question will cause their distances from *bcab* to decrease. Note, however, that if the intervals *ca* and *cac* had been taken from *abcacac* at their first appearance, rather than second, then each one could have been widened by putting *b* at the left end, so as to decrease its distance from *bcab*. Taken this way, $d(ca, bcab)$ and $d(cac, bcab)$ are not local minima.

Proof of Theorem 2. Let I be an interval $b_p b_{p+1} \ldots b_q$ which has been constructed according to the theorem. It must be shown for any $J \subset I$ that

$$d(A, I) \leq d(A, J)$$

and for any H, such that $I \subset H$, that

$$d(A, I) \leq d(A, H).$$

Let us prove the second case, since both are proved in the same way.

The matrices $(e(i,j))$ and $(f(i,j))$ are both constructed by filling the squares shown in Fig. 2. The e-values are filled inductively, starting in the

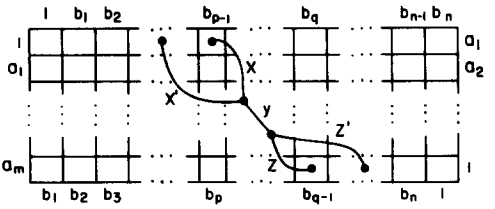

FIGURE 2

first row and column, and the f-values are filled, starting in the last row and column. The path drawn from the column headed by b_{p-1} to the column headed by b_q symbolizes the sequence of adjoining matrix positions, by which I was defined, and the other path, starting further to the left and ending further to the right, symbolizes a sequence of adjoining matrix positions, which are determined by the inductive procedure for the distance between $a_1a_2 \ldots a_m$ and any interval H, which contains I. The two paths must cross. Each path is made up of segments, which go from one matrix position to an adjoining one, and each segment has a number associated with it, which is the absolute value of the difference between the values in the two matrix positions, joined by the segment. This number is the increment corresponding to an inductive step and is the same, whether we are speaking of the e-matrix, the f-matrix, or the distance matrix. The parts of each path in the diagram are given numbers, x, x', y, z, and z', which are the totals of the numbers, belonging to the segments in them. Therefore, if $a_1a_2 \ldots a_m$ is denoted by A, we have

$$d(A,I) = x + y + z$$

and

$$d(A,H) = x' + y + z'.$$

The position at the lower end of the part marked x would have the value x in it in the e-matrix. Of all paths, which start in the top row and end at this position, x is the smallest number which can be associated with the path. Therefore, $x \leq x'$. Likewise, by the nature of the f-matrix we have $z \leq z'$. Therefore,

$$d(A,I) \leq d(A,H).$$

Similarly, for $J \subset I$

$$d(A,I) \leq d(A,J).$$

Therefore, A most resembles $I \subset B$ locally.

It remains to be shown that every $I \subset B$ which A most resembles may be found by the selection procedure defined in part (iii) of the theorem.

Suppose A most resembles $b_p \dots b_q$ locally. Then, if the interval is lengthened or shortened at the right end to produce $b_p \dots b_s$,

$$d(A, b_p \dots b_q) \leq d(A, b_p \dots b_s).$$

This holds for any s, which implies that

$$d(A, b_p \dots b_q) = e(m, q).$$

Likewise,

$$d(A, b_p \dots b_q) = f(1, p).$$

But, $d(A, b_p \dots b_q)$ is the value associated with some path in the e-matrix from $(0, p-1)$ to (m, q). This path conforms with the selection described in part (iii) of the theorem.

4. Finding Patterns Common to A and B

A and B are finite sequences. The results of the last section can be adapted to the problem of finding intervals, $I \subset A$ and $J \subset B$, such that I most resembles J locally. The first step in this direction is to redefine the e-matrix and the f-matrix.

Let $A = a_1 a_2 \dots a_m$ and $B = b_1 b_2 \dots b_n$. The $(m+1) \times (n+1)$ matrices $(e(i,j))$ and $(f(i,j))$ are defined by the same inductive steps as before, but now all the initial values are zero.

$$e(i,j) = 0, \qquad \text{whenever } i = 0 \text{ or } j = 0,$$

$$f(i,j) = 0, \qquad \text{whenever } i = m+1 \text{ or } j = n+1.$$

These matrices have the following meaning in terms of evolutionary distances: $e(i,j)$ is the shortest evolutionary distance between $a_1 a_2 \dots a_i$ and a partial sequence of $b_1 b_2 \dots b_j$ or between $b_1 b_2 \dots b_j$ and a partial sequence of $a_1 a_2 \dots a_i$, whichever is less. That is,

$$e(i,j) = \min\{d(a_g \dots a_i, b_h \dots b_j) : g = 1 \text{ or } h = 1\},$$

and, correspondingly,

$$f(i,j) = \min\{d(a_i \dots a_g, b_j \dots b_h) : g = m \text{ or } h = n\}.$$

Now let us consider how these two functions will enter into the definition of a "local resemblance" between two intervals, $a_p \ldots a_r$ and $b_q \ldots b_s$ in A and B, respectively. If both intervals are extended to the left until a_1 or b_1 is reached, then the smallest possible distance

$$d(a_g \ldots a_{p-1}, b_h \ldots b_{q-1}),$$

between the left extensions, where $g = 1$ or $h = 1$, is equal to $e(p-1, q-1)$, and the corresponding distance between their right extensions is $f(r+1, s+1)$. Accordingly, a pair of intervals has three distances associated with it, the *evolutionary distance*

$$d(a_p \ldots a_r, b_q \ldots b_s)$$

and two *complementary distances* $e(p-1, q-1)$ and $f(r+1, s+1)$. In the forthcoming definition of two intervals which "most resemble each other locally" a necessary condition will be that, if one or both intervals are changed at the left end, such that the complementary distance at that end remains fixed, then the evolutionary distance between them will increase or remain the same. The same is true if one or both intervals are changed at the right end. To be more explicit,

$$d(a_p \ldots a_r, b_q \ldots b_s) \leq d(a_i \ldots a_r, b_j \ldots b_s), \tag{1}$$

provided i and j are chosen so that

$$e(i-1, j-1) = e(p-1, q-1), \tag{2}$$

and

$$d(a_p \ldots a_r, b_q \ldots b_s) \leq d(a_p \ldots a_i, b_q \ldots b_j) \tag{3}$$

provided i and j are chosen so that

$$f(i+1, j+1) = f(r+1, s+1). \tag{4}$$

The above condition is not sufficient for our definition of local resemblance, because the values of the complementary distances belong to a discrete set, so that there may be no way or very few ways to vary two intervals and have the complementary distances take exactly the same values. Therefore, it becomes necessary to add corrective terms on the right side of inequalities (1) and (3), which will vanish when eqs. (2) and (4) hold. Suppose we wish to change the left ends of the intervals, but there are no values, i and j, such that

$$e(p-1, q-1) = e(i-1, j-1).$$

Then, inequality (1) can be extended as follows:

$$d(a_p \ldots a_r, b_q \ldots b_s) \leq d(a_i \ldots a_r, b_j \ldots b_s) + [e(i-1, j-1) - e(p-1, q-1)].$$

Then Eq. (2) is no longer needed, because the corrective term will vanish whenever it holds.

DEFINITION 3. Let d denote evolutionary distance, and for any two sequences

$$a_1 a_2 \ldots a_m \,\&\, b_1 b_2 \ldots b_n$$

let

$$e(i,j) + \min\{d(a_g \ldots a_i, b_h \ldots b_j) : g = 1 \text{ or } h = 1\},$$

and

$$f(i,j) = \min\{d(a_i \ldots a_g, b_j \ldots b_h) : g = m \text{ or } h = n\}.$$

Then *interval* $a_p \ldots a_r$ *most resembles interval* $b_q \ldots b_s$ *locally*, if

$$d(a_p \ldots a_r, b_q \ldots b_s) \leq d(a_i \ldots a_r, b_j \ldots b_s) + [e(i-1, j-1) - e(p-1, q-1)]$$

for $i = 1, 2, \ldots, r+1$ and $j = 1, 2, \ldots, s+1$, and

$$d(a_p \ldots a_r, b_q \ldots b_s) \leq d(a_p \ldots a_i, b_q \ldots b_j) + [f(i+1, j+1) - f(r+1, s+1)]$$

for $i = p-1, p, \ldots, m$ and $j = q-1, q, \ldots, n$. (It is a notational convention that $a_i \ldots a_r$ is the empty sequence, denoted by 1, when $i = r+1$. Likewise, $b_{s+1} \ldots b_s, a_p \ldots a_{p-1}$, and $b_q \ldots b_{q-1}$ equal 1.)

The following theorem provides an algorithm for determining all intervals $I \subset A$ and $J \subset B$ which most resemble each other locally.

THEOREM 3. *Let d denote evolutionary distance, and for any two sequences*

$$a_1 a_2 \ldots a_m \,\&\, b_1 b_2 \ldots b_n$$

let the matrices $(e(i,j))$ and $(f(i,j))$ be constructed as follows: If $i = 0$ or $j = 0$, then $e(i,j) = 0$, and for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$ the value of

$e(i,j)$ is the smallest of

$$e(i-1,j)+d(a_i,1).$$

$$e(i-1,j-1)+d(a_i,b_j).$$

$$e(i,j-1)+d(1,b_j)$$

If $i=m+1$ or $j=n+1$, then $f(i,j)=0$, and for $i=1,2,\ldots,m$ and $j=1,2,\ldots,n$ the value of $f(i,j)$ is the smallest of

$$f(i+1,j)+d(a_i,1).$$

$$f(i+1,j+1)+d(a_i,b_j).$$

$$f(i,j+1)+d(1,b_j).$$

Then interval $a_p\ldots a_r$ most resembles interval $b_q\ldots b_s$ locally if, and only if, there exists a monotone sequence

$$(p-1,q-1),\ldots,(r,s)$$

of pairs such that

(i) *each term except the first in*

$$e(p-1,q-1),\ldots,e(r,s)$$

is derivable from the term immediately preceding it by one of the inductive steps allowed to be used in the construction, described above, of the matrix $(e(i,j))$; and

(ii) *with every number in the sequence of pairs increased by* 1, *each term except the last in*

$$f(p,q),\ldots,f(r+1,s+1)$$

is derivable from the term immediately following it by one of the inductive steps allowed to be used in the construction, described above, of the matrix $(f(i,j))$.

The above theorem furnishes an algorithm for determining every pair of intervals, one from $a_1\ldots a_m$ and the other from $b_1\ldots b_n$, which most resemble each other locally. The procedure is to compute all the matrix values, which takes on the order of mn steps. As the matrices are being evaluated, a pair of e-matrix positions

$$(i_1,j_1),(i_2,j_2)$$

is earmarked provided it is linked by an allowable inductive step in the construction of the e-matrix and provided also the corresponding pair

$$(i_1+1,j_1+1),(i_2+1,j_2+1)$$

of f-matrix positions is linked by an allowable inductive step in the construction of the f-matrix. When this procedure is over, if

$$(p-1,q-1),\ldots,(r,s)$$

is any sequence of positions, in which every successive pair has been earmarked, then $a_p \ldots a_r$ and $b_q \ldots b_s$ is a pair of intervals which most resemble each other locally.

Proof of Theorem 3. It is evident that $e(i,j)$, as constructed in the present theorem, equals the minimum value associated with any path from (g,h) to (i,j) for which $g = 0$ or $h = 0$.

For a particular choice of (g,h) the minimum value associated with any path from (g,h) to (i,j) is equal to

$$d(a_{g+1} \ldots a_i, b_{h+1} \ldots b_j).$$

Therefore,

$$e(i,j) = \min\{d(a_{g+1} \ldots a_i, b_{h+1} \ldots b_j) : g = 0 \text{ or } h = 0\},$$

which shows that the characterizations of $e(i,j)$ in Definition 3 and Theorem 3 are equivalent.

Let us now assume that the sequence

$$(p-1, q-1), \ldots, (r,s)$$

exists, as described in the theorem, and show that this implies the conditions of Definition 3.

The sequence implies that $e(r,s)$ is equal to the value associated with a path which passes through $(p-1, q-1)$, which means that it may be separated into two parts:

$$e(r,s) = e(p-1, q-1) + d(a_p \ldots a_r, b_q \ldots b_s).$$

The fact that $e(r,s)$ is the minimum value associated with any path to (r,s) implies that for any (i,j)

$$e(r,s) \leq e(i-1, j-1) + d(a_i \ldots a_r, b_j \ldots b_s).$$

This, combined with the equation above, gives the first inequality of Definition 3. The second inequality follows from the symmetry of the theorem with respect to the e-matrix and the f-matrix.

To prove the converse of the theorem, we assume that the inequalities of Definition 2 hold. In the first one let $i = r + 1$ and $j = s + 1$, and in the second let $i = p - 1$ and $j = q - 1$. Then

$$e(p-1, q-1) + d(a_p \ldots a_r, b_q \ldots b_s) \leq e(r,s)$$

and

$$d(a_p \ldots a_r, b_q \ldots b_s) + f(r+1, s+1) \leq f(p,q).$$

Equality holds in both cases, because $e(r,s)$ and $f(p,q)$ are minima. Therefore, there exist paths

$$\ldots,(p-1,q-1)\ldots,(r,s),$$
$$(p,q),\ldots,(r+1,s+1),\ldots,$$

the values associated with them being $e(r,s)$ and $f(p,q)$, respectively. This means that the positions

$$(p-1,q-1),\ldots,(r,s)$$

satisfy conditions (i) and (ii) of the theorem.

REFERENCES

1. R. A. WAGNER AND M. J. FISCHER, The string-to-string correction problem, *J. Assoc. Comput. Mach.* **21** (1974), 168–173.
2. P. H. SELLERS, On the theory and computation of evolutionary distances, *SIAM J. Appl. Math.* **26** (1974), 787–793.
3. P. H. SELLERS, An algorithm for the distance between two finite sequences, *J. Combinatorial Theory* **16** (1974), 253–258.
4. S. M. ULAM, Some combinatorial problems studied experimentally on computing machines, in "Applications of Number Theory to Numerical Analysis" (S. K. Zaremba, Ed.), pp. 1–3, Academic Press, New York, 1972.
5. M. S. WATERMAN, T. F. SMITH, AND W. A. BEYER, Some biological sequence metrics, *Advances in Math.* **20** (1976), 367–387.