Scan A5373
A5264

Felsenstein

add to £4 sop

THE NUMBER OF EVOLUTIONARY TREES

JOSEPH FELSENSTEIN

Abstract

Felsenstein, J. (Department of Genetics, University of Washington, Seattle, Washington 98195) 1978. The number of evolutionary trees. Syst. Zool. 27:27–33.—A simple method of counting the number of possible evolutionary trees is presented. The trees are assumed to be rooted, with labelled tips but unlabelled root and unlabelled interior nodes. The method allows multifurcations as well as bifurcations. It makes use of a simple recurrence relation for T(n,m), the number of trees with n labelled tips and m unlabelled interior nodes. A table of the total number of trees is presented up to n = 22. There are 282,137,824 different trees having 10 tip species, and over 8.87×10^{23} different trees having 20 tip species. The method is extended to count trees some of whose interior nodes may be labelled. The principal uses of these numbers will be to double-check algorithms and notation systems, and to frighten taxonomists. [Evolutionary trees; cladistic methods; combinatorial algorithms; graphs.]

Estimating the phylogeny of a group involves selecting one evolutionary tree from among a large number of possibilities. Most taxonomists probably do not stop to think how many possibilities there are. The calculation of this number is hardly the most vital task facing contemporary taxonomy, but it may be of some educational value to persons proposing phylogenetic methods. In particular, if a computer procedure involves evaluation of all possible trees, knowledge of their number is useful as a check on whether all possibilities are in fact being considered.

The counting of trees has been a mathematician's diversion ever since the pioneering work of Cayley (1856). More recent work has been reviewed by Moon (1970). Harding (1971) has considered the shape of unlabelled trees generated by a random speciation process. Cavalli-Sforza and Edwards (1967) have counted the number of bifurcating trees having n unlabelled tips, as have Dobson (1974a, b) and Phipps (1976a). Dobson (1974a, b) has computed the number of unrooted bifurcating trees with n unlabelled tips.

The feature of evolutionary trees which makes them unusual among the "trees" of graph theory is that some (usually all) of their interior nodes (forks) are unlabelled. Figure 1 illustrates the terminology. When all interior nodes must have labels, in addition to the tip species, one can adapt Cayley's (1889) enumeration of the number of labelled unrooted trees to enumerate the number of labelled rooted trees. Harper (1976) has done this: he finds that there are n^{n-1} such trees having n species. But this case is not the one most commonly encountered. Usually, the tip species are labelled but the interior nodes are not.

Two such cases were treated by Schröder (1870). The first is the case in which only bifurcations are allowed at each interior node. Edwards and Cavalli-Sforza (1964; Cavalli-Sforza and Edwards, 1967) have used a much simpler method to compute the same quantities. It will be of interest to present their method, as the results in this paper are generalizations of it.

COUNTING BIFURCATING TREES

Edwards and Cavalli-Sforza's argument is recursive: it computes the number of bifurcating trees with n tip species in terms of the number with $n - 1$ tip species. They make use of the fact that there is exactly one way to construct a given n-species tree by successively adding new species to the tree in the order 1, 2, 3, . . . , n, starting from the single one-species tree containing only species 1.

This should be reasonably obvious, but in case it is not, it can be justified as follows. There is only one result possible

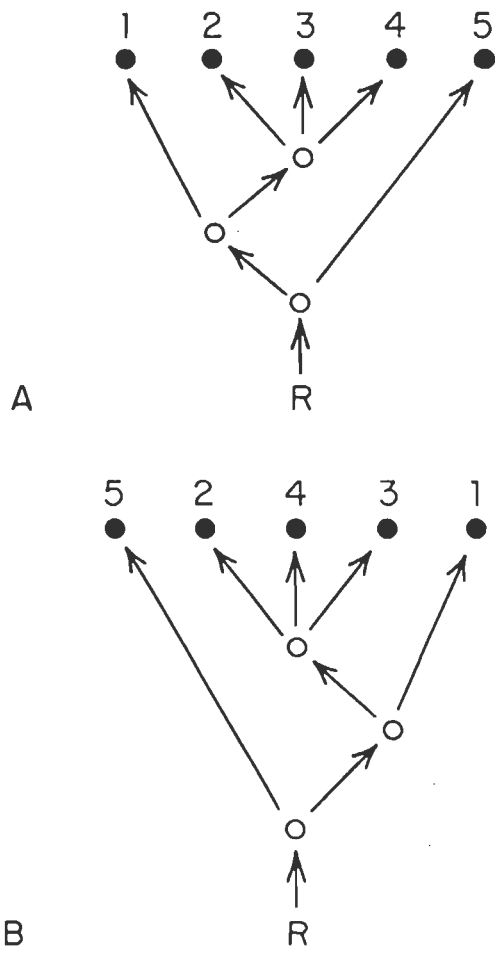

FIG. 1.—(A) A rooted tree with labelled tips and unlabelled interior nodes. (B) The same tree, drawn differently to show the sorts of rearrangements (changing the order of branches at a fork) which for the purposes of this paper do not result in a different tree. The labelled nodes are indicated by open circles, and the root of the tree (which is not considered to be a node) by R.

(for a given tree) if we remove species *n* from the tree. Furthermore the operation of adding species *n* to this tree (in the proper place) exactly reverses this removal. These operations are exact inverses of one another. There is one sequence of removal events which leads from a tree with species 1, 2, . . . , n down to a tree with only species 1, provided that we remove species n, n − 1, n − 2, . . . , 2 in that order. So there is at least one way to add species 2, 3, . . . , n to obtain the given tree with *n* species. Furthermore there is only one such way. If there were two, then for some *k*, two different trees containing species 1, 2, . . . , k would exist such that we could add species k + 1, k + 2, . . . , n to both and obtain the given tree. But this is impossible, since removal of species n, n − 1, . . . k + 1 would then result in two different trees of size *k*. Since that is impossible, we conclude that there is a one-to-one correspondence between the ways of adding species 2, 3, 4, . . . , n (in that order) and the resulting trees.

Edwards and Cavalli-Sforza simply noted that when one was adding the k-th species, to have the result be a bifurcating tree, the new species had to be added by creating a new unlabelled interior node in the middle of a segment of the tree, and the k-th species had to arise from that node. Figure 2 shows all possible additions of a fourth species to a three-species tree giving bifurcating trees as the result. When the k-th species is being added, there are 2k − 3 places it can be added. This is because each segment of the tree has a node at its upper end. There are thus k − 1 segments leading to tips, and k − 2 leading to interior nodes, for a total of 2k − 3. Then the number of different bifurcating trees with *n* labelled tip species is, for n ≥ 2,

$$1 \cdot 3 \cdot 5 \cdot \cdots (2n-3) = \prod_{k=2}^{n} (2k-3) = \frac{(2n-3)!}{2^{n-2}(n-2)!} \quad (1)$$

Edwards and Cavalli-Sforza originally presented (1964) a calculation of the number of *unrooted* bifurcating trees with *n* labelled tips; the above result is given by them in a later paper (Cavalli-Sforza and Edwards, 1967). It can also be obtained from a result of Moon (1970:6) concerning the number of completely labelled (unrooted) trees with a given sequence of degrees of nodes. Phipps (1976b) computed this number by direct

enumeration methods, and guessed formula (1) from the results.

COUNTING MULTIFURCATING TREES

The tree shown in Fig. 1 does not fit into the class of bifurcating trees, because one of its nodes is a trifurcation. Schröder (1870) developed methods for enumerating the number of trees with n labelled tip species when multifurcations were allowed. His methods were somewhat complex. In this section I present a considerably simpler method which makes computation straightforward. The key to this approach lies in noticing that when a tree contains a multifurcation, it has fewer than n − 1 interior nodes. Let us classify n-species trees by the number of their interior nodes. They may have between 1 and n − 1 interior nodes. Let T(n, m) be the number of distinct trees having n (labelled) tip species and m (unlabelled) interior nodes. If we can compute the T(n, m), then the total number of trees with n tip species will be the sum of the T(n, m) over all values of m.

The method will be a direct extension of that of Edwards and Cavalli-Sforza. We will compute the T(n, m) from the T(n − 1, m) by counting the number of ways the n-th species can be added to the tree. There will be a one-to-one correspondence between the ways of adding species 2, 3, . . . , n and the n-species trees, since the argument to this effect in the previous section did not apply only to bifurcating trees. Clearly T(1, 0) = 1 and all other T(1, i) = 0. Suppose that we know all the T(n − 1, i) and wish to compute T(n, m).

If we add species n to a tree and obtain a tree with m interior nodes, this could happen in two ways:

(i) We could take a tree with n − 1 tip species and m interior nodes and have species n arise from one of those interior nodes. For each of the T(n − 1, m) trees of this sort there are then m places at which species n could be added.

(ii) We could take a tree with n − 1 tip species and m − 1 interior nodes, place a new interior node in the midst of one of its segments, and have species n arise from this new node. There are T(n − 1, m − 1) such trees, and each has n + m − 2 interior segments (since each interior segment has at its upper end either a tip species or an interior node). This second way of adding species n is only possible when the resulting m is greater than one.

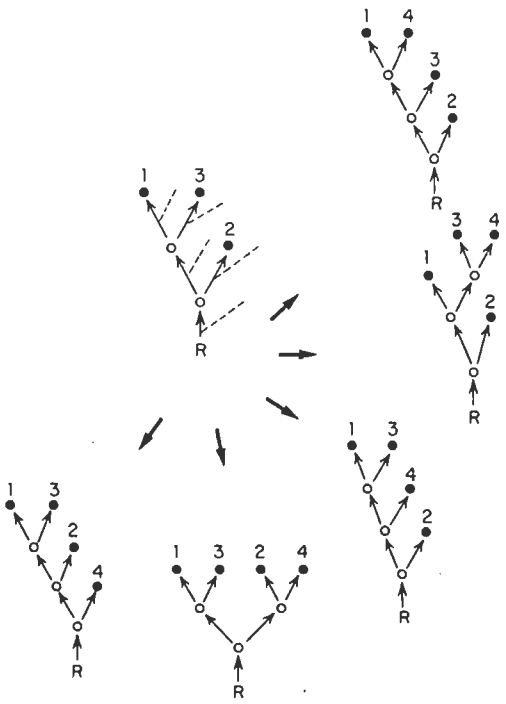

FIG. 2.—All ways in which a fourth species can be added to a given 3-species bifurcating tree so as to result in a bifurcating tree.

By the one-to-one correspondence between ways of adding species n and resulting trees, we find the following recurrence relation for the T(n,m):

$$T(n, m) = \begin{cases} m\,T(n-1, m) \\ \quad + (n + m - 2) \\ \quad \cdot T(n-1, m-1) & (m > 1) \\ T(n-1, 1) & (m = 1) \end{cases} \qquad (2)$$

for n > 1.

This algorithm is simple enough to compute by hand for small n. For larger n a computer is necessary. I have computed the T(n, m) up to n = 22 using a double-precision FORTRAN program on

the CDC 6400 computer at the University of Washington Academic Computing Center. Above n = 22 the total number of trees exceeds 2^{96}, and hence cannot be exactly represented by a double-precision number in this computer.

Table 1 shows the numbers of bifurcating and multifurcating trees for different values of *n*. The bifurcating trees are the T(n, n − 1), and are the same as given by Cavalli-Sforza and Edwards. The numbers of multifurcating trees are the sum of the T(n, m) over all values of *m*, so that the bifurcating trees are included. The number of multifurcating trees rises faster than the number of bifurcating trees: for n = 20 there are over 100 times as many of the former as of the latter. Using the individual T(n, m), which are not given here, one can also compute the average number of interior nodes in randomly constructed trees with *n* species. It is

$$\sum_{m} m\,T(n, m) \Big/ \sum_{m} T(n, m).$$

TREES WITH ANCESTORS PARTLY LABELLED

Harper (1976) has shown that there are n^{n-1} trees having a total of *n* labelled nodes, with no unlabelled ancestors. It is the more usual situation in paleontology to know the identities of some, but not all, of the ancestors in the tree. It may therefore be of interest to compute the number of trees with *n* labelled species, where each species may be either at a tip or at an interior node. In such trees, all of the tips must be labelled, but not all of the interior nodes need be labelled.

Let us denote by U(n, m) the number of different *n*-species trees with *m* unlabelled interior nodes (allowing some of the *n* species to be labelled interior nodes). For n = 1, we clearly have U(1, 0) = 1 and all other U(1, i) = 0. We can add the *n*-th species to a tree, so as to end up with *n* species and *m* unlabelled nodes in four different ways:

(i) By inserting it into any of the n − 1 + m segments of any of the U(n − 1, m) trees with *m* unlabelled nodes. The result is that species *n* becomes a labelled interior node.

(ii) By placing an unlabelled node in any of the (n − 1) + (m − 1) segments of any of the U(n − 1, m − 1) trees having m − 1 unlabelled nodes, and by having the *n*-th species arise from this new unlabelled node. This is only possible if m > 0.

(iii) By labelling any of the m + 1 unlabelled nodes in any of the U(n − 1, m + 1) trees having m + 1 unlabelled nodes. This is only a possibility if n − 1 > m + 1, as otherwise there would be no such trees available.

(iv) By adding species *n* as the immediate descendant of any one of the (n − 1) + m nodes in any of the U(n − 1, m) trees having *m* unlabelled interior nodes.

The resulting recurrence relation is, for n > 1,

$$\begin{aligned} U(n, m) = {} & (n + m - 2)\,U(n - 1, m - 1) && (m > 0) \\ & + 2\,(n + m - 1)\,U(n - 1, m) \\ & + (m + 1)\,U(n - 1, m + 1). && (n > m + 2) \end{aligned} \qquad (3)$$

The conditions to the right of each term specify when the term is included in the expression. When each is not satisfied, that term must be taken to be zero.

Table 2 shows the result of applying (3), up to n = 19. Notice how much larger the numbers of trees are than the numbers in Table 1. They are also much larger than Harper's (1976) value of n^{n-1}, which would be U(n, 0).

BIFURCATING TREES WITH ANCESTORS PARTLY LABELLED

To restrict attention to bifurcating trees, the computation must be made somewhat more complex. There are still the same four ways of adding species *n* to the tree, but one of them, number (iv), must be restricted. To avoid multifurcations, the new species can be allowed to

TABLE 1. THE NUMBERS OF ROOTED TREES WITH n LABELLED TIPS AND WITH UNLABELLED INTERIOR NODES. THE LEFT COLUMN COUNTS ALL TREES, THE RIGHT COLUMN ONLY BIFURCATING TREES.

n	All trees	Bifurcating trees
1	1	1
2	1	1
3	4	3
4	26	15
5	236	105
6	2,752	945
7	39,208	10,395
8	660,032	135,135
9	12,818,912	2,027,025
10	282,137,824	34,459,425
11	6,939,897,856	654,729,075
12	188,666,182,784	13,749,310,575
13	5,617,349,020,544	316,234,143,225
14	181,790,703,209,728	7,905,853,580,625
15	6,353,726,042,486,112	213,458,046,676,875
16	238,513,970,965,250,048	6,190,283,353,629,375
17	9,571,020,586,418,569,216	191,898,783,962,510,625
18	408,837,905,660,430,516,224	6,332,659,870,762,850,625
19	18,522,305,410,364,568,764,416	221,643,095,476,699,771,875
20	887,094,711,304,094,583,095,296	8,200,794,532,637,891,559,375
21	44,782,218,857,751,551,087,214,592	319,830,986,772,877,770,815,625
22	2,376,613,641,928,796,906,249,519,104	13,113,070,457,687,988,603,440,625

arise directly from a pre-existing interior node only if that node has only one immediate descendant. It must therefore always be a labelled interior node (although some of the labelled interior nodes will have two immediate descendants). There will be different numbers of ways of adding species n to a tree, depending on how many of these eligible labelled interior nodes there are. Letting n = number of labelled nodes, m = number of labelled interior nodes with two descendant, and p = number of labelled interior nodes with exactly one descendant, we wish to compute V(n, m, p), so that by addition over all m and p we can obtain the total number of bifurcating trees with partially labelled interior nodes (as before, including the case where none is labelled). The recurrence relation is:

$$
\begin{aligned}
V(n, m, p) = \\
&(2n - 2m - p - 2)\, V(m - 1, m, p - 1) && (p > 0) \\
+\ &(2n - 2m - p - 3)\, V(n - 1, m, p) && (2n - 2m - p - 3 > 0) \\
+\ &(n - 2m - p)\, V(n - 1, m - 1, p) && (m > 0) \\
+\ &(p + 1)\, V(n - 1, m - 1, p + 1) && (m > 0) \\
+\ &(n - m - p)\, V(n - 1, m, p - 1). && (p > 0)
\end{aligned}
\tag{4}
$$

TABLE 2. THE NUMBERS OF ROOTED TREES WITH n LABELLED SPECIES, ALLOWING MULTIFURCATIONS AND ALLOWING SOME INTERIOR NODES TO BE LABELLED.

n	Number of trees
1	1
2	3
3	22
4	262
5	4,336
6	91,984
7	2,381,408
8	72,800,928
9	2,566,606,784
10	102,515,201,984
11	4,575,271,116,032
12	225,649,908,491,264
13	12,187,240,730,230,208
14	715,392,567,595,384,832
15	45,349,581,052,868,558,848
16	3,087,516,727,770,917,896,192
17	224,691,760,916,824,988,844,032
18	17,406,010,163,636,762,337,869,824
19	1,430,047,520,046,896,777,021,882,368

TABLE 3. THE NUMBERS OF ROOTED TREES WITH n LABELLED SPECIES, ALLOWING SOME INTERIOR NODES TO BE LABELLED, BUT ALLOWING NO MORE THAN TWO IMMEDIATE DESCENDANTS OF EACH INTERIOR NODE.

n	Number of trees
1	1
2	3
3	21
4	231
5	3,495
6	67,455
7	1,584,765
8	43,897,455
9	1,400,923,755
10	50,619,052,575
11	2,042,745,514,425
12	91,066,568,444,775
13	4,444,738,893,770,175
14	235,731,740,255,186,175
15	13,499,365,993,279,291,125
16	830,161,812,269,496,081,375
17	54,564,569,247,212,367,217,875
18	3,817,309,552,613,869,238,301,375
19	283,213,212,610,863,528,421,052,625

The five terms here each is associated with the condition under which it is taken to be nonzero. The recursion starts, of course, with V(1, 0, 0) = 1 and all other V(1, i, j) = 0. The rationale for these five terms will not be intuitively obvious. The terms correspond to the four ways (i)–(iv) of adding species n, except that the last way is subdivided according to whether the new species, is made to arise directly from an "eligible" interior node or from a tip. If the number of tip species is taken to be t, and the number of unlabelled interior nodes to be u, then the total number of labelled nodes, n = t + m + p, and the total number of segments in the tree is the total number of nodes (labelled and unlabelled), n + u. Furthermore, as we go up the tree from its root, each bifurcating interior node (and there are m + u of them), adds one new lineage to the tree. Since we must end up with t lineages, t = m + u + 1.

Solving these equations for t and u,

$$t = n - m - p,$$

and

$$u = n - 2m - p - 1,$$

so that there are a total of 2n − 2m − p − 1 segments in the tree. The five coefficients of the terms in (4) are thus respectively: one less than the number of segments, two less than the number of segments, one more than the number of unlabelled nodes, one more than the number of "eligible" nodes, and the number of tips. It does not seem worthwhile to go over the argument leading to (4) more exhaustively; interested readers can reconstruct it themselves. Table 3 shows the results of applying (4).

EXTENSIONS AND APPLICATIONS

Many extensions of the present approach are possible. The availability of the individual T(m, n) and their equivalents makes it straightforward to compute the mean and variance of the number of interior nodes, given n. One could also consider the order of speciation events on trees. Of particular interest would be the number of different rooted trees (bifurcating or multifurcating) which are consistent with a set of fossil species ordered in time, plus a certain number of contemporary species. Of course, as soon as one allowed times of branching to characterize a tree, there are an infinite number of possible trees, corresponding to the infinite number of possible values of each such continuous variable.

There seems to me to be little point in following up these possibilities, as the enumeration of evolutionary trees has somewhat restricted interest. There are three possible applications. First, one may have a computer algorithm which is intended to examine all possible evolutionary trees of a certain kind, or all possible hierarchies of clustering events. Computation of the numbers of such trees allows us a check on whether the algorithm works, or on whether it is feasible at all to attempt to use it. Second, one may have a proposed notation system for a particular category of trees. By considering the ratio between the number of different trees and the number of different configurations of the notation system, one has a measure of the efficiency of the

notation system. Third, from time to time a taxonomist will propose a method of finding evolutionary trees in which one proposed step is examining all possible trees to see whether some criterion is satisfied. Enumeration of evolutionary trees may then be a powerful argument for adopting some procedure either less ambitious or more powerful.

ACKNOWLEDGMENTS

I am indebted to Charles W. Cotterman for pointing out the work of Schröder on this subject, to Annette J. Dobson, Wayne Moss, and Luca Cavalli-Sforza for other references, and to Walter Fitch for catching a blunder. This work was supported by ERDA contract AT(45-1)2225 TA 5 with the University of Washington.

REFERENCES

CAYLEY, A. 1856. Note sur une formule pour la réversion des séries. Journal für die reine und angewandte Mathematik 52:276–284. Collected Papers, Cambridge 4(1897):30–37.

CAYLEY, A. 1889. A theorem on trees. Quart. J. Math. 23:376–378. Collected Papers, Cambridge 13(1897):26–28.

CAVALLI-SFORZA, L. L., AND A. W. F. EDWARDS. 1967. Phylogenetic analysis. Models and estimation procedures. Amer. J. Human Genet. 19:233–257; Evolution 21:550–570.

DOBSON, A. J. 1974a. Unpublished Ph.D. thesis, James Cook University of North Queensland.

DOBSON, A. J. 1974b. Unrooted trees for numerical taxonomy. J. Appl. Prob. 11:32–42.

EDWARDS, A. W. F., AND L. L. CAVALLI-SFORZA. 1964. Reconstruction of evolutionary trees. *In* Heywood, W. H., and J. McNeill (eds.), Phenetic and phylogenetic classification. Systematics Association Publication No. 6, London, pp. 67–76.

HARDING, E. F. 1971. The probabilities of rooted tree-shapes generated by random furcation. Adv. Appl. Prob. 3:44–77.

HARPER, C. W. 1976. Phylogenetic inference in paleontology. J. Paleont. 50:180–193.

MOON, J. W. 1970. Counting labelled trees. Canadian Mathematical Monographs No. 1, Canadian Mathematical Congress.

PHIPPS, J. B. 1976a. Dendrogram topology: capacity and retrieval. Canadian J. Bot. 54:679–685.

PHIPPS, J. B. 1976b. The numbers of classifications. Canadian J. Bot. 54:686–688.

SCHRÖDER, E. 1870. Vier combinatorische Probleme. Z. Math. und Phys. 15:361–376.

Manuscript received January 1977
Revised November 1977