Combinatorics of trees

Introduction

The basic problem:

When we draw trees depicting the specific evolutionary relationships among a set of taxa, we may do this in many different ways (slanted, rectangular, circular, etc). But while the drawings are different, they are meant to depict the same thing. Thus the idea of a tree is really an abstract one, and by giving it an abstract definition we can legitimately see that two trees are 'the same' even though the diagram we draw may have different forms. If we also look ahead to trying to write software to perform phylogenetic analyses, a computer must contain an internal representation of a tree, even though it has no visual abilities.

Graphs and Trees

Definition of a tree:

A **_tree T = (V, E)_** is a connected (simple) graph with no cycles.

> A **_graph_** G is pair G = (V, E) where V = V (G) is a set of *vertices* or *nodes*, and E = E(G) is a set of *edges*. Each edge e ∈ E is a two-element set e = $\{v_1, v_2\}$ of vertices $v_1, v_2 \in V$.

> When e = $\{v_1, v_2\} \in E$, we say v_1 and v_2 are the ends of e, that e joins v_1 and v_2, that v_1 and v_2 are incident to e, and that v_1 and v_2 are adjacent. The *degree* or *valence* of a vertex is the number of edges to which it is incident.

> In a graph, a **_path of length n_** from vertex v_0 to vertex v_n is a sequence of distinct vertices $v_0, v_1, v_2, \ldots, v_n$ such that each v_i is adjacent to v_{i+1} . A graph is **_connected_** if there is a path between any two distinct vertices.

> **A cycle** is a sequence of vertices $v_0, v_1, \ldots, v_n = v_0$ which are distinct (other than $v_n = v_0$) with n ≥ 3 and v_i adjacent to v_{i+1}.

If a vertex lies in two or more distinct edges of a tree, we say it is an **_interior vertex_**. If it lies in only one edge, then we call it a **_terminal vertex or a leaf_**. Similarly, an edge joining two interior vertices is an **_interior edge_**, while one incident to a leaf is called a **_pendant edge_**.

A basic property of trees:

Theorem 1.[*] **If v_0 and v_1 are any two vertices in a tree T, then there is a unique path from v0 to v_1.**

Rooted vs. unrooted trees:

The trees we have defined are sometimes called unrooted trees or undirected trees. Although, time provides a direction that is of course central to evolutionary processes, most mathematical models and inference methods are more naturally associated with undirected trees, and so we make them our basic objects.

---

[*] You don't have to worry about the proofs of these theorems, unless you are into this kind of things.

However, sometimes we pick a particular vertex ρ in a tree and distinguish it as the root of the tree. More often, we introduce a new vertex to be the root, by subdividing an edge, {u, v}, into two {u, ρ} and {ρ, v}.

Biologically, a root represents the most recent common ancestor (MRCA) of all the taxa in the tree. Sometimes, however, roots are chosen for mathematical convenience rather than biological meaning, so one must be careful with such an interpretation.

Polytomies:
An unrooted tree is said to be *binary* if each interior vertex has **degree three**. This terminology of course fits with the biological view of one lineage splitting into two, but this leads to the rather odd situation that a synonym for binary is *trivalent*. We call a rooted tree T^ρ binary if all interior vertices other than ρ are of degree three, while ρ is of degree two.

When non-binary trees arise in phylogenetic applications, they usually have vertices of degree greater than 3 (also called multifurcations) to indicate ambiguity arising from ignorance of the true binary tree. This is sometimes referred to as a ***soft polytomy***. In contrast, a ***hard polytomy*** refers to a multifurcation representing positive knowledge of many lineages arising at once, such as a radiation of species. Since even radiations are likely to be modeled well by a succession of bifurcations with short times between them, in most applications, biologists generally prefer to find 'highly resolved' trees, with few multifurcations; a binary tree is the goal.

The leaves:
The trees used in phylogenetics have a final distinguishing feature — the leaves represent known taxa, which are typically currently extant and are the source of the data used to infer the tree. The internal vertices, in contrast, usually represent taxa that are no longer present, and from which we have no direct data. This is formalized by labeling the leaves of the tree with the names of the known taxa, while leaving unlabeled the interior vertices. Thus for either rooted or unrooted trees we are interested primarily in the following objects.

Let X denote a finite set of taxa, or labels. Then a *phylogenetic X-tree* is a tree together with a one-to-one correspondence $\varphi : X \to L$, where $L \subseteq V$ denotes the set of leaves of the tree. We call φ the *labeling map*. Such a tree is also called a leaf-labeled tree.

More informally, the labeling map simply assigns each of the taxa to a different leaf of the tree, so that every leaf receives a label.

Phylogenetic trees can be distinguished from one another either due to different 'shapes' of the underlying trees, or merely by a different labeling of the leaves.

Counting Binary Trees
How many trees?
Two phylogenetic X-trees are isomorphic if there is a one-to-one correspondence between their vertices that respects adjacency, and their leaf labelings.

Theorem 2. An unrooted binary tree with n ≥ 2 leaves has 2n − 2 vertices and 2n − 3 edges.
Proof. The statement is certainly true for n = 2, which is the base case for induction.
Suppose T has n ≥ 3 leaves, and assume the statement is true for a tree with n−1 leaves. If v_1 is one of the leaves of T, then v_1 lies on a unique edge $\{v_1, v_2\}$, while v_2 lies additionally on two other edges $\{v_2, v_3\}$ and $\{v_2, v_4\}$. Removing these three edges and the vertices v_1, v_2 from T, and

introducing a new edge {v₃, v₄} gives a binary tree T′ with n − 1 leaves. Since both the number of vertices and the number of edges have been decreased by 2, for T the number of vertices must have been $(2(n-1)-2)+2 = 2n-2$, and the number of edges $(2(n-1)-3) + 2 = 2n-3$.

Theorem 3. If X has n ≥ 3 elements, there are b(n) = (2n − 5)!! = 1 · 3 · 5 · · · (2n − 5) distinct unrooted binary phylogenetic X-trees.
Proof. Again we use induction, with the base case of n = 3 clear.
Suppose then that T has n leaves, and let T' be the (n − 1)-leaf tree constructed from T as in Theorem 2 by 'removing' the leaf v_1 and adjusting edges appropriately. Then with v_1 fixed, the map $T \mapsto (T', \{v_3, v_4\})$ is a bijection from n-leaf trees to pairs of (n − 1)-leaf trees and edges. In this pair, we think of the edge {v₃, v₄} as the one onto which we 'graft' a new edge to v_1 to recover T. Counting these pairs shows

$b(n) = b(n - 1) \cdot (2(n - 1) - 3)$, where b(n-1) is # of trees for n-1 taxa and 2(n-1)-3 is their number of branches. Simplifying,
$b(n) = b(n - 1) \cdot (2(n - 1) - 3) = b(n - 1) \cdot (2n - 5)$.

But since we inductively assume $b(n - 1) = (2(n - 1) - 5)!! = (2n - 7)!!$, we obtain the desired formula: $(2n - 7)!! \cdot (2n - 5) = (2n - 5)!!$

In real life we have 3 trees for 4 taxa, 2,027,025 trees for 10 taxa, 221,643,095,476,699,771,875 trees for 20 taxa (it's 221+ quintillion) 27,529,213,532,835,651,545,259,729,751,524,430,639,300,973,035, 816,196,098,326,553,772,152,587,890,625 for 50 taxa (it's 27+ quattuorvigintillion), more trees than atoms in the universe for 53 taxa, more trees than the volume of Universe in cubic Ångstroms for 67 taxa, and more than the number with 70,000,000 zeroes for the tree of life (Googol "only" has 100 zeroes)

Metric trees
A metric tree (T,w) is a rooted or unrooted tree T together with a function $w : E(T) \rightarrow \mathbb{R}^{\geq 0}$ assigning non-negative numbers to edges. We call w(e) the length or weight of the edge e.

When we wish to emphasize that edge lengths are not specified for a tree T, we will sometimes refer to T as a topological tree. We can obtain a topological tree from a metric one by simply ignoring the edge lengths.

A metric tree leads to a way of measuring distances between its vertices. For any $v_1, v_2 \in V(T)$, define $d(v_1, v_2) = \sum w(e)$, on the path from v_1 to v_2.

Proposition 5. For a metric tree (T, w), the function $d : V \times V \rightarrow \mathbb{R}^{\geq 0}$ satisfies
(i) $d(v_1, v_2) \geq 0$ for any v_1, v_2 (non-negativity),
(ii) $d(v_1, v_2) = d(v_2, v_1)$ for any v_1, v_2 (symmetry),
(iii) $d(v_1, v_3) \leq d(v_1, v_2) + d(v_2, v_3)$ for any v_1, v_2, v_3 (triangle inequality).
If all edges of T have positive length, then, in addition, $d(v_1, v_2) = 0$ if, and only if, $v_1 = v_2$.

If all edges of T have positive length, then the proposition above shows d is a metric on V (T) in the usual sense of the word in topology or analysis. In this case, we call d a tree metric.

Ultrametric trees and molecular clock

Definition. A rooted metric tree is said to be ultrametric if all its leaves are equidistant from its root using the tree metric: For any two leaves a, b and the root ρ, $d(\rho, a) = d(\rho, b)$.

Theorem 6. Suppose T^ρ is a rooted ultrametric tree with positive edge lengths. Let v_1, v_2 be any two leaves with $d(v_1, v_2)$ maximal among distances between leaves. Then ρ is the unique vertex along the path from v_1 to v_2 with $d(\rho, v_1) = d(\rho, v_2) = d(v_1, v_2)/2$. Thus the placement of ρ can be determined from an unrooted metric version of T.

Sometimes when a root must be added to an unrooted metric tree, either as a rough approximation of the MRCA of the taxa or simply for convenience in drawing the tree, the root location is chosen as the midpoint of the path between the most distant pair of leaves. Called "midpoint rooting"

Rooting trees with outgroup

Though using an outgroup is the standard way of rooting phylogenetic trees when a molecular clock hypothesis is not reasonable, and the method usually seems to work well, two features should be noted. First, this approach requires prior knowledge that the outgroup is distantly related to the other taxa. This may be obvious in some situations, such as that mentioned with ducks. But in others, particularly when one is attempting to infer very ancient relationships between very different taxa, it may not be clear what can serve as an outgroup. Second, even if it is clear how to pick an outgroup, there may be difficulties in inferring a tree that has it correctly placed. By definition, an outgroup is distantly related, and so whatever method is used to infer the tree may have more difficulties placing it correctly than the other taxa.

Newick notation

Newick tree is a way of representing graph-theoretical trees with edge lengths using parentheses and commas. It was adopted by James Archie, William H. E. Day, Joseph Felsenstein, Wayne Maddison, Christopher Meacham, F. James Rohlf, and David Swofford, at two meetings in 1986, the second of which was at Newick's restaurant in Dover, New Hampshire, US.

Newick notation uses parenthetical grouping of taxon labels to specifying the clustering pattern of a tree. There are several ways to describe any tree given this notation. This non-uniqueness of the Newick specification of a tree can make it hard to recognize by eye when two large trees are the same.