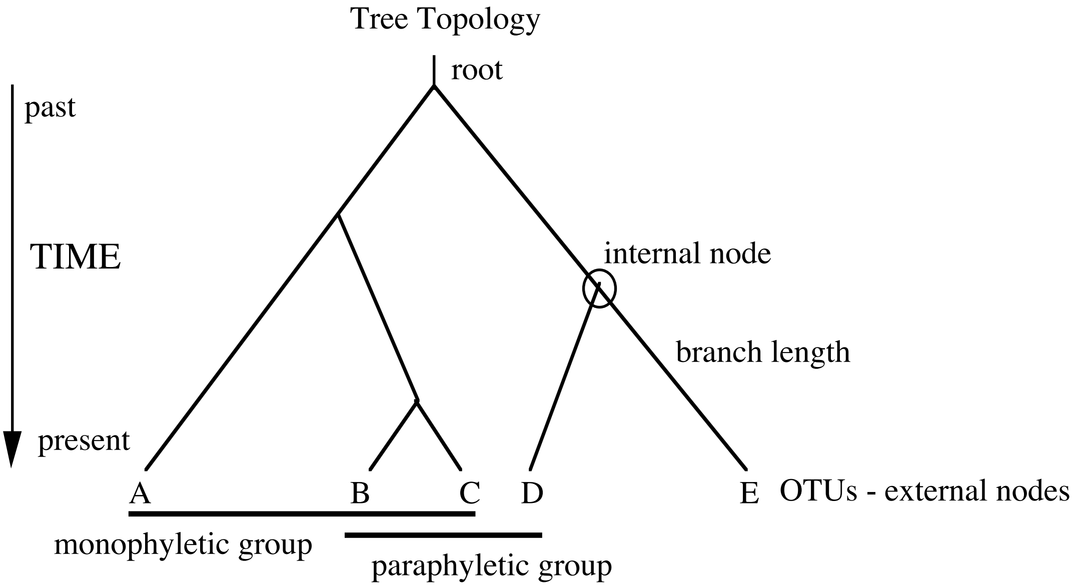
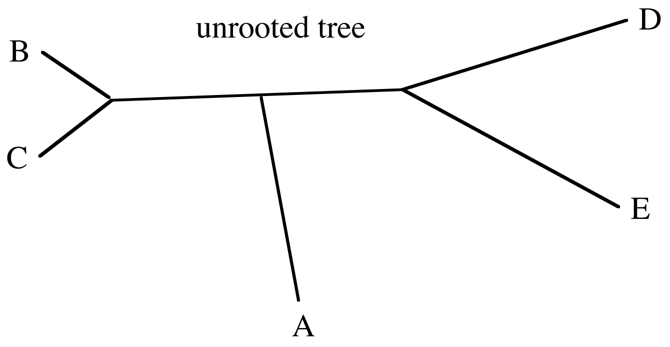
Molecular Phylogenetics

Molecular phylogenetics (or systematics) is the study of evolutionary relationships among organisms or genes using molecular data (typically protein or DNA sequences) and statistical techniques. It is sometimes referred to as “tree building”.

1. The Wonderful World of Trees

a) Nomenclature





Additional notes:

- OTU = Operational Taxonomic Unit. These are often (but not always!) species.
- A monophyletic group is also known as a clade.
- trees may also be drawn with the OTUs at the top, left, or right. It is the branching pattern (topology) that is important!

b) Species Trees *vs*. Gene Trees

Species trees show the evolutionary relationships among species. Gene trees show evolutionary relationships based on single genes, or groups of homologous genes. They are not necessarily the same as species trees. Example: the Drosophila *janusA*, *janusB*, and *ocnus* genes – gene duplication predates speciation.

Homologous genes (homologs) = similar sequence due to common ancestry

- orthologous genes (orthologs) = homologs due to speciation
- paralogous genes (paralogs) = homologs due to gene duplication.

c) Number of Possible Trees

The number of possible unrooted trees for n OTUs ($n \geq 3$):
$N_U = (2n - 5)! / [2^{n-3} (n - 3)!]$

The number of possible rooted trees for n OTUs ($n \geq 3$)
$N_R = (2n - 3)! / [2^{n-2} (n - 2)!]$

Note that $N_R (n) = N_U (n + 1)$

The number of possible trees increases rapidly with n, yet only 1 is the "true" tree.

For $n = 35$, the number of possible rooted trees is $\approx 10^{47}$. This is greater than the number of water molecules on earth!

For $n = 50$, the number of possible rooted trees is $\approx 10^{76}$. This is greater than the number of hydrogen atoms in the universe!

2. General Methods

There are two major approaches to classification:

Phenetics – organisms grouped by *overall similarity*. All characters are considered. Allows paraphyletic groups.

Cladistics – organisms grouped by evolutionary relationship using *shared*, *derived* Characters (also known as "synapomorphies"). Accepts only monophyletic groups (clades).

Example: Do reptiles exist? Birds are not considered to be reptiles. However, birds are more closely related to some reptile groups than the reptile groups are to each other. Thus, one would have to either accept reptiles as a paraphyletic group or eliminate the category altogether.

3. Specific Methods

Distance – based on number of sequence differences (overall divergence). This is a phenetic approach.

a) UPGMA – unweighted pair-group method with arithmetic mean

1. Generate distance matrix
2. Identify the 2 OTUs (A and B) that are most similar to each other and cluster them
3. Compute new distance matrix using composite A and B (average)
4. Repeat until only 2 OTUs are left

b) Neighbor-joining

1. Generate distance matrix
2. Begin with "star" tree
3. Try every pairwise clustering of sequences, accept the pairing (A and B) which minimizes TOTAL tree branch length

4. Compute new distance matrix using composite A and B (average)
5. Repeat until all interior branches are found

Pro – fast and easy to implement on computer
Con – considers only similarity, not evolutionary history (assumes molecular clock)

Parsimony – based on character states (it considers the entire sequence, not just a summary statistic such as distance); Selects the tree requiring the fewest changes ("most parsimonious" tree). This is a cladistic approach.
Pro – simple, evolutionary model
Con – must examine all possible trees to find the most parsimonious, which takes much time or is impossible with large datasets; Systematically underestimates divergence between species; There may be many "ties" – different trees with equal number of changes; Does not account for multiple hits ("homoplasies").

Maximum Likelihood (ML) – similar to parsimony, but employs a more complex probabilistic model for the substitution process. For example, it may include a parameter for the transition/transversion ratio, or transition probabilities for pairs of amino acids. Selects the tree most likely under the given model.
Pro – most realistic evolutionary model, fewer "ties", can account for multiple hits
Con – must examine all trees (long time or impossible). Results depend on the quality and appropriateness of the model.

Bayesian – uses a probabilistic model like ML, but samples only a proportion of possible trees with Markov chain Monte Carlo (MCMC) method. Estimates the posterior probability that a particular tree (or node) is correct.
Pro – same evolutionary model as ML, but faster
Con – may give misleading results if only a small range of possible trees is sampled. Not guaranteed to find the "best" tree.

4. Bootstrapping

This is a statistical method that can be applied to the above distance, parsimony, and ML methods. The sequence data are re-sampled (with replacement) and a new tree is built. This is repeated many times. How often does a particular tree (or node) appear from re-sampled data? This gives you an idea of how well the tree (or node) is supported. Typically, bootstrap values are given for each node of a tree. If the bootstrap percentage is high, then the node is supported by many different sites in the alignment and we have greater confidence that it is correct.

5. Application

Ape phylogeny
Darwin suggested that African apes (chimpanzee and gorilla) were man's closest relatives. However, most taxonomists of the time disagreed and thought humans were distantly related to the African apes and to orangutans.

Serological precipitation of blood samples in the 1960's established that chimpanzee, human, and gorilla form a natural clade, with orangutan as an outgroup. However, it was impossible to sort out the chimp-human-gorilla tree until DNA sequences became available in the 1980s. These results place humans and chimps together, with gorilla as an outgroup.