SIAM J. APPL. MATH.
Vol. 28, No. 1, January 1975

MINIMAL MUTATION TREES OF SEQUENCES*

DAVID SANKOFF†

Abstract. Given a finite tree, some of whose vertices are identified with given finite sequences, we show how to construct sequences for all the remaining vertices simultaneously, so as to minimize the total edge-length of the tree. Edge-length is calculated by a metric whose biological significance is the mutational distance between two sequences.

1. Introduction. The problem solved in this paper may be summarized as follows. Given a finite tree T, some of whose vertices are identified with points in a certain metric space (S, d), locate the remaining vertices in S so as to minimize the total length of the edges of T. For $S = \mathbf{R}^n$, this is a well-known generalization of Steiner's problem, but in the present paper S will be the set of finite sequences over some alphabet A, and for two such sequences $\mathbf{x} = (x(1), \cdots, x(n_x))$, $\mathbf{y} = (y(1), \cdots, y(n_y))$, if $n_x \leqq n_y$,

$$d(\mathbf{x}, \mathbf{y}) = n_x + n_y - \max_{\substack{0 \leqq \lambda \leqq n_x \\ 1 \leqq i_1 < \cdots < i_\lambda \leqq n_x \\ 1 \leqq j_1 < \cdots < j_\lambda \leqq n_y}} \sum_{k=1}^{\lambda} [1 + \delta(x(i_k), y(j_k))], \tag{1}$$

where $\delta(x(i), y(j)) = 1$ if $x(i)$ and $y(j)$ have the same value in A; otherwise $\delta(x(i), y(j)) = 0$.

The metric d arises in the study of molecular evolution as discussed by Ulam [1] and Sellers [2]. The integer $d(\mathbf{x}, \mathbf{y})$ equals the minimum number of *mutations* required to transform sequence $\mathbf{x}$ into $\mathbf{y}$, or $\mathbf{y}$ into $\mathbf{x}$, where a mutation may be either a change (replacement) of the value in A of a single term $x(i)$ to correspond with the value of some $y(j)$, or else the deletion from, or insertion into sequence $\mathbf{x}$, of a single term. E.g., $\mathbf{x} = (1, 1, 0, 1)$, $\mathbf{y} = (1, 0, 0, 1, 1)$, and $d(\mathbf{x}, \mathbf{y}) = 2$ as may be computed by changing $x(2)$ to 0 and inserting a 1 between $x(3)$ and $x(4)$.

For a given $\mathbf{x}$ and $\mathbf{y}$, we may write $d(i, j)$ for the distance between the subsequences consisting of the first i terms of $\mathbf{x}$ and the first j terms of $\mathbf{y}$. Setting $d(k, 0) = d(0, k) = k$ for $k \geqq 0$, it is not difficult to show that

$$d(i, j) = 1 + \min \begin{cases} d(i-1, j) \\ d(i-1, j-1) - \delta(x(i), y(j)) \\ d(i, j-1) \end{cases} \tag{2}$$

for $1 \leqq i \leqq n_x$ and $1 \leqq j \leqq n_y$. The use of recursions like (2) for comparing pairs of sequences has been explored by a number of authors [2]–[5]. Our main algorithm described in § 5 generalizes (2) for the simultaneous comparison of three or more sequences, each one identified with a vertex of a given tree.

To obtain this result, we prove, in the next two sections, two theorems about the total edge-length of a tree with vertices in S. In § 4, we describe a rapid method

* Received by the editors June 25, 1973.

† Mathematics Research Center, University of Montreal, Montreal, Quebec, Canada. This work was supported in part by a grant from the Ministry of Education, Government of Quebec.

for decomposing a tree with some of its vertices colored, into a minimum number of subtrees with disjoint vertex sets in the original tree, such that each subtree contains no two differently colored vertices. This method, the essentials of which are attributable to Fitch [6] and Hartigan [7], facilitates the calculation of the incremental term of the recursion which generalizes (2).

2. The frame sequence of a tree of sequences. In this section and the next we consider trees for which the locations of all vertices are specified in S; that is, all the sequences are given.

We shall use the notation $x(i)$ mainly to refer to the ith term, or position, in sequence $\mathbf{x}$, and not the value of this term in A, but if terms $x(i)$ and $y(j)$ from different sequences have the same value, we may write, without ambiguity, $x(i) = y(j)$.

Let $V(T) = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ be the vertex set of T, and $E(T)$ be the edge set. Suppose $\mathbf{xy} \in E(T)$. In definition (1) we can find a λ, and subsequences $i_1 < \cdots < i_\lambda$ and $j_1 < \cdots < j_\lambda$ which maximize the summation. We say that the sequence $M_{\mathbf{xy}} = \{(x(i_1), y(j_1)), \cdots, (x(i_\lambda), y(j_\lambda))\}$ is a maximal match of length λ. The pairs are unordered, so that each $M_{\mathbf{xy}}$ is also a $M_{\mathbf{yx}}$.

LEMMA 1. *Suppose* $(x(p), y(q)) \in M_{\mathbf{xy}}$ *and* $(x(r), y(s)) \in M_{\mathbf{xy}}$. *Then*

$$r < p \Leftrightarrow s < q.$$

Proof. For some h and k, $p = i_h$, $q = j_h$, $r = i_k$, $s = j_k$, and $r < p \Leftrightarrow i_k < i_h \Leftrightarrow k < h \Leftrightarrow j_k < j_h \Leftrightarrow s < q$.

Now, independently for each edge $\mathbf{xy} \in E(T)$, choose an $M_{\mathbf{xy}}$, and for each $\mathbf{x} \in V(T)$, for $1 \leqq i \leqq n_x$, define $\Omega(x(i))$ as follows:

$$\left.\begin{array}{l} x(i) \in \Omega(x(i)), \\ y(j) \in \Omega(x(i)) \text{ and } (z(k), y(j)) \in M_{\mathbf{zy}} \end{array}\right\} \Rightarrow z(k) \in \Omega(x(i)). \tag{3}$$

The sets $\Omega(x(i))$ are clearly pairwise disjoint or equal; hence they form a partition of $\Omega = \{x_1(1), \cdots, x_1(n_1), \cdots, x_N(1), \cdots, x_N(n_N)\}$.

THEOREM 1. *For a given set of maximal matches for* $E(T)$, *the different* $\Omega(x(i))$ *can be enumerated as* $\Omega_1, \cdots, \Omega_v$ *so that*

$$x(i) \in \Omega_h \quad \textit{and} \quad x(j) \in \Omega_k \quad \textit{implies} \quad i < j \Leftrightarrow h < k \tag{4}$$

for any $\mathbf{x} \in V(T)$.

Proof. Suppose $\Omega(x(i)) = \Omega(y(p))$. By (3), there must exist a sequence of edges in $E(T)$, $\mathbf{xu}_1, \mathbf{u}_1\mathbf{u}_2, \cdots, \mathbf{u}_t\mathbf{y}$, and pairs $(x(i), u_1(r_1)) \in M_{\mathbf{xu}_1}, \cdots, (u_t(r_t), y(p)) \in M_{\mathbf{u}_t\mathbf{y}}$. Similarly if $\Omega(x(j)) = \Omega(y(q))$, then for the same sequence of edges there must exist pairs $(x(j), u_1(s_1)) \in M_{\mathbf{xu}_1}, \cdots, (u_t(s_t), y(q)) \in M_{\mathbf{u}_t\mathbf{y}}$ since there is a unique sequence of edges in $E(T)$ between any two vertices $\mathbf{x}$ and $\mathbf{y}$ in $V(T)$. Repeated applications of Lemma 1 imply that $i < j \Leftrightarrow p < q$.

This assures that if different elements Λ and ψ of the partition both contain positions of one or more sequences in common, then for, say, $x(q_x) \in \Lambda$, $x(p_x) \in \psi$, either $q_x < p_x$ for all such $\mathbf{x}$ simultaneously, or $q_x > p_x$ for all such $\mathbf{x}$ simultaneously. In the first case we may write $\Lambda \ll \psi$, and in the second case $\psi \ll \Lambda$. If $\psi \ll \Lambda$ and $\Lambda \ll \theta$ by this definition, we also write $\psi \ll \theta$. It is easily verified that "$\ll$" is a partial ordering of the elements of the partition, satisfying $x(i) \in \Lambda$, $x(j) \in \psi$ implies $i < j \Leftrightarrow \Lambda \ll \psi$. But any finite partially ordered set may be

enumerated in a manner consistent with the partial order [8, p. 40]. This enumeration will then satisfy (4).

We call *any* $\Omega_1, \cdots, \Omega_\nu$ a *frame sequence* associated with the given set of maximal matches, as long as it is a partition satisfying (3) and (4).

3. The incremental function. Let $\Omega_1, \cdots, \Omega_\nu$ be a frame sequence associated with a given set of maximal matches for $E(T)$. For each $k = 1, \cdots, \nu$ we define a coloring of T as follows. If $x(i) \in \Omega_k$, color the vertex $\mathbf{x}$ with a color representing the value of $x(i)$ in A. (By Theorem 1 there can be at most one i for which $x(i) \in \Omega_k$.) If $x(i) \in \Omega_k$ for no $i = 1, \cdots, n_x$, then color the vertex with a color representing ϕ, where $\phi \notin A$. Then let $f(\Omega_k)$ be the number of edges in the tree whose two endpoints are colored differently (*bicolored edges*). We shall refer to f as the *incremental function.*

THEOREM 2. *Let* $\Omega_1, \cdots, \Omega_\nu$ *be a frame sequence associated with a given set of maximal matches for* $E(T)$. *Then the total edge-length of* T *is*

$$\sum_{\mathbf{xy} \in E(T)} d(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{\nu} f(\Omega_k). \tag{5}$$

Conversely, if $\Omega_1, \cdots, \Omega_\nu$ *is any partition satisfying* (4), *it is a frame sequence associated with some set of maximal matches for* $E(T)$ *if and only if* $\sum f(\Omega_k)$ *is minimal.*

Proof. If $M_{\mathbf{xy}}$ is of length λ, then by (1),

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) &= n_x + n_y - \sum_{k=1}^{\lambda} [1 + \delta(x(i_k), y(j_k))] \\ &= n_x - \lambda + n_y - \lambda + \sum_{k=1}^{\lambda} [1 - \delta(x(i_k), y(j_k))]. \end{aligned} \tag{6}$$

For how many h is $\mathbf{xy}$ bicolored in calculating $f(\Omega_h)$? There are three ways in which $\mathbf{xy}$ may be bicolored:

(i) $x(i_k) \neq y(j_k)$. Since $(x(i_k), y(j_k)) \in M_{\mathbf{xy}}$, these two terms are in the same Ω_h, by (3), and hence $\mathbf{xy}$ is bicolored in calculating $f(\Omega_h)$. There are clearly $\sum_{k=1}^{\lambda} [1 - \delta(x(i_k), y(j_k))]$ such pairs in $M_{\mathbf{xy}}$.

(ii) $(x(i), y(j)) \in M_{\mathbf{xy}}$ for no $j = 1, \cdots, n_y$. Then if $x(i) \in \Omega_h$, no $y(j) \in \Omega_h$ and hence $\mathbf{xy}$ is bicolored in calculating $f(\Omega_h)$, with colors representing $x(i)$ and ϕ, respectively. There are $n_x - \lambda$ such $x(i)$.

(iii) $(x(i), y(j)) \in M_{\mathbf{xy}}$ for no $i = 1, \cdots, n_x$. There are $n_y - \lambda$ such $y(j)$.

Summing all cases of type (i), (ii) and (iii), the edge $\mathbf{xy}$ contributes, by (6), exactly $d(\mathbf{x}, \mathbf{y})$ bicolored edges in computing $\sum f(\Omega_h)$. Summing over $E(T)$ proves (5).

Conversely, suppose $\Omega_1, \cdots, \Omega_\nu$ is a partition satisfying (4). Then for each $\mathbf{xy} \in E(T)$ we note which Ω_h contain both an $x(i)$ and a $y(j)$. This determines subsequences $1 \leqq i_1 < \cdots < i_\mu \leqq n_x$ and $1 \leqq j_1 < \cdots < j_\mu \leqq n_y$. If

$$n_x + n_y - \sum_{k=1}^{\mu} [1 - \delta(x(i_k), y(j_k))] = d(\mathbf{x}, \mathbf{y}),$$

then the pairs $(x(i_1), y(j_1)), \cdots, (x(i_\mu), y(j_\mu))$ constitute a maximal match. By (1), the only other possibility is

$$n_x + n_y - \sum_{k=1}^{\mu} [1 - \delta(x(i_k), y(j_k))] > d(\mathbf{x}, \mathbf{y})$$

in which case

$$\sum_{h=1}^{v} f(\Omega_h) > \sum_{\mathbf{xy} \in E(T)} d(\mathbf{x}, \mathbf{y}).$$

In other words, $\sum f(\Omega_h)$ is minimal if and only if $\Omega_1, \cdots, \Omega_v$ is associated with a set of maximal matches for $E(T)$.

Returning to our original problem, only part of $V(T)$, say $\mathbf{x}_1, \cdots, \mathbf{x}_{N'}$, consists of known sequences. In § 5 we shall show how to construct the remaining sequences $\mathbf{x}_{N'+1}, \cdots, \mathbf{x}_N$ by simultaneously constructing a partition $\Omega_1, \cdots, \Omega_v$ satisfying (4), such that $\sum f(\Omega_h)$ is minimal compared to any other set of sequences and any other such partition. By Theorem 2, this will solve the problem.

This procedure will involve being able to minimize $f(\Omega_h)$, given colors only for vertices $\mathbf{x}_1, \cdots, \mathbf{x}_{N'}$, and required to find an optimal assignment of colors for the remaining vertices. In the next section we present a rapid algorithm for accomplishing this.

4. Minimizing f. The problem of minimizing the incremental function may be stated as follows. For a tree T with some of its vertices colored, find a coloring for all the remaining vertices which minimizes the number of edges having two differently colored endpoints (bicolored edges). An equivalent formulation is to find a partition of the vertices of T among m subtrees, such that each subtree contains only vertices with a common color (plus uncolored vertices) and such that m is minimal.

If the removal of any edge $\mathbf{xy} \in E(T)$ decomposes T into two subtrees $T_{\mathbf{x}}$ and $T_{\mathbf{y}}$, where $\mathbf{x} \in V(T_{\mathbf{x}})$, $\mathbf{y} \in V(T_{\mathbf{y}})$, and where $V(T_{\mathbf{y}})$ contains no colored vertices, then it suffices to find a suitable coloring for $T_{\mathbf{x}}$ first, and then color all vertices in $V(T_{\mathbf{y}})$ by the color of $\mathbf{x}$. Thus we need only solve the coloring problem for trees where the colors are given for all terminal vertices at least.

We first define the *depth* of a vertex as follows. All colored vertices, including all terminal vertices, are of depth zero. An uncolored vertex having all incident edges, except at most one, connected to vertices of depth zero is defined to be of depth one. A vertex not of depth $0, 1, \cdots, D-1$ having all incident edges, except at most one, connected to vertices of depth $0, 1, \cdots$, or $D-1$ is defined to be of depth D.

LEMMA 2. *Every vertex in T has finite depth.*

Proof. Suppose there is a vertex $\mathbf{x}$ for which the depth is not defined. Then $\mathbf{x}$ must be collinear with at least two other vertices of undefined depth. These must each be collinear with one other such vertex besides $\mathbf{x}$, and so on. By the finiteness of T, there must exist a cycle of vertices of undefined depth, which is impossible since T contains no cycles. Therefore every vertex in T has a well-defined depth.

Next, we define the *candidate* colors for a vertex $\mathbf{x}$. If $\mathbf{x}$ has depth zero, there is a single candidate, namely its color. If $\mathbf{x}$ has depth $D > 0$, consider all vertices collinear with $\mathbf{x}$ and of lesser depth. Those colors which are candidates for a maximum number of these vertices are the candidate colors for $\mathbf{x}$.

THEOREM 3. *To minimize the number of bicolored edges, the following procedure suffices. Choose any vertex $\mathbf{x}$ of maximal depth, and color it by one of its*

candidate colors, say i. If **x** *is collinear with another vertex* **y** *of equal depth, i must be chosen from the intersection of their two sets of candidate colors, if this exists. For any other vertex* **y** *collinear with* **x**, *if i is a candidate color for* **y**, *then color* **y** *by i. Otherwise color* **y** *according to any of its candidate colors. After coloring all* **y** *collinear with* **x**, *consider any other vertex* **z** *collinear with any such* **y**. *If the color of* **y** *is a candidate for* **z**, *then color* **z** *accordingly. Otherwise color* **z** *by any of its candidate colors, and so on. When this process is exhausted, choose any remaining vertex of maximal depth, and repeat.*

Proof. We proceed by induction on D. Assume that for any coloring of all vertices of depth at least D, the procedure applied to the rest of the vertices produces a minimum of additional bicolored edges. Then consider any coloring of all vertices of depth at least $D + 1$, and let **x** be a vertex of depth D, valence v, and such that

$$v_1 = v_2 = \cdots = v_m > v_{m+1} \geqq \cdots$$

are the number of vertices of depth less than D, collinear with **x**, and with candidate colors $1, 2, \cdots$, respectively. Note that $1, \cdots, m$ are the candidate colors for **x**. By the induction hypothesis, for the aforementioned vertices collinear with **x**, it suffices to consider only colorings by their candidate colors. Now suppose the coloring of **x** by $m + 1$ could lead to a minimum of additional bicolored edges. Then, by the induction hypothesis, **x** will be incident to $v - v_{m+1} - 1$ or $v - v_{m+1}$ bicolored edges, depending on whether or not **x** is collinear with a vertex of depth greater than D colored $m + 1$. But $v - v_1 \leqq v - v_{m+1} - 1$ which means that 1 is also a minimizing color, as are $2, \cdots, m$. This stems from the fact implicit in the induction hypothesis and theorem statement that any coloring by candidate colors of the vertices collinear with **x** and of lesser depth, is as good as any other such coloring, except possibly with respect to the edges coincident with **x**.

Hence it suffices to consider only candidate colors for **x**. If **x** is collinear with a vertex of depth greater than D colored i, where i is a candidate for **x**, then i is the only minimizing color for **x** since $v - v_i - 1 < v - v_j$, for $j = 1, \cdots, m$. If, on the other hand, **x** is collinear with a vertex **y** of depth D, then any candidate color for both **x** and **y** is a minimizing color when applied to both.

Thus the procedure in the theorem statement produces a minimum number of additional bicolored edges when applied to all vertices of depth less than or equal to D. It remains to prove that for any coloring of all vertices of depth at least 2, the procedure applied to the vertices of depth 1 produces a minimum of bicolored edges. Suppose **x** is connected to

$$v_1 = v_2 = \cdots = v_m > v_{m+1} \geqq \cdots$$

vertices colored $1, 2, \cdots$, respectively (all of depth zero, of course), and possibly to one vertex **y** of depth greater than or equal to 1. If **y** is of depth greater than 1 and is colored $1, \cdots$, or m, or if **y** is of depth 1 and shares candidates with **x**, then clearly **x** and **y** must be of the same color to produce a minimum of bicolored edges. In all other cases, it is equally clear that any of $1, 2, \cdots, m$ will do for **x**.

Since an uncolored vertex may be connected to at most one vertex of greater or equal depth, by the definition of depth, no vertex will be colored more than

once by this procedure, and Lemma 2 ensures that all vertices will be colored at least once. This completes the proof.

5. The main algorithm. For a tree T with given sequences for vertices $\mathbf{x}_1, \cdots, \mathbf{x}_{N'}$, we are required to find sequences for the remaining vertices $\mathbf{x}_{N'+1}, \cdots, \mathbf{x}_N$ and a partition $\Omega_1, \cdots, \Omega_v$ satisfying (4) such that $\sum f(\Omega_k)$ is minimized. For each $k = 1, \cdots, v$, the input into the coloring procedure in the previous section consists of the colors for all $\mathbf{x}_i$, $1 \leqq i \leqq N'$. If for some j, $\mathbf{x}_i(j) \in \Omega_k$, then the color of $\mathbf{x}_i$ will represent the value of $x_i(j)$ in A. Otherwise $\mathbf{x}_i$ will be colored by ϕ. For $N' < i \leqq N$, on the other hand, the procedure *output* determines everything, even whether or not Ω_k should contain any position $x_i(j)$, depending on whether the vertex $\mathbf{x}_i$ is colored some color other than ϕ, or colored by ϕ, respectively. Let $\psi = \{x_1(1), \cdots, x_1(n_1), \cdots, x_{N'}(1), \cdots, x_{N'}(n_{N'})\}$ be the set of positions in $\mathbf{x}_1, \ldots, \mathbf{x}_{N'}$. To minimize $\sum f$ over all possibilities for sequences $\mathbf{x}_{N'+1}, \cdots, \mathbf{x}_N$ and all partitions of Ω satisfying (4), first consider any partition of ψ, say $\psi_1, \cdots, \psi_v$, satisfying (4), apply Theorem 3 for $k = 1, \cdots, v$ to calculate optimal colorings, and define, for each $\mathbf{x}_i$, where $N' < i \leqq N$, the value of $x_i(j)$ to be the color of vertex $\mathbf{x}_i$ in the jth coloring in which it is not ϕ. Then define Ω_k to contain all the elements of ψ_k plus all positions of the variable vertices colored non-ϕ in the kth optimal coloring. Then $\Omega_1, \cdots, \Omega_v$ satisfies (4) and no other partition $\Omega'_1, \cdots, \Omega'_v$, where $\Omega'_k \supseteq \psi_k$ for $k = 1, \cdots, v$ could have a lesser $\sum f(\Omega'_k)$, without contradicting Theorem 3. Therefore, instead of examining all partitions of Ω satisfying (4) to find a minimum, it suffices to examine only all partitions of ψ which satisfy (4).

Let $\mathbf{j} = (j_1, \cdots, j_{N'})$ and $\mathbf{e}$ be any N'-vector of zeros and ones, where $0 \leqq e_i \leqq j_i \leqq n_i$ for $i = 1, \cdots, N'$. Let $f(\mathbf{j} \cdot \mathbf{e})$ be the number of bicolored edges determined by Theorem 3 when vertex $\mathbf{x}_i$ is colored by $x_i(j_i)$ if $e_i = 1$, and by ϕ if $e_i = 0$. Then writing $d_T(\mathbf{j})$ for $\min \sum_{E(T)} d(\mathbf{x}, \mathbf{y})$ when only the first j_i terms of sequence $\mathbf{x}_i$ are considered, $i = 1, \cdots, N'$, and setting $d_T(\mathbf{0}) = 0$, we have the following theorem.

THEOREM 4.

$$d_T(\mathbf{j}) = \min_{\mathbf{e}} \{f(\mathbf{j} \cdot \mathbf{e}) + d_T(\mathbf{j} - \mathbf{e})\}. \tag{7}$$

Proof. For some frame sequence $\Omega_1, \cdots, \Omega_v$ for the given and the variable sequences, $d_T(\mathbf{j}) = f(\Omega_v) + \sum_{k=1}^{v-1} f(\Omega_k)$.

Now, Ω_v can contain only positions of the form $x_i(j_i)$, $1 \leqq i \leqq N'$, plus variable sequence positions; for if $x_i(h) \in \Omega_v$, $h < j_i$, then $x_i(j_i) \in \Omega_k$, where $k < v$, contradicting (4). If $x_i(j_i) \in \Omega_v$ set $e_i = 1$, otherwise $e_i = 0$, for $i = 1, \cdots, N'$. Then $f(\Omega_v) = f(\mathbf{j} \cdot \mathbf{e})$.

Now the partition $\Omega_1, \cdots, \Omega_{v-1}$ is either a frame sequence for $d_T(\mathbf{j} - \mathbf{e})$, or else $\sum_{k=1}^{v-1} f(\Omega_k) > d_T(\mathbf{j} - \mathbf{e})$, by the converse of Theorem 2. If the inequality holds, then for any frame sequence $\Omega'_1, \cdots, \Omega'_\mu$ of $d_T(\mathbf{j} - \mathbf{e})$,

$$\begin{aligned} d_T(\mathbf{j}) &= f(\Omega_v) + \sum_{k=1}^{v-1} f(\Omega_k) \\ &> f(\Omega_v) + \sum_{k=1}^{\mu} f(\Omega'_k) \end{aligned}$$

which contradicts the minimality of $\sum_{k=1}^{v} f(\Omega_k)$ and hence the fact that $\Omega_1, \cdots, \Omega_v$ is a frame sequence for $d_T(\mathbf{j})$.

Therefore,

$$d_T(\mathbf{j}) = f(\mathbf{j} \cdot \mathbf{e}) + d_T(\mathbf{j} - \mathbf{e}),$$

and for no $\mathbf{e}'$ could $f(\mathbf{j} \cdot \mathbf{e}') + d_T(\mathbf{j} - \mathbf{e}')$ be less than $d_T(\mathbf{j})$ without contradicting either or both of Theorem 2 and 3. Hence (7) follows.

MAIN ALGORITHM. Enumerate the vectors

$$\mathbf{0} = (0, \cdots, 0), \cdots, \mathbf{n} = (n_1, \cdots, n_{N'})$$

so that each component is nondecreasing with the enumeration. Setting $d_T(\mathbf{0}) = 0$, we can calculate $d_T(\mathbf{j})$ for each successive vector in the enumeration by searching each vector $\mathbf{e}$ consisting of zeros and ones, $0 \leqq e_i \leqq j_i \leqq n_i$, for

$$\min_{\mathbf{e}} \{f(\mathbf{j} \cdot \mathbf{e}) + d_T(\mathbf{j} - \mathbf{e})\},$$

where each such $d_T(\mathbf{j} - \mathbf{e})$ has already been calculated, thanks to the componentwise nondecreasing property of the enumeration.

Once $d_T(\mathbf{n})$ has been calculated, we find a frame sequence and hence the variable sequences as follows. Search for an $\mathbf{e}$ satisfying Theorem 4 when $\mathbf{j} = \mathbf{n}$. This defines ψ_v, the last element of some optimal partition of ψ, and hence Ω_v, the last term of a frame sequence for $d_T(\mathbf{n})$. Setting $\mathbf{j} = \mathbf{n} - \mathbf{e}$, we search for an $\mathbf{e}'$ satisfying the theorem, and so on, each step providing an additional term of the frame sequence.

Both the calculation of d_T and the construction of the frame sequence require a number of steps proportional to $(2^{N'} - 1)n_1 \cdots n_{N'}$ since there are $2^{N'} - 1$ vectors $\mathbf{e}$ to search for most $\mathbf{j}$. This figure is approximately $(2n)^{N'}$ if all the sequences have about n terms.

6. Applications. The construction we have developed has immediate application to problems of macromolecular evolution. The given sequences are the nucleotide base sequences of homologous RNA or DNA molecules from different species, and T represents the phylogenetic (family tree) relationships of these species. The variable vertices of T represent hypothetical ancestral species, so that the solution of the minimal mutation tree constitutes a probable reconstruction of the RNA or DNA molecules of these ancestors.

The main problem in practical applications is the strong dependence of computing time and memory on N', the number of given sequences. There are two ways of evading this problem. First, it is frequently a biologically reasonable assumption that $d_T(\mathbf{j})$ need not be computed for any $\mathbf{j}$ for which $|j_k - j_h| > C$, for any k, h and some suitable C. Second we can restrict ourselves to locally optimal solutions by using only the $N' = 3$ version of the algorithm, analogous to the Steiner three-point problem with one unknown vertex. See [9] for the type of results obtained by these methods.

7. Acknowledgments. I am indebted to R. J. Cedergren for discussions of the biological significance of this problem, to P. H. Sellers for his encouragement and for making available his manuscript [2], to Anton Kotzig for a discussion of the coloring problem in § 4, and to Cristiane Morel for computer implementations of the $N' = 2$ and $N' = 3$ cases of the algorithm (n up to 120).

REFERENCES

[1] 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, pp. 1–10.

[2] P. H. SELLERS, *An algorithm for the distance between two sequences*, J. Comb. Theory, 16 (1974), pp. 253–258.

[3] 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 Biol., 48 (1970), pp. 443–453.

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

[5] D. SANKOFF AND P. H. SELLERS, *Shortcuts, diversions and maximal chains in partially ordered sets*, Discrete Math., 4 (1973), pp. 287–293.

[6] W. M. FITCH, *Towards defining the course of evolution: Minimum change for a specific tree topology*, Systematic Zoology, 20 (1971), pp. 406–416.

[7] J. A. HARTIGAN, *Minimum mutation fits to a given tree*, Biometrics, 29 (1973), pp. 53–65.

[8] G. BIRKHOFF AND T. C. BARTREE, *Modern Applied Algebra*, McGraw-Hill, New York, 1970.

[9] D. SANKOFF, C. MOREL AND R. J. CEDERGREN, *Evolution of 5S RNA and the non-randomness of base replacement*, Nature New Biology, 245 (1973), pp. 232–234.