1 Subjects

- Different distance measures between evolutionary trees
- Computing distance between trees

2 Notes

2.1 Phylogenetic trees

There is strong evidence that all life on earth is decended from a single common ancestor. Over the course of at least 3.8 billion years, that life form has changed and split itself into new and independent lineages. The evolutionary relationships among these species is referred to as their phylogeny, phylogenetic reconstruction is concerned with inferring the phylogeny of groups of organisms.

We call these groups *taxa* (or singular taxon), the splitting of lineages is called *speciation* (think of splitting it into different species). Usually speciation happens if one population is split into two that can no longer interbreed (e.g. if a river splits them apart). We can use methods of phylogenetic to contemplate over the tree of life, or simply to inferr the phylogeny of different populations within a species.

In a rooted phylogenetic tree T, the root node r corresponds to the last common ancestor of all species in T, we then call a path from the root to a leaf an *evolutionary path*. If an equal amount of change occurs on every evolutionary path (i.e. each species have the same amount of "change") then the evolutionary change occur in a more-or-less clocklike fashion and we then say that this tree satisfies the *molecular clock hypothesis*, we can then assign a time $t(v)$ to every internal node v in the tree and a length of $t(v) - t(w)$ to an edge (v, w) in the tree.

Every extant (still alive) species corresponds to time 0, and a speciation event (internal node v) occured in the tree $t(v)$ time ago. We then have that the length of an edge represents the amount of time that lies between two speciation events.

Furthermore, the last common ancestor (root) lived $t(r)$ and all evolutionary paths have the same length $t(r)$ where the length of a path is the sum of the lengths of all edges along the path.

Once the *molecular clock hypotheses* was widely accepted, but has since been disproven so now time instead refers to the expected amount of evolution.

2.1.1 Distance methods

Distance methods construct a phylogenetic tree from a distance matrix that contains the evolutionary distances between all pairs of taxa (groups of organisms). If we ignore edge weights, we speak of the topology (shape) of a tree, and it turns out that we have methods that *provably* find the correct tree if the distance matrix is ultrametric, and some that works if the distance matrix is

additive. However the data we can record is usually approximation of the true (additive) data, and thus we can't rely on the distance matrix being additive. However, it turns out that the methods can still find the correct *topology* of the tree under certain criteria.

If we just have the topology, then we must assign weights to the edges of the tree that best fit the data, which can be done with the least squares methods.

2.1.2 Basic definitions

Phylogenies are usually represented as binary trees, because generially speciation happens when one lineage splits into two independent lineages. This is not entirely correct, sometimes horizontal gene-transfer happens and hybrid speciation but it is rare. So for simplicity we just look deal with phylogenetic trees, however let's not insist they have to be binary for the moment:

Let $S = \{s_1, \ldots, s_n\}$ be a set of taxa. A phylogenetic tree on S is a triple $T = (V, E, \alpha)$ where:

- V is the set of nodes, E is the set of undirected edges
- (V, E) is a an acyclic connected graph, in which there might be a distinguished root node of degree ≥ 2 and all other internal nodes have a degree ≥ 3. So either rooted or unrooted. We will denote the set of leaves by V_L and the set of internal nodes by V_I
- α is a bijection $\alpha : S \to V_L$ between the set of taxa and the set of leaves

An edge $(v, w) \in E$ is an external edge if either v or w is a leaf. Otherwise it is an internal edge.

We haven't included edge-weights here, they will be introduced later.

2.2 Distance between tree-topologies

Distance between trees is about measuring how far they are from each other, that is, how different are they. There are a number of different ways that have been proposed to this end, we will start by looking at the measures that compute the distance between the topologies of trees.

2.2.1 The symmetric difference

If we ignore the edge weights, then each edge can be seen as a branch which divides the species into a partition with two sets. If we make a list of the partitions each tree implies (e.g. $AB|CD$ and $A|B$ and $C|D$) we will then simply count how many partitions are in one list which is not in the other and that will be the symmetric difference.

The symmetric distance is very easy to compute, but is very sensitive to all differences between trees. Therefore a tree which has partial similarities, might still score the maximum possible distance.

This can also be named the Robinson-Fould distance

2.2.2 The quartets distance

The *quartets distance* is more sensitive to partial similarities of structures between trees. A quartet is four named species in an unrooted tree. Quartet topology is the topology of the quartet induced by the tree.

The quartet distance is then simply the number of quartets that don't have the same topology in the two trees. Computing this distance can be done in $\mathcal{O}\left(n^2\right)$. It has been shown that it can be done in $\mathcal{O}\left(n(\ln n)^2\right)$

2.2.3 Nearest-Neighbour interchange distance

Simply, compute the minimum number of nearest-neighbour interchange rearrangements that are needed to go from one tree to the other. Does not have teh same issue with tree-distance symmetric distance, however it is NP-Complete.

2.2.4 Computing RF-Distance

There are various ways to compute the Robinson Fouls distance (also called the symmetric difference). The $\mathcal{O}(n)$ algorithm proposed by Day. So here come's Day's algorithm for computing the RF-Distance:

- Root both trees by the same leaf
- Do a DFS numbering of T_1
- Assign the same numbers to T_2
- Find all the "valid" intervals in T_1 and T_2
- Count the number of intervals that don't exist in both T_1 and T_2

2.3 Distance between trees

Now, there are some distance-measures that take the branch-weights into account.

The Robinson Foulds distance can be taken into account by, for each edge, compute the difference of the weights. If an edge exists in one and not the other (i.e. the case where we would nomrally count one up) we add the complete weight of the edge.

2.3.1 Computing the Quartet distance

The triplet and quartet distance are very close, I will focus on the Quartet distance.

First let's look at some of the cases, the Quartet distance counts the number of quartets where the two trees disagree. To this end there are "resolved" and "unresolved" cases, the unresolved case, is when all the leaves are connected through a single node (all same parent).

If both T_1 and T_2 deems the quartet to be unresolved, then they agree if either one or the other think it's unresolved then they disagree. If they both think it's resolved then they agree if they have the same topology. So we count the number of times they disagree. The number of quartets is $\binom{n}{4}$.

Let's just look at binary trees, if we are dealing with binary trees, then the unresolved case is not possible, thus we just look at the quartets where the topologies differ.

A dynamic programming approach, conceptually we simply say that each edge is oriented both ways, we then say that $F_1 \xrightarrow{e_1} (F_2, F_3)$ in T_1, *claims* all quartets, $ij|kl$ with $i, j \in F_1, k \in F_2$ and $l \in F_3$ in T_1, where F_1 is the tree behind the edge and F_2, F_3 is the two subtrees in front of the edge. Every shared quartet is counted twice, so we divide by two. Giving us the formula:

$$Count(e_1, e_2) = \binom{|F_1 \cap G_1|}{2} \left(|F_2 \cap G_2| \cdot |F_3 \cap G_3| + |F_2 \cap G_3| \cdot |F_3 \cap G_2|\right)$$

$$d_{quartet} = \binom{n}{4} - \frac{1}{2} \sum_{e_1, e_2} Count(e_1, e_2)$$

and then we count the how many of these edges $ij \to kl$ are claimed by both trees. This can be done in $\mathcal{O}\left(n^2\right)$.

There is another way that is based on tree coloring that can do it in $\mathcal{O}\left(n \log n\right)$