Phylogenetic tree

: Evolutionary history of a group of organisms or sequences

Finding out what phylogenetic trees can do for you
Making the right multiple sequence alignment for the right tree
Estimating distances between your sequences
Building a neighbor joining tree
Evaluating the quality of a tree by bootstraping

Finding out what phylogenetic trees can do for you

The purpose of phylogeny

reconstruct the history of life and explain the present diversity of living creatures

The underlying principles

group living creatures according to their level of similarity

The molecular stories that shaped the world as it is today: the destiny of gene families

- mutations and deletions
- duplications or speciation (diversification)
- loss and gain of function
- inactivation

Why you want to use phylogenetics

- determining the closest relatives of the organism that you’re interested in
- discovering the function of a gene
- retracing the origin of a gene

Preparing your phylogenetic data

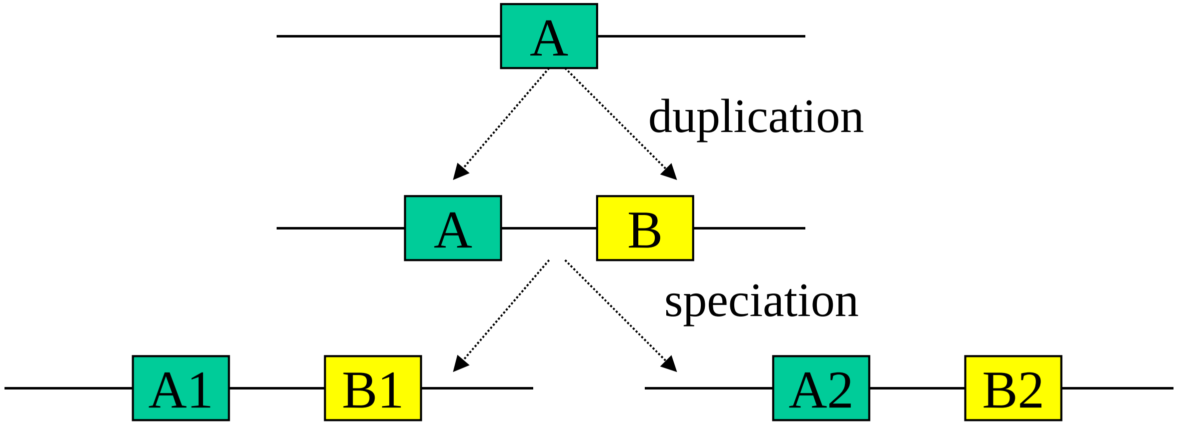
The quality of data: Choose the right sequences for the right tree

DNA or protein sequences

whenever possible use protein sequence
unless your sequences are identical, it is easier to keep working at the protein level

Homologous genes: derived from a common ancestor

orthologs: separated by speciation, similar functions and structure
paralogs: separated by a duplication event, different but related functions
xenologs: lateral transfer between two organisms



Preparing a set of sequences for making a phylogenetic tree

Avoid sequence fragments
Avoid xenologs
Avoid recombinant sequences

- proteins resulting from the combination of several proteins
- common in viruses

Avoid complex families

- various domains and repeats

Try to make a small set

Add an outgroup to your dataset

- an out group is a sequence that you know has diverged long ago
- symbolizes the first common ancestor of all your sequences
- (the most evolutionarily remote sequence)

Preparing your multiple sequence alignment

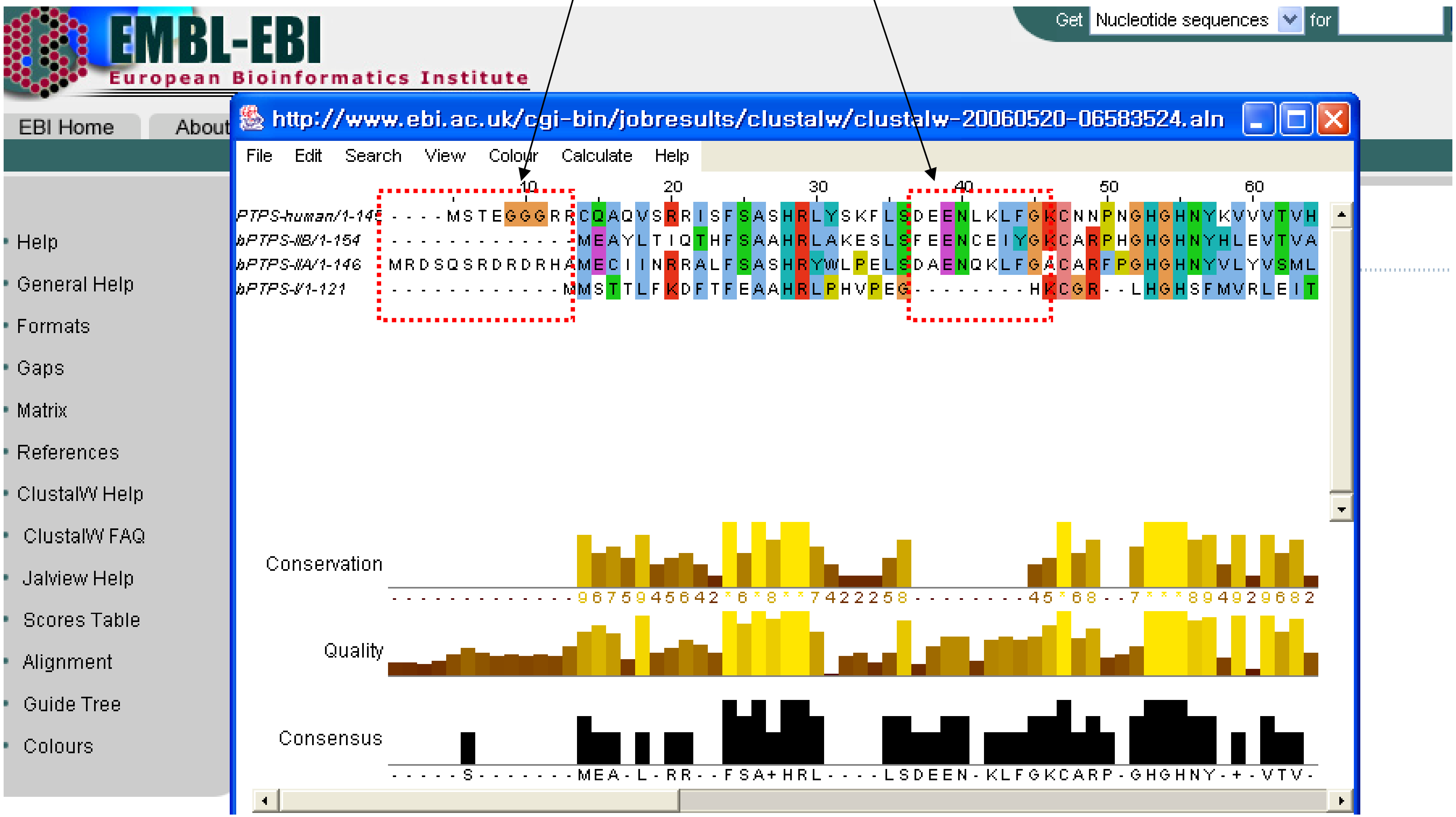
Computing your multiple alignment

- ClustalW: www.ebi.ac.uk/clustalw/index.html
- Dialign: bibiserv.techfak.uni-bielefeld.de/dialign/
- Tcoffee: igs-server.cnrs-mrs.fr/Tcoffee

Make sure you have the right multiple sequence alignment

- make sure there are as many gap-free columns as possible
- remove the extremities of your multiple alignment
  - the N- and C-terminus
- remove the gap-free regions of your alignment
  - internal gap-free regions often correspond to loops
- be sure to keep the most informatic blocks
  - blocks of 20-30 amino acids having a few conserved positions
  - (you can do it at the Jalview or with MS-word)

Using Jalview you can remove gapped sequences



```
PTPS-human/1-145  ----MSTEGGGRRCQAQVSRRISFSASHRLYSKFLSDEENLKLFGKCNNPNGHGHNYKVVVTVH
bPTPS-IIB/1-154   -------------MEAYLTIQTHFSAAHRLAKESLSFEENCEIYGKCARPHGHGHNYHLEVTVA
bPTPS-IIA/1-146   MRDSQSRDRDRHAMECIINRRALFSASHRYWLPELSDAENQKLFGACARFPGHGHNYVLYVSML
bPTPS-I/1-121     ------------MMSTTLFKDFTFEAAHRLPHVPEG--------HKCGR--LHGHSFMVRLEIT

Conservation      -------------9675945642*6*8**7422258--------45*68--7***89492968 2
Consensus         -----S-------MEA-L-RR--FSA+HRL----LSDEEN-KLFGKCARP-GHGHNY-+-VTV-
```

Building the kind of tree you need

Three major groups of methods you can use to build a tree

- distance methods
- parsimony methods: maximum parsimony
- likelihood methods: maximum likelihood
- *none of them are almighty

similarity

- Strong: maximum parsimony
- Weak: distance methods
- Very weak: maximum likelihood

check validity (bootstrapping analysis)

Distance-based methods

- easy to set up
- consider the four major ingredients
  - your multiple alignment: which method? Which sequences?
  - the distance measure: which substitution matrix?
  - the tree reconstruction algorithm: UPGMA, NJ, Fitch, or Kitsch?
  - the type of tree you want to display:
    - a cladogram without distances (evolutionary relationship)
    - a phylogram showing distances
    - a rooted or an unrooted tree

- Maximum parsimony

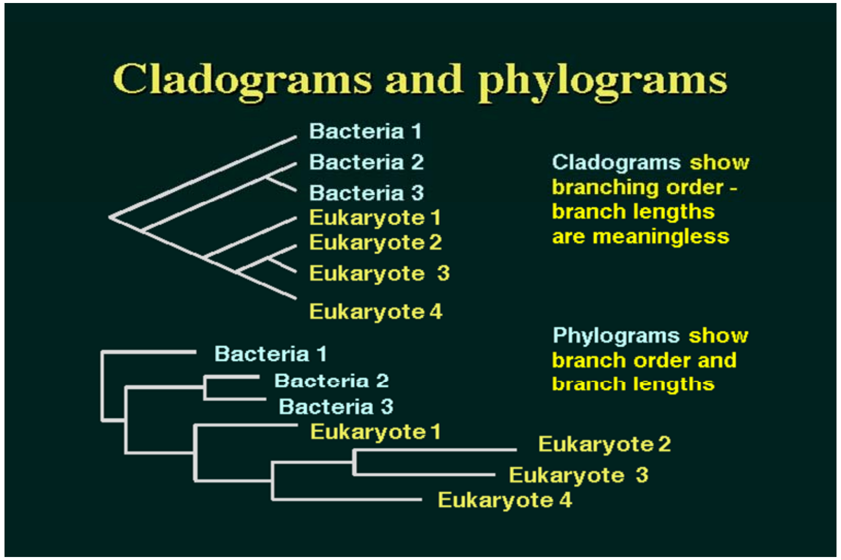
A tree that requires the least evolutionary change to explain some observed data

- Maximum likelihood

A tree with the best score is selected

- Distance based methods (Neighbour Joining)

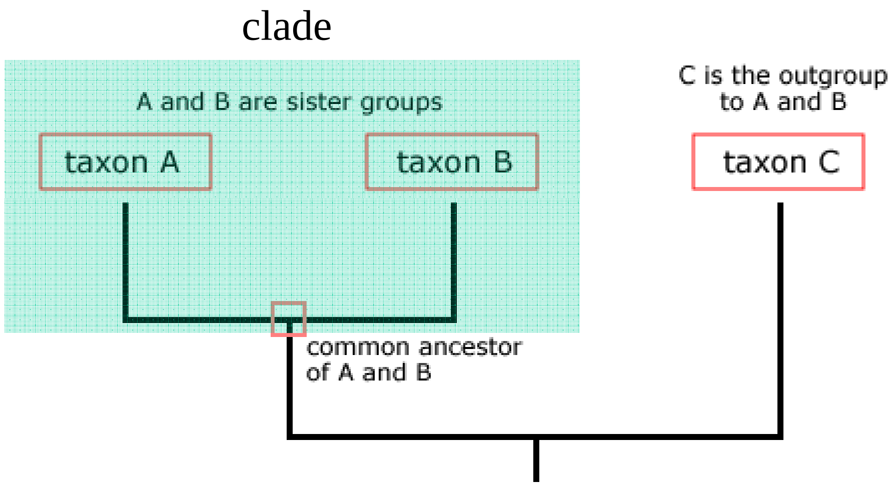
A tree such that branch lengths of paths between sequences fit a matrix of pairwise distances between sequences



Cladograms and phylograms

Cladograms show branching order - branch lengths are meaningless

Phylograms show branch order and branch lengths



Branch length: an indicator of either evolutionary timing or quantity of change

What's the difference between a phylogeny, an evolutionary tree, a phylogenetic tree, and a cladogram? For general purposes, not much. This site, along with many biologists, use these terms interchangeably — all of them essentially mean a tree structure that represents the evolutionary relationships within a group of organisms. The context in which the term is used will tell you more details about the representation (e.g., whether the tree's branch lengths represent nothing at all, genetic differences, or time; whether the phylogeny represents a reconstructed hypothesis about the history or the organisms or an actual record of that history; etc.) However, some biologists do use these words in more specific ways. To some biologists, use of the term "cladogram" emphasizes that the diagram represents a hypothesis about the actual evolutionary history of a group, while "phylogenies" represent true evolutionary history. To other biologists, "cladogram" suggests that the lengths of the branches in the diagram are arbitrary, while in a "phylogeny," the branch lengths indicate the amount of character change. The words "phylogram" and "dendrogram" are also sometimes used to mean the same sort of thing with slight variations. These vocabulary differences are subtle and are not consistently used within the biological community. For our purposes here, the important things to remember are that organisms are related and that we can represent those relationships (and our hypotheses about them) with tree structures. http://evolution.berkeley.edu/evolibrary/article/phylogenetics_02

1. Computing your tree

Producing an (almost) instant phylogenetic tree:

http://www.ebi.ac.uk/Tools/phylogeny/clustalw2_phylogeny/

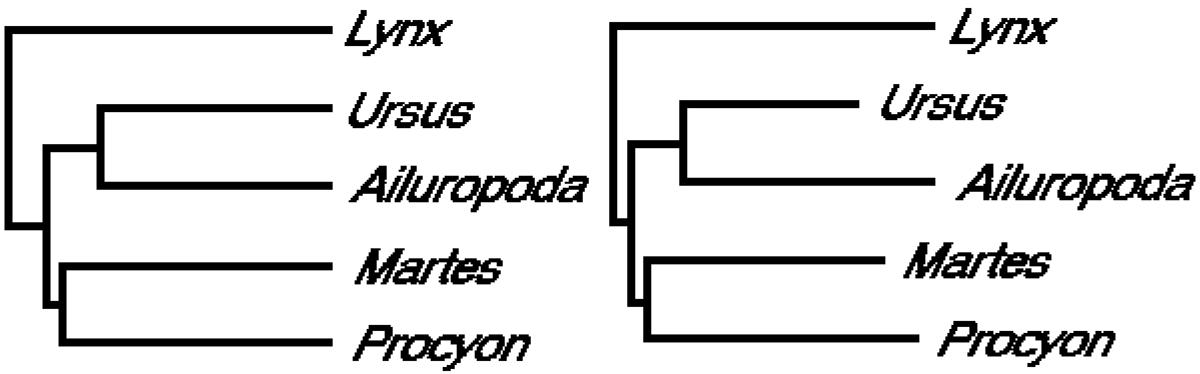
Carry out ClustalW to get ***.aln file

Upload the file and submit

The result will be shown in a new window

You can change options (use different clustering methods)

Clustering method



Alternative UPGMA & NJ phenetic analyses of Panda relationships

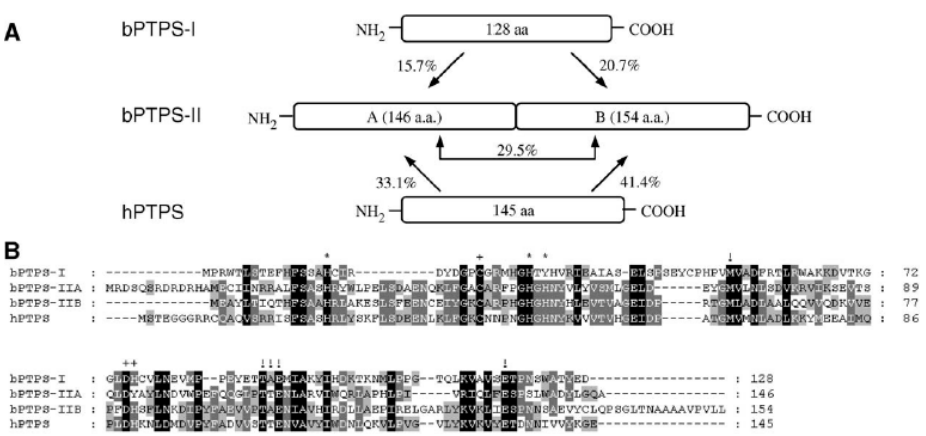
The **Unweighted Pair-Group Method** with Arithmetic Averaging (**UPGMA**) algorithm (left) assumes equal rates of evolution, so that branch tips come out equal. The **Neighbor-Joining** (**NJ**) (right) algorithm allows for unequal rates of evolution, so that branch lengths are proportional to amount of change. If rates on different branches are not markedly unequal, the branching orders produced by the two methods will not differ, as in this example with Carnivora.

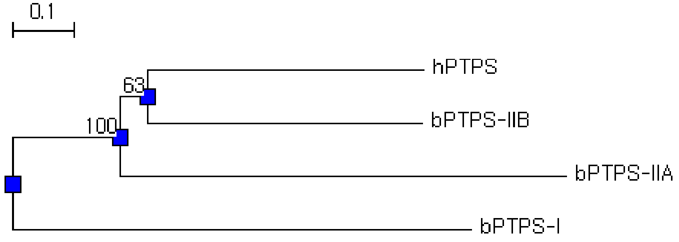
http://www.mun.ca/biology/scarr/Panda_UPGMA_&_NJ.html

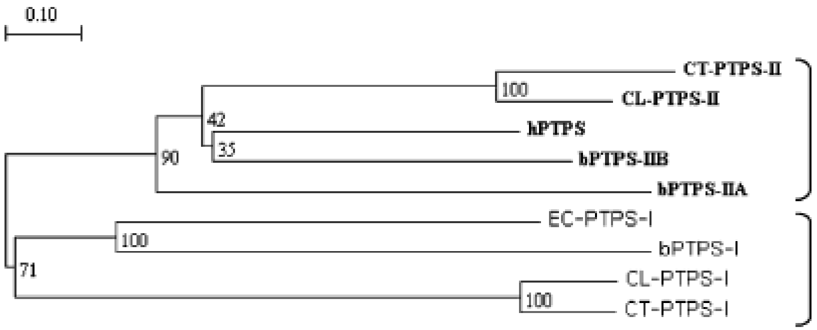
2. Bootstrapping analysis

A method for testing how good a dataset fits a evolutionary model.
Can check the branch arrangement (topology) of a phylogenetic tree.

In bootstrapping, the program re-samples columns in a multiple aligned group of sequences, and creates many new alignments. These new sets represent the population.







Assignment

Construction of a phylogeny tree in EMBL-EBI site
Check evolutionary relationship
Copy the phylogram and add common names

Perform bootstrapping analysis using ClutsalW
(Splits Tree: http://www.embl.de/~seqanal/courses/commonCourseContent/usingSplitstree.html)