Scan A5172

Morris & Zaslavsky

add to 4 seqs

A5172- A5175

finished 5/91

# The Number of Cladistic Characters

F. R. McMORRIS
*Department of Mathematics, Bowling Green State University, Bowling Green, Ohio 43403*

AND

THOMAS ZASLAVSKY
*Department of Mathematics, The Ohio State University, Columbus, Ohio 43210*

*Received 13 August 1980*

ABSTRACT

A cladistic character can be viewed as a type of set-labeled tree. This representation is used to derive a recurrence equation giving the number $t(n,r)$ of cladistic characters on $n$ species having $r$ states. Values for $t(n,r)$ are given for $r$ up to 5 and $n$ up to 30.

## INTRODUCTION

Felsenstein [5] gave nice recursion formulas that enabled him to compute the numbers of various sorts of evolutionary trees. A companion problem to the one of counting evolutionary trees is that of counting the number of taxonomic characters that are possible for a particular study collection of evolutionary units (EUs). These problems, as noted by Felsenstein, certainly are not the most pressing for taxonomy, but we feel that they present interesting challenges and may be useful in probabilistic investigations.

## TREES OF SUBSETS

The type of taxonomic character that we consider here has been called a *cladistic character* (Estabrook, Johnson and McMorris [2]). Recently (Estabrook and McMorris [3]) the definition of cladistic character was slightly changed so that we could perform a more elegant analysis using the concept of *trees of subsets*. Each cladistic character on $S$ corresponds in a natural way to precisely one tree of subsets of $S$, and each tree of subsets of $S$ corresponds to precisely one cladistic character. To illustrate this before we give a formal definition we refer to Fig. 1.

*MATHEMATICAL BIOSCIENCES* 54: 3–10 (1981)





0025-5564/81/030003+8$02.50

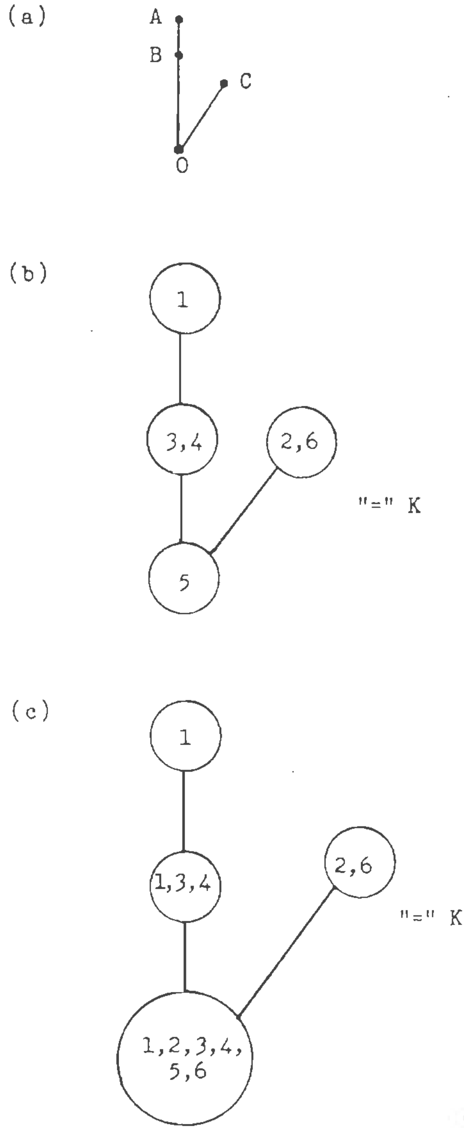


FIG. 1. (a) The character state tree of $K$ (see text). (b, c) Two equivalent ways of representing $K$.

Suppose each EU in $S$ is identified with a positive integer. In this example, $S=\{1,2,3,4,5,6\}$. The character state tree of the cladistic character $K$ is given in Fig. 1(a). Notice that it is required that the character state tree be

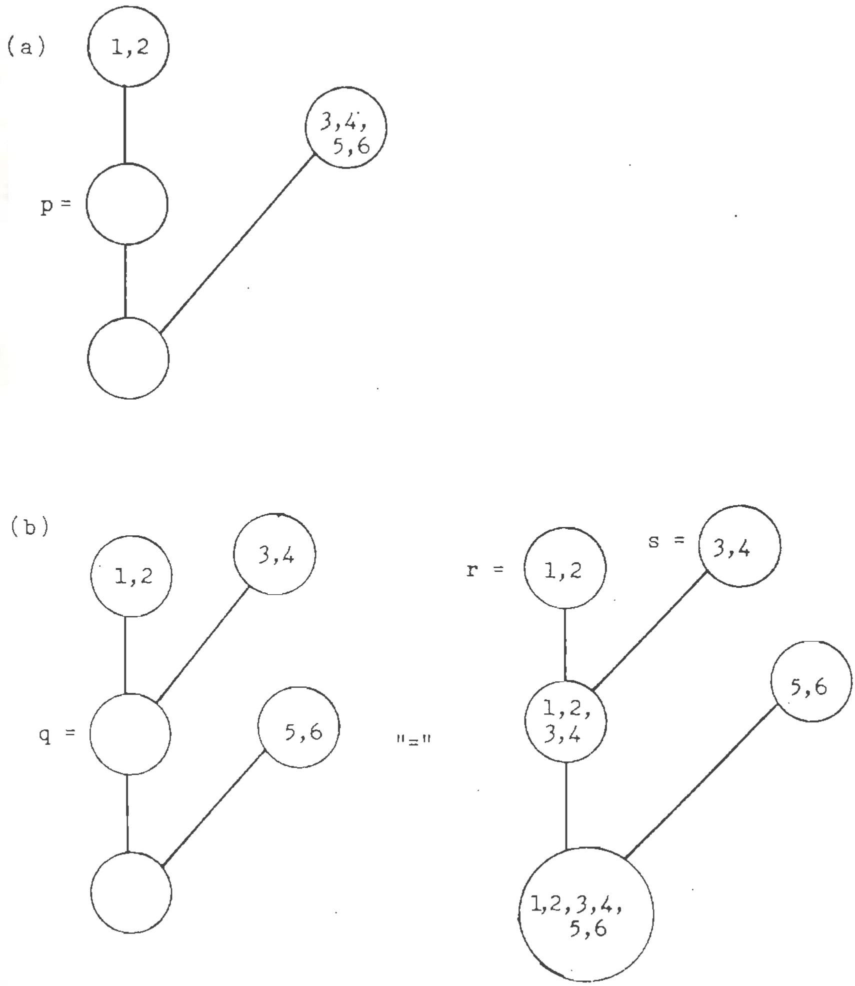


FIG. 2. (a) This tree is not the character state tree of a cladistic character and therefore does not have a representation as a tree of subsets. The problem is that the state $p$ is not the greatest lower bound of occupied states (see Estabrook and McMorris [3]). (b) The empty state $q$ of this character state tree is the greatest lower bound of the occupied states $r$ and $s$.

directed from primitive to advanced. Figure 1(b) indicates that EU 1 possesses state A of character $K$, EUs 3 and 4 state B, EUs 2 and 6 state C, and EU 5 the most primitive state 0. Figure 1(c) illustrates how $K$ can then be considered as the tree of subsets $K=\{\{1\},\{1,3,4\},\{2,6\},\{1,2,3,4,5,6\}\}$. A detailed account of this process can be found in Estabrook and McMorris [3]. The reader should also look at Fig. 2 for another example.

A precise definition is as follows: If $S=\{1,2,\ldots,n\}$, then a *tree of subsets* of $S$ is a collection $\mathfrak{T}$ of nonempty subsets of $S$ such that

(i) $S\in\mathfrak{T}$, and
(ii) if $A,B\in\mathfrak{T}$ and $A\cap B\neq\varnothing$, then $A\subseteq B$ or $B\subseteq A$.

Thus the following set of subsets of $\{1,2,3,4,5,6\}$ is not a tree of subsets and could not be obtained from a character state tree: $\{\{1\},\{1,3,4\},\{2,6\},\{4,5\},\{1,2,3,4,5,6\}\}$. Here $\{1,3,4\}\cap\{4,5\}=\{4\}\neq\varnothing$, but neither set is a subset of the other one.

Since every cladistic character on $S$ is, in effect, a tree of subsets of $S$, the problem of counting the number of possible cladistic characters on $S$ reduces to that of counting the number of trees of subsets of $S$. The method of character compatibility analysis (Estabrook and Anderson [1]; Estabrook, Strauch, and Fiala [4]; Strauch [7]) uses cladistic chararacters to construct estimates of the evolutionary history of a group $S$ of EUs. These evolutionary trees can be considered equivalent to cladistic characters on $S$. Hence we will be counting the number of distinct estimates of the evolutionary history of $S$ that are of this type.

## THE NUMBERS

We will say that a tree of subsets $T$ has $r$ nodes if $T$ contains $r$ elements (i.e., subsets of $S$). Let $t(n,r)$ denote the number of trees of subsets of an $n$-element set $S$ having $r$ nodes, and let $t(n)$ denote the number of trees of subsets of $S$. Then clearly

$$t(n)=\sum_{r=1}^{2n-1} t(n,r).$$

A recurrence equation for the calculation of $t(n,r)$ is obtained by counting the number of ways to enlarge a tree $T^*$ of subsets of an $(n-1)$-element set $S^*$ to a tree of subsets of $S^*$ with an additional element, which we label $n$, adjoined. For a set $X$ we use the notation $nX$ for $X$ with $n$ adjoined; and $S=nS^*$.

Suppose $T^*$ has $r$ nodes. Figure 3 shows the ways $n$ can be added to $T^*$. First notice that in adding $n$ to $T^*$ there is exactly one greatest node ("highest" on the tree) to which it is added. (This is equivalent to adding $n$ to an already existing node on the original character state tree.) Call this node $C$. The simplest tree of subsets of $S$ is obtained by just adding $n$ to $C$ and every node below it. This tree, called $T_1$, has $r$ nodes. A tree $T_2$ with $r+1$ nodes is derived from $T_1$ by adding $n$ as an extra node above $nC$. Two more trees, $T_3$ with $r+1$ nodes and $T_4$ with $r+2$ nodes, are obtained respectively from $T_1$ and $T_2$ by keeping $C$ itself as a node just above $nC$.

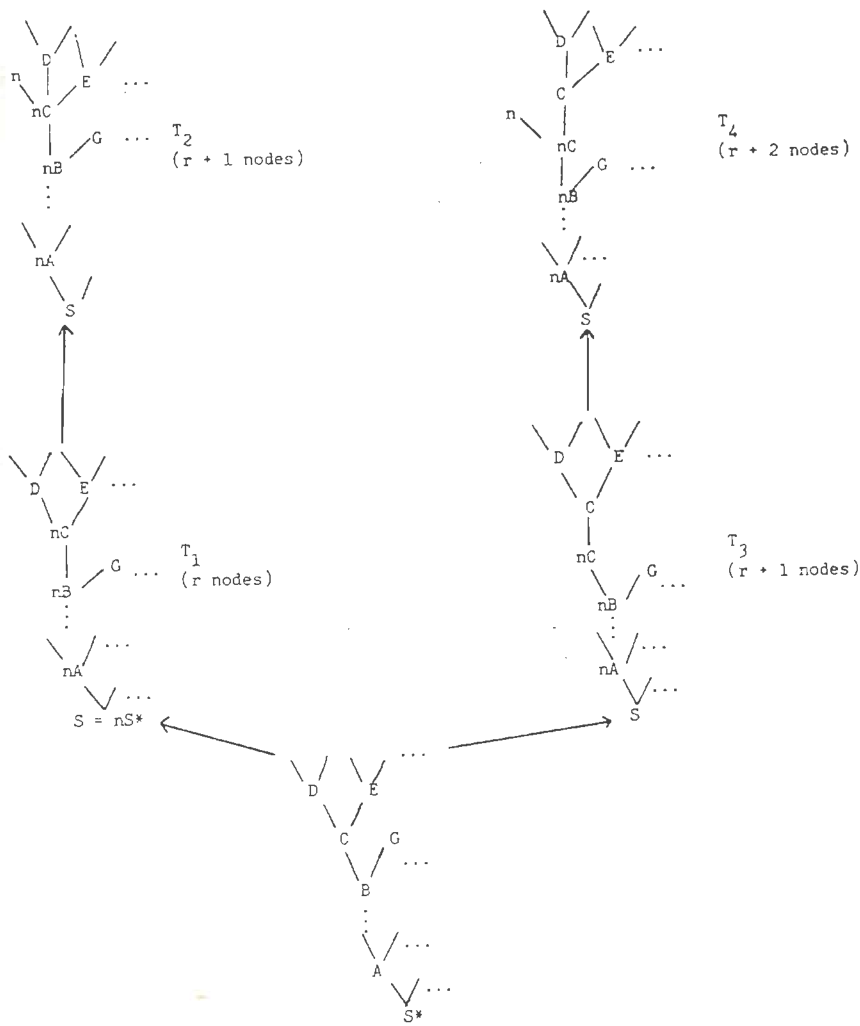


FIG. 3. The four ways that the element $n$ can be added to the tree of subsets $T^*$ of a set with $(n-1)$-elements $S^*$ to produce a tree of subsets of $nS^*$.

Each of these four trees of subsets of $S$ has the property that, if $n$ is deleted from every node (throwing away the empty set and the duplicate $C$ node if necessary), one gets $T^*$ back again. And the only trees of subsets of $S$ which give $T^*$ back upon deleting $n$ are $T_1$, $T_2$, $T_3$, and $T_4$. Thus if $n$ is added as we describe to all possible trees of subsets of $S^*$, we will get all trees of subsets of $S$ without duplication.

Let us count up the trees of subsets of the $n$-element set $S$ having $r$ nodes according to the size of the tree $T^*$ from which they were derived. We find that

$r$ of type $T_1$ come from each $T^*$ with $r$ nodes (one tree $T_1$ for each choice of the node $C$ in $T^*$),
$r-1$ of type $T_2$ come from each $T^*$ with $r-1$ nodes,
$r-1$ of type $T_3$ come from each $T^*$ with $r-1$ nodes,
$r-2$ of type $T_4$ come from each $T^*$ with $r-2$ nodes.

This gives

$$t(n,r)=rt(n-1,r)+2(r-1)t(n-1,r-1)+(r-2)t(n-1,r-2) \quad (1)$$

with the initial conditions $t(n,1)=1$ for all $n \geqslant 1$ and $t(1,r)=0$ for all $r \geqslant 2$.

In an actual study, each cladistic character will usually have less than five states. (For example, see Gardner and LaDuke [6], Strauch [7], Estabrook and Anderson [1].) We give the exact values of $t(n,r)$ for $r$ up to 5. They can be proved by induction on $n$ using (1):

$$t(n,1)=1,$$

$$t(n,2)=2^n-2,$$

$$t(n,3)=\tfrac{3}{2}\cdot 3^n-4\cdot 2^n+\tfrac{7}{2},$$

$$t(n,4)=\tfrac{8}{3}\cdot 4^n-9\cdot 3^n+11\cdot 2^n-\tfrac{17}{3},$$

$$t(n,5)=\tfrac{125}{24}\cdot 5^n-\tfrac{64}{3}\cdot 4^n+\tfrac{135}{4}\cdot 3^n-\tfrac{76}{3}\cdot 2^n+\tfrac{209}{24}.$$

TABLE 1

The Number of Trees of Subsets of a Set of $n$ Elements

| $n$ | Number of trees of subsets |
|---|---:|
| 1 | 1 |
| 2 | 4 |
| 3 | 32 |
| 4 | 416 |
| 5 | 7,552 |
| 6 | 176,128 |
| 7 | 5,018,624 |
| 8 | 168,968,192 |
| 9 | 6,563,282,944 |
| 10 | 288,909,131,776 |
| 11 | 14,212,910,809,088 |
| 12 | 772,776,684,683,264 |
| 13 | 46,017,323,176,296,448 |
| 14 | 2,978,458,881,388,183,550 |
| 15 | 208,198,894,956,559,677,000 |

New
A5172

AS173 AS174 AS175

TABLE 2

The Number of Trees of Subsets of an *n*-Element Set Having *r* Nodes

| *n* \ *r* | 2 ($2^n-2$) | 3 | 4 | 5 |
|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 |
| 2 | 2 | 1 | 0 | 0 |
| 3 | 6 | 12 | 10 | 3 |
| 4 | 14 | 61 | 124 | 131 |
| 5 | 30 | 240 | 890 | 1,830 |
| 6 | 62 | 841 | 5,060 | 16,990 |
| 7 | 126 | 2,772 | 25,410 | 127,953 |
| 8 | 254 | 8,821 | 118,524 | 851,361 |
| 9 | 510 | 27,480 | 527,530 | 5,231,460 |
| 10 | 1,022 | 84,481 | 2,276,020 | 30,459,980 |
| 11 | 2,046 | 257,532 | 9,613,010 | 170,761,503 |
| 12 | 4,094 | 780,781 | 40,001,324 | 931,484,191 |
| 13 | 8,190 | 2,358,720 | 164,698,170 | 4,979,773,890 |
| 14 | 16,382 | 7,108,921 | 672,961,380 | 26,223,530,970 |
| 15 | 32,766 | 21,392,292 | 2,734,531,810 | 136,522,672,653 |
| 16 | 65,534 | 64,307,941 | 11,066,546,524 | 704,553,794,621 |
| 17 | 131,070 | 193,185,960 | 44,652,164,810 | 3,611,494,269,120 |
| 18 | 262,142 | 580,082,161 | 179,768,037,140 | 18,415,268,221,960 |
| 19 | 524,286 | 1,741,295,052 | 722,553,165,810 | 93,516,225,653,403 |
| 20 | 1,048,574 | 5,225,982,301 | 2,900,661,482,124 | 473,366,777,478,651 |
| 21 | 2,097,150 | 15,682,141,200 | 11,634,003,919,450 | 2,390,054,857,197,150 |
| 22 | 4,194,302 | 47,054,812,201 | 46,630,112,719,300 | 12,043,393,363,764,950 |
| 23 | 8,388,606 | 141,181,213,812 | 186,802,788,139,010 | 60,590,148,885,015,753 |
| 24 | 16,777,214 | 423,577,195,861 | 748,058,256,616,124 | 304,445,590,273,832,281 |
| 25 | 33,554,430 | 1,270,798,696,440 | 2,994,774,523,194,090 | 1,528,213,688,153,677,980 |
| 26 | 67,108,862 | 3,812,530,307,041 | 11,986,722,952,063,860 | 7,665,030,449,350,031,940 |
| 27 | 134,217,726 | 11,437,859,356,572 | 47,969,767,124,315,410 | 38,421,057,467,824,787,900 |
| 28 | 268,435,454 | 34,314,114,940,621 | 191,947,695,921,836,524 | 192,489,079,784,152,131,000 |
| 29 | 536,870,910 | 102,943,418,563,680 | 767,996,668,913,860,730 | 963,981,083,457,036,435,000 |
| 30 | 1,073,741,822 | 308,832,403,174,681 | 3,072,604,337,240,566,820 | 4,826,049,699,337,424,750,000 |

If $n$, the number of EUs, is large, the first term is clearly dominant. Thus we have, for $r \leqslant 5$, the asymptotic estimate

$$t(n,r) \sim \frac{r^{r-1}}{r!} \cdot r^n \quad \text{as} \quad n \to \infty .$$

(This is valid for all $r$, but we omit the proof.)

The formula (1) makes it easy to compute $t(n,r)$ recursively. Table 1 shows $t(n)$ for $n$ up to 15, and Table 2 gives $t(n,r)$ for $n$ up to 30 and $r$ up to 5. These values were computed using a double-precision FORTRAN program written by Tim Margush on an IBM 360/75 at Bowling Green State University. The figures are significant only to 18 digits.

Because a bifurcating tree with $n$ labeled tip EUs has $n-1$ interior nodes, we get the number of bifurcating trees by letting $r=2n-1$. From (1) we have by an easy induction the well-known formula

$$t(n,2n-1)=(2n-3)(2n-5)\cdots(5)(3)(1).$$

*We thank T. Margush for programming assistance, and an unknown conference housing director at Florida Atlantic University for originally getting us together. Thanks also to Christopher Meacham and an anonymous reviewer for helpful comments.*

REFERENCES

1. G. F. Estabrook and W. R. Anderson, An estimate of phylogenetic relationships within the genus *Crusea* (Rubiaceae) using character compatibility analysis, *Syst. Bot.* 3:179–196 (1978).
2. G. F. Estabrook, C. S. Johnson, Jr., and F. R. McMorris, An idealized concept of the true cladistic character, *Math. Biosci.* 29:263–272 (1975).
3. G. F. Estabrook and F. R. McMorris, When is one estimate of evolutionary relationships a refinement of another?, *J. Math. Biology*, 10: 367–373 (1980).
4. G. F. Estabrook, J. G. Strauch, Jr., and K. L. Fiala, An application of compatibility analysis to the Blackith's data on Orthopteroid insects, *Syst. Zool.* 26:269–276 (1977).
5. J. Felsenstein, The number of evolutionary trees, *Syst Zool.* 27:27–33 (1978).
6. R. C. Gardner and J. C. LaDuke, Phyletic and cladistic relationships in *Lipochaeta* (Compositae), *Syst. Bot.* 3:197–207 (1978).
7. J. G. Strauch, Jr., The phylogeny of the *Charadriiformes* (Aves): a new estimate using the method of character compatibility analysis, *Trans. Zool. Soc. Lond.* 34:263–345 (1978).