J. theor. Biol. (1987) **124**, 213–218

Computational Complexity of Inferring Phylogenies from Chromosome Inversion Data†

WILLIAM H. E. DAY‡ AND DAVID SANKOFF

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

(*Received* 14 *April* 1986, *and in revised form* 24 *August* 1986)

In systematics, parsimony methods construct phylogenies, or evolutionary trees, in which characters evolve with the least evolutionary change. The chromosome inversion, or polymorphism, parsimony criterion is used when each character of a population may exhibit homozygous or heterozygous states, but when the heterozygous state must evolve uniquely. Variations of the criterion concern whether or not the ancestral states of characters are specified. We establish that problems of inferring phylogenies by these criteria are NP-complete and thus are so difficult computationally that efficient optimal algorithms for them are unlikely to exist.

1. Introduction

Since constructing phylogenies is so often computationally expensive, an important research problem is whether efficient polynomial-time algorithms exist to infer them. 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, workers 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 & Johnson (1979) give a fine 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, Camin-Sokal, and Dollo parsimony have been shown to be NP-complete (Foulds & Graham, 1982; Graham & Foulds, 1982; Day, 1983; Day & Sankoff, 1986; Day *et al.*, 1986). Since the chromosome inversion (Farris, 1978), or polymorphism (Felsenstein, 1979), parsimony 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 this hope: we will establish that problems of inferring phylogenies by chromosome inversion parsimony criteria are also NP-complete.

† 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, Newfoundland A1C 5S7, Canada.

0022-5193/87/020213+06 $03.00/0

We adopt the biological model described by Farris (1978, p. 275).

> "I shall suppose that any one inversion site of an individual chromosome may show just one of two alternative inversion types A, D, of which A is the ancestral type and D the derivative type. Evolution of such a site occurs when a chromosome showing the new type D arises in an individual belonging to a population originally homozygous for A. The population is then in a heterozygous condition denoted H. Once in heterozygous condition H, a population may proceed to fixation of the considered site in either of two ways, yielding populations homozygous either for the A or the D inversion type. In a real phylogenetic inference problem it may of course occur that the worker is uncertain as to which of two alternative inversion types is the ancestral . . ."

When reconstructing phylogenies from chromosome inversion data, workers may be reluctant to envisage the same inversion type evolving more than once at a site; instead they may prefer to explain apparent convergence or parallelism in a phylogeny by parsimonious persistence of the heterozygous condition. We show in the next section that such problems of phylogenetic inference have natural graph-theoretic formulations.

2. Definitions

We use these terminologies and conventions to formulate problems of inferring phylogenies from chromosome inversion data. There exist finite nonempty sets of objects (e.g., populations) and of characters that describe the objects. Each character has $\{A, H, D\}$ as its set of possible states. A character is called *qualitative* (Estabrook & McMorris, 1977) if its states are an unordered set on which no further structure is imposed; it is called *cladistic* (Estabrook & McMorris, 1980) if one of its states is designated as *ancestral* relative to the others. 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_k is the state of character k for object x.

Phylogenies in chromosome inversion parsimony problems can be modeled as subtrees of an appropriate graph. (The reader can consult Harary (1969) for graph-theoretic terms whose meanings are not obvious.) Let $\{A, H, D\}^n$ denote the set of all vectors, or n-tuples, whose symbols are selected from $\{A, H, D\}$. Let $C = (\{A, H, D\}^n, E)$ be the graph with the vectors of $\{A, H, D\}^n$ as vertices and an edge between two vertices if and only if they differ in exactly one vector position. C is called a *hypercube* of dimension n. If X is a subset of $\{A, H, D\}^n$, a *Steiner tree* for X is a minimal connected subgraph $T = (X', E')$ of C 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.

Usually, an edge of a phylogeny has associated with it either: a number whose value represents an interval of time; or descriptions of character-state changes between the objects incident with the edge. Since we use the latter convention, a *phylogeny* for a subset X of $\{A, H, D\}^n$ is simply a rooted Steiner tree for X. The phylogeny's root is a vector describing the character-state values (i.e. ancestral states) of a putative ancestor of the objects in X. The phylogeny's edges can be oriented away from the root; thus 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 at position k, has an associated *transition* $u_k \Rightarrow v_k$ of character k from states u_k to v_k.

Chromosome inversion parsimony problems restrict permissible transitions in phylogenies to conform with the biological model. Since transitions must involve a heterozygous state, $A \Rightarrow D$ and $D \Rightarrow A$ are forbidden. If A (respectively, D) is a character's ancestral state, then $A \Rightarrow H$ (respectively, $D \Rightarrow H$) can appear at most once and $D \Rightarrow H$ (respectively, $A \Rightarrow H$) is forbidden. If H is its ancestral state, then $A \Rightarrow H$ and $D \Rightarrow H$ are forbidden.

Although chromosome inversion parsimony problems can be stated as optimization problems, we formulate them 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.

Cladistic Chromosome Inversion (CCI)

Instance: Positive integers n and B; subset X of $\{A, H, D\}^n$.
Question: Is there a chromosome inversion phylogeny for X that has size at most B and is rooted at $\langle A, \ldots, A \rangle$?

Qualitative Chromosome Inversion (QCI)

Instance: Positive integers n and B; subset X of $\{A, H, D\}^n$.
Question: Is there a chromosome inversion phylogeny for X that has size at most B?

To prove that a decision problem P is NP-complete, we establish both that P is in the class NP (of problems having polynomial-time verification algorithms), and that a known NP-complete problem transforms to P in the following sense. Let D_1 (respectively, D_2) denote the set of all instances of a decision problem P_1 (respectively, P_2). A *polynomial transformation* from P_1 to P_2 is a map f from D_1 to D_2 such that: f is computable by a polynomial-time algorithm; for each instance λ in D_1, the P_1 solution for λ is "yes" if and only if the P_2 solution for $f(\lambda)$ is "yes". Our reference NP-complete problem is the following graph-theoretic one.

Vertex Cover (*Garey & Johnson*, 1979, *p.* 46)

Instance: Graph $G = (V, E)$; positive integer $K \leq |V|$.
Question: Is there a vertex cover of size K or less for G, i.e., a subset V' of V such that $|V'| \leq K$ and, for each edge in E, at least one of its incident vertices is in V'?

3. Results

Our results depend on a characterization of structure in Steiner trees for the n-cube $C' = (\{0, 1\}^n, E)$. Let the *rank* of any vector in $\{0, 1\}^n$ be simply the number of ones it has; for example, the rank-zero vector $\langle 0, \ldots, 0 \rangle$ is denoted by $\emptyset$.

LEMMA 1 (*Day et al.*, 1986). *If a subset X of $\{0, 1\}^n$ is such that $X = Y \cup \{\emptyset\}$, 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 $\emptyset$.*

We extend this result to the hypercube C. Let the *rank* of any vector in $\{A, H, D\}^n$ be a 3-tuple (n_A, n_H, n_D), each n_k counting the occurrences of k in the vector; for example, the rank-$(0, n, 0)$ vector $\langle H, \ldots, H\rangle$ is denoted by ϕ_H.

LEMMA 2. *If a subset X of $\{A, H, D\}^n$ is such that $X = Y \cup \{\phi_H\}$, where all vertices in Y are rank-$(0, n-2, 2)$, then a minimum-sized Steiner tree for X exists in which all Steiner vertices are rank-$(0, n-1, 1)$ and are adjacent to ϕ_H.*

Proof. Let T be a minimum-sized Steiner tree for X that has the fewest Steiner vertices violating the Lemma. Suppose T has Steiner vertices with A's at position k. Then T contains a subtree T' whose terminal vertices have H's or D's at position k and whose interior vertices have A's at position k. If any terminal vertices of T' have H's at position k, transform T to T'' by changing all A's at position k of T' to H's; otherwise change them to D's. In T'', at least one edge has become a loop and so many may be removed; thus T'' has smaller size than T and contradicts our assumption that T was minimum-sized. Thus we may assume that state A is absent from all vertices of T. The result then follows from Lemma 1. ■

The proof of our main result uses an argument developed by Day *et al.* (1986) to establish the NP-completeness of inferring phylogenies by the Dollo parsimony criteria.

THEOREM 3. *CCI and QCI are NP-complete problems.*

Proof. Since the problems clearly are in NP, we next exhibit a polynomial transformation from Vertex Cover to CCI and QCI simultaneously. Let graph $G = (V, E)$ and positive integer $K \leq |V|$ be an arbitrary instance λ of Vertex Cover. The corresponding instance $f(\lambda) = (n, X, B)$ of CCI or QCI is defined on hypercube C of dimension $n = 2K + |V|$, its last $|V|$ vector positions corresponding to vertices in V. Let Y denote the set of vectors $x(e)$ corresponding to the edges $e = \{u, v\}$ in E, where $x(e)$ has D's in the positions for u and v, and H's elsewhere. Let Z denote the set of vectors p_k, $1 \leq k \leq n$, with p_k having A's in positions 1 through k, and H's elsewhere. To specify $f(\lambda)$, let $X = Y \cup Z \cup \{\phi_H\}$ and $B = n + K + |E| + 1$.

To complete the proof we show that G has a vertex cover of size at most K if and only if X has a phylogeny of size at most B. Suppose the Vertex Cover solution to λ is "yes" by virtue of vertex cover V', and construct a phylogeny T as follows. For each v in V', T has a rank-$(0, n-1, 1)$ Steiner vertex $s(v)$ with D in the position for v, and H's elsewhere. T has an edge $\{s(v), \phi_H\}$ for each v in V'; T has for every e in E an edge $\{x(e), s(v)\}$, v being an endpoint of e that is in V'; and T has edges $\{\phi_H, p_1\}$ and $\{p_k, p_{k+1}\}$ for $1 \leq k < n$. The size of T is $n + |V'| + |E| + 1 \leq B$; when T is rooted at p_n, both CCI and QCI solutions to $f(\lambda)$ are "yes" by virtue of T.

Conversely, suppose the desired phylogeny exists; but ignore its root and consider it just as a Steiner tree. Since it has at most K Steiner vertices, there must be for the vertices in X a minimum-sized (unrestricted) Steiner tree T with at most K Steiner vertices. 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 ϕ_H and p_n 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 ϕ_H and p_n. Replacing such an edge by $\{u, v\}$ yields a new tree with one fewer edge missing from the path between ϕ_H and p_n. By induction we can thus assume that none are missing.

Next we may assume that no Steiner vertex is adjacent to any p_k where $1 \leq k \leq n$. 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 ϕ_H and p_n. Notice that no bad vertex p can be adjacent to any p_k where $2K < k \leq n$. By the minimality of T, p would lie on a path between p_k 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. (Without loss of generality, we may assume 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_k where $1 \leq k \leq 2K$. Suppose to the contrary that p is adjacent to p_m where $1 \leq m \leq 2K$. Consider the set of all vertices of T that are connected to p_m through p, together with p_m itself. If we project the subtree T' of T induced by these vertices onto the subgraph of C in which the first $2K$ positions are all H's, we obtain a connected subgraph T'' that has no more vertices than T' and has, as distinct vertices, ϕ_H 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 $Y \cup \{\phi_H\}$. It is easy to show that, because of minimality, all Steiner vertices in T' must have H's in the first $2K$ positions. Thus we can project T' on the last $|V|$ positions to obtain a Steiner tree whose K Steiner vertices, by Lemma 2, are all rank-$(0, n-1, 1)$. Since the corresponding set V' of vertices in G must be a vertex cover, the Vertex Cover solution to λ is "yes" by virtue of V'. ■

4. Discussion

Farris (1978) and Felsenstein (1979) described chromosome inversion problems that are equivalent to our formulation. Their phylogenetic model, in which phylogenies bifurcate and edges have zero or more associated transitions, is equivalent to our rooted Steiner tree. If N_{IJ} denotes the number of $I \Rightarrow J$ transitions in a phylogeny T, then the optimizing versions of CCI and QCI minimize an objective function $L(T) = N_{AH} + N_{DH} + N_{HA} + N_{HD}$. Farris used an objective function $L'(T) = N_{HA} + N_{HD}$ for a variant of CCI in which every character exhibits the D or H state in at least one of the n objects; thus $N_{DH} = 0$, $N_{AH} = n$, and $L'(T) = L(T) - n$ so that minimizing $L(T)$ is equivalent to minimizing $L'(T)$. Farris also established that minimizing $L'(T)$ is equivalent to Felsenstein's minimization of the number of heterozygous states in ancestral populations of T.

Felsenstein (1979) specialized his probabilistic model of character evolution to obtain, as special cases, phylogenies satisfying the Camin-Sokal, Dollo, chromosome inversion, and Wagner parsimony criteria, as well as the compatibility criterion. Problems of phylogenetic inference based on all these criteria are now known to be NP-complete. Such results symbolize our inability to design efficient algorithms that guarantee optimal solutions; they challenge us to develop efficient approximation algorithms, and to characterize restricted classes of problems for which optimal solutions can be obtained efficiently by algorithms that are inefficient in the worst case.

REFERENCES

DAY, W. H. E. (1983). Computationally difficult parsimony problems in phylogenetic systematics. *J. theor. Biol.* **103,** 429.

DAY, W. H. E. & SANKOFF, D. (1986). Computational complexity of inferring phylogenies by compatibility. *Syst. Zool.* **35,** 224.

DAY, W. H. E., JOHNSON, D. S. & SANKOFF, D. (1986). The computational complexity of inferring rooted phylogenies by parsimony. *Math. Biosci.* **81,** 33.

ESTABROOK, G. F. & MCMORRIS, F. R. (1977). When are two qualitative taxonomic characters compatible? *J. math. Biol.* **4,** 195.

ESTABROOK, G. F. & MCMORRIS, F. R. (1980). When is one estimate of evolutionary relationships a refinement of another? *J. math. Biol.* **10,** 367.

FARRIS, J. S. (1978). Inferring phylogenetic trees from chromosome inversion data. *Syst. Zool.* **27,** 275.

FELSENSTEIN, J. (1979). Alternative methods of phylogenetic inference and their interrelationship. *Syst. Zool.* **28,** 49.

FOULDS, L. R. & GRAHAM, R. L. (1982). The Steiner problem in phylogeny is NP-complete. *Adv. Appl. Math.* **3,** 43.

GAREY, M. R. & JOHNSON, D. S. (1979). *Computers and Intractability: a Guide to the Theory of NP-completeness.* San Francisco: W. H. Freeman.

GRAHAM, R. L. & FOULDS, L. R. (1982). Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computational time. *Math. Biosci.* **60,** 133.

HARARY, F. (1969). *Graph Theory.* Reading, Massachusetts: Addison-Wesley.