Reprinted from ADVANCES IN MATHEMATICS
All Rights Reserved by Academic Press, New York and London
Vol. 20, No. 3, June 1976
Printed in Belgium

Some Biological Sequence Metrics*

M. S. WATERMAN

Idaho State University, Pocatello, Idaho 83209

T. F. SMITH

Northern Michigan University, Marquette, Michigan 49855

AND

W. A. BEYER

Los Alamos Scientific Laboratory, Los Alamos, New Mexico 87545†

Some new metrics are introduced to measure the distance between biological sequences, such as amino acid sequences or nucleotide sequences. These metrics generalize a metric of Sellers, who considered only single deletions, mutations, and insertions. The present metrics allow, for example, multiple deletions and insertions and single mutations. They also allow computation of the distance among more than two sequences. Algorithms for computing the values of the metrics are given which also compute best alignments. The connection with the information theory approach of Reichert, Cohen, and Wong is discussed.

1. INTRODUCTION

A protein is specified by the sequence of amino acids composing the protein. Since twenty kinds of amino acids are used in proteins, a protein is a finite word over an alphabet of twenty letters. Biology seeks to discover the evolutionary relations among the set of proteins. It postulates that proteins have evolved in time past, and this evolution has produced a tree (or trees) whose terminal nodes are the extant

* Work performed under an NSF Faculty Research Participation Program at Los Alamos Scientific Laboratory and also supported by the Atomic Energy Commission.
† Theoretical Division.

proteins. It is a task of biology to decide what the mechanisms and probabilities of the various protein evolutionary processes are. It is a mathematical task then to produce from existing protein sequences and the postulated mechanisms and probabilities of evolutionary processes a probable tree of the extant proteins. These matters are discussed in the references, but there is no adequate survey. The closest reference to a survey is Dayhoff [3]. These tree construction methods can also be regarded as part of the theory of clusters or of numerical taxonomy. See Jardine and Sibson [4].

One of the major techniques used in taxonomic tree construction depends on the introduction of a measure of dissimilarity of sequences (see [1, 2, or 3]). Beginning with Fitch [10], there is a growing literature concerned with the distance between two protein sequences [5–17]. The work on distances or metrics on protein sequences has grown more sophisticated and is no longer entirely concerned with producing measures of dissimilarity for the construction of evolutionary trees. One of the main concerns is to discover what genetic mutations are required to change one sequence into another. Perhaps the most mathematically satisfactory treatment to date is by Sellers [6]. The present paper presents some new metrics on sequences and some of their mathematical and algorithmic properties.

A metric $\rho(s_1, s_2)$ is a nonnegative function on the set $S \times S$ of pairs (s_1, s_2) of finite sequences s_i over a fixed alphabet. It is to be thought of as a measure of the amount of evolutionary change from the sequence s_1 to s_2. (1) It is zero if and only if $s_1 = s_2$. (2) It is assumed that evolutionary changes are reversible[1] and therefore that $\rho(s_1, s_2) = \rho(s_2, s_1)$ for all $s_i \in S$.[2] (3) It is further assumed to measure the fewest number of evolutionary changes (weighted according to probability) from s_1 to s_2. Hence ρ should satisfy the triangle inequality:

$$\rho(s_1, s_3) \leqslant \rho(s_1, s_2) + \rho(s_2, s_3)$$

for all s_1, s_2, and s_3 in S.

Sequence metrics have been introduced in the past in references [1], [2], [6], and [7]. The metric in [2] was suggested by T. Smith. While it seems to yield satisfactory results, its interpretation in terms of

[1] By reversible is meant that evolutionary changes and their inverses are equally probable.

[2] However, nonsymmetric ρ's can be handled within the theory of this paper and are discussed in Section 7.

evolutionary changes is not clear. The interpretation of the metric due to Sellers [6] in terms of evolutionary changes is more clear. This metric is referred to as the s-metric. In this paper the s-metric is generalized to allow for more general evolutionary changes.

2. The τ-metric

Let A be a finite set such that $\Delta \in A$. Δ is called the neutral element. As in [6], the following definitions are made.

(i) $\mathbf{a} = a_1 a_2 \cdots$ is an A-sequence if the sequence is nonterminating, only finitely many of the a_i are $\neq \Delta$, and $a_i \in A$.

(ii) Two A-sequences are equivalent if the subsequences of nonneutral terms are identical.

(iii) An evolutionary sequence $\bar{\mathbf{a}} = \overline{a_1 a_2 \cdots}$ is the set of all A-sequences equivalent to $\mathbf{a} = a_1 a_2 \cdots$.

Every evolutionary sequence $\overline{a_1 a_2 \cdots}$ has a member such that if $a_n \neq \Delta$, then $a_k \neq \Delta$ for all $k \leqslant n$. Let $\bar{S}$ be the set of A-sequences such that this property holds. $\bar{S}$ includes $\Delta\Delta \cdots$.

A finite set τ of weighted transformations T: $\bar{S} \to \bar{S}$ such that each $T \in \tau$ has as its domain $\mathscr{D}(T)$ and range $\mathscr{R}(T)$ nonempty subsets of $\bar{S}$ is considered which satisfies the following two conditions.

(i) The identity transformation I is in τ.

(ii) Each $T \in \tau$ has an associated nonnegative number $w(T)$ called the weight of T which is zero if and only if $T = I$.

Since τ is finite, (ii) implies

$$\min_{\substack{T \in \tau \\ T \neq I}} w(T) > 0. \tag{1}$$

Fix a $j \geqslant 1$. Suppose $a_1 a_2 \cdots$ is an A-sequence and $a_j a_{j+1} \cdots \in \mathscr{D}(T)$. Then T^j is defined by $T^j(a_1 a_2 \cdots) = a_1 \cdots a_{j-1}\, T(a_j a_{j+1} \cdots)$ where, if $j = 1$, $a_1 \cdots a_{j-1}$ is omitted. If $T \in \tau$, then define $w(T^j) = w(T)$.

Suppose a finite set τ of transformations satisfying (i) and (ii) is given. For $\mathbf{a}, \mathbf{b} \in \bar{S}$, define

$$\{\mathbf{a} \to \mathbf{b}\}_\tau = \{T_{i_l}^{j_l} \cdots T_{i_1}^{j_1} \mid T_{i_l}^{j_l} \cdots T_{i_1}^{j_1} \mathbf{a} = \mathbf{b}\},$$

where $T_{i_k} \in \tau$.

Thus $\{\mathbf{a} \to \mathbf{b}\}_\tau$ is the set of all finite transformation sequences from τ which map $\mathbf{a}$ into $\mathbf{b}$. The set $\{\mathbf{a} \to \mathbf{b}\}_\tau$ may be the empty set $\varnothing$.

For $\mathbf{a}, \mathbf{b} \in \bar{S}$, define

$$\rho(\mathbf{a}, \mathbf{b}) = \max \left\{ \min_{\{\mathbf{a}\to\mathbf{b}\}_\tau} \sum_{k=1}^{l_1} w(T_{i_k}), \min_{\{\mathbf{b}\to\mathbf{a}\}_\tau} \sum_{k=1}^{l_2} w(T_{i_k}) \right\}.$$

If $\{\mathbf{a} \to \mathbf{b}\}_\tau \neq \varnothing$, then, by the finiteness of τ, the above minima exist. If $\{\mathbf{a} \to \mathbf{b}\}_\tau = \varnothing$, define

$$\min_{\{\mathbf{a}\to\mathbf{b}\}_\tau} \sum_{k=1}^{l} w(T_{i_k}) = +\infty.$$

It is obvious that the relation $\rho(\mathbf{a}, \mathbf{b}) < \infty$ is reflexive, symmetric, and transitive and therefore this relation partitions $\bar{S}$ into equivalence classes $\{\bar{S}_i\}$. In each $\bar{S}_i$ the value of ρ between any two elements of S_i is finite.

THEOREM 1. *Each equivalence class $\bar{S}_i$ of $\bar{S}$ together with the function ρ (called a τ-metric) is a metric space.*

Proof. It is obvious that the function ρ is symmetric and nonnegative. Because of (1), $\rho(\mathbf{a}, \mathbf{b}) = 0$ if and only if $\mathbf{a} = \mathbf{b}$. One has

$$\rho(\mathbf{a}, \mathbf{b}) + \rho(\mathbf{b}, \mathbf{c}) \geqslant \min_{\{\mathbf{a}\to\mathbf{b}\}_\tau} \sum_{k=1}^{l_1} w(T_{i_k}) + \min_{\{\mathbf{b}\to\mathbf{c}\}_\tau} \sum_{k=1}^{l_2} w(T_{i_k}). \tag{2}$$

Suppose the first min on the right of (2) is attained by $T_{i_{l_1}}^{j_{l_1}} \cdots T_{i_1}^{j_1}$; i.e.

$$\sum_{k=1}^{l_1} w(T_{i_k}) = \min_{\{\mathbf{a}\to\mathbf{b}\}_\tau} \sum_{k=1}^{l_1} w(T_{i_k}),$$

and

$$T_{i_{l_1}}^{j_{l_1}} \cdots T_{i_1}^{j_1}(\mathbf{a}) = \mathbf{b}.$$

Likewise, suppose

$$\sum_{k=1}^{l_2} w(T_{i_k}) = \min_{\{\mathbf{b}\to\mathbf{c}\}_\tau} \sum_{k=1}^{l_2} w(T_{i_k}).$$

Then

$$\rho(\mathbf{a}, \mathbf{b}) + \rho(\mathbf{b}, \mathbf{c}) \geqslant \sum_{k=1}^{l_1} w(T_{i_k}) + \sum_{k=1}^{l_2} w(T_{i_k}) \geqslant \rho(\mathbf{a}, \mathbf{c}).$$

This completes the proof of the theorem.

3. The Relationship between the τ-metric and the s-metric

Sellers [6] has extended a metric d on A to a metric d on A-sequences and a metric $\bar{d}$ on the set of evolutionary sequences by the formulae:

$$d(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{\infty} d(a_i, b_i),$$

and

$$\bar{d}(\bar{\mathbf{a}}, \bar{\mathbf{b}}) = \min_{\substack{\mathbf{a}\in\bar{\mathbf{a}} \\ \mathbf{b}\in\bar{\mathbf{b}}}} d(\mathbf{a}, \mathbf{b}).$$

The $\bar{d}$ metric is referred to as the s-metric.

If one relates the weights of mutations and deletions to the metric d, then $\bar{d}$ gives the smallest total of the weights of the sets of mutations and deletions which make $\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$ identical.

A set τ of transformations will be defined for which

$$\bar{d}(\bar{\mathbf{a}}, \bar{\mathbf{b}}) = \rho(\mathbf{a}, \mathbf{b}),$$

where $\mathbf{a}(\mathbf{b})$ is the member of $\bar{S}$ which is in the evolutionary sequence $\bar{\mathbf{a}}(\bar{\mathbf{b}})$, thus showing that the τ-metric is at least as general as $\bar{d}$. Define T_{a-} by

$$\mathscr{D}(T_{a-}) = \{aa_2a_3 \cdots \mid a_2a_3 \cdots \in \bar{S}\}$$

and

$$T_{a-}(aa_2a_3 \cdots) = a_2a_3 \cdots .$$

Define T_{a+} by letting $\mathscr{D}(T_{a+})$ be $\bar{S}$ and

$$T_{a+}(a_1a_2 \cdots) = aa_1a_2 \cdots .$$

Define $T_{(a,b)}$, where $a \neq b$, $a \neq \Delta$, $b \neq \Delta$, by

$$\mathscr{D}(T_{(a,b)}) = \{aa_2a_3 \cdots \mid a_2a_3 \cdots \in \bar{S}\}$$

and

$$T_{(a,b)}(aa_2a_3 \cdots) = ba_2a_3 \cdots .$$

Put

$$w(T_{a+}) = w(T_{a-}) = d(a, \Delta)$$

and

$$w(T_{(a,b)}) = d(a, b).$$

The notion of the "path" of a site through a sequence of transformations is now introduced for use in the proof of Theorems 2 and 3. Suppose

$$\mathbf{b} = T_{i_l}^{j_l} \cdots T_{i_1}^{j_1}(\mathbf{a})$$

For an $a_i \in \mathbf{a}$, let the sequence $p_0, p_1, \ldots, p_l$ record the sequence of subscripts of the successors of a_i and let $c_0, c_1, \ldots, c_l$ record the sequence of successors of a_i under the l transformations $T_{i_k}^{j_k}$. One puts $p_0 = i$ and $c_0 = a_i$. If $c_j = \Delta$, then p_j is designated by $p_j = -$. If $p_j = -$, then $p_k = -$ and $c_k = \Delta$ for all $k \geqslant j$. For $n < l$ suppose c_n and p_n are known. If $T_{i_{n+1}}^{j_{n+1}}$ is such that $j_{n+1} < p_n$, let $c_{n+1} = c_n$ and

$$p_{n+1} = \begin{cases} p_n & \text{if } i_{n+1} = (a, b), \\ p_n + 1 & \text{if } i_{n+1} = a+, \\ p_n - 1 & \text{if } i_{n+1} = a-. \end{cases}$$

If $j_{n+1} = p_n$, let

$$p_{n+1} = p_n \quad \text{and} \quad c_{n+1} = a \quad \text{if} \quad i_{n+1} = (c_n, a),$$

$$p_{n+1} = - \quad \text{and} \quad c_{n+1} = \Delta \quad \text{if} \quad i_{n+1} = c_n -,$$

$$p_{n+1} = p_n + 1 \quad \text{and} \quad c_{n+1} = c_n \quad \text{if} \quad i_{n+1} = a+.$$

If $j_{n+1} > p_n$, then $p_{n+1} = p_n$ and $c_{n+1} = c_n$.

Let $N = \max\{j: b_j \neq \Delta\}$. Let P be the set of all p_l associated with $a_i \neq \Delta$ which are members of $\mathbf{a}$. If $j \in \{1, 2, \ldots, N\} \sim P$, we create a p and c sequence associated with b_j as follows. Let $p_l = j$ and $c_l = b_j$. In an obvious way, we can trace p_k and c_k backwards. If $p_n = -$, then $p_k = -$ and $c_k = \Delta$ for all $k \leqslant n$ and every such sequence has $p_0 = -$ and $c_0 = \Delta$.

This scheme gives one the "path" of a site through the sequence of transformations. Each a_i becomes a b_j or is deleted. Each b_j comes from an a_i or is inserted.

THEOREM 2. *If τ and $w(T)$ are the above class of transformations and weights, then $\rho(\mathbf{a}, \mathbf{b}) = \bar{d}(\bar{\mathbf{a}}, \bar{\mathbf{b}})$.*

Proof. Recall that

$$\bar{d}(\bar{\mathbf{a}}, \bar{\mathbf{b}}) = \min_{\substack{\mathbf{a}\in\bar{\mathbf{a}} \\ \mathbf{b}\in\bar{\mathbf{b}}}} \sum_{i=1}^{\infty} d(a_i, b_i).$$

We first show that for any $\mathbf{a} \in \bar{\mathbf{a}}$ and $\mathbf{b} \in \bar{\mathbf{b}}$, there is a $T_1 \in \{\mathbf{a} \to \mathbf{b}\}_\tau$ and $T_2 \in \{\mathbf{b} \to \mathbf{a}\}_\tau$ such that:

(α) The sum of weights $\sum w(T)$ associated with T_1 is $d(\mathbf{a}, \mathbf{b})$.

(β) The sum of weights $\sum w(T)$ associated with T_2 is $d(\mathbf{b}, \mathbf{a})$.

(Recall that T_i may denote compositions of transformations from τ.) Since $d(\mathbf{a}, \mathbf{b}) = d(\mathbf{b}, \mathbf{a})$ and $\mathbf{a}$ and $\mathbf{b}$ are arbitrary, we need only show case α.

There is an N such that $a_i = b_i = \Delta$ for $i \geqslant N$. For the largest $j < N$ such that $a_j \neq b_j$, define $T_k{}^j$ by:

$$k = a_j - \text{(a deletion)} \quad \text{if} \quad b_j = \Delta,$$

$$k = b_j + \text{(an insertion)} \quad \text{if} \quad b_j \neq \Delta \quad \text{and} \quad a_j = \Delta,$$

and

$$k = (a_j, b_j)\ \text{(a mutation)} \quad \text{if} \quad b_j \neq \Delta \quad \text{and} \quad a_j \neq \Delta.$$

Then go to the next smallest $j < j$ such that $a_j \neq b_j$. A finite number of applications of this rule yields a sequence of transformations in $\{\mathbf{a} \to \mathbf{b}\}_\tau$. The corresponding sum of weights is

$$\sum w(T) = \sum_{a_j \neq b_j} d(a_j, b_j) = \sum_{i=1}^{\infty} d(a_i, b_i).$$

This result implies

$$\rho(\mathbf{a}, \mathbf{b}) \leqslant d(\mathbf{a}, \mathbf{b})$$

and therefore

$$\rho(\mathbf{a}, \mathbf{b}) \leqslant \bar{d}(\bar{\mathbf{a}}, \bar{\mathbf{b}}).$$

We next prove that for

$$T_{i_t}^{j_t} \cdots T_{i_1}^{j_1} \mathbf{a} = \mathbf{b},$$

there are sequences $\mathbf{a}' \in \bar{\mathbf{a}}$ and $\mathbf{b}' \in \bar{\mathbf{b}}$ such that

$$\sum_{k=1}^{l} w(T_{i_k}^{j_k}) \geqslant d(\mathbf{a}', \mathbf{b}'), \tag{3}$$

which implies that

$$\rho(\mathbf{a}, \mathbf{b}) \geqslant \bar{d}(\bar{\mathbf{a}}, \bar{\mathbf{b}})$$

and will complete the proof.

As observed in the discusion preceeding Theorem 2, each a_i either becomes a b_j or is deleted. In the alignment visualized by writing the **a** sequence above the **b** sequence, put b_j under a_i in the first case and Δ under a_i in the second. In case any b_j are not listed, insert them in their proper positions in the **b** sequence with a Δ above them. The new **a** sequence is called $\mathbf{a}'$ and the new **b** sequence is called $\mathbf{b}'$. One must show that (3) holds.

The application of the transformation corresponds to changes in the sequence $c_0\, c_1 \cdots c_l$. If $c_0 = c_1 = \cdots = c_l$, then no transformations were applied which affected the position corresponding to that c-sequence. Each transformation causes exactly one c-sequence to change. The triangle inequality for $d(\cdot, \cdot)$ implies the sum of weights of the transformation associated with a sequence is at least $d(c_0, c_l)$. Therefore one has (3), which completes the proof.

Theorem 2 shows that the τ-metric is at least as general as the s-metric. That the τ-metric is more general than the s-metric can be seen from the following example. Let $A = \{\Delta, a\}$, $w(T_{a+}) = w(T_{a-}) = 1$, and $w(T_{aa+}) = w(T_{aa-}) = 1.1$ where

$$T_{aa+}(a_1 a_2 \cdots) = aaa_1 a_2 \cdots$$

and

$$T_{aa-}(aaa_1 a_2) = a_1 a_2 \cdots .$$

Then $\rho(a, \Delta) = 1$. So an equivalent s-metric must have $\bar{d}(a, \Delta) = d(a, \Delta) = 1$. But then $\bar{d}(aa, \Delta) = 2$ which does not agree with $\rho(aa, \Delta) = w(T_{aa-}) = 1.1$.

The weights used with the τ-metric need not be related to the s-metric on the alphabet A. For example, let $A = \{\Delta, a, b\}$. Let $w(T_{a-}) = w(T_{a+}) = 3$, $w(T_{b-}) = w(T_{b+}) = 1$, and $w(T_{(a,b)}) = w(T_{(b,a)}) = 1$. These w's do not provide a metric on A if we choose $d(a, \Delta) = w(T_{a-})$,

$d(b, \Delta) = w(T_{b-})$, and $d(a, b) = w(T_{(a,b)})$, since $d(a, b) + d(b, \Delta) = 2$, but $d(a, \Delta) = 3$. One can also allow $w(T_{(a,b)}) \neq w(T_{(b,a)})$ and $w(T_{a+}) \neq w(T_{a-})$, which is not possible with the s-metric.

4. Deletions and Insertions of More Than One Letter

A motivation for considering τ-metrics is to allow a more general class of transformations than the s-metric. Specifically one wants to allow deletions and insertions of letter sequences of length greater than 1. This can be accomplished in the s-metric formalism by successive deletions. However, this may be a greater distance than one might want to permit.

A major result of [6] is an efficient algorithm to compute distance. Below we define a τ-metric which includes longer deletions and insertions and give a Sellers-like algorithm for this τ-metric under certain conditions.

Let α be a fixed positive integer. For $1 \leqslant m \leqslant \alpha$, define the domain of $T_{b_1 \cdots b_m -}$ to be those members of $\bar{S}$ whose first m letters are $b_1 \cdots b_m$. Put

$$T_{b_1 \cdots b_m -(b_1 \cdots b_m a_{m+1} \cdots)} = a_{m+1} \cdots .$$

Define the domain of $T_{b_1 \cdots b_m +}$ as $\bar{S}$ and

$$T_{b_1 \cdots b_m +}(a_1 a_2 \cdots) = b_1 \cdots b_m a_1 a_2 \cdots .$$

Define $T_{(a,b)}$ as before.

We say that $\{\mathbf{a} \to \mathbf{b}\}_\tau$ satisfies condition **M** if there is a minimum sequence of transformations in $\{\mathbf{a} \to \mathbf{b}\}_\tau$ (i.e., $T_1 \cdots T_k \in \{\mathbf{a} \to \mathbf{b}\}_\tau$ such that

$$\sum_{i=1}^{k} w(T_i)$$

is not greater than any other

$$\sum_{i=1}^{k'} w(T_i')$$

for which $T_1' \cdots T_k' \in \{\mathbf{a} \to \mathbf{b}\}_\tau$) such that $T_1 \cdots T_k$ contains no multiple hits. That is, if an element is deleted, inserted, or undergoes mutation, it is not changed again. If one considers the associated $c_0\, c_1\, c_2 \cdots c_l$ for each position in **a** and in **b**, then there are no multiple hits if $\{c_0, c_1, \ldots, c_l\}$ has at most two distinct elements.

For brevity let

$$A_k = a_1 a_2 \cdots a_k \Delta \Delta \cdots$$

and

$$B_j = b_1 b_2 \cdots b_j \Delta \Delta \cdots$$

where $A_0 = \Delta \Delta \ldots$ and $B_0 = \Delta \Delta \ldots$. Let

$$\rho_1(\mathbf{a}, \mathbf{b}) = \min_{\{\mathbf{a} \to \mathbf{b}\}_\tau} \sum_{i=1}^{k} w(T_i),$$

$$\rho_2(\mathbf{a}, \mathbf{b}) = \rho_1(\mathbf{b}, \mathbf{a}).$$

The length of $\mathbf{a} \in \bar{S}$ is the number of $a_i \neq \Delta$. The following theorem is analogous to Theorem 2 of [6]. (Put $\bar{A} = A \sim \Delta$.)

THEOREM 3. *Suppose* $\mathbf{a}$ *and* $\mathbf{b}$ *in* $\bar{S}$ *and the length of* $\mathbf{a}(\mathbf{b})$ *is* $M(N)$. *Let* τ *be the set of transformation consisting of the following types*: (1) $T_{(a,b)}$ *for every* a, $b \in \bar{A}$, (2) $T_{b_1 \cdots b_m +}$, $T_{b_1 \cdots b_m -}$ *for all sequences* $b_1 \cdots b_m$, $1 \leqslant m \leqslant \alpha$. *Suppose* $\{A_k \to B_j\}_\tau$ $(0 \leqslant k \leqslant M;\ 0 \leqslant j \leqslant N)$ *satisfies condition* **M**. *Obviously* $\rho_1(A_0, B_0) = 0$. *Then* $\rho_1(A_i, B_j)(0 \leqslant i \leqslant M,$ $0 \leqslant j \leqslant N,\ i + j \neq 0)$ *is the minimum of the following values*:

$$\rho_1(A_{i-k}, B_j) + w(T_{a_{i-k+1} \cdots a_i -})(1 \leqslant k \leqslant \min(\alpha, i)),$$

$$\rho_1(A_{i-1}, B_{j-1}) + w(T_{(a_i, b_j)}),$$

$$\rho_1(A_i, B_{j-k}) + w(T_{b_{j-k+1} \cdots b_j +})(1 \leqslant k \leqslant \min(\alpha, j)),$$

where $\rho_1(A_p, B_q)$ *is ignored if* p *or* q *is negative and if an index set is empty the corresponding quantities are not computed.*

Proof. Fix i and j. It is evident that if $T_{i_l}^{j_l} \cdots T_{i_1}^{j_1} \in \{A_i \to B_j\}_\tau$ has no multiple hits, then the T's may be reordered in any order and the j's appropriately changed so that the resulting transformation sequence is also in $\{A_i \to B_j\}_\tau$.

There are three cases to be considered with regard to the fate of a_i under a minimum mapping in $\{A_i \to B_j\}_\tau$. (i) a_i is deleted. (ii) a_i is unchanged or undergoes a mutation to b_j . (iii) b_j is inserted at the end of the sequence.

We place the T that carries out the operation in (i), (ii), or (iii) in the first (right-most) position.

In case (i), $T_{i_1}^{j_1}(a_1 \cdots a_i) = a_1 \cdots a_{i-k}$, $1 \leqslant k \leqslant \min(\alpha, i)$. Thus $T_{i_1}^{j_1} \cdots T_{i_2}^{j_2} \in \{A_{i-k} \to B_j\}_\tau$ and, if $T = T_{i_1}^{j_1} \cdots T_{i_1}^{j_1}$ has minimum weight in $\{A_i \to B_j\}_\tau$, then $T_{i_1}^{j_1} \cdots T_{i_2}^{j_2}$ has minimum weight in $\{A_{i-k} \to B_j\}_\tau$ and

$$\rho_1(A_i, B_j) = \rho_1(A_{i-k}, B_j) + w(T_{a_{i-k+1} \cdots a_i -}).$$

Case (iii) is handled in a similar fashion.

In case (ii), let $T_{i_1} = I$ if a_i is left unchanged. Reasoning as in case (i), one sees that $T_{i_1}^{j_1} \cdots T_{i_2}^{j_2} \in \{A_{i-1} \to B_{i-1}\}_\tau$, so that

$$\rho_1(A_i, B_j) = \rho_1(A_{i-1}, B_{j-1}) + w(T_{(a_i, b_j)}),$$

which completes the proof.

Next consider the number of steps required in the algorithm of Theorem 3. If $\alpha \ll \min\{M, N\}$, there are essentially $2\alpha + 1$ values to find the minimum of for each i and j. Thus the algorithm runs in approximately $(2\alpha + 1)\ MN$ steps. If $\alpha = \max\{M, N\}$, the number of steps is approximately

$$\sum_{i=1}^{M} \sum_{j=1}^{N} (i + j - 1) = N \frac{M(M+1)}{2} + M \frac{N(N+1)}{2} - MN$$

or of the order $MN \max\{M, N\}$ steps.

Theorem 3 can be used, in an obvious way, to construct an induction which will calculate $\rho_1(\mathbf{a}, \mathbf{b})$. The value of $\rho_2(\mathbf{a}, \mathbf{b})$ is then calculated from $\rho_2(\mathbf{a}, \mathbf{b}) = \rho_1(\mathbf{b}, \mathbf{a})$ and

$$\rho(\mathbf{a}, \mathbf{b}) = \max(\rho_1(\mathbf{a}, \mathbf{b}),\ \rho_2(\mathbf{a}, \mathbf{b}))$$

COROLLARY. *If* $w(T_{(a,b)}) = w(T_{(b,a)})$ *and* $w(T_{b_1 \cdots b_k -}) = w(T_{b_1 \cdots b_k +})$, $(1 \leqslant k \leqslant \alpha)$, *then*

$$\rho_1(\mathbf{a}, \mathbf{b}) = \rho_2(\mathbf{a}, \mathbf{b}) = \rho(\mathbf{a}, \mathbf{b}).$$

Proof. The transformations $T_{(a,b)}$ and $T_{(b,a)}$ are inverses of each other, as are $T_{b_1 \cdots b_k -}$ and $T_{b_1 \cdots b_k +}$. Thus if a sequence of T's is in $\{\mathbf{a} \to \mathbf{b}\}_\tau$, there is a corresponding sequence of inverse mappings in $\{\mathbf{b} \to \mathbf{a}\}_\tau$ and the weights, by the symmetry assumptions, are equal.

THEOREM 4. *The hypotheses of Theorem* 3 *are satisfied if we define* $w(T_{(a,b)}) = \lambda_0$, $w(T_{a_1 \cdots a_k \pm}) = \lambda_{k\pm}$ $(1 \leqslant k \leqslant \alpha)$ *and require*

$$\lambda_{1+} \leqslant \lambda_{2+} \leqslant \lambda_{3+} \leqslant \cdots \leqslant \lambda_{\alpha+},$$

and

$$\lambda_{1-} \leqslant \lambda_{2-} \leqslant \lambda_{3-} \leqslant \cdots \leqslant \lambda_{\alpha-} .$$

Proof. We show that there is some minimum sequence in $\{\mathbf{a} \to \mathbf{b}\}_\tau$ which has no multiple hits. It is sufficient to consider pairs of transformations acting on the same site. We shall show they can, in each case, be replaced by equivalent transformations, which do not act on the same site and whose weight is less than or equal to the sum of the weights of the pairs of transformations.

The three types of transformations (insertion, deletion, and mutation) allows nine categories of pairs of transformations. However any transformation followed by an insertion cannot act on the same site and three of the nine categories are eliminated.

A mutation followed by a mutation, both on the same site, must be of the form $T_{(b,c)}T_{(a,b)}$ which has the weight $2\lambda_0$ and this pair can be replaced by $T_{(a,c)}$ which has weight λ_0.

A deletion followed by a mutation or deletion cannot act at the same site. An insertion followed by a mutation must be of the form $T_{(a_i,b)}$ $T_{a_1\cdots a_k+}$ (where $1 \leqslant i \leqslant k$) which has the weight $\lambda_0 + \lambda_{k+}$. This pair can be replaced by $T_{a_1\cdots a_{i-1}ba_{i+1}\cdots a_k+}$ which as weight λ_{k+}.

A mutation followed by a deletion has the form $T_{a_1\cdots a_k-}T_{(b,a_i)}$ (where $1 \leqslant i \leqslant k$) and weight $\lambda_{k-} + \lambda_0$. We replace this pair by $T_{a_1\cdots a_{i-1}ba_{i+1}\cdots a_k-}$ which has weight λ_{k-}.

The remaining category is an insertion followed by a deletion. Suppose this pair is of the form $T_{a_j\cdots a_l-}T_{a_i\cdots a_k+}$ where $i \leqslant j \leqslant l \leqslant k$, or $i \leqslant j \leqslant k \leqslant l$, or $j \leqslant i \leqslant l \leqslant k$, or $j \leqslant i \leqslant k \leqslant l$. This pair has weight $\lambda_{l-j+1-} + \lambda_{k-i+1+}$. Suppose $i \leqslant j \leqslant l \leqslant k$. Replace the previous pair by $T_{a_i\cdots a_{j-1}a_{l+1}\cdots a_k-}$, which has weight $\lambda_{(k-l)+(j-i)+} = \lambda_{k-i-(l-j)+} \leqslant \lambda_{k-i+1+}$. Suppose $i \leqslant j \leqslant k \leqslant l$. Replace the original pair by $T_{a_{k+1}\cdots a_l-}$ $T_{a_i\cdots a_{j-1}+}$ with weight $\lambda_{l-k-} + \lambda_{j-i+} \leqslant \lambda_{l-j+1-} + \lambda_{k-i+1+}$. The remaining two subcases are handled in the same way.

5. An Example

Suppose $A = \{\Delta, a, b, c\}$. Suppose all single mutations have weight 1, all single deletions and insertions have weight 1, and all double deletions and insertions have weight 1.1. Theorem 3, the Corollary, and Theorem 4 all apply. Suppose $\mathbf{a} = abccaaa$ and $\mathbf{b} = abaaa$. (We omit writing the

terminal Δ's.) Table 1 is a matrix of values from Theorem 3. The first row has the values

$$\rho(\Delta, \Delta), \quad \rho(\Delta, a), \quad \rho(\Delta, ab), \quad \rho(\Delta, abc) \quad ,\ldots$$

TABLE 1

Double Deletion and Insertion Distance Calculation

	Δ	a	b	c	c	a	a	a
Δ	0	1	1.1	2.1	2.2	3.2	3.3	4.3
a	1	0	1	1.1	2.1	2.2	3.2	3.3
b	1.1	1	0	1	1.1	2.1	2.2	3.2
a	2.1	1.1	1	1	2	1.1	2.1	2.2
a	2.2	2.1	1.1	2	2	2	1.1	2.1
a	3.2	2.2	2.1	2.1	3	2	2	1.1

The second row has the values

$$\rho(a, \Delta), \quad \rho(a, a), \quad \rho(a, ab), \quad \rho(a, abc),\ldots$$

To illustrate:

$$\begin{aligned}\rho(ab, a) &= \min\{\rho(ab, \Delta) + w(T_{a-}),\\ &\qquad \rho(a, \Delta) + w(T_{(b,a)}),\\ &\qquad \rho(a, a) + w(T_{b-}), \rho(\Delta, a) + w(T_{ab-})\}\\ &= \min\{1.1 + 1, 1 + 1, 0 + 1, 1 + 1.1\} = 1\end{aligned}$$

TABLE 2

Single Deletion and Insertion Distance Calculation

	Δ	a	b	c	c	a	a	a
Δ	0	1	2	3	4	5	6	7
a	1	0	1	2	3	4	5	6
b	2	1	0	1	2	3	4	5
a	3	2	1	1	2	2	3	4
a	4	3	2	2	3	2	2	3
a	5	4	3	3	3	2	2	2

For comparison, Table 2 is a matrix for the distance if double deletions and insertions are not allowed. The double deletion distance is $\rho(\mathbf{a}, \mathbf{b}) = 1.1.$ (the value in the lower right-hand corner). The single deletion distance is $\bar{d}(\mathbf{a}, \mathbf{b}) = 2.$

Finally, we remark that the matrix of Table 1 gives an alignment or best matching between sequences in the same manner that Sellers obtains best matching in the single insertion and deletion case. Our alignment is:

a b	*c c*	*a a a*
a b	*Δ*	*a a a*

6. Connection with Information Theory

Reichert, Cohen, and Wong [5, 8, 9] have studied an application of information theory for determining the quality of alignment of two biosequences. Their papers [5, 8] are related to this work, although they do not obtain any mathematical results for their model. The transformations they consider are insertion, deletion, substitution, and unequal replacement.

Suppose one associates a probability p_i with each transformation T_i and with any sequence of transformations $T_{i_l}^{j_l} \cdots T_{i_1}^{j_1}$, the probability $\prod_{k=1}^{l} p_{i_k}$. Reichert *et al.* choose the element of $\{\mathbf{a} \to \mathbf{b}\}_\tau$ such that the associated information,

$$I_{a \to b} = -\sum_{k=1}^{l} \log p_{i_k}$$

is minimized, which is equivalent to maximizing $\prod_{k=1}^{l} p_{i_k}$. To associate this with what we have considered, take $w_i > 0$ (identity transformation is not allowed) and put

$$w_i = -\log p_i .$$

The distance between two sequences and the resulting alignment can then be computed as in Sections 3 and 4.

The model of Reichert *et al.* includes a location cost which is omitted in the above correspondence. It should be possible to change the weights of the transformations to include the consideration of location costs, but this has not been carried out. The model of Reichert *et al.* also

includes the concept of a "mutation machine" whose motion and action is specified by the sequence $T_{i_l}^{j_l} \cdots T_{i_1}^{j_1}$.

On page 45 of [8] multiple hits are excluded so that the set of possible alignments is finite. In our treatment, the concept of the path of a site through a sequence of transformations makes possible a mathematically precise definition of multiple hits. Then, in Theorem 4, these concepts are used to exclude such occurrences in our algorithm.

Finally, the efficiency of the algorithm in [8] is that, for insertions and deletions of any length, it runs in approximately $M^2N^2/4$ steps. Recall that our algorithm runs in approximately $MN \max\{M, N\}$ steps. If insertions and deletions of length $\alpha \ll \min\{M, N\}$ are allowed, our algorithm runs in approximately $(2\alpha + 1)\, MN$ steps. Reichert *et al.*, do not consider limiting the length of insertions and deletions, but, for that case their algorithm should run in approximately $\alpha^2\, MN$ steps.

7. Quasimetrics

In the introduction, metric measures of dissimilarity were presented. Of the three requirements for a metric measure of dissimilarity, perhaps the requirement of symmetry is the least realistic. The assumption that $\rho(s_1, s_2) = \rho(s_2, s_1)$ for all $s_1, s_2 \in S$ implies that evolutionary changes are reversible. Here it will be shown that the assumption of symmetry can be dropped in the previous work.

A quasimetric space $\rho(s_1, s_2)$ is a nonnegative function on $S \times S$ which satisfies

(1) $\rho(s_1, s_2) = 0$ if and only if $s_1 = s_2$

(2) $\rho(s_1, s_3) \leqslant \rho(s_1, s_2) + \rho(s_2, s_3)$ for all $s_1, s_2, s_3 \in S$.

To define a τ-quasimetric, let

$$\rho_1(\mathbf{a}, \mathbf{b}) = \min_{\{\mathbf{a}\to\mathbf{b}\}_\tau} \sum_{k=1}^{l} w(T_{i_k}).$$

Then it is easily seen that Theorem 1 becomes: Each equivalence class S_i of S together with ρ_1 is a quasimetric space.

The computation of $\rho_1(a, b)$ has already been handled by Theorem 3 and Theorem 4, which gives sufficient conditions for the hypotheses of Theorem 3 to be satisfied, does not require symmetry.

8. The n-distance

This section extends the notion of distance between two sequences to a distance among n sequences: the n-distance. An algorithm which computes this distance is given. The algorithm also gives the alignment of the n sequences which has least weight. These alignments can then be used in ancestral reconstruction in which an alignment is interpreted as a common ancestor of the n sequences. Evolutionary tree construction can be based on the method.

Suppose ρ_n maps A^n to the nonnegative real numbers and $\rho_n(\Delta, \Delta,..., \Delta) = 0$. Then extend ρ_n to n A-sequences by the formula:

$$\rho_n(\mathbf{a}^{(1)}, \mathbf{a}^{(2)},..., \mathbf{a}^{(n)}) = \sum_{i=1}^{\infty} \rho_n(a_i^{(1)}, a_i^{(2)},..., a_i^{(n)}).$$

Then ρ_n is extended to n evolutionary sequences by

$$\tilde{\rho}_n(\bar{\mathbf{a}}^{(1)}, \bar{\mathbf{a}}^{(2)},..., \bar{\mathbf{a}}^{(n)}) = \min \rho_n(\mathbf{a}^{(1)}, \mathbf{a}^{(2)},..., \mathbf{a}^{(n)})$$

$$= \min \sum_{i=1}^{\infty} \rho_n(a_i^{(1)}, a_i^{(2)},..., a_i^{(n)}),$$

where the minimum is taken over all $\mathbf{a}^{(1)}, \mathbf{a}^{(2)},..., \mathbf{a}^{(n)}$ in the respective equivalence classes. Due to the requirement of an A-sequence to have only a finite number of terms not equal to Δ and $\rho(\Delta, \Delta,..., \Delta) = 0$, we have the existence of the minimum. Also

$$0 \leqslant \tilde{\rho}_n(\bar{\mathbf{a}}^{(1)}, \bar{\mathbf{a}}^{(2)},..., \bar{\mathbf{a}}^{(n)}) < \infty,$$
$$\tilde{\rho}_n(\bar{\Delta}, \bar{\Delta},..., \bar{\Delta}) = 0.$$

For $1 \leqslant i \leqslant n$, consider

$$\mathbf{a}^{(i)} = a_1^{(i)} a_2^{(i)} \cdots a_{l_i}^{(i)} \Delta \Delta \cdots$$

where $a_{l_i}^{(i)} \neq \Delta$ and $\mathbf{a}^{(i)} \in \bar{S}$. Each such sequence in $\bar{S}$ will be represented by

$$a_1^{(i)} a_2^{(i)} \cdots a_{l_i}^{(i)}.$$

If $\mathbf{a}^{(i)} = \bar{\Delta}$, we represent $\mathbf{a}^{(i)}$ by Δ.

Then $\bar{\rho}_n$ can be defined on such sequences by

$$\bar{\rho}_n(a_1^{(1)} \cdots a_{l_1}^{(1)}, a_1^{(2)} \cdots a_{l_2}^{(2)}, \ldots, a_1^{(n)} \cdots a_{l_n}^{(n)})$$

$$= \bar{\rho}_n(\overline{a_1^{(1)} \cdots a_{l_1}^{(1)}}, \overline{a_1^{(2)} \cdots a_{l_2}^{(2)}}, \ldots, \overline{a_1^{(n)} \cdots a_{l_n}^{(n)}}).$$

The next theorem yields an algorithm to compute this number.

THEOREM 5. *The quantity*

$$\bar{\rho}_n(a_1^{(1)} \cdots a_{l_1}^{(1)}, a_1^{(2)} \cdots a_{l_2}^{(2)}, \ldots, a_1^{(n)} \cdots a_{l_n}^{(n)}),$$

where not all arguments are Δ, can be found by taking the minimum of the following $2^n - 1$ quantities:

$$\bar{\rho}_n(a_1^{(2)} \cdots a_{l_1-\epsilon_1}, a_1^{(2)} \cdots a_{l_2-\epsilon_2}^{(2)} \cdots, a_1^{(n)} \cdots a_{l_n-\epsilon_n}^{n}) + \rho_n(a_{l_1\cdot\epsilon_1}^{(1)}, a_{l_2\cdot\epsilon_2}^{(2)}, \ldots, a_{l_n\cdot\epsilon_n}^{(n)}) \tag{*}$$

where $\epsilon_i = 0$ or 1, $\epsilon_1 = \epsilon_2 = \cdots = \epsilon_n = 0$ is excluded, and $a_0^{(i)} \equiv \Delta$. If $a_1^{(i)} \cdots a_{l_i}^{(i)} = \Delta$, then omit computations of (*) which involve $\epsilon_i = 1$.

Proof. Since all sequences in $\bar{S}$ are Δ after a certain point, there is an $l > 0$ such that, for some $b_1^{(i)} \cdots b_l^{(i)} \in \overline{a_1^{(i)} \cdots a_{l_i}^{(i)}}$,

$$\bar{\rho}_n(a_1^{(1)} \cdots a_{l_1}^{(1)}, a_1^{(2)} \cdots a_{l_2}^{(2)}, \ldots, a_1^{(n)} \cdots a_{l_n}^{(n)})$$

$$= \rho_n(b_1^{(1)} \cdots b_l^{(1)}, b_1^{(2)} \cdots b_l^{(2)}, \ldots, b_1^{(n)} \cdots b_l^{(n)})$$

$$= \sum_{i=1}^{l-1} \rho_n(b_i^{(1)}, b_i^{(2)}, \ldots, b_i^{(n)}) + \rho_n(b_l^{(1)}, b_l^{(2)}, \ldots, b_l^{(n)})$$

$$= \bar{\rho}_n(b_1^{(1)} \cdots b_{l-1}^{(1)}, b_1^{(2)} \cdots b_{l-1}^{(2)}, \ldots, b_1^{(n)} \cdots b_{l-1}^{(n)}) + \rho_n(b_l^{(1)}, b_l^{(2)}, \ldots, b_l^{(n)}).$$

The last equality is true because if the sequences of length $l - 1$ were not minimal, then the sequences of length l would not be minimal.

Take l to be the smallest integer such that not all $b_l^{(1)}, \ldots, b_l^{(n)}$ are equal to Δ. Therefore the possibilities in our last equation are identical with the $2^n - 1$ possibilities described in Theorem 5. Moreover, each of $2^n - 1$ possible numbers in Theorem 5 is greater or equal to the minimum value. This completes the proof of the theorem.

For $n = 2$, there are $2^2 - 1 = 3$ possibilities:

$$\begin{aligned}\bar{\rho}_2(a_1 \cdots a_{l_1}, b_1 \cdots b_{l_2}) \\ = \min\{&\bar{\rho}_2(a_1 \cdots a_{l_1}, b_1 \cdots b_{l_2-1}) + \rho_2(\Delta, b_{l_2}), \\ &\bar{\rho}_2(a_1 \cdots a_{l_1-1}, b_1 \cdots b_{l_2-1}) + \rho_2(a_{l_1}, b_{l_2}), \\ &\bar{\rho}_2(a_1 \cdots a_{l_1-1}, b_1 \cdots b_{l_2}) + \rho_2(a_{l_1}, \Delta)\}.\end{aligned}$$

This is the algorithm given by Sellers.

In general we can use Theorem 5 to compute $\bar{\rho}_n(a_1^{(1)} \cdots a_{l_1}^{(1)}, \ldots, a_1^{(n)} \cdots a_{l_n}^{(n)})$ in $(l_1 + 1)(l_2 + 1) \cdots (l_n + 1)$ steps where each step consists in finding the minimum of (at most) $2^n - 1$ numbers. For implementation on computers we note that it is not necessary to store an n-dimensional array with dimensions $l_1 + 1$, $l_2 + 1, \ldots, l_n + 1$. Entries can be discarded after they will not be needed in computing future minima.

Before discussing the computation in more detail, consider the problem of defining ρ_3 on A^3. Let $A = \{a, b, c, d, \Delta\}$, and define ρ_3 on A^3 by

$$\rho_3(\alpha, \beta, \gamma) = \begin{cases} 0 & \text{if } \alpha = \beta = \gamma \\ 1 & \text{if exactly 2 of } \alpha, \beta, \gamma \text{ are equal} \\ 2 & \text{if none of } \alpha, \beta, \gamma \text{ are equal.} \end{cases}$$

This definition is motivated by

$$\rho_3(\alpha, \beta, \gamma) = \min_{\lambda \in A} \{d(\lambda, \alpha) + d(\lambda, \beta) + d(\lambda, \gamma)\}$$

where

$$d(x, y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{if } x = y \end{cases}.$$

The definitions of ρ_3 coincide and the second definition is the sum of the pairwise distances to the "nearest" $\lambda \in A$.

Let $\mathbf{a} = abcd$, $\mathbf{b} = bcc$, $\mathbf{c} = abc$. Below are the four matrices which compute $\bar{\rho}_3(\mathbf{a}, \mathbf{b}, \mathbf{c})$.

(i)

(Δ)	Δ	a	b	c	b
Δ	0	1	2	3	4
b	1	2	2	3	4
c	2	3	3	3	4
c	3	4	4	4	5

(ii)

(a)	Δ	a	b	c	b
Δ	1	1	2	3	4
b	2	1	2	3	4
c	3	2	3	3	4
c	4	3	4	4	5

(iii)

(b)	Δ	a	b	c	b
Δ	2	2	2	3	4
b	2	2	1	2	3
c	3	3	2	2	3
c	4	4	3	3	4

(iv)

(c)	Δ	a	b	c	b
Δ	3	3	3	3	4
b	3	3	2	2	3
c	3	3	2	1	2
c	4	4	3	2	3

Therefore $\bar{\rho}_3(\mathbf{a}, \mathbf{b}, \mathbf{c}) = 3$ which results from the tri-alignment

a	*b*	*c*	*b*
Δ	*b*	*c*	*c*
a	*b*	*c*	Δ

Of course, for this alignment,

$$\sum_{i=1}^{4} \rho_3(a_i^{(1)}, a_i^{(2)}, a_i^{(3)}) = 1 + 0 + 0 + 2 = 3.$$

The following scheme is used in reconstructing ancestral sequences. For each minimal alignment $a^{(1)}$, $a^{(2)}$, $a^{(3)}$ consider $a_i^{(1)}$, $a_i^{(2)}$, $a_i^{(3)}$. If at least two of $a_i^{(1)}$, $a_i^{(2)}$, $a_i^{(3)}$ are equal, then that is the ith element in the ancestral sequence. Otherwise all three are distinct and we use $\{a_i^{(1)}, a_i^{(2)}, a_i^{(3)}\}$ as the ith element in the ancestral sequence.

For our alignment, then, the ancestral sequence is

$$abc\{b, c, \Delta\}.$$

9. Conclusions

These new metrics should help in the investigation of the evolutionary relationship between two proteins by allowing more realistic evolutionary steps. It would be of interest to compare the new metrics for various lengths of insertions and deletions with existing metrics for a set of proteins. The problem of which metric and tree construction technique to use in the construction of evolutionary trees is a very difficult problem that may never be satisfactorily solved.

In connection with the construction of evolutionary trees it is possible that the method of reconstructing ancestral sequences with the 3-metric will be of value. We feel such an investigation should be carried out.

In conclusion, it should be remarked that one of the authors (M.S.W.) has used the multiple insertion and deletion metric as a tool to solve

the problem of prediction of RNA secondary structure. The secondary structure problem is quite distinct from the problems handled in this paper, but these metrics are fundamental in the solution. The work on secondary structures will appear elsewhere.

Note Added in Proof. Section 4 discusses multiple insertions and deletions. The notion of the path of site through a sequence of the more general transformations has been omitted by oversight. If multiple insertions or deletions overlap a site previously altered, then the transformations cannot be reordered and the argument given in Theorem 3 fails. To correct this alter the definition on p. 6, so that every A has a position number. Then there are no multiple hits if each position has been acted on or overlapped by at most one transformation.

References

1. W. A. Beyer, T. F. Smith, M. L. Stein, and S. M. Ulam, Metrics in biology, an introduction, Los Alamos Scientific Laboratory report, LA-4973, 1972.
2. W. A. Beyer, T. F. Smith, M. L. Stein, and S. M. Ulam, A molecular sequence metric and evolutionary trees, *Math. Biosci.* **19** (1974), 9–25.
3. M. O. Dayhoff *et al.*, "Atlas of Protein Sequence and Structure," National Biomedical Research Foundation, Silver Spring, Maryland.
4. N. Jardine and R. Sibson, "Mathematical Taxonomy," John Wiley and Sons, New York, 1971.
5. T. A. Reichert, D. N. Cohen, and A. K. C. Wong, An application of information theory to genetic mutations and the matching of polypeptide sequences, *J. Theor. Biol.* **42** (1973), 245–261.
6. P. H. Sellers, On the theory and computation of evolutionary distances, *SIAM J. Appl. Math.* **26** (1974), 787–793.
7. S. M. Ulam, Some ideas and prospects in biomathematics, *Ann. Rev. Biophys. Bioeng.* **1** (1972), 277–292.
8. A. K. C. Wong, T. A. Reichert, D. N. Cohen, and B. O. Aygun, A generalized method for matching informational macromolecular code sequences, *Comput. Biol. Med.* **4** (1974), 43–57.
9. D. N. Cohen, T. A. Reichert, and A. K. C. Wong, Matching code sequences utilizing context free quality measures, *Math. Biosci.* **24** (1975), 25–30.
10. W. M. Fitch, An improved method of testing for evolutionary homology, *J. Mol. Biol.* **16**, 9, (1966).
11. 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. Mol. Biol.* **48** (1970), 443–453.
12. D. Sankoff, Matching sequences under deletion/insertion constraints, *Proc. Nat. Acad. Sci.* **69**(1), (1972), 4–6.
13. J. E. Haber and D. E. Koshland, An evaluation of the relatedness of proteins based on comparison of amino acid sequences, *J. Mol. Biol.* **50** (1970), 617–639.
14. M. J. Sackin, Crossassociation: a method of comparing protein sequences, *Biochem. Genet.* **5** (1970), 287–313.

15. D. SANKOFF AND R. J. CEDERGREN, A test for nucleotide sequence homology, Technical Report 122, Centre de Recherches Mathématiques, Université de Montréal, Montreal, Canada.
16. A. J. GIBBS AND G. A. MCINTYRE, The diagram, a method for comparing sequences, *Eur. J. Biochem.* **16** (1970), 1–11.
17. W. M. FITCH, Locating gaps in amino acid sequences to optimize the homology between two proteins, *Biochem. Genet.* **3** (1969), 99.

Printed by the St Catherine Press Ltd., Tempelhof 37, Bruges, Belgium.