Journal of Classification 3:349-356 (1986)

Journal of
Classification

Tree Enumeration Modulo a Consensus

Mariana Constantinescu
Université de Montréal

David Sankoff
Université de Montréal

Abstract: The number of trees with n labeled terminal vertices grows too rapidly with n to permit exhaustive searches for Steiner trees or other kinds of optima in cladistics and related areas Often, however, structured constraints are known and may be imposed on the set of trees to be scanned These constraints may be formulated in terms of a consensus among the trees to be searched We calculate the reduction in the number of trees to be enumerated as a function of properties of the imposed consensus

Keywords: Consensus; Tree enumeration

The number of trees with n labeled terminal vertices grows so rapidly that exhaustive searches for optimal trees in classification, numerical taxonomy, cladistics and related areas becomes unfeasible for relatively small values of n. Often, however, structural constraints are known or may be imposed on the set of trees to be scanned for optimality. For example, Sankoff, Cedergren and McKay (1982) and Gray, Sankoff and Cedergren (1984), in the study of molecular evolution, required that a tree consist of several given subtrees, each one fixed *a priori*, and joined together through a number of additional edges and vertices Aho *et al.* (1981) searched for rooted trees obeying given combinations of constraints on configurations of triples or quadruples of terminal vertices. In this paper, we impose constraints on configurations of arbitrary subsets of the terminal vertices. These constraints may be considered to constitute a consensus among subtrees of the trees to be searched.

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada through operating grant A8867 to D Sankoff and infrastructure grant A3092 to D Sankoff, R J Cedergren and G Lapalme We are grateful to William H E Day for much encouragement and many helpful suggestions

Authors' Address: Mariana Constantinescu and David Sankoff, Centre de recherches mathématiques, Université de Montréal, C P 6128, Succ A, Montréal (Québec) H3C 3J7, Canada

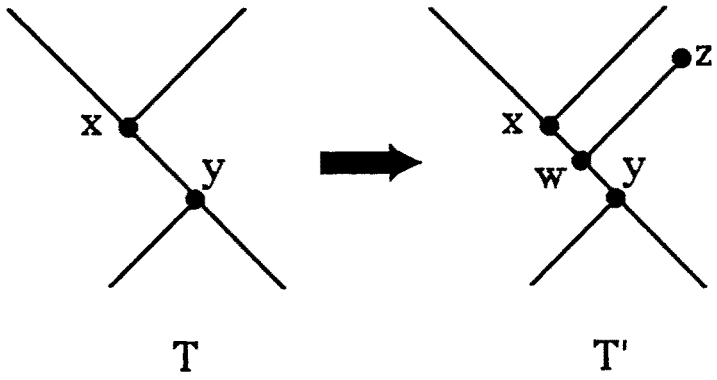

Figure 1 Edge addition

One objective will be to calculate the reduction in the number of trees to be generated as a function of the properties of the imposed subtree consensus. First we will enumerate free (unrooted) binary trees, i.e., trees restricted to nonterminal vertices of degree three, though we will allow the consensus to have nonterminal vertices of arbitrary degree (not less than three). We will show that the reduction in the number of trees depends only on general characteristics of the subtree consensus, and not on its detailed structure We will then show that this independence from structural detail also holds in the enumeration of nonbinary, or multifurcating, trees.

We make the following definitions

1. Let T be a free tree with n labeled terminal vertices and m nonterminal vertices An *edge addition* transforms T into T', a tree with $n + 1$ labeled terminal vertices. This is carried out, as in Figure 1, by replacing any of the b edges of T, say the edge between x and y (where x or y may be labeled or not) by two new vertices w and z and three new edges, between x and w, between w and y, and between w and z. Vertex z, which is a terminal vertex, is given a new label. Note that the new tree T' has $b + 2$ edges, $m + 1$ nonterminal vertices, and has been constructed in one of b different ways.

2. A *vertex addition* transforms T into T'', which also has $n + 1$ labeled terminal vertices. This is carried out, as in Figure 2, by adding a new labeled terminal vertex z, plus a new edge between z and any of the preexisting nonterminal vertices T'' has $b + 1$ edges, m nonterminal vertices and has been constructed in one of m ways.

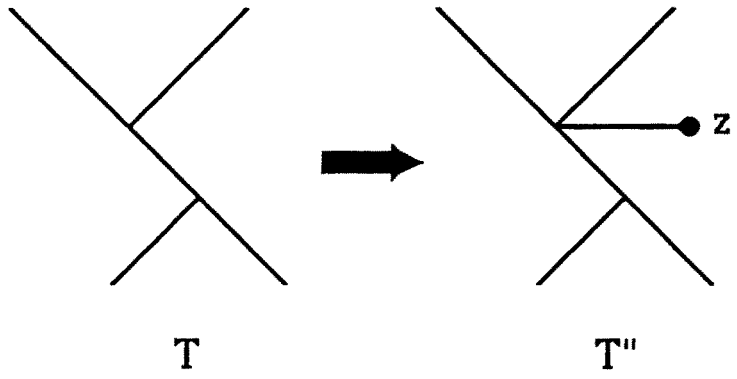

Figure 2 Vertex addition

3 We say of two trees T and T' with labeled terminal vertices, that T is a *subtree* of T' if $T' = T$ or if T' can be constructed from T by a series of edge additions and/or vertex additions as in Figure 1 or Figure 2 Note that if T and U are different trees with the same labeled terminal vertices, not both of them can be subtrees of T'

4 An *edge contraction* is an operation which transforms a tree U to a tree T by replacing two adjacent nonterminal vertices v_1 and v_2, of degrees d_1 and d_2 respectively, as well as their common edge, by a single vertex v whose adjacent vertices are the remaining vertices adjacent to either v_1 or v_2 The degree of v is $d_1 + d_2 - 2$ This is illustrated in Figure 3.

5 A tree U' is *compatible* with a tree T if U' has a subtree U with the same terminal vertex labels as T, where $U = T$ or U can be transformed to T through a series of edge contractions, as in Figure 4

The notion of consensus of free trees is not often encountered in the literature Nevertheless it is as appropriate in the context of Steiner trees (e.g., Graham and Foulds 1982), Wagner trees (Farris 1970), or other unrooted, nonhierarchically constructed trees, as is the more familiar consensus in the context of Lance-Williams (1967), Camin-Sokal (1965), Dollo (Le Quesne 1974, Farris 1977) and other inherently rooted trees. The (strict) consensus of $T_1, \ldots, T_h$, free trees on n labeled terminal vertices, is the unique tree on n labeled terminals with which all of $T_1, \ldots, T_h$ are compatible, and which has the maximum number of nonterminal vertices possible, i e , conserves as much as possible the branching information contained in the individual $T_1, \ldots, T_h$ (Uniqueness is easily proved using the

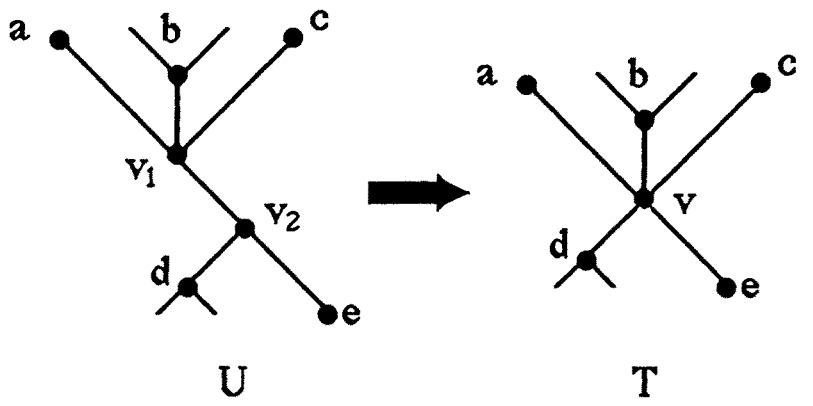

Figure 3 Edge contraction

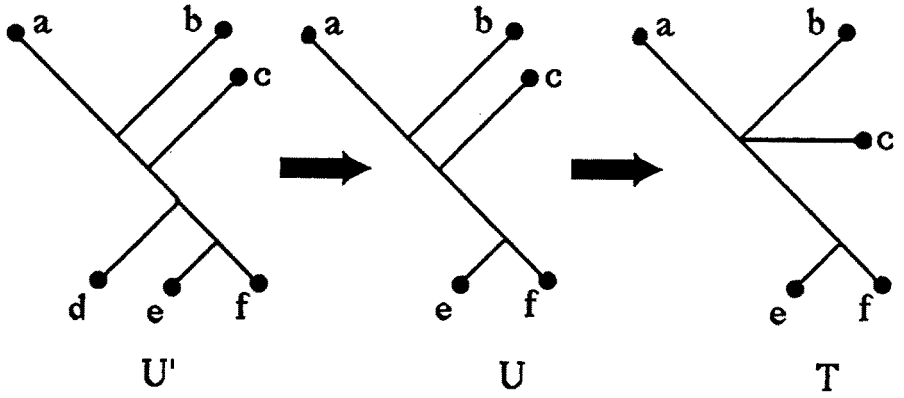

Figure 4 U' compatible with T

bipartition notation for trees, cf. Buneman 1971, Waterman and Smith 1978.)

We first ask how many binary trees on n terminal vertices are compatible with a given tree T on $k \leqslant n$ terminal vertices. Let $N(n) = 1 \cdot 3 \cdot 5 \cdot \cdots \cdot (2n-5) = (2n-4)!/\,2^{n-2}(n-2)!$. Essentially the following result was stated by Rohlf (1982).

Theorem 1 *Let T be a free tree with k labeled terminal vertices and nonterminal vertices $v_1, \ldots, v_m$, where $m \leqslant k-2$, of degree $d_1, \ldots, d_m$ respectively (all $d_i \geqslant 3$). Then the number of free binary trees whose $n \geqslant k$ labeled terminal vertices include the k terminals of T, and which are compatible with T, is $[N(n)/N(k)] \prod_{i=1}^{m} N(d_i)$.*

Proof We first consider T to be a binary tree. In this case it is necessary to prove only that the number of binary trees with n terminal vertices, and with T as a subtree, is $N(n)/N(k)$. If $n = k$, only T contains itself and $N(n)/N(k) = 1$ For $n > k$, how many ways can we construct a tree with n labeled terminal vertices from one with only k, using a series of edge additions only? (Vertex additions would destroy the binary character of the tree.) Supposing $N(n)/N(k)$ correctly counts compatible trees for a certain value of n, then $N(n+1)/N(k)$ is correct for trees on $n+1$ terminals, since each edge addition can be carried out in $2n-3$ different ways, there being $2n-3$ edges in a binary tree, and $(2n-3)\,N(n) = N(n+1)$. By induction, $N(n)/N(k)$ is the number of binary trees with n terminals and with T as a subtree, for all n.

Now suppose T contains a vertex v which has degree $d > 3$. Let $v_1, \ldots, v_d$ be the vertices adjacent to v. Consider any unrooted binary tree V on d labeled terminal vertices. We can create a new tree W by replacing v and its incident edges in T by V, as in Figure 5, where the d labeled terminal vertices of V are stripped of their labels while being identified with $v_1, \ldots, v_d$. Each such V, of which there are $N(d)$, yields a different W And each such W is compatible with T since it may be transformed back into T by contracting all the edges in V, in any order, except those incident with $v_1, \ldots, v_d$

This procedure may be repeated on any remaining vertex in W which has degree greater than three. Continuing in this way we must eventually arrive at a binary tree U on k terminal vertices which is compatible with T. Note that the order in which the high degree vertices are replaced is immaterial, since the replacement of v by V does not change the degree of any other vertices There are clearly $\prod_{i=1}^{m} N(d_i)$ different trees U which can be obtained by this process (recall $N(3) = 1$).

Any tree on n terminal vertices containing any of these trees U as a subtree is also compatible with T, by definition, and there are $N(n)/N(k)$ binary trees specific to each such subtree U. •

Note that T is the strict consensus of all the subtrees U constructed in the proof above

This result tells us that the number of trees compatible with T is not dependent on the detailed shape of T; all that matters is n, k and the degrees of the nonterminal vertices

To apply Theorem 1 in the context of searching for optimal trees, we may calculate how large k must be to compensate for the rapid increase in $N(n)$. For example, suppose our computational capacity is limited in terms of the total number of trees which can be evaluated for optimality, and suppose for each n we will consider a fixed binary subtree ($d_i = 3$) on $k = k(n)$ terminal vertices Then Theorem 1 assures us (via Stirling's approximation) that if $1 - k/n$ approaches 0 as fast as $(n \log n)^{-1}$, then

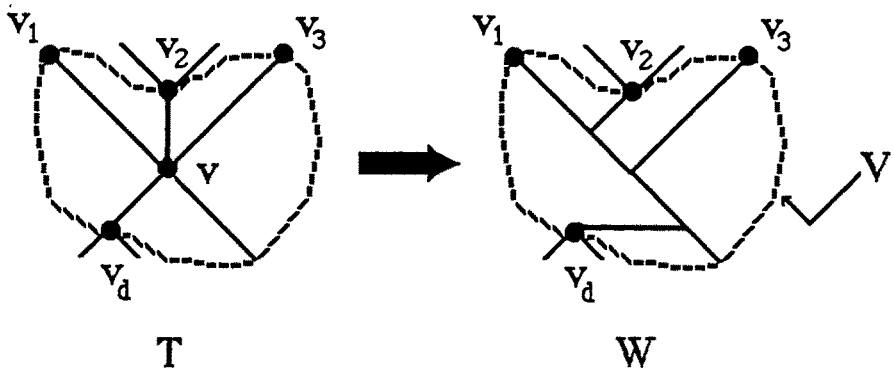

Figure 5 Replacing vertex ($d>3$) by binary tree

the number of trees to be searched is bounded; i e , the search is feasible. On the other hand, if $1 - k/n$ approaches 0 only as fast as n^{-1}, the number of eligible trees increases without bound.

We next consider the enumeration of nonbinary trees Though there is no closed form formula analogous to $N(n)$ for counting multifurcating trees, the basic arguments of Theorem 1 carry through nonetheless in the more general case.

Let $F(n)$ be the number of free trees with n labeled terminal vertices (all nonterminal vertices of degree greater than two) Let $F(n,m)$ be the number of these which have m nonterminal vertices

Theorem 2 *Let T be a free tree with k labeled terminal vertices and nonterminal vertices $v_1, \ldots, v_m$, of degree $d_1, \ldots, d_m$ respectively (all $d_i \geqslant 3$) Then the number of free trees whose $n \geqslant k$ labeled terminals include the k terminals of T, and which are compatible with T, depends only on n, m and $d_1, \ldots, d_m$*

Proof The first step is to determine how many trees on k terminal vertices are compatible with T In a way analogous to our constructions of Theorem 1, we can replace in T any nonterminal vertex v of degree $d>3$ with one of $F(d)$ different trees V to create a new tree W compatible with T This may be done independently for all m nonterminal vertices of T so that there are a total of $\Pi_{i=1}^{m} F(d_i)$ different trees U on k terminal vertices compatible with T This independence means that the number of edges b (and the number of nonterminal vertices $m' = b + 1 - k$), though not the same for all such trees U, occurs with frequency $f_T(b)$, determined entirely by $d_1, \ldots, d_m$

It remains to enumerate the trees on n labeled terminal vertices which contain each U as a subtree. This may be done through counting the different combinations of edge and vertex additions with the well-known recursion

$$F(n,m) = mF(n-1,m) + (n+m-3)\,F(n-1,m-1)$$

Under the initial conditions $f(k,1) = 1$ and $f(3,j) = 0$ for all $k \geqslant 3$ and all $j > 1$, this recursion is used often in counting multifurcating trees (e g , Felsenstein 1978) However, under the initial conditions $F(k,b+1-k) = 1$ and $F(k,j) = 0$ for $j \neq b+1-k$, it can be seen to count all trees on n labeled terminal vertices containing a specific subtree U on k labeled terminal vertices, where U has $m' = b + 1 - k$ nonterminal vertices.

Since the recursion depends only on n, k, and m', and since $f_T(b)$ the number of different subtrees U with m' terminals is determined entirely by $d_1, \ldots, d_m$, the theorem is proved ●

Results analogous to Theorems 1 and 2 may be proved for rooted binary and nonbinary trees as well

Can these theorems be generalized to the case of two or more sets of constraints? The form of Theorem 1 suggests that the number of binary trees with n terminals labels containing binary subtrees $T_1, \ldots, T_r$ with disjoint sets of $k_1, \ldots, k_r$ terminal vertices respectively, might be $N(n)/\prod_{i=1}^{r} N(k_i)$ Indeed for the smallest cases, where $r = 2$, $n \leqslant 10$ and $k_i \leqslant 5$, this suggestion is valid. In general, however, it is not For example, if $r = 2$, $n = 12$, $k_1 = k_2 = 6$, it is easily verified that $N(n)$ is not even divisible by $N(k_1)\,N(k_2)$. Thus if there are two or more sets of constraints, even if these pertain to disjoint sets of terminal vertices, and even if they are all representable by binary structures, the number of trees satisfying them depends on more than just the number of vertices involved

The situation is even more difficult when there is more than one set of constraints and these are not disjoint, as in the problem discussed by Gordon (1986) He defines a "strict consensus supertree" of two or more trees which is compatible (if possible) with all of them, and gives an algorithm for constructing this supertree Our results tie in with his in that the total number of compatible supertrees satisfies our Theorem 2, when applied to the strict consensus supertree

References

AHO, A V , SAGIV, Y , SZYMANSKI, T G , and ULLMAN, J D (1981), "Inferring a Tree from Lowest Common Ancestors with an Application to the Optimization of Relational Expressions," *SIAM Journal on Computing, 10*, 405-421

BUNEMAN, P (1971), "The Recovery of Trees from Measures of Similarity," in *Mathematics in the Archaeological and Historical Sciences*, eds F R Hodson, D G Kendall, and P Tautu, Edinburgh: Edinburgh University Press, 387-395

CAMIN, J H , and SOKAL, R R (1965), "A Method for Deducing Branching Sequences in Phylogeny," *Evolution, 19*, 311-326

FARRIS, J S (1970), "Methods for Computing Wagner Trees," *Systematic Zoology, 19*, 83-92

FARRIS, J S (1977), "Phylogenetic Analysis under Dollo's Law," *Systematic Zoology, 26*, 77-88

FELSENSTEIN, J (1978), "The Number of Evolutionary Trees," *Systematic Zoology, 27*, 27-33

GORDON, A D (1986), "Consensus Supertrees: The Synthesis of Rooted Trees Containing Overlapping Sets of Labeled Leaves," *Journal of Classification, 3*, xx-xx

GRAHAM, R L, and FOULDS, L R (1982), "Unlikelihood that Minimal Phylogenies for a Realistic Biological Study can be Constructed in Reasonable Computational Time," *Mathematical Biosciences, 60*, 133-142

GRAY, M W, SANKOFF, D, and CEDERGREN, R J (1984), "On the Evolutionary Descent of Organisma and Organelles: A Global Phylogeny Based on a Highly Conserved Structural Core in Small Subunit Ribosomal RNA," *Nucleic Acids Research, 12*, 5837-5852

LANCE, G N, and WILLIAMS, W T (1967), "A General Theory of Classificatory Sorting Strategies I Hierarchical Systems," *Computer Journal, 9*, 231-239

LE QUESNE, W J (1974), "The Uniquely Evolved Character Concept and its Cladistic Application," *Systematic Zoology, 23*, 513-517

ROHLF, F J (1982), "Consensus Indices for Comparing Classifications," *Mathematical Biosciences, 59*, 131-144

SANKOFF, D, CEDERGREN, R J, and MCKAY, W (1982), "A Strategy for Sequence Phylogeny Research," *Nucleic Acids Research, 10*, 421-431

WATERMAN, M S, and SMITH, T F (1978), "On the Similarity of Dendrograms," *Journal of Theoretical Biology, 73*, 789-800