Phylogenetic Trees

Phylogenetic Trees

https://highered.mheducation.com/sites/9834092339/student_view0/chapter23/animation_-_phylogenetic_trees.html

Tree of Life Web Project

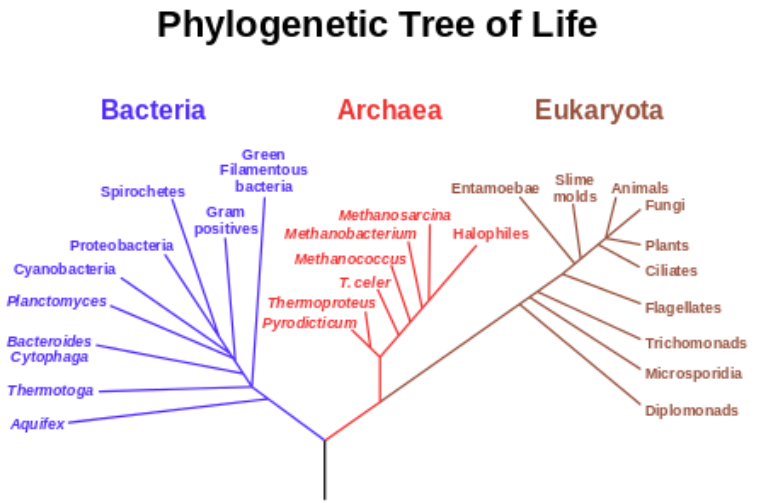
The Tree of Life Web Project (ToL) is a collaborative effort of biologists and nature enthusiasts from around the world

Using trees to learn about the order of evolution: The spider's web

Phylogenies can be used to test hypotheses about evolution. Before phylogenies became a standard tool within biology, many biologists assumed that the **orb-weaving spiders, with their intricate and orderly webs, had evolved from spiders with disorderly cobweb-like webs**. However, the cladistic analysis of these spiders showed that, in fact, orb-weaving was the ancestral state, and that cobweb-weaving had evolved from spiders with more orderly webs. The phylogeny caused the biologists to reject their original hypothesis about orb-weaving evolution.

Phylogenetic trees

A **phylogenetic tree** or **evolutionary tree** is a branching diagram or "tree" showing the inferred evolutionary relationships among various biological species or other entities—their **phylogeny**—based upon similarities and differences in their physical or genetic characteristics.



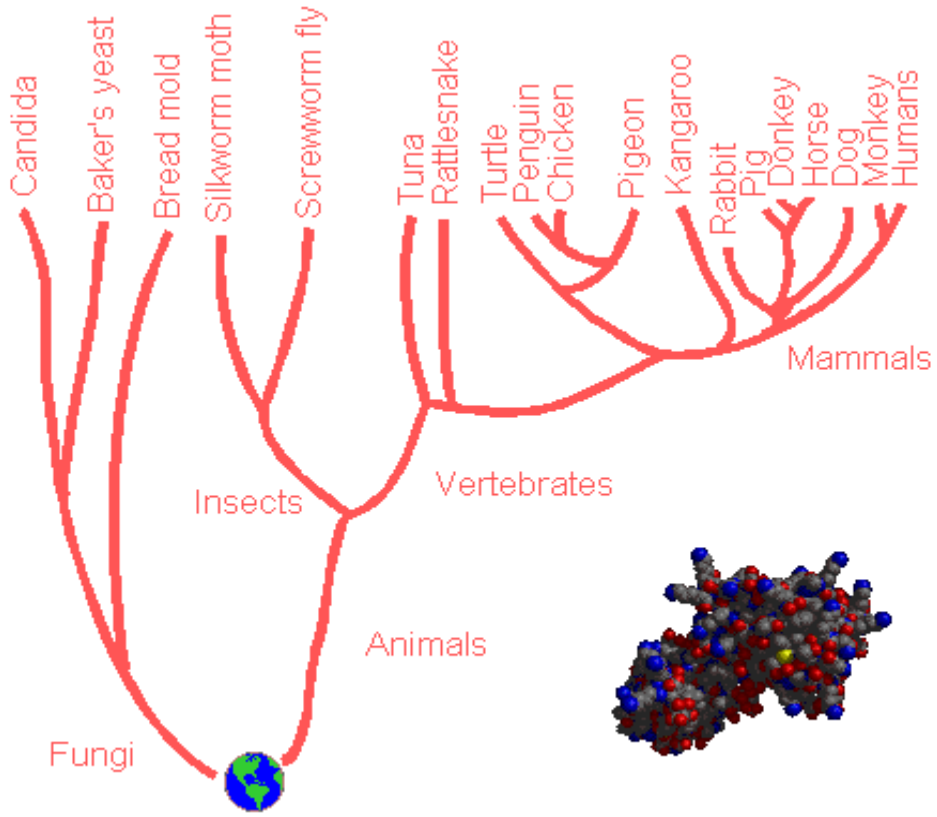
Phylogenetic trees

- A phylogenetic tree is a graph reflecting the approximate distances between a set of objects in a hierarchical fashion. There are different types of trees:

Unrooted versus rooted trees: A rooted tree has an additional node representing the origin, in molecular phylogeny the last common ancestor of the sequences analyzed

Recall

- In Phylogenetic trees
  - Leaves represent present day species
  - Interior nodes represent hypothesized ancestors



Why Build a Phylogenetic Tree ?

- Phylogenetic trees reconstruct the evolutionary history of your sequences
- They tell you who is closer to whom in the big tree of life
- Phylogenetic trees are based on sequence similarity rather than morphologic characters

3 Ways to Use Your Tree

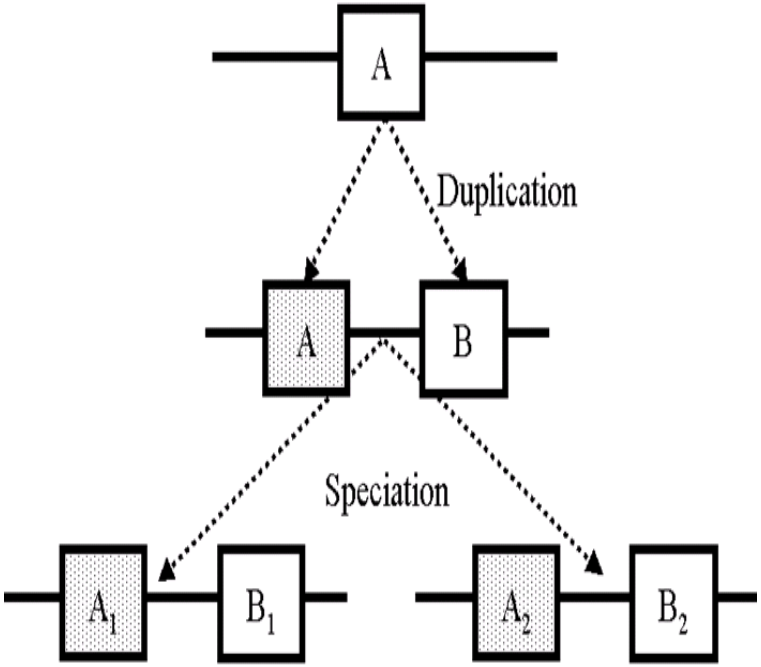
- Finding the closest relative of your organism
  - Usually done with a tree based on the ribosomal RNA

3 Ways to Use Your Tree ...

- Discovering the function of a gene
  - Finding the orthologues of your gene
- Finding the origin of your gene
  - Finding whether your gene comes from another species

Orthology and Paralogy

- Orthologous genes
  - Separated by speciation
  - Often have the same function
- Paralogous genes
  - Separated by duplications
  - Can have different functions
- In the graph:
  - A is paralogous with B
  - A1 is orthologous with A2



Working on the Right Data

- The quality of your tree depends on the quality of the data
- Your first task is to assemble a very accurate MSA

DNA or Proteins

- Most phylogenetic methods work on **Proteins** and **DNA** sequences
- If possible, always compute a multiple-sequence alignment on the protein sequences

DNA or Proteins …

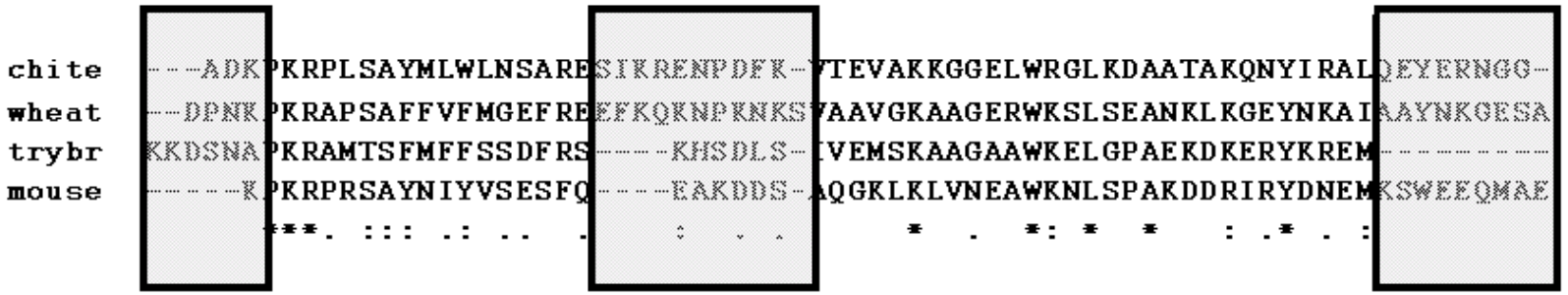
- If your DNA sequences are coding and have more than 70% identity . . .
  - Compute the **tree** on the **DNA** multiple-sequence alignment
- If your DNA sequences are coding and have less than 70% identity . . .
  - Compute the **tree** on the **protein** multiple-sequence alignment

Which Sequences ?

- Orthologous sequences
  - Produce a species tree
  - Show how the considered species have diverged
- Paralogous sequences
  - Produce a gene tree
  - Show the evolution of a protein family

Building the Right MSA

- Your MSA should have as few gaps as possible.
- Some variability but not too much!
- Some conservation but not too much!



Building the Right Tree

- There are two types of tree-reconstruction methods
  - Distance-based methods
  - Statistical methods

Building the Right Tree …

- Statistical methods are the most accurate
- Statistical methods take more time
  - Limited to small datasets

Distance-based Methods for Tree Reconstruction

- Distance-based methods are the most popular
  - Neighbor Joining (NJ)
  - UPGMA

Distance-based Methods for Tree Reconstruction …

- Distance-based methods involve 2 steps:
  - Measure the distances between pairs of sequences in the MSA
  - Transform the distance matrix into a tree

Distance-based Methods for Tree Reconstruction

- The two most popular packages for making trees are
  - **Clustalw**: very simple, not very sophisticated
  - **Phylip**: very powerful, less convivial

Computational challenge: There is an enormous number of different topologies even for a relatively small number of sequences:

- 3 sequences: 1 (unrooted)
- 4 sequences: 3
- 5 sequences: 15
- 10 sequences: 2,027,025
- 20 sequences: 221,643,095,476,699,771,875

Consequence: Most tree construction algorithm are heuristic methods not guaranteed to find the optimal topology.

Input data for two major classes of algorithms:

1. Input data distance matrix, examples UPGMA, neighbor-joining
2. Input data multiple alignment: parsimony, maximum likelihood

Distance matrix methods use distances computed from pairwise or multiple alignments as input.

- Distance measures respect the following constraints: $d = 0$ if the sequences are identical, $d > 0$ if the sequences are different
- Distances between molecular sequences are computed from pair-wise alignment scores.
- For closely related DNA sequences, one could simply use f , the fraction of non-identical residues (readily computed from the % identity value returned by an alignment program).

Unweighted pair-group method with arithmetic means (UPGMA)

Step 1

Initialization:

- assign each sequence to its own cluster
- define one leaf node for each single-sequence cluster
- put all leaf nodes at height zero

Step 2

Iteration.

- determine the two clusters for which the distance is minimal and combine them in a new cluster.
- compute the distance between the new cluster and all other clusters by averaging over all pair-wise distances between cluster elements
- define a new node for the new cluster and place it at height corresponding to the average distance between the cluster elements

Step 3

Termination:

- When no new nodes remains to be added, the process terminates