Mathematical Programming 9 (1975) 240–246.
North-Holland Publishing Company

LOCATING THE VERTICES OF A STEINER TREE IN AN ARBITRARY METRIC SPACE

David SANKOFF and Pascale ROUSSEAU
Université de Montréal, Montréal, Canada

Received 20 March 1974
Revised manuscript received 22 March 1975

Given a tree each of whose terminal vertices is associated with a given point in a compact metric space, the problem is to optimally associate a point in this space to each nonterminal vertex of the tree. The optimality criterion is the minimization of the sum of the lengths, in the metric space, over all edges of the tree. This note shows how a dynamic programming solution to this problem generalizes a number of previously published algorithms in diverse metric spaces, each of which has direct and significant applications to biological systematics or evolutionary theory.

Given T a tree with vertex set $V(T) = \{X_1, ..., X_m, Y_1, ..., Y_n\}$ and edge set $E(T)$. Let (S, d) be a compact metric space where to each vertex $X_i \in V(T)$ is associated a given point $x_i \in S$. The problem is to associate to each $Y_j \in V(T)$ some point $y_j \in S$ so as to minimize the *edge-length* of T: $\Sigma_{WZ \in E(T)} d(w, z)$, where $w, z \in S$ are associated with vertices $W, Z \in V(T)$. This arises in connection with the larger and generally intractable Steiner problem, in which the only information given is the position of $x_1, ..., x_m \in S$ and where both the structure of T *and* the positions of the $y_1, ..., y_n \in S$ must be determined. This type of problem arises frequently in numerical taxonomy where the X_i represent different organisms, the x_i their positions in some space (S, d) of characters or features, and the Y_j represent hypothetical ancestral organisms. The minimality criterion corresponds to the economy or likelihood of the evolutionary explanation represented by the tree T.

Here we present a dynamic programming solution for the more restricted problem of finding the y_i's, given T. Though a solution to a full Steiner problem can always be considered to involve a tree with vertices of degree one or three only (allowing for edges of length zero), our method also applies in contexts where the given tree has vertices of higher degree. Our formulation of the problem and its solution subsumes and generalizes a number of previously published versions, in particular metric spaces, and these will furnish some of our examples.

It suffices to consider the case where the X_i are all terminal vertices (i.e. of degree one) of T, and the Y_j are all non-terminal (degree $\geqslant 3$). Other cases are easily reducible to this one, e.g. by decomposing T into the subtrees coincident with any non-terminal X_i and optimizing each one separately, and by constraining y_j for any Y_j of degree <3 to take on the same value as one of its neighbours.

Choose any Y_r, $r = 1, ..., n$ to be the *root* of the tree. Then for any vertex Z of T, all vertices on the unique path between Z and Y_r, including Z and Y_r, are said to *dominate* Z. The vertices dominated by any vertex Z determine the subtree T_Z dominated by Z. Those vertices $Z_1, ..., Z_{p(Y)}$ which are both dominated by Y and share an edge with Y are said to be *immediately* dominated by Y.

We then construct a number of functions f_Z on S, one for each vertex $Z \in V(T)$. For $i = 1, ..., m$ we set $f_{X_i}(x_i) = 0$ and $f_{X_i}(x) \equiv \infty$ if $x \neq x_i$. The f_Y are then defined so that $f_Y(x)$ is the minimal edge-length of T_Y given that Y is associated with x. From the principle of optimality it follows that

$$f_Y(x) = \min_{(z_1, ..., z_{p(Y)})} \sum_{i=1}^{p(Y)} [f_{Z_i}(z_i) + d(z_i, x)].$$

Then $\min_{x \in S} f_{Y_r}(x) = f_{Y_r}(y_r^*)$ is the minimal edge-length of T and the usual backtrack process starting from Y_r determines optimum locations y_i^* of all Y_i.

It might seem that in some metric spaces, this dynamic programming approach would be of little use, due to the difficulty or impossibility of computing the f_Y explicitly. However, in a diverse selection of spaces, such as those described below, it can be shown that the f_Y must be calculated only over small subsets of S, leading to a feasible and rapid algorithm.

The Manhattan metric space

$$S = \mathbf{R}^N, \qquad d(u, v) = \sum_{J=1}^{N} |u(J) - v(J)|.$$

In this space, the coordinates $y_1(J), ..., y_n(J)$ may be found separately for each $J = 1, ..., N$, so that it suffices to consider $S = \mathbf{R}^1$. Suppose Y_i dominates immediately $X_1, ..., X_p$ where $x_1 < ... < x_p$. If p is even, set $q = \frac{1}{2}p - 1$, and $I(Y_i) = [x_{\frac{1}{2}p}, x_{\frac{1}{2}p+1}]$. If p is odd set $q = \frac{1}{2}(p-1) - 1$, and $I(Y_i) = \{x_{\frac{1}{2}(p+1)}\}$. Then for $y \in I(Y_i)$

$$f_{Y_i}(y) = \sum_{j=0}^{q} (x_{p-j} - x_{j+1})$$

and it will not be necessary to calculate f_{Y_i} elsewhere. Now suppose Y_i immediately dominates $Z_1, ..., Z_p$. Let $r_1 \leqslant ... \leqslant r_p$ be the right-hand end-points of $I(Z_{u(1)}), ..., I(Z_{u(p)})$, and $t_1 \leqslant ... \leqslant t_p$ the left-hand end-points of $I(Z_{v(1)}), ..., I(Z_{v(p)})$, where u and v are suitable permutations of $(1, ..., p)$. If $I(Z_k)$ is a single point, it is listed both as a t and an r. Suppose $r_1 < t_p$, $r_2 < t_{p-1}, ..., r_{1+q} < t_{p-q}$, but the remaining r_j are all greater than or equal to the remaining t_k. Define

$$I(Y_i) = [t_{p-q-1}, r_{q+2}]$$

and for $y \in I(Y_i)$

$$f_{Y_i}(y) = \sum_{j=0}^{q} [(t_{p-j} - r_{j+1}) + f_{Z_{u(j+1)}}(r_{j+1}) + f_{Z_{v(p-j)}}(t_{p-j})] + \sum_{j=q+1}^{p-q-2} f_{Z_{u(j+1)}}(y).$$

The y_i^* can then be chosen, since they will all be in the $I(Y_i)$, $i = 1, ..., n$. A version of this algorithm, specific to the case where all non-terminal vertices have degree 3, was published by Farris [2], and is routinely applied in the study of evolution using continuous characters. Other results on Steiner trees in Manhattan space of dimension two are given in [5,7].

The space of qualitative characters

$$S = \{1, ..., N\}, \qquad d(I, J) = 1 \text{ if } I \neq J.$$

The case of 3-valent non-terminal nodes was investigated by Fitch [3], and the more general case by Hartigan [6]. Suppose Y_i dominates $X_1, ..., X_p$. If $J \in S$ occurs as frequently or more frequently than any other $K \in S$ among $x_1, ..., x_p$, say α times,

$$f_{Y_i}(J) = p - \alpha.$$

If $K \in S$ occurs $\alpha - 1$ times,

$$f_{Y_i}(K) = p - \alpha + 1$$

and it will not be necessary to calculate the remaining values of f_{Y_i}.We say J is a best value and K a next best value.

Consider a Y_i which dominates just $Z_1, ..., Z_p$. Suppose J is a best value most frequently among the $Z_1, ..., Z_p$, say α times. Then

$$f_{Y_i}(J) = p - \alpha$$

and if K is a best value only $\alpha - 1$ times, then

$$f_{Y_i}(K) = p - \alpha + 1.$$

Then for each Y_i it will be possible to choose y_i^* from among the best or next best values.

This algorithm is widely used in studying protein and DNA evolution and has been somewhat generalized in [11].

A space of finite sequences

Let S be the set of $(N-1)$-ary seqnences, and for $\boldsymbol{a} = (a(1), ..., a(r))$, $\boldsymbol{b} = (b(1), ..., b(s))$, where $r \leqslant s$

$$d(\boldsymbol{a}, \boldsymbol{b}) = r + s - \max_{\substack{0 \leqslant \lambda \leqslant r \\ 1 \leqslant i_1 < ... < i_\lambda \leqslant r \\ 1 \leqslant j_1 < ... < j_\lambda \leqslant s}} \sum_{k=1}^{\lambda} [1 + \delta(a(i_k), b(j_k))],$$

where $\delta(u, v) = 1$ if $u = v$, and $\delta(u, v) = 0$ otherwise. This metric, representing the *mutational distance* between the two sequences, arises in the study of molecular evolution as discussed by Ulam [16] and Sellers [15]. See [17] for another interpretation of this metric. In [12] we show that the search for $y_1^*, ..., y_n^*$ can be reduced to a large number (about $(2r)^m$, where r is the length of the longest sequence among the $\boldsymbol{x}_i = (x_i(1), ..., x_i(r_i))$, $i = 1, ..., m$) of applications of the algorithm in the simple metric space ($S = \{1, ..., N\}$, $d(I, J) = 1 - \delta(I, J)$) described in the preceding section. An application of these methods to infer the structure of RNA in ancestral organisms is presented in [13, 14].

Euclidean space

$$S = \mathbf{R}^N, \qquad d(u, v) = \left(\sum_{j=1}^{N} (u(J) - v(J))^2 \right)^{1/2}.$$

It is, of course, in this metric space, especially the case $N = 2$, that the Steiner problem is classically posed. Making use of the fact that in a Steiner tree the Y_i must all be of degree 3 and the y_i must all be vertices of three angles of $\frac{2}{3}\pi$ radians, Melzak [9, 10] and Gilbert and Pollak [4] have given an inductive algorithm which produces at most 2^n possible configurations for the set of n non-terminal vertices of a given tree topology.

For arbitrary T, however, where the Y_i may have degree >3, no non-iterative algorithm is known. For $n = 1$, the problem is a version of the Weber problem, or Fermat's problem and Kuhn [8] has shown that a well-known gradient-based iteration converges to the unique solution for almost all starting approximations. This algorithm can be extended to apply to our problem when $n > 1$, though proofs of convergence would seem to be more difficult.

The dynamic programming formulation can also be employed in an iterative manner in Euclidean space. Although not as simple as the gradient method, we sketch briefly the procedure for $N = 2$, for the sake of comparison with the other metric spaces we have studied.

The principle of the algorithm is to find regions R_i which necessarily contain the y_i^*, $i = 1, ..., n$, and to shrink these regions as much as possible in successive iterations.

(0) Divide a rectangle R containing all of the x_i, $i = 1, ..., m$, into squares of side $\sqrt{2}s$, and set $R_1 = ... = R_n = R$.

(i) For Y dominating only terminal nodes $X_{k_1}, ..., X_{k_p}$, $\bar{f}_Y$ and $\underline{f}_Y$ are calculated just at the center point c of each square as follows

$$\bar{f}_Y(c) = \sum_{j=1}^{p} d(x_{k_j}, c) + ps$$

$$\underline{f}_Y(c) = \sum_{j=1}^{p} d(x_{k_j}, c) - ps.$$

For Y immediately dominating $Z_1, ..., Z_p$, set

$$\bar{f}_Y(c) = \min_{(c_1, \dots, c_p)} \sum_{k=1}^{p} [\bar{f}_{Z_k}(c_k) + d(c_k, c)] + 2ps,$$

$$\underline{f}_Y(c) = \min_{(c_1, \dots, c_p)} \sum_{k=1}^{p} [\underline{f}_{Z_k}(c_k) + d(c_k, c)] - 2ps,$$

where the minima are taken over all those c_i falling in the appropriate R_j. The functions $\bar{f}_Y$ and $\underline{f}_Y$ represent upper and lower bounds for f_Y valid not only at c but throughout the square of which c is the center.

(ii) All squares for which $\underline{f}_{Y_r}(c) > \min_c \bar{f}_{Y_r}(c)$ could not possibly contain y_r^*, i.e., a y which minimizes f_{Y_r}, so they are ignored in all further calculations of $\bar{f}_{Y_r}$ and $\underline{f}_{Y_r}$. Let R_r' be the remaining region.

(iii) If Y_j immediately dominates Y_k and a new region R_j' has already been delimited, then define R_k' as follows. Any square $\in R_k$ with center c for which

$$\underline{f}_{Y_k}(c) + \min_{e \in R_j'} d(e, c) - 2s > \min_{c \in R_k} [\bar{f}_{Y_k}(c) + \max_{e \in R_j'} d(e, c) + 2s]$$

could not contain y_k^*. The remainder constitute R_k'. This step is repeated until all Y_i are exhausted, $i = 1, \dots, n$.

(iv) Redefine R_i to be R_i' for $i = 1, \dots, n$. If all the R_i are now smaller than some critical area, stop. Otherwise divide the R_i into squares of side $\sqrt{2}s$, where s is suitably smaller than in the present cycle, and return to step (i).

See [1] for a biological interpretation of Steiner trees in Euclidean space.

References

[1] L.L. Cavalli-Sforza and A.W.F. Edwards, "Phylogenetic analysis: models and estimation procedures", *American Journal of Human Genetics* 19 (1967) 233–257.

[2] J.S. Farris, "Methods for computing Wagner trees", *Systematic Zoology* 19 (1970) 83–92.

[3] W.M. Fitch, "Towards defining the course of evolution: minimum change for a specific tree topology", *Systematic Zoology* 20 (1971) 406–416.

[4] E.N. Gilbert and H.O. Pollak, "Steiner minimal trees", *SIAM Journal on Applied Mathematics* 16 (1968) 1–29.

[5] M. Hanan, "On Steiner's problem with rectilinear distance", *SIAM Journal on Applied Mathematics* 14 (1966) 255–265.

[6] J.A. Hartigan, "Minimum mutation fits to a given tree", *Biometrics* 29 (1973) 53–65.

[7] F.K. Hwang, "On Steiner minimal tree with rectilinear distance", manuscript.

[8] H.W. Kuhn, "A note on Fermat's problem", *Mathematical Programming* 4 (1973) 98–107.

[9] Z.A. Melzak, "On the problem of Steiner", *Canadian Mathematical Bulletin* 4 (1961) 143–148.

[10] Z.A. Melzak, *Companion to concrete mathematics* (Wiley, New York, 1973) Chapter 4, Section 3.

[11] G.W. Moore, J. Barnabas and M. Goodman, "A method for constructing maximum parsimony ancestral amino acid sequences on a given network", *Journal of Theoretical Biology* 38 (1973) 459–485.

[12] D. Sankoff, "Minimal mutation trees of sequences", *SIAM Journal on Applied Mathematics* 28 (1975) 35–42.

[13] D. Sankoff, R.J. Cedergren and G. Lapalme, "Frequency of insertion-deletion, transversion and transition in the evolution of 5S ribosomal RNA", *Journal of Molecular Evolution,* to appear.

[14] D. Sankoff, C. Morel and R.J. Cedergren, "Evolution of 5S RNA and the non-randomness of base replacement", *Nature New Biology* 245 (1973) 232–234.

[15] P.H. Sellers, "An algorithm for the distance between two sequences", *Journal of Combinatorial Theory* 16 (1974) 253–258.

[16] S.M. Ulam, "Some combinatorial problems studied experimentally on computing machines", in: S.K. Zaremba, ed., *Applications of number theory to numerical analysis* (Academic Press, New York, 1973) pp. 1–10.

[17] R.A. Wagner and M.J. Fischer, "The string-to-string correction problem", *Journal of the Association for Computing Machinery* 21 (1974) 168–173.