

Molecular phylogenetics

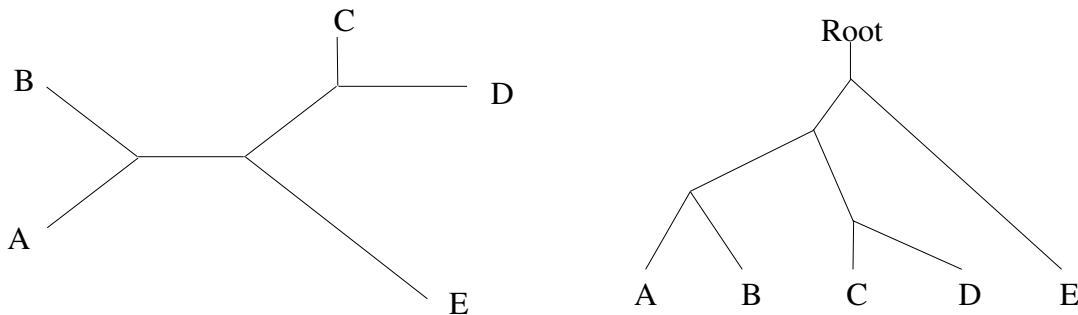
Phylogenetics - the reconstruction of the evolution history of genes or species.

Phylogenetic tree - a graphical representation of organisms evolutionary relationships.

Topology - the branching pattern in a phylogenetic tree.

Root - a common ancestor to all taxa.

Figure 1. Rooted and unrooted phylogenetic trees



Methodology: Blast, alignment, tree reconstruction, biological interpretation

1. Important issues related to interpretation of phylogenetic trees

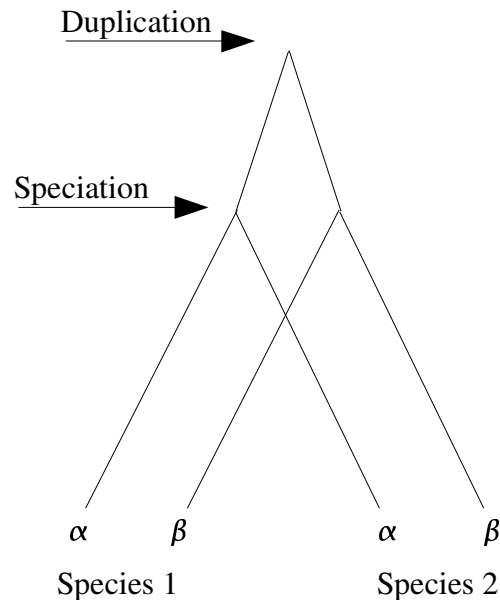
1. Orthology
2. Independence (no concerted evolution or horizontal transfer)

Orthologs and paralogs

Orthologs are genes created by speciation events.

Paralogs are genes created by duplication events.

Homologs are genes that are similar because of shared ancestry.



Orthologues and paralogues can be distinguished by i) synteny or ii) phylogeny.

Evolution by gene duplication: The evolution of new genes by gene duplication may be responsible for numerous innovations during evolution (Ohno 1970). Once duplicate gene arises in a population, it can be lost or fixed. If fixed, it can become nonfunctional, subfunctional or neofunctional with respect to the original function.

Nonfunctionalization is the process by which a gene becomes a pseudogene.

Neofunctionalization is the process by which a gene procures a new function. This was the original model proposed by Ohno (1970): a duplicate gene is unconstrained since the other copy guarantees the original function is present, or vice-versa. Sequence divergence allows the duplicate gene to obtain a new function. Once the new function is attained the protein becomes constrained and does not further diverge.

Subfunctionalization is the process by which two duplicate genes lose complementary functions present in the original gene. One variant of the subfunctionalization model is that the duplication event allows a protein with many functions to become two proteins that are better specialized at performing a subset of the original functions.

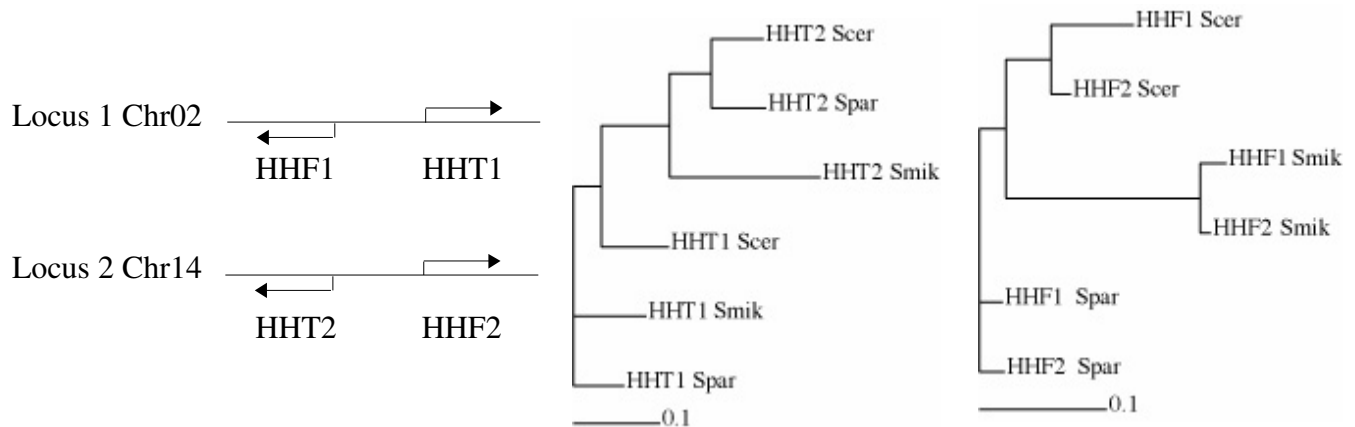
IMPORTANT (and a little off the topic) NOTE: Although duplicate genes result in functionally redundant sequences in a genome, they're presence cannot be explained by selection for genetic or environmental canalization, i.e. selection for buffering through redundancy.

Concerted evolution

The most common causes of concerted evolution are unequal crossing-over and ectopic gene conversion. For both, concerted evolution can be detected by sequences that are more similar within the same genome compared to their orthologues in another species.

Unequal crossing-over results in a duplication or deletion event. Thus, unequal crossing-over can homogenize tandem repeats, such as the rDNA. Ectopic gene conversion is gene conversion between non-orthologous genes.

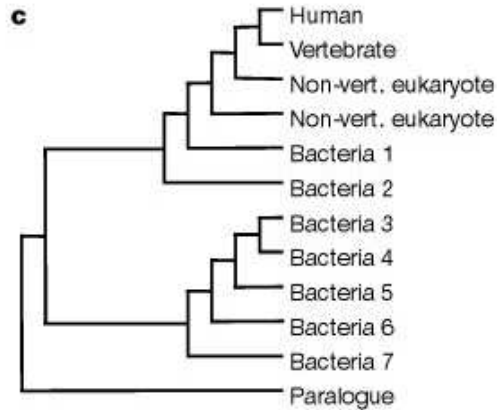
Phylogenetic relationship of the HHF1, HHF2, HHT1 and HHT2 genes from *Saccharomyces*.



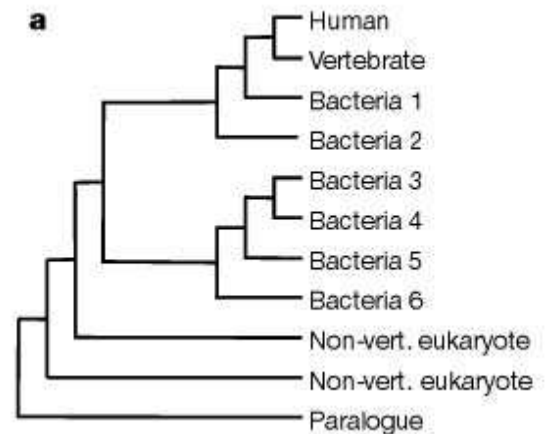
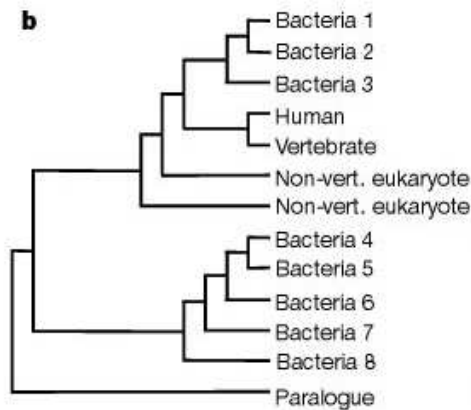
Horizontal Gene Transfer

Definition - The movement of DNA from one organism to another without reproduction. Examples include antibiotic resistance genes in prokaryotes. These are called xenologues.

Mechanisms - Include transformation (uptake of naked DNA), transduction (viral or bacteriophage mediated DNA transfer) and conjugation. Horizontal gene transfer in plants and eukaryotes has implications for genetically modified plant and animal products.



A phylogenetic approach can be used to identify horizontal gene transfer by an incongruence between gene trees (a and b) and species trees (c).



Human genome paper (2001) - “An interesting category is a set of 223 proteins that have significant similarity to proteins from bacteria, but no comparable similarity to proteins from yeast, worm, fly and mustard weed, or indeed from any other (nonvertebrate) eukaryote”

2. Reconstructing phylogenetic trees

Table 1. Number of possible rooted and unrooted tr

Number of sequences	Number of rooted trees	Number of unrooted trees
2	1	1
3	3	1
4	15	3
5	954	105
10	34,459,425	2,027,025

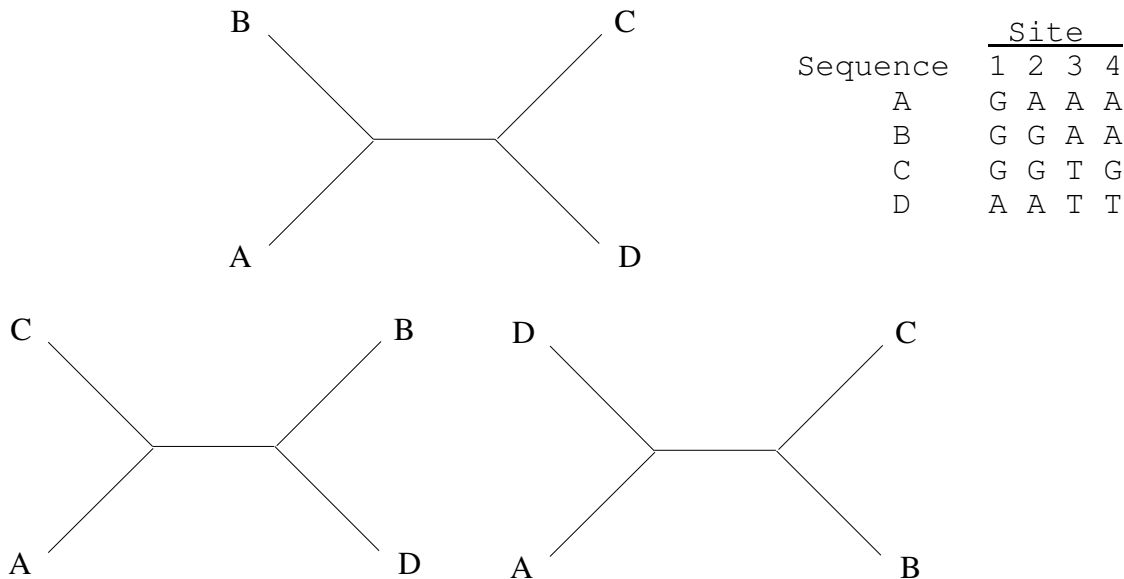
Taxonomists have long debated phylogenetic methods. There are two types of methods: character state methods, like parsimony, and distance or similarity based methods, like UPGMA. Both have advantages and disadvantages.

```

M5      TGGCTCGGTGTCAAAACATAGTTTGC GTGATAACAGCGTGTGTGCTCTC
M22     .....
M32     .....C.....A.....
M34     .....T.....
M56     .....C.....T.....
M88     .....C.....T.....T.G..
M100    .C.....C.....G.T...G.....G..
    
```

Parsimony method

Parsimony uses the minimum number of evolutionary changes to explain character states (nucleotides, peptides or morphological character states).



The maximum parsimony method searches for the best tree that can explain the data. Non-informative sites such as #1 are not used. This is different from distance based methods which use all the sites.

Distance method

Distance based methods use the distances between each pair of sequences to create a phylogenetic tree. Substitutions rates are the most commonly used measure for distance, but insertion/deletion rates or other characters can also be used. The data used in distance methods forms a distance matrix:

Table 2. Distance matrix.

Sequence	A	B	C
A			
B	d(AB)		
C	d(AC)	d(BC)	
D	d(AD)	d(BD)	d(CD)

Each d is the distance (substitution rate) between pairs of sequences

UPGMA - unweighted pair-group method with arithmetic mean (assumes a molecular clock). This method assumes a molecular clock. The method iteratively joins the two most similar sequences. Thus, if $d(AB)$ is the smallest value in the matrix, the matrix is recalculated between AB, C and D. The distance between AB and C is the average between AB and AC. More generally the distance between any two clusters of sequences X and Y is:

$$d_{XY} = \sum_{ij} \frac{d_{ij}}{n_x n_y}$$

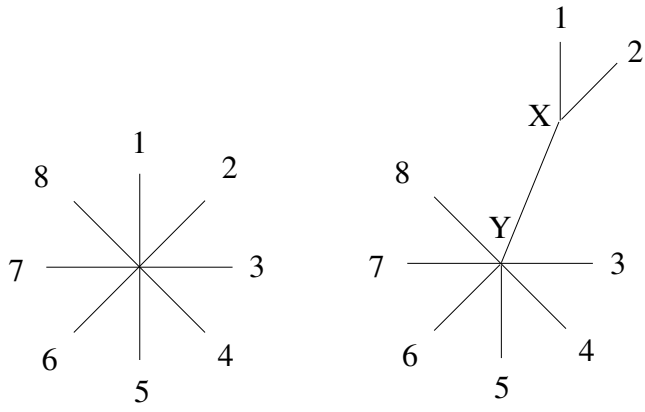
where the summation is over every i in cluster X and every j in cluster Y, and n_x and n_y are the number of sequences in cluster X and Y.

Neighbor-Joining - iteratively finds closest neighbors so as to minimize the total length of the tree (no molecular clock assumption). Starting with a star phylogeny, iteratively join the two sequences that minimize the sum of the branch lengths, given by:

$$S_{ij} = \frac{1}{2(N-2)} \sum_{k=3}^N (d_{ik} + d_{jk}) + \frac{1}{2} d_{ij} + \frac{1}{N-2} \sum_{3 \leq l \leq m \leq N} d_{ml}$$

For the minimal value of S_{ij} , i and j are joined and their mean distance is used in calculating a new distance matrix.

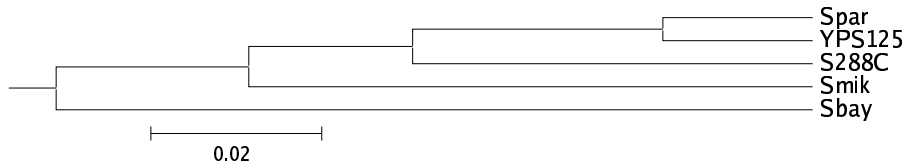
For $i = 1$ and $j = 2$, the sum of the branch lengths can be found by adding the distances over:



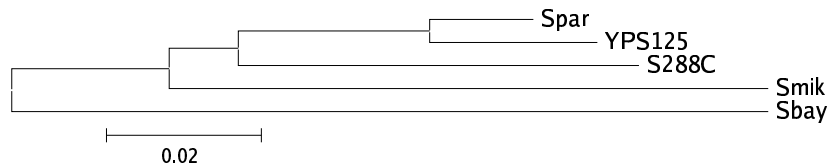
If 1 and 2 are closely related the sum of the branch lengths will be smaller than if they are distantly related.

Branch lengths can be estimated using a method similar to that used in UPGMA or using a least-squares method.

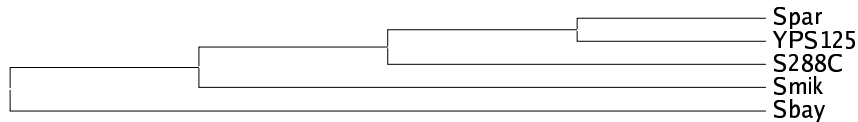
UPGMA tree



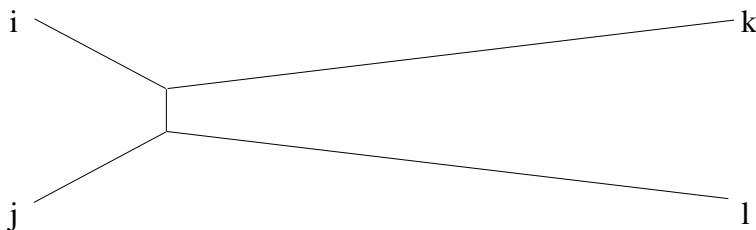
NJ tree



Parsimony



Long branch attraction, or “the Felsenstein zone”, occurs when there are two very long branches:



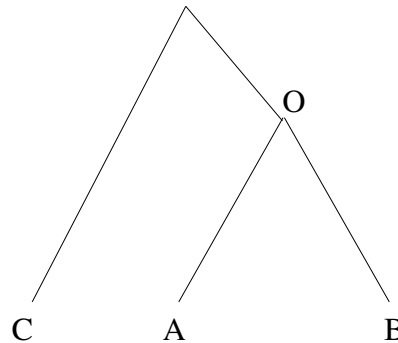
If i and $j = A$ and k and l both have $A \rightarrow G$ substitutions, k and l may be erroneously joined.

3. Relative rates test

If A and B are more closely related, the molecular clock predicts that the substitution rate between A and O equals that between B and O, i.e. $K_{OA} = K_{OB}$. Because $K_{OA} + K_{OB} = K_{AB}$,

$$K_{OA} = (K_{AB} + K_{AC} - K_{BC})/2$$

$$K_{OB} = (K_{AB} + K_{BC} - K_{AC})/2$$



4. Bootstrap and jackknife methods

Bootstrap and jackknife methods can be used to get an idea of confidence in a reconstructed tree.

Bootstrap - resampling the data with replacement, i.e. sites not sequences are sampled.

Jackknife - dropping one observation at a time from the sample, i.e. sites are dropped not samples.

5. Species trees vs gene trees

There are a number of conditions which together can result in differences between gene trees and species trees.

- 1) The branch lengths for gene trees can vary depending on the time to the last common ancestor. This time can be variable depending on the ancestral population size.
- 2) The topology for gene trees can vary depending on the time between speciation events. This can occur when the time is less than $4N$ generations.
- 3) The topology and branch lengths can differ for different genes if speciation is not instantaneous or if only a portion of the genome is incompatible between two species.

