The Computational Complexity of Inferring Rooted Phylogenies by Parsimony*

WILLIAM H. E. DAY†
Centre de recherches mathématiques, Université de Montréal, Case Postale 6128, Succursale A, Montréal, Québec H3C 3J7, Canada

DAVID S. JOHNSON
AT&T Bell Laboratories 2C-355, 600 Mountain Avenue, Murray Hill, New Jersey 07974

AND

DAVID SANKOFF
Centre de recherches mathématiques, Université de Montréal, Case Postale 6128, Succursale A, Montréal, Québec H3C 3J7, Canada

Received 20 March 1986; accepted 11 April 1986

ABSTRACT

In systematics, parsimony methods construct phylogenies, or evolutionary trees, in which characters evolve with the least evolutionary change. The Camin-Sokal and Dollo parsimony criteria are used to construct phylogenies from discrete characters. Variations of these problems depend on whether characters are: cladistic (rooted) or qualitative (unrooted); binary (two states) or multistate (more than one state). The computational cost of known algorithms that guarantee optimal solutions to these problems increases exponentially with problem size; practical computational considerations restrict the use of such algorithms to analyzing problems of small size. We establish that the basic variants of these problems are all NP-complete and thus are so difficult computationally that efficient optimal algorithms are unlikely to exist for them.

1. INTRODUCTION

Systematists in recent years have witnessed a dramatic increase in the availability of computer resources for research. As a consequence, theoretical and practical investigations of numerical methods to infer phylogenies have

*The Natural Sciences and Engineering Research Council of Canada partially supported this research through individual operating grants to W. H. E. Day (A4142) and D. Sankoff (A8867), as well as through an infrastructure grant to D. Sankoff, R. J. Cedergren, and G. Lapalme (A3092).

†Permanent address: Department of Computer Science, Memorial University of Newfoundland, St. John's, NF A1C 5S7, Canada.

 0025-5564/86/$03.50

flourished (Felsenstein [11] and references therein). Major strategies for inferring phylogenies have been developed from basic concepts of compatibility and parsimony. For a given set of objects, compatibility criteria are used to seek phylogenies on which a largest set of characters is perfectly compatible (Le Quesne [18, 19]), whereas parsimony criteria are used to seek phylogenies on which characters evolve with the least evolutionary change. Camin-Sokal parsimony (Camin and Sokal [1]), Dollo parsimony (Le Quesne [20, 21], Farris [6, 7]), and Wagner parsimony (Kluge and Farris [17], Farris [5]) are well-known parsimony criteria for discrete character data. The logical and philosophical bases of parsimony have been much discussed (Farris [9], Felsenstein [12], Sober [22]).

An important research problem is whether efficient polynomial-time algorithms exist for inferring phylogenies. An algorithm is called *polynomial-time* if its execution requires a number of steps that can be bounded in advance by a polynomial function of the problem's size. For many problems of phylogenetic inference, researchers have been unable to design polynomial-time algorithms that always obtain optimal solutions. Results concerning computational complexity sometimes help to explain why polynomial-time algorithms seem difficult to develop. Garey and Johnson [14] give an exposition of problems which are known to be identical with respect to whether or not polynomial-time algorithms could exist to solve them. These problems are called NP-complete; although they have not been proved intractable, solving any (and thus all) of them by polynomial-time algorithms is unlikely. Recently problems of inferring phylogenies by compatibility and by Wagner parsimony have been shown to be NP-complete (Foulds and Graham [13], Graham and Foulds [15], Day [2], Day and Sankoff [3]). Since the Camin-Sokal and Dollo criteria are restricted variants of the Wagner criterion, one might still hope to use them to infer phylogenies by polynomial-time algorithms. Our results essentially dash these hopes: we will establish that problems of inferring phylogenies by the Camin-Sokal and Dollo parsimony criteria are also NP-complete.

2. DEFINITIONS

Although rooted phylogenies arise in systematics and the biological sciences, here we shall use basically graph-theoretic terminology and conventions. There exist finite nonempty sets of objects (e.g., terminal taxa, operational taxonomic units) and of characters that describe the objects. Each character has two states and so is called *binary*. A binary character is called *qualitative* if its states are an unordered set on which no further structure is imposed; it is called *cladistic* if its states are ordered so that one is *ancestral* and the other *derived*. The n character states of an object x are described by a vector $v(x) = \langle v_1, \ldots, v_n \rangle$, in which v_i is the state of character i for object x.

Phylogenies in Camin-Sokal and Dollo parsimony problems can be represented as subtrees of an appropriate hypercube. (The reader can consult Harary [16] for graph-theoretic terms whose meaning is not obvious.) Let $\{0,1\}^n$ denote the n-fold Cartesian product of $\{0,1\}$. Let $H = (\{0,1\}^n, E)$ be the graph with the elements (i.e., vectors) of $\{0,1\}^n$ as vertices and an edge between two vertices if and only if they differ in exactly one vector position. H is called a *hypercube* of dimension n. If X is a subset of $\{0,1\}^n$, a *Steiner tree* for X is a minimal connected subgraph $T = (X', E')$ of H with X a subset of X'; minimality implies that T is acyclic and thus a tree. The *size* of T is the number $|X'|$ of vertices in X'; the vertices in X', but not in X, are called *Steiner* vertices.

If X is a subset of $\{0,1\}^n$, a *phylogeny* for X is a rooted Steiner tree for X. The phylogeny's root is a vector describing the character state values, or *ancestral states*, of a putative ancestor of the objects in X. The edges of a phylogeny can be oriented away from the root; an edge from $u = \langle u_1, \ldots, u_n \rangle$ to $v = \langle v_1, \ldots, v_n \rangle$, where u and v differ only in position i, represents a *transition* of character i from u_i to v_i. Camin-Sokal and Dollo parsimony problems further restrict permissible transitions in phylogenies. Camin-Sokal problems require that no character transition be from derived state to ancestral state. Dollo problems require that every character be uniquely derived, i.e., that every character have exactly one transition from ancestral state to derived state.

Although Camin-Sokal and Dollo parsimony problems are often stated as optimization problems, we formulate them in Table 1 as decision problems in which each solution is either "yes" or "no." Each decision problem is equivalent to its optimization problem, since one has a polynomial-time algorithm if and only if the other has a polynomial-time algorithm. To prove that a decision problem X is NP-complete, we establish both that X is in the class NP (of problems having polynomial-time verification algorithms) and that a known NP-complete problem transforms to X in the following sense. Let D_1 (D_2) denote the set of all instances of a decision problem X_1 (X_2). A *polynomial transformation* from X_1 to X_2 is a map f from D_1 to D_2 such that: f is computable by a polynomial-time algorithm; for each instance I in D_1, the X_1 solution for I is "yes" if and only if the X_2 solution for $f(I)$ is "yes."

Our reference NP-complete problem concerns the existence of a vertex cover in a graph.

VERTEX COVER (*Garey and Johnson* [14, *p.* 46])

Instance: A graph $G = (V, E)$ and a positive integer $K \leqslant |V|$.

Question: Is there a *vertex cover* of size K or less for G, i.e., a subset V' of V such that $|V'| \leqslant K$ and, for each edge in E, at least one of its incident vertices is in V'?

TABLE 1

Camin-Sokal and Dollo Decision Problems

Cladistic Camin-Sokal (CCS)

Instance: Positive integer n; a subset X of $\{0,1\}^n$; and a positive integer B.
Question: Is there a Camin-Sokal phylogeny for X that is rooted at O and has size at most B?

Qualitative Camin-Sokal (QCS)

Instance: Same as for CCS.
Question: Is there a Camin-Sokal phylogeny for X that has size at most B?

Cladistic Dollo (CDO)

Instance: Same as for CCS.
Question: Is there a Dollo phylogeny for X that is rooted at the rank-n vertex and has size at most B?

Qualitative Dollo (QDO)

Instance: Same as for CCS.
Question: Is there a Dollo phylogeny for X that has size at most B?

3. RESULTS

Theorems 2 and 3 report NP-completeness results for Camin-Sokal and Dollo parsimony problems. Both proofs depend on a characterization, in Lemma 1, of structure in Steiner trees for a special family of problems. To name vertices of the hypercube, we rely in Lemma 1 on the correspondence between a vector in $\{0,1\}^n$ and a subset of the integers in $\{1,\ldots,n\}$, where the vector $\langle v_1,\ldots,v_n\rangle$ corresponds to the set $\{i: v_i = 1\}$. For example, $\{a,b\}$ (or $\langle ab\rangle$, for short) is the vertex with ones in positions a and b and zeros elsewhere. The *rank* of any vector in $\{0,1\}^n$ is simply the number of ones it has; for example, the rank-zero vector $\langle 0,\ldots,0\rangle$ has no ones and is denoted by O.

LEMMA 1

If a subset X of $\{0,1\}^n$ is such that $X = Y \cup \{O\}$, where all vertices in Y are rank-two, then a minimum-sized Steiner tree for X exists in which all Steiner vertices are rank-one and are adjacent to O.

Proof. Let T be a minimum-sized Steiner tree for X that has the fewest Steiner vertices violating the Lemma. Such Steiner vertices are of two types: they are rank-one but are not adjacent to O; or their rank is greater than one. We show that T can have no such violations.

Suppose in T there were a rank-one Steiner vertex p not adjacent to O. Exactly one of the edges incident on p in T must be in the path in T between p and O. But that edge could be replaced by $\{p, O\}$ to obtain a Steiner tree of equal size but with one fewer violation. Thus we may assume that all rank-one Steiner vertices in T are adjacent to O.

Let x be a vertex of Y at maximum distance in T from O. If x is adjacent to a rank-one Steiner vertex, then by the argument above x is at distance two from O; hence all vertices of Y must be adjacent to rank-one Steiner vertices and T is as claimed. Therefore we may assume x is adjacent to a rank-three vertex $p = \langle abc \rangle$. Edges incident on p are strongly restricted. Since T is minimal, every Steiner vertex must be in a path between O and a vertex of Y. Thus any edge incident on p and a Steiner vertex not in the path between p and O would imply the existence of a vertex of Y whose distance from O was greater than that of x, a contradiction of our choice of x. Therefore the only edges in T incident on p are an edge e in the path between p and O, and one, two, or three edges incident on vertices of Y that are leaves of T.

These leaves must be in $\{\langle ab \rangle, \langle ac \rangle, \langle bc \rangle\}$. If there are one or two of them, p can be replaced by a single rank-one Steiner vertex (Figure 1) to obtain a Steiner tree of no greater size but with one fewer violation. Thus we may assume that $\langle ab \rangle$, $\langle ac \rangle$, and $\langle bc \rangle$ are vertices of Y and are leaves adjacent to p in T. Hence e must be incident in T with both p and a rank-four vertex $q = \langle abcd \rangle$.

Now q cannot be adjacent in T with any rank-three vertex other than p. Of the three possibilities $(\langle abd \rangle, \langle acd \rangle, \langle bcd \rangle)$, suppose q were adjacent to, say, $\langle bcd \rangle$. By the argument above for p, we may assume that $\langle bc \rangle$, $\langle bd \rangle$, and $\langle cd \rangle$ are vertices of Y and are leaves in T adjacent to q; but this is impossible, since $\langle bc \rangle$ cannot be a leaf if it is adjacent to both p and $\langle bcd \rangle$.

Thus the path between q and O must include an edge incident on q and a rank-five vertex $\langle abcde \rangle$. Furthermore, q can be adjacent in T to no other rank-five vertex, since if it were, Y would contain a vertex whose distance in T from O was greater than that of x from O, a contradiction of our choice of x. Hence the vicinity of x in T must be as in Figure 2; but the transformation of T shown there would yield a new tree of no greater size but with two fewer violations, a contradiction. ■

THEOREM 2

CCS *and* QCS *are* NP-*complete problems.*

Proof. The problems are clearly in NP. To complete the proof, we exhibit a polynomial transformation from VERTEX COVER to CCS and QCS simultaneously. Let the graph $G = (V, E)$ and the positive integer $K \leqslant |V|$ be an arbitrary instance I of VERTEX COVER. The corresponding instance

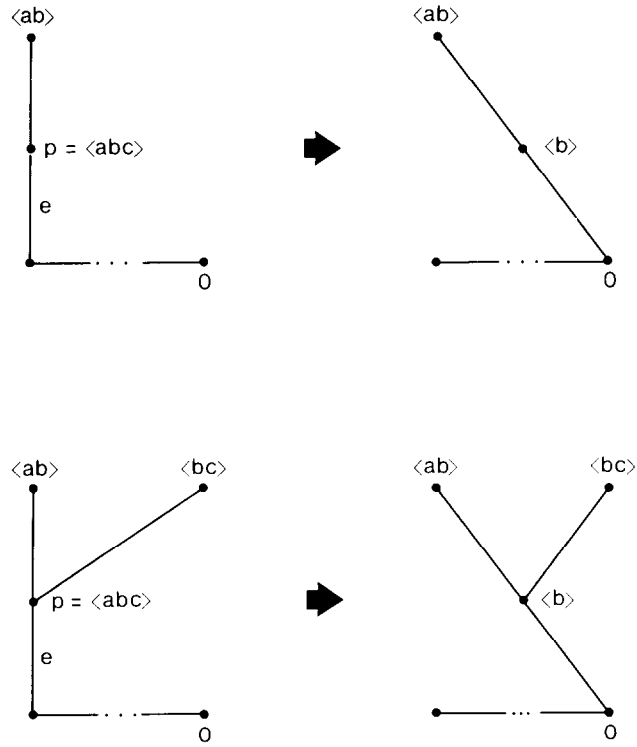

FIG. 1. Transformations in the proof of Lemma 1. They delete vertex p, and they insert $\langle b \rangle$ if it is not already in the tree.

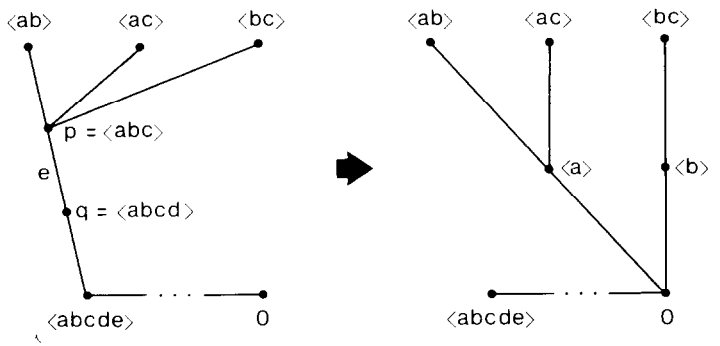

FIG. 2. Another transformation in the proof of Lemma 1. It deletes vertices p and q, and it inserts $\langle a \rangle$ and $\langle b \rangle$ if they are not already in the tree.

$f(I) = (n, X, B)$ of CCS or QCS is defined on the hypercube of dimension $n = |V|$, its vector positions corresponding to vertices in V. X contains O together with rank-two vectors $x(e)$ for every edge $e = \{u, v\}$ in E, where $x(e)$ has ones in the positions for u and v, and zeros elsewhere. With $B = K + |E| + 1$, $f(I)$ is a valid instance of CCS and QCS that can be constructed in polynomial time.

To complete the proof we must show that G has a vertex cover of size at most K if and only if X has a phylogenetic tree of size at most B. Suppose the VERTEX COVER solution to I is "yes" by virtue of a vertex cover V', and construct the desired tree T as follows. For each vertex v in V', T has a rank-one Steiner vertex $s(v)$ with a one in the position for v and zeros elsewhere. T has an edge $\{s(v), O\}$ for each v in V'; T also has for every e in E an edge $\{x(e), s(v)\}$, v being an endpoint of e that is in V'. The size of T is $|V'| + |E| + 1 \leqslant B$; when T is rooted at O, both CCS and QCS solutions to $f(I)$ are "yes" by virtue of T.

Conversely, suppose the CCS or QCS solution to $f(I)$ is "yes" by virtue of a phylogenetic tree T. Since T can have at most K Steiner vertices, there must exist an unrestricted Steiner tree with at most K Steiner vertices, and by Lemma 1 we may assume that these Steiner vertices are all rank-one. Since the corresponding set V' of vertices in G must be a vertex cover, the VERTEX COVER solution to I is "yes" by virtue of V'. ■

THEOREM 3

CDO *and* QDO *are* NP*-complete problems.*

Proof. Since the problems are clearly in NP, we next exhibit a variant of the transformation in the proof of Theorem 2. The dimension of the underlying hypercube is increased to $n = 2K + |V|$, the $2K$ new vector positions coming *before* the original $|V|$ positions and being set to zero in the vectors of the original construction. Let Y denote the set of rank-two vectors corresponding to the edges of G. In addition to O and Y, X contains $2K + |V|$ new vectors p_i, $1 \leqslant i \leqslant 2K + |V|$, with p_i having ones in positions 1 through i, and zeros elsewhere. Finally, we set $B = 3K + |V| + |E| + 1$.

Suppose the desired vertex cover V' exists. To construct the phylogenetic tree T, we use rank-one Steiner vertices, as in the proof of Theorem 2, to link O with the vectors corresponding to edges of G; in addition T has edges $\{O, p_1\}$ and $\{p_i, p_{i+1}\}$ for $1 \leqslant i < L = 2K + |V|$. It is easy to verify that T has the desired size and satisfies the CDO and QDO requirements when rooted at vertex p_L.

Conversely, suppose the desired phylogenetic tree exists; but let us ignore its phylogenetic rooting and consider it just as a Steiner tree. Since it has at most K Steiner vertices, there must be for the given required vertices a

minimum-sized (unrestricted) Steiner tree with at most K Steiner vertices. Let T be such a tree. We shall show that T can be transformed to a Steiner tree of equal size but with the form described in the previous paragraph.

First, we may assume that all edges in the path between O and p_L are present in T. Suppose one such edge $\{u, v\}$ is missing. Since u and v are required vertices, T must have a path between u and v. That path must contain at least one edge not in the path between O and p_L. Replacing such an edge by $\{u, v\}$ yields a new tree with one fewer edge missing from the path between O and p_L. By induction we can thus assume that none are missing.

Next we may assume that no Steiner vertex is adjacent to any p_i where $1 \leqslant i \leqslant L$. Suppose T has the minimum number of such "bad" Steiner vertices, subject to the requirements that it be a minimum-sized Steiner tree and contain all edges between O and p_L. Notice that no bad vertex p can be adjacent to any p_i where $2K+1 \leqslant i \leqslant L$. By the minimality of T, p would be in a path between p_i and a vertex in the set Y. Since p disagrees with all vertices in Y in at least $2K-1$ of the first $2K$ positions, that path must contain at least $2K-1$ Steiner vertices and so contradicts the fact that T has at most K Steiner vertices. (We may assume, without loss of generality, that K is greater than one.) Thus there are no bad Steiner vertices of this type.

Notice also that no bad vertex p can be adjacent to any p_i where $1 \leqslant i \leqslant 2K$. Suppose to the contrary that p is adjacent to p_k where $1 \leqslant k \leqslant 2K$. Consider the set of all vertices of T that are connected to p_k through p, together with p_k itself. If we project the subtree T' of T induced by these vertices onto the subgraph of the hypercube in which the first $2K$ positions are all zero, we obtain a connected subgraph T'' that has no more vertices than T' and contains, as distinct vertices, O and every vertex from Y that is in T'. Replacing T' in T by a spanning tree for T'' would yield a Steiner tree with no more vertices than T but with one fewer bad Steiner vertex, a contradiction of the minimality of T with respect to such bad vertices.

We conclude that T must contain a minimum-sized Steiner tree T' for the set $Y \cup \{O\}$. It is easy to show that, because of minimality, all Steiner vertices in T' must have zeros in the first $2K$ positions. Thus we can project T' on the last $|V|$ positions to obtain a Steiner tree from which the desired vertex cover for G can be derived as in the proof of Theorem 2. ■

4. DISCUSSION

Although the decision problems in Table 1 require that every character be binary, they can also be respecified to permit multistate characters. Since every binary character also satisfies the definition of a multistate character,

the binary version of each decision problem transforms to its multistate version. Thus, as a consequence of Theorems 2 and 3, the multistate versions of the Camin-Sokal and Dollo phylogeny problems are NP-complete as well.

The construction in the proof of Theorem 2 is suitable for obtaining NP-completeness results in a range of problems of inferring phylogenies. Since the proof does not assume that phylogenies are rooted, it provides an alternative demonstration that decision problems based on the Wagner parsimony criterion (Kluge and Farris [17], Farris [5]) are NP-complete; indeed, it provides a simple proof of Graham and Foulds's result [13, 15] that the Steiner problem for the N-cube Q_N is NP-complete. Day and Sankoff [4] use the construction in the proof of Theorem 3 to establish that decision problems of inferring rooted phylogenies from chromosome inversion data (Farris [8], Felsenstein [10]) are NP-complete.

REFERENCES

1 J. H. Camin and R. R. Sokal, A method for deducing branching sequences in phylogeny, *Evolution* 19:311–326 (1965).

2 W. H. E. Day, Computationally difficult parsimony problems in phylogenetic systematics, *J. Theoret. Biol.* 103:429–438 (1983).

3 W. H. E. Day and D. Sankoff, The computational complexity of inferring phylogenies by compatibility, *Syst. Zool.* 35 (1986), to appear.

4 W. H. E. Day and D. Sankoff, The computational complexity of inferring phylogenies from chromosome inversion data, Rep. CRM-1366, Centre de recherches mathématiques, Université de Montréal, C.P. 6128, Succ. A, Montréal, PQ H3C 3J7, Canada, 1986.

5 J. S. Farris, Methods for computing Wagner trees, *Syst. Zool.* 19:83–92 (1970).

6 J. S. Farris, Phylogenetic analysis under Dollo's law, *Syst. Zool.* 26:77–88 (1977).

7 J. S. Farris, Some further comments on Le Quesne's methods, *Syst. Zool.* 26:220–223 (1977).

8 J. S. Farris, Inferring phylogenetic trees from chromosome inversion data, *Syst. Zool.* 27:275–284 (1978).

9 J. S. Farris, The logical basis of phylogenetic analysis, in *Advances in Cladistics, Volume 2: Proceedings of the Second Meeting of the Willi Hennig Society* (N. I. Platnick and V. A. Funk, Eds.), Columbia U. P., New York, 1983, pp. 7–36.

10 J. Felsenstein, Alternative methods of phylogenetic inference and their interrelationship, *Syst. Zool.* 28:49–62 (1979).

11 J. Felsenstein, Numerical methods for inferring evolutionary trees, *Quart. Rev. Biol.* 57:379–404 (1982).

12 J. Felsenstein, Parsimony in systematics: Biological and statistical issues, *Ann. Rev. Ecol. Syst.* 14:313–333 (1983).

13 L. R. Foulds and R. L. Graham, The Steiner problem in phylogeny is NP-complete, *Adv. Appl. Math.* 3:43–49 (1982).

14 M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-completeness*, Freeman, San Francisco, 1979.

15 R. L. Graham and L. R. Foulds, Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computational time, *Math. Biosci.* 60:133–142 (1982).

16 F. Harary, *Graph Theory*, Addison-Wesley, Reading, Mass., 1969.

17 A. G. Kluge and J. S. Farris, Quantitative phyletics and the evolution of anurans, *Syst. Zool.* 18:1–32 (1969).

18 W. J. Le Quesne, A method of selection of characters in numerical taxonomy, *Syst. Zool.* 18:201–205 (1969).

19 W. J. Le Quesne, Further studies based on the uniquely derived character concept, *Syst. Zool.* 21:281–288 (1972).

20 W. J. Le Quesne, The uniquely evolved character concept and its cladistic application, *Syst. Zool.* 23:513–517 (1974).

21 W. J. Le Quesne, The uniquely evolved character concept, *Syst. Zool.* 26:218–220 (1977).

22 E. Sober, A likelihood justification of parsimony, *Cladistics* 1:209–233 (1985).