ADVANCES IN APPLIED MATHEMATICS 3, 43–49 (1982)

The Steiner Problem in Phylogeny Is NP-Complete

L. R. FOULDS

University of Canterbury, Christchurch, New Zealand

AND

R. L. GRAHAM

Bell Laboratories, Murray Hill, New Jersey 07974

The problem of determining a phylogeny (evolutionary tree) for a given set of species using protein sequences is introduced and defined as the Steiner problem in phylogeny (SPP). In this note we show that the SPP is *NP*-complete, even when restricted to the special case of just two amino acid triples (in which case the SPP is the ordinary Steiner problem in $\{0, 1\}^n$). This reinforces the recent emphasis on the development of heuristic techniques for the problem.

INTRODUCTION

According to current theories of evolution, existing biological species have been linked in the past by common ancestors. Following the Darwinian school, many scientists have represented postulated ancestral relationships by trees, called *phylogenies*. The ancestors of certain groups of species, such as vertebrates, have left a rich fossil record of their existence which can be used to make comparisons with similar existing species. This has led to a fair degree of agreement on the structure of phylogenies of these groups. Unfortunately, for most groups the record is inadequate and in some cases, unknown or nonexistent. In these cases there is often considerable disagreement over the nature of the phylogenies which describe their histories.

Over the past two decades attempts have been made to overcome these problems by using techniques which construct tentative phylogenies from protein sequence data rather than using classical numerical taxonomy. (Some of these methods are discussed in [1, 3, 11, 14].) These methods typically construct a phylogeny for a particular set of species given a unique protein sequence for each of the member species. It is assumed that all the sequences represent the same protein (typically the respiratory protein

0196-8858/82/010043-07$05.00/0

cytochrome c, hemoglobin α or β, or fibrinopeptide α or β) and are of the same length. The symbols in each sequence normally represent amino acids. Each amino acid can be represented by an ordered triple of nucleotides. Since there are four nucleotides, A, C, G and U, there are a priori 64 possible triples. However, only 20 of these have been found to occur in nature. Thus, each sequence consists of a string of symbols from an alphabet of size 20.

Several workers have found it advantageous to convert amino acid sequences into nucleotide sequences for the construction of phylogenies. This is because differences between different pairs of amino acids can vary whereas differences between different pairs of nucleotides can conveniently all be assigned unit weight. (Descriptions of this process can be found in [3, 11, 13] and especially in Watson [14].) Henceforth, we will assume that the data have undergone such a transformation.

The basic objective of this approach is to construct a phylogeny in which each given species and its sequence is represented. It is usual first to construct an unrooted phylogeny which does not have the point representing the common ancestor of all the given species distinguished. A common ancestor is then specified by *directing* the phylogeny, i.e., by giving each link in the tree an orientation directed away from the common ancestral point (e.g., see [2, 10, 12]).

In this note we will be concerned with the construction of undirected phylogenies. The endpoints of each link in the phylogeny represent nucleotide sequences which can be examined at each site for differences. The number of sites at which differences occur is associated with the link. A commonly used optimality criterion (which we also use) is to minimize the sum of these numbers taken over all links. A phylogeny selected under this criterion is said to be of *maximum parsimony*.

We remark that it is not assumed that the evolutionary history of the given species necessarily followed the path laid out by the phylogeny of maximum parsimony. This tree is merely a minimal solution to an extremal problem in this model, a criterion which is often used to describe natural phenomena. In the next section we make these notions more precise.

Preliminaries

For a metric space (S, d), define a weighted graph[1] $G = G(S, d)$ with vertex set S so that each edge $\{s, t\}$ has weight $d(s, t)$. For a finite subset $X \subseteq S$, a *minimum spanning tree* $T(X)$ for X is a tree (i.e., connected, acyclic subgraph) with vertex set X such that the sum of edge weights of $T(X)$ is a minimum. Finally, a Steiner minimal tree $S(X)$ for X is a tree having the

[1] For undefined graph-theoretic terminology, see [9].

minimum possible length over all trees in G which contain X in their vertex sets.

It is well known that for arbitrary weighted graphs, finding a Steiner minimal tree (SMT) is in general, an *NP*-complete problem (see [6] for a discussion of these concepts). More recently, it has been shown that for graphs whose edge weights come from certain metric structures, such as the Euclidean plane or the L_1 plane, finding SMTs is also *NP*-complete (see [5, 7]).

The problem we are considering, i.e., that of constructing phylogenies, is easily seen to have the following formalization: For a fixed alphabet A, let d denote the Hamming distance on A^N, i.e., $d((a_1,\dots,a_N), (a'_1,\dots,a'_N))$ is equal to be the number of indices i such that $a_i \neq a'_i$. In the metric space (A^N, d), the Steiner problem for phylogeny (SPP) is:

(SPP): Given a set $X \subseteq A^N$, find a Steiner minimal tree $S(X)$ for X.

What we will show in the next section is that even when A consists of two elements, the SPP for A^N is *NP*-complete.[2]

The Main Result

Let $A = \{0, 1\}$. For a fixed positive integer N, denote A^N by Q_N. The graph $G = G(Q_N, d)$ is just the 1-skeleton of the N-cube. To show that the Steiner problem for Q_N is *NP*-complete (which we will denote by SPQ), we will reduce the known [6] *NP*-complete problem Exact 3-Cover to SPQ. A general instance of Exact 3-Cover (X3C) has the following form:

INPUT: $\mathcal{F} = \{F_1, F_2,\dots,F_n\}$, where $|F_i| = 3$ and
$F_i \subseteq \{1, 2,\dots,3m\} \equiv I_{3m}$, $1 \le i \le n$;

X3C: Does $\mathcal{F}$ contain m sets $F_{i_1},\dots,F_{i_m}$ whose union is I_{3m}?

Note that if $\mathcal{F}$ does contain such F_{i_k} then they must be disjoint.

We now give the details for the construction of the desired corresponding instance of SPQ. To begin with we set $N = 4m(n + 3m + 1)$. A point $q = (q_1,\dots,q_N) \in Q_N$ can be thought of as consisting of $n + 3m + 1$ blocks, each of length $4m$:

$$q = (X_0, X_1,\dots,X_{3m}; Y_1,\dots,Y_n). \tag{1}$$

To each integer $i, 0 \le i \le 3m$, define a point x_i by taking in (1)

$$X_i = \overbrace{(1, 1,\dots,1)}^{4m} \equiv \bar{1}$$

[2] Strictly speaking, we really should be considering the problem of deciding whether X has a Steiner tree with length at most some prespecified value L.

and all other X_j and all Y_k to be

$$\overbrace{(0,0,\dots,0)}^{4m} \equiv \bar{0}.$$

Similarly, for $1 \leq j \leq n$, define s_j to have $X_i = \bar{0}$ for all i, $Y_j = \bar{1}$ and $Y_k = \bar{0}$ for all $k \neq j$. Intuitively for $i \neq 0$, x_i will correspond to the integer i and y_j will correspond to the 3-set F_j.

Next, if $i \in F_j$ we define a sequence of points $x_{i,j}(k), 0 \leq k \leq 8m - 1$, as follows:

$$x_{i,j}(k) = \left(X_0,\dots,X_i(k),\dots,X_{3m};\, Y_1,\dots,Y_j(k),\dots,Y_n\right),$$

where $X_u = \bar{0}$, $u \neq i$, $Y_v = \bar{0}$, $v \neq j$ and

$$\begin{cases} X_i = \bar{1},\ Y_j^{(k)} = (\overbrace{0,0,\dots,0}^{4m-k}, \overbrace{1,1,\dots,1}^{k}), & 0 \leq k \leq 4m, \\ X_i^{(k)} = (\overbrace{0,0,\dots,0}^{k-4m}, \overbrace{1,1,\dots,1}^{8m-k}), & 4m < k \leq 8m - 1,\ Y_j = \bar{1}. \end{cases}$$

Also, define $x_{0,j}(k)$ as above for all j, k, where $1 \leq j \leq n, 0 \leq k \leq 8m - 1$. Note that $x_{i,j}(0) = x_i$ for $1 \leq i \leq 3m$. Observe (for future reference) that the $x_{i,j}(k)$ form "chains" from x_i to s_j, where consecutive points on the chain have distance 1.

The set $X = X(\mathfrak{F})$ will consist of the $8m(3m + 1)n$ points $\{x_{i,j}(k)\colon 0 \leq i \leq 3m, 1 \leq j \leq n, 0 \leq k \leq 8m - 1\}$. We point out that X has a *spanning* tree with maximum edge length 2. This implies (see [5]) that any edge in an SMT for X has length at most 2. Define

$$L_0 = 4n(8m - 1) + 4m$$

and let $L_S(X)$ denote the length of an SMT for X.

Fact 1. If $\mathfrak{F}$ has an $X3C$ then $L_S(X) \leq L_0$.

Proof. Let $F_{j_1},\dots,F_{j_m}$ be an exact 3-cover of I_{3m}. For $1 \leq k \leq m$, adjoin to X the Steiner points $s_k = (X_0,\dots,X_{3m}, Y_1,\dots,Y_n)$ with $X_i = \bar{0}, 1 \leq i \leq 3m$, $Y_{j_k} = \bar{1}$, $Y_j = \bar{0}$, $j \neq j_k$. An easy calculation shows that $X^+ = X \cup \{s_k\colon 1 \leq k \leq m\}$ has a spanning tree of length L_0; just form the spanning tree consisting of all length 1 edges for X^+. □

Let us call an SMT for X *greedy* if it uses all the length 1 edges between points in X.

FACT 2. X has a *greedy* SMT with length $L_S(X)$.

Proof. Suppose T is an SMT for X with length strictly less than any greedy SMT. Thus, some edge e of length 1 does not occur in T. Adjoin e to T, thereby forming a cycle C. Some edge e' in C must be incident to a Steiner point s of T (since the length 1 edges in X do not form any cycles). Form the SMT T' by deleting the edge e'. Since

$$\text{length}(e') \geq 1 = \text{length}(e)$$

then

$$\text{length}(T') \leq \text{length}(T).$$

However, T is by hypothesis an SMT so that in fact

$$\text{length}(L') = \text{length}(T) = L_S(X).$$

Note that T' contains one more length 1 edge between points of X than T has. Fact 2 now follows by induction. □

FACT 3. If $L_S(X) \leq L_0$ then $\mathfrak{F}$ has an $X3C$.

Proof. By Fact 2 we can assume X has a greedy SMT T' with length $(T') \leq L_0$. As usual, we can assume without loss of generality that every Steiner point of T' has degree at least 3. For $0 \leq i \leq 3m$, define the subtree T_i of T' to be the tree induced by the points $x_{i,j}(k)$, $1 \leq j \leq n, 0 \leq k \leq 8m - 1$. By construction, all edges of T_i have length 1. We can think of T' as being formed by connecting the T_i together with a set E of edges, each of which is incident to some Steiner point. Observe that since

$$\sum_i \text{length}(T_i) = 4n(8m - 1) \tag{2}$$

then

$$\sum_{e \in E} \text{length}(e) \leq 4m \tag{3}$$

A key fact to be noted is this: For any

$$i_1 < i_2 < i_3 < i_4 < i_5 \text{ and any } q \in Q_N,$$

$$\sum_{k=1}^{5} d(q, T_{i_k}) > 4m, \tag{4}$$

where $d(q, T_i)$ denotes the minimum distance from q to a point of T_i. (This is the reason that subblocks of length $4m$ are used in the definition of the $x_{i,j}(k)$.)

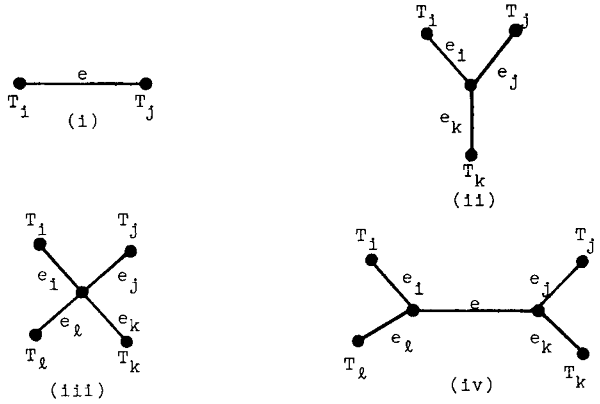

FIG. 1. Possible components of E.

A consequence of this observation is that *no component of E can have more than two Steiner points*. Otherwise, some Steiner point of T' would be connected by paths in E to (at least) five different T_i's which, by (3), would force

$$\text{length}(T') > 4n(8m - 1) + 4m = L_0.$$

Thus, there are at most four types of connected components E_k which can be formed by edges from E (also called full Steiner subtrees; see [5] or [8]). In Table I we list for each case a lower bound on the length of E_k, the *decrease* $\Delta(E_k)$ in the number of components due to E_k, and ρ, a lower bound on the ratio $\text{length}(E_k)/\Delta(E_k)$. Note that for all of E, $\text{length}(E) \leq 4m$ and $\Delta(E) = 3m$; thus, $\rho(E) \leq 4/3$.

Therefore, the only possibility is that case (iii) holds (with equality) in *all* cases. In other words, T' must have m Steiner points, each of degree 4, with all connecting edges of length 1. However, this is only possible if these

TABLE I

Case	lower bound on length(E_k)	$\Delta(E_k)$	ρ
(i)	2	1	2
(ii)	3	2	3/2
(iii)	4	3	4/3
(iv)	5	3	5/3

Steiner points are m of the s_j's, i.e., with $X_i = \bar{0}$ for all i, $Y_{j_k} = \bar{1}$, $Y_j = \bar{0}$, $j \neq j_k$. Consequently, the corresponding F_j's form an exact 3-cover of I_{3m}. □

The preceding facts have as immediate consequences the following results.

THEOREM. *The Steiner problem for the N-cube Q_N is NP-complete.*

COROLLARY. *The Steiner problem in phylogeny is NP-complete.*

REFERENCES

1. R. V. ECK AND M. O. DAYHOFF, "Atlas of Protein Sequence and Structure 1966," National Biomedical Research Foundation, Silver Springs, Md, 1966.
2. J. S. FARRIS, *Amer. Natur.* **106** (1972), 645–668.
3. W. M. FITCH AND E. MARGOLIASH, *Science* **155** (1967), 279–284.
4. L. R. FOULDS, M. D. HENDY, AND DAVID PENNY, *J. Mol. Evol.* **13** (1979), 127–149.
5. M. R. GAREY, R. L. GRAHAM, AND D. S. JOHNSON, The complexity of computing Steiner minimal trees, *SIAM J. Appl. Math.* **32** (1977), 835–859.
6. M. R. GAREY AND D. S. JOHNSON, "Computers and Intractability: A Guide to the Theory of *NP*-Completeness," Freeman, San Francisco, 1979.
7. M. R. GAREY AND D. S. JOHNSON, The rectilinear Steiner problem is *NP*-complete, *SIAM J. Appl. Math.* **32** (1977), 826–834.
8. E. N. GILBERT AND H. O. POLLAK, Steiner minimal trees, *SIAM J. Appl. Math* **16** (1968), 1–29.
9. F. HARARY, "Graph Theory," Addison–Wesley, Reading, Mass. 1969.
10. M. D. HENDY, DAVID PENNY, AND L. R. FOULDS, *J. Theor. Biol.* **71** (1978), 441–452.
11. G. W. MOORE, J. BARNABAS, AND M. GOODMAN, *J. Theor. Biol.* **38** (1973), 459–485.
12. DAVID PENNY, *J. Mol. Evol.* **8** (1976), 95–116.
13. DAVID PENNY, M. D. HENDY, AND L. R. FOULDS, *Biochem. J.* **187** (1980), 65–74.
14. J. D. WATSON, "Molecular Biology of the Gene," 3rd Ed. Benjamin, Menlo Park, Calif., 1975.