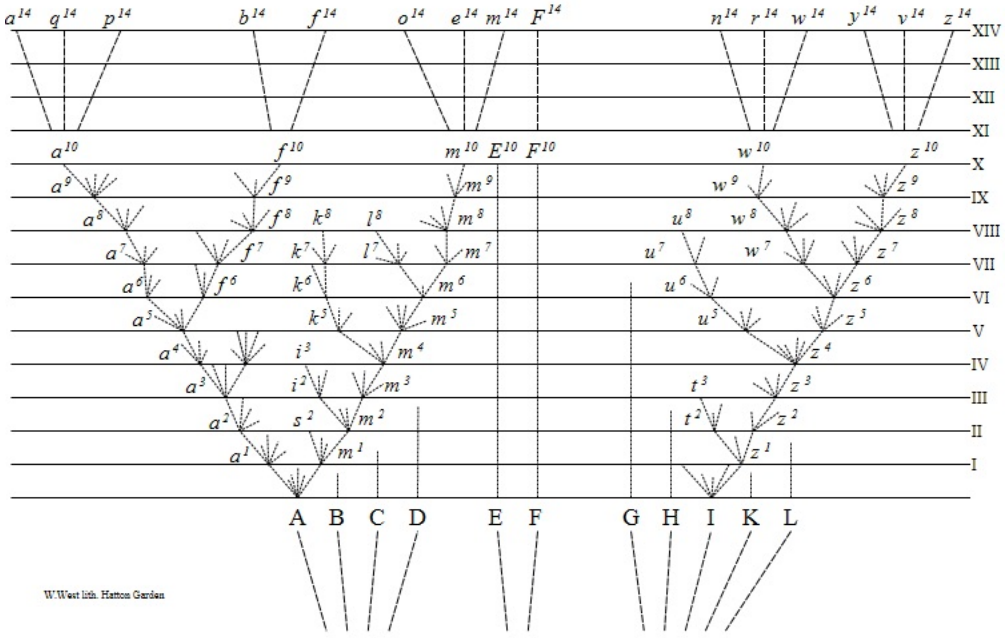
1. Phylogenetic trees

EEOB563

Spring 2021

"The time will come I believe, though I shall not live to see it, when we shall have fairly true genealogical (phylogenetic) trees of each great kingdom of nature". Darwin, 1857.

1 Introduction



This "Tree of Life" figure is the only illustration in the Darwin's Origin of Species.

The central claim of the theory of evolution proposed by Charles Darwin in the Origin of Species is that all life is related by descent with modification from a common ancestor. He also suggested that these relationships could be depicted by a tree ("Tree of Life"). Interestingly, a drawing of a tree (presented in Fig.1) was the only figure in the book. Ernst Haeckel, the greatest popularizer of the theory of evolution in XIX century, made Darwin's metaphorical description of the "tree of life" literal by applying his artistic skills (Fig. S1).

2 Graphs and Trees

The basic problem: When we draw trees depicting 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.

Definition of a tree: A tree ***T = (V, E)*** is a connected (simple) graph with no cycles.

This is a very concise definition, but it requires several additional definitions:

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/branches. Each edge $e \in E$ 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 $v_0, v_1, v_2, ..., v_n$ 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 $v_0, v_1, ..., v_n = v_0$ which are distinct (other than $v_n = v_0$) with $n \geq 3$ and vi 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*** or ***terminal branch***.

Note that the mathematical definition of the tree provides two important insights into biological inference that we are trying to make. First, because any tree is a connected graph, we will always get a single tree even if the data (sequences) we used are not evolutionarily related. Hence, it is our responsibility to make sure that it makes sense to build a tree for the data we have. Second, because we are not allowing cycles, phylogenetic trees will be a poor model for relationships that include hybridization and/or horizontal gene transfer.

Alternatives : Note, that there are other way to represent phylogenetic trees, which may be preferable in some circumstances. For example, a binary tree for 5 taxa can be represented by a set of two splits (or bipartitions): $\{\{A, B\}, \{C, D, E\}\}$ and $\{\{A, B, C\}, \{D, E\}\}$.

Rooted vs. unrooted trees: The trees we have defined are sometimes called unrooted trees or undirected trees. Although, time provides a direction that is 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.

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 $\{u, \rho\}$ and $\{\rho, v\}$.

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

Rooting trees with an outgroup: *Though using an outgroup is the standard way of rooting phylogenetic trees, two features should be noted. First, this approach requires prior knowledge that the outgroup is equidistantly related to the other taxa. This may be obvious in some situations, but not so obvious in others. 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 as the result forms a long branch on a phylogenetic tree. Whatever method is used to infer the tree may have more difficulties placing it correctly than the other taxa.*

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.

3 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 $n \geq 2$ leaves has $2n-2$ vertices and $2n-3$ edges. Check the book for a formal proof by induction. Here is an approximation by a biologist: Draw a phylogenetic tree. Add a new taxon to it. Note how adding a new taxon adds two vertices (a new leaf for the new taxon you add and an internal node where you connect the branch). Also notice that adding a new taxon adds two branches (you add a new terminal branch and subdivide one existing branch). So you'd expect that both the number of nodes and the number of branches will be $2n$. However, there are two exceptions: with just one taxon, the "tree" has only one vertex and no edges (+1, +0), and with the second taxon, you add one vertex and one edge (+1, +1). After that it's always (+2, +2). Overall, $2n-2$ and $2n-3$ and

Theorem 3. If X has $n \geq 3$ elements, there are $b(n) = (2n-5)!! = 1 \times 3 \times 5 \times ... \times (2n-5)$ distinct unrooted binary phylogenetic X-trees. Again, consult the book for a formal proof. For a biologist, it is important to realize that we add a new taxon by connecting a new terminal branch to one of the existing branches. Since there are $2(n-1)-3 = 2n-5$ existing branches there are $2n-5$ times as many trees for n taxa comparing to $n-1$. So when you add your 100th sequence to your dataset, your tree space increases 95 times.

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)

4 Metric trees

A metric tree (T, w) is a rooted or unrooted tree T together with a function $w : E(T) \to R \geq 0$ 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 $d : V \times V \to R \geq 0$ 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.

5 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 $v1, v2$ be any two leaves with $d(v1, v2)$ 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"

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.

PEDIGREE OF MAN.