Molecular phylogenetics

Dániel Gerber & Eszter Ari
ELTE TTK Department of Genetics
https://genetics.elte.hu
username & password: genetika2019

Scheme

What is phylogenetics and whit is it good for

What is the input and the output

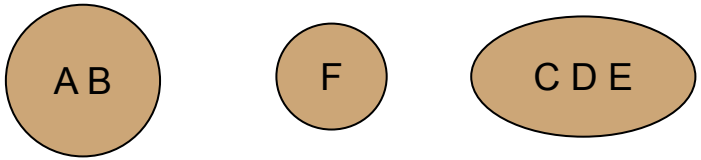
Phylogenetic terms and the phylogenetic tree

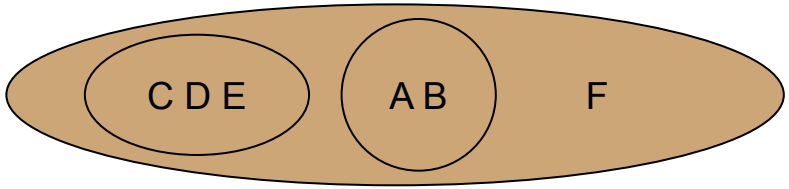
Molecular phylogeny methods

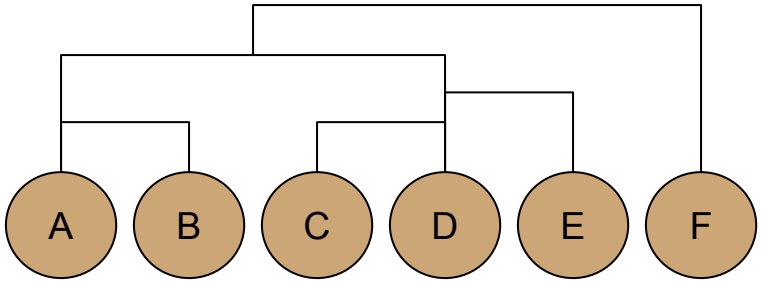
Tree drawing, rooting and assessing

Consensus trees and bootstrapping

Classification of elements (organisms, genes, etc.)







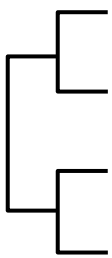
Non-hierarchical classification: groupings based on similarities without any further ranking between groups

Hierarchical classification: groupings based on similarities

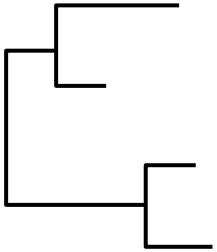
exclusive hierarchy: where groups are ranked

inclusive hierarchy: where all elements are ranked

Cladistics and phylogeny



Cladistics: inclusive ranking of elements (organisms, organs, genes, etc.), where the relationship between them equals to biological (= evolutional) relationship. The farther is one element from another, the bigger the **relative** evolutionary distance between them



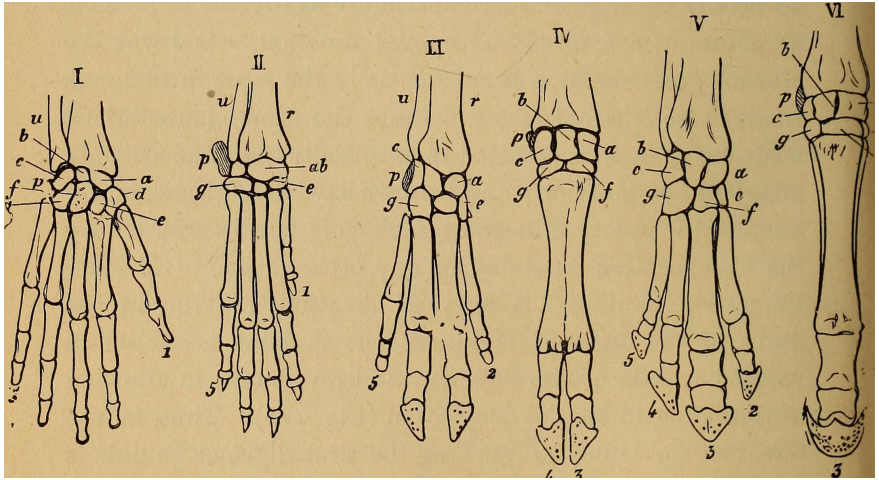
Phylogeny: cladistics, where we are aware not only of the relationships, but the **absolute** evolutionary distances (and changes) as well.

What is phylogeny good for?

A number of scientific analyses are based on the evolutionary relationships:

- taxonomy (eg.: evolution based grouping of taxa)
- epidemology (eg.: recover the origin of a certain strain)
- evolutionary biology (eg.: inferring ancestral state of a certain element)
- conservation biology (eg.: inferring population size)
- bioinformatics (eg.: gene function prediction)
- forensics (eg.: inferring kinship)
- archaeogenetics (eg.: inferring population geographical origins)
- etc.

Input data




```
Rhesus_CHID1      IHMLTHLAEALHQARLLALLVIPPAITPGTDQLGMFTHKEFEQLAPVLDGFSLMTYDYST
Hsa_CHID1_isoB    IHMLTHLAEALHQARLLALLVIPPAITPGTDQLGMFTHKEFEQLAPVLDGFSLMTYDYST
Horse_CHID1       IHMLTHVAEALHQARLLAILVIPPAVTPGTDQLGMFTHKEFEQLAPVLDGFSLMTYDYPT
Cat_CHID1         IHMLTHLAEALHQARLLAILVIPPAVTPGTDQLGMFTHKEFEQLAPVLDGFSLMTYDYPT
Finch_CHID1       IHLLTHLSEALHEAQLKLILVIPPAVAAGTNQPGMFTKKEFDQLASAIDSFSLMTYDYSA
Catfish_CHID1     VHLVTHLCKYLKACKFSCILVIPPSVIPGTNQPGMFGREDFEKLAPVVDAFSLMTYDYSG
Honeybee_CHID1L   VALIQFIAHQLKNNNLDTILAVPPSRGP---NIQLFNRDQFDQLSPYLKAFSLMTYDYSS
consensus         ihmlthlaeaLhqarllaiLviPPavtpgtdqlgmFthkeFeqLapvldgFSLMTYDYst
```

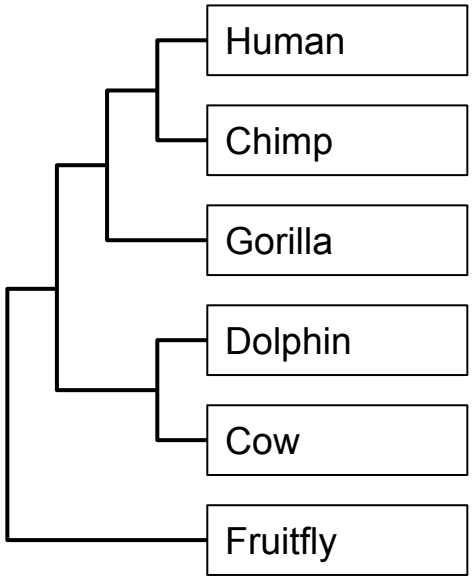
Inferring phylogenies can be done by basically any phylogenetically informative elements of the subject under analysis.

Phylogenetically informative element: any feature that can be compared between or within organisms and possess information about it's evolutionary history

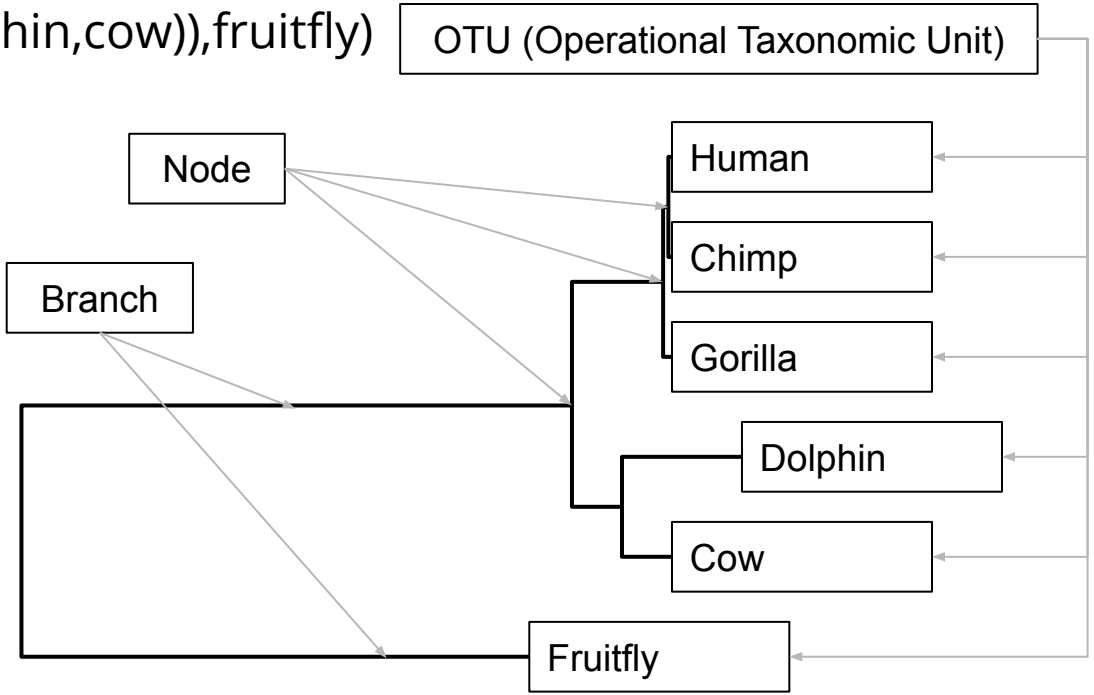
- organs (wings, legs, skulls, certain bones, number of fingers, and so...)
- biomolecules (DNA, RNA, Protein sequences)

Output: phylogenetic tree (coded in newick file)

((((human,chimp),gorilla),(dolphin,cow)),fruitfly)

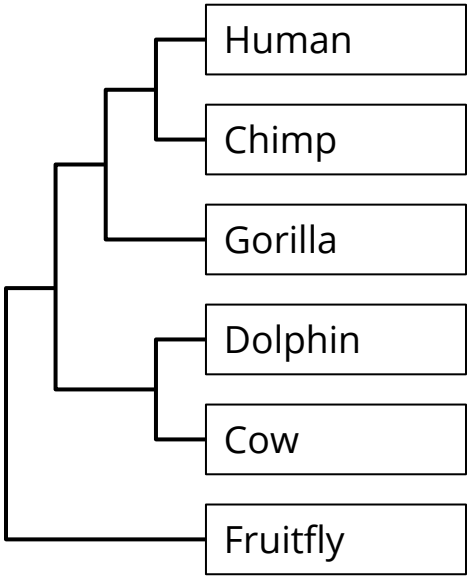


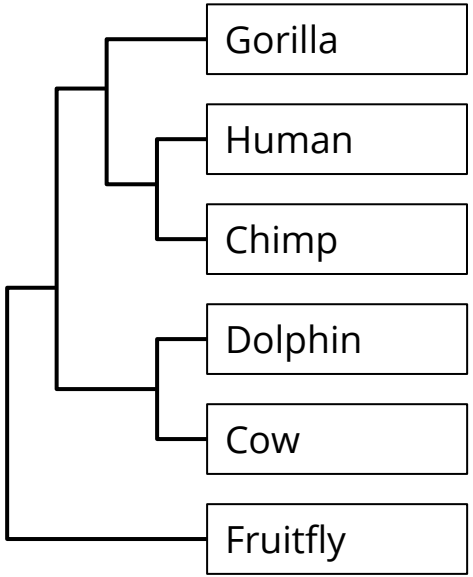
Cladogram

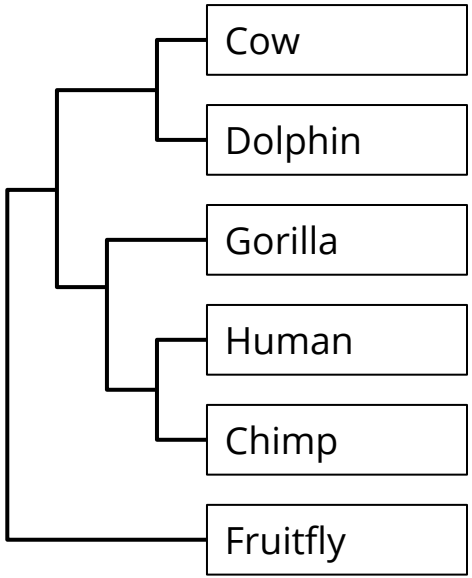


Phylogram

Rotating the tree: these are exactly the same!!!





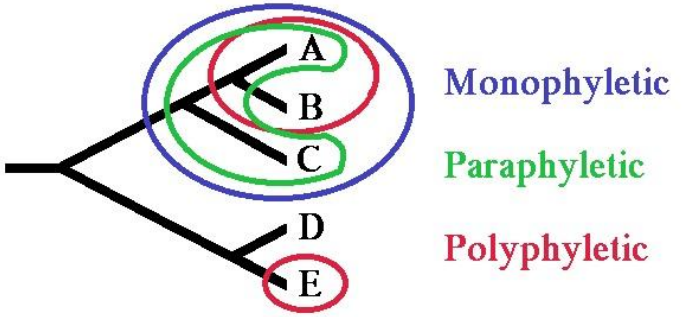


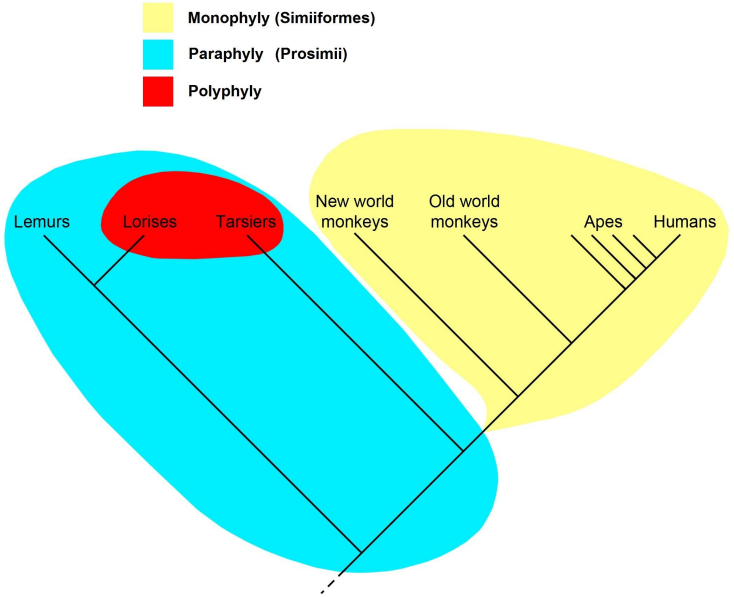
Phylogenetic terms

Monophyletic: evolutionary relationship, common origin and inclusion of all known taxa

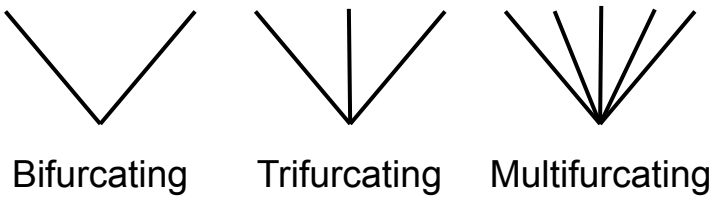
Paraphyletic: evolutionary relationship, common origin but some taxa are excluded

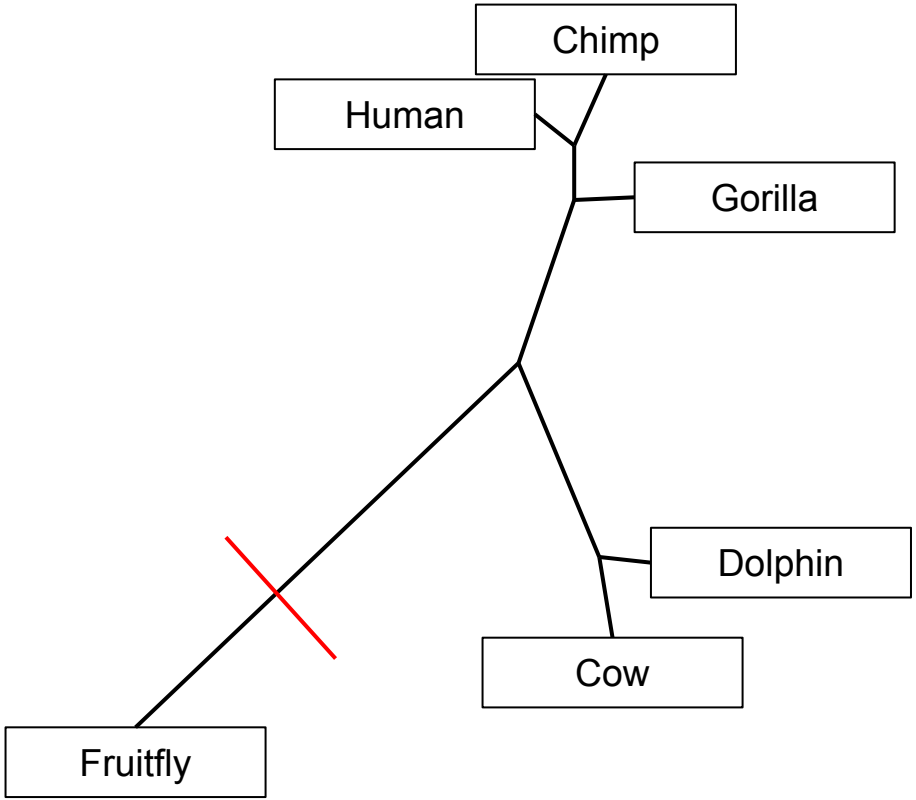
Polyphyletic: no evolutionary relationship, only grouped by arbitrary common features (capable of flying) or by mistake (Lorises and Tarsiers)

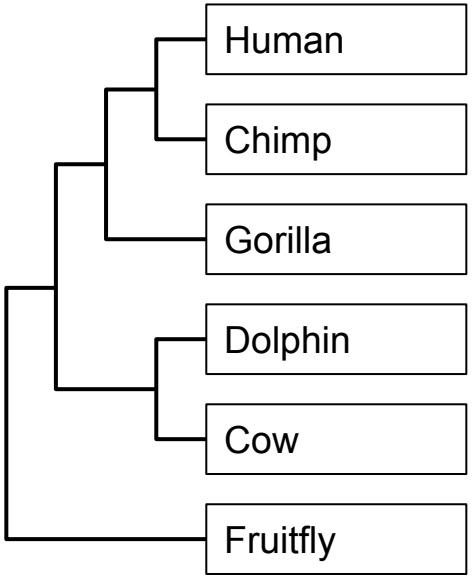




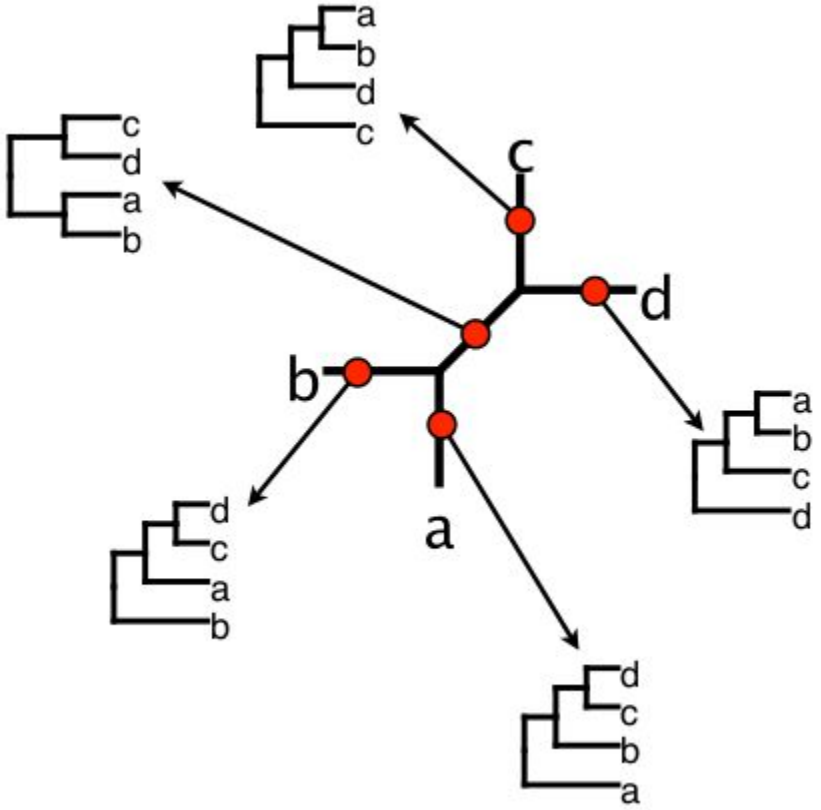
Tree shapes and rooting







Tree rooting



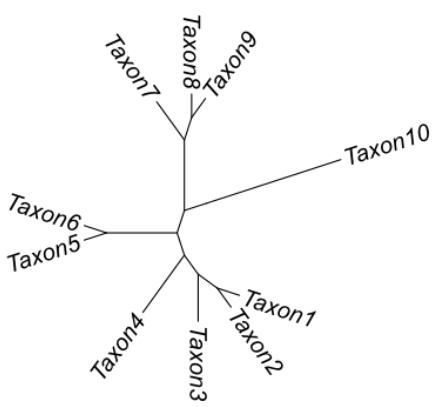
By rooting a tree, we can observe the way of evolution and assess the evolutionary relationships easier (also required for a lot of further analysis)

Three main way to root a tree:

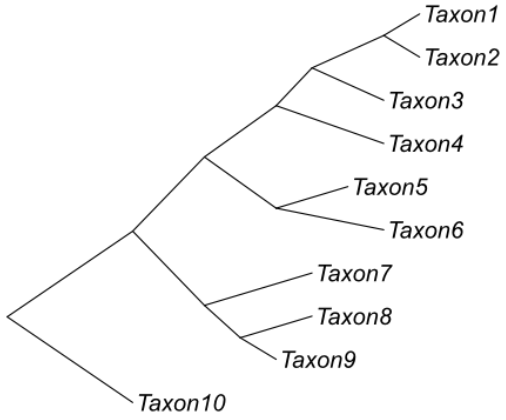
- midpoint (consider the basal node as the halfway between the two most distant taxa; lazy and not really reliable)
- using outgroup (has to be chosen really carefully)
- directed from character states (tricky and unnecessary in most cases)

Real trees with multiple shapes

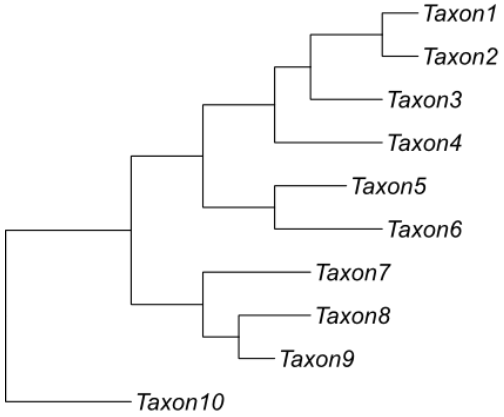
a) radial (unrooted) tree
b) rectangular tree I
c) rectangular tree II
d) polar tree
e) cladogram type



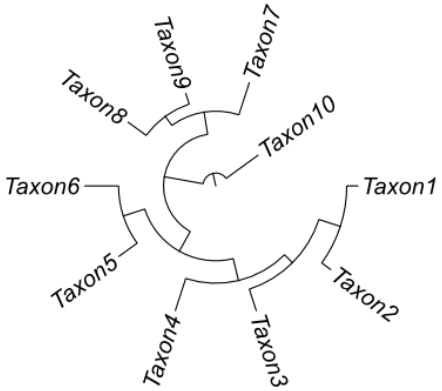
a



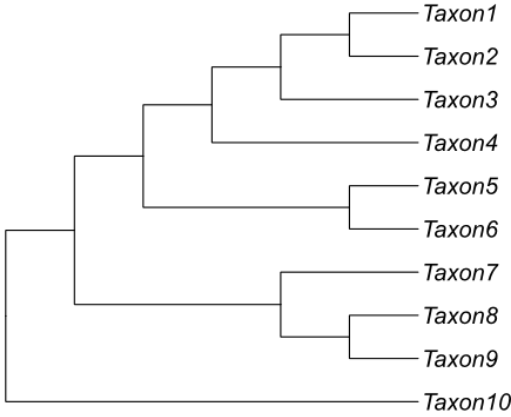
b



c



d



e

Molecular phylogeny workflow

Step 1: creating input data, in this case a multiple sequence alignment (DNA, RNA, Protein)

Step 2: Substitution model (necessary in predicting the timing of evolutionary changes, filtering out noise, etc.)

Step 3: Phylogenetic method choosing (distance or character based) and calculating (or finding) the actual tree

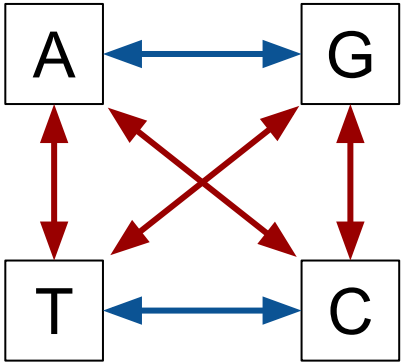
Step 4: Bootstrapping (statistical validation of the final tree)

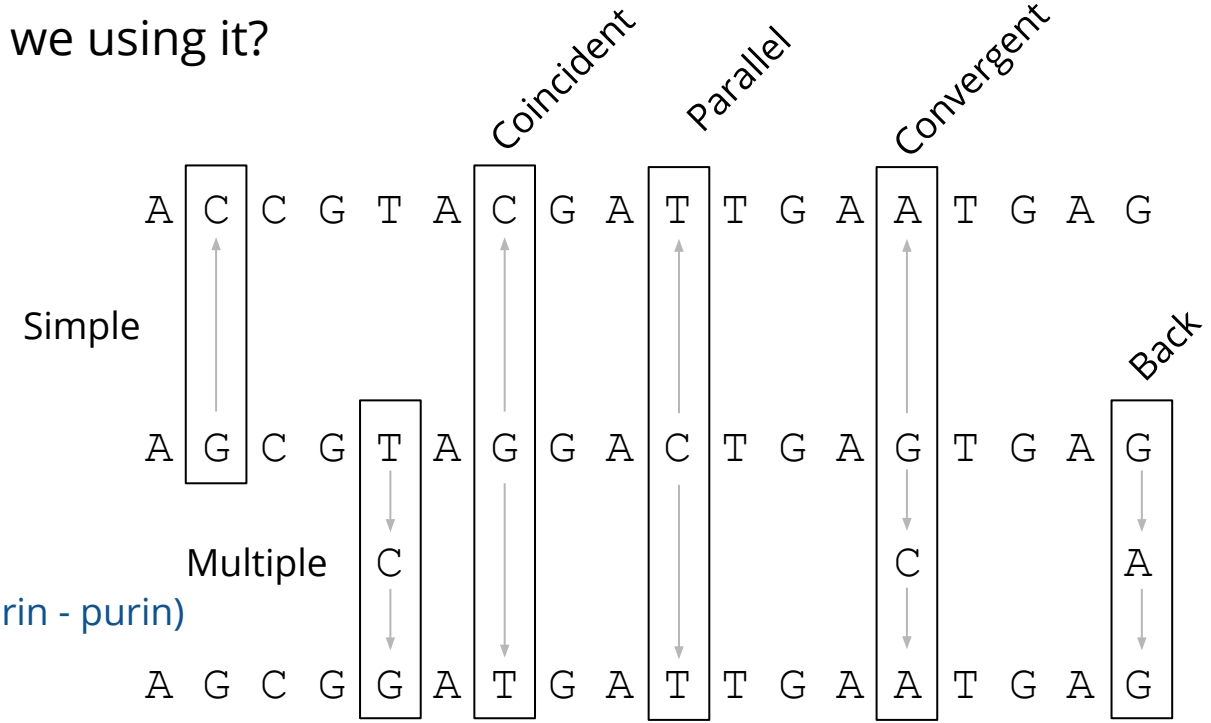
Step 1: creating multiple sequence alignment

See previous lesson!

Step 2: Choosing the best substitution model I.

What is that and why are we using it?





Transitions (pirimidin - pirimidin, purin - purin)

Transversions (purin - pirimidin)

Step 2: Choosing the best substitution model II.

Aim: predict the rate of base changes (= substitutions) during evolution (time)

Hamming distance: calculating everything without substitution models, but only used for testing in molecular datasets

Various substitution models exist which help to predict more accurate tree topologies and evolutionary distances:

Most basic: **Jukes-Cantor** model (JC69), only calculates transitions and transversions (2 substitution classes)

Most sophisticated: General Time reversible (**GTR**), every substitution type is calculated in (6 substitution classes)

A number of other models exist between them (F84, K2P, HKY85, etc.)

Model choice is critical for the analysis, however, online tools or downloadable softwares exist which predicts the best model according to data composition

Step 3: Choosing the phylogenetic method I.

Distance based methods is when we calculate the (pairwise) differences between sequences, where we get a distance matrix. After this, we can calculate the phylogenetic tree from this distance matrix.

A number of distance based methods can be found today, like median joining network (MJN) or minimum spanning network (MSN), but the most well known is the Neighbour Joining (NJ) method.

There are two main steps for calculating an NJ tree:

- calculate a distance matrix
- calculate the tree itself

Calculating a distance matrix (here by using JC69)

Distance (d) between sequences A and B:

$d_{AB} = -(3/4) \ln (1 - 4/3 D)$

D: ratio of differences (differences / number of sites)

if D = 0,05, that means that just by counting the differences (=Hamming distance) there is 5% differences between sequences, which could mean, that the distance between them is 0,05 or 5 (does not really matter, just be on the same scale)

BUT! due to the previously described substitution types, the real evolutionary difference between these sequences is slightly higher. Calculating in a two class substitution model (3 possible changes, 4 nucleotides (AGTC)) alters the output:

$d_{AB} = -(3/4) \ln (1- 4/3 * 0,05) = 0,0517$

0,17% extra difference masked by noise (=unseen evolutionary changes)

Now consider a 50% difference between two sequences:

$d_{AB} = -(3/4) \ln (1- 4/3 * 0,5) =$ **0,824**

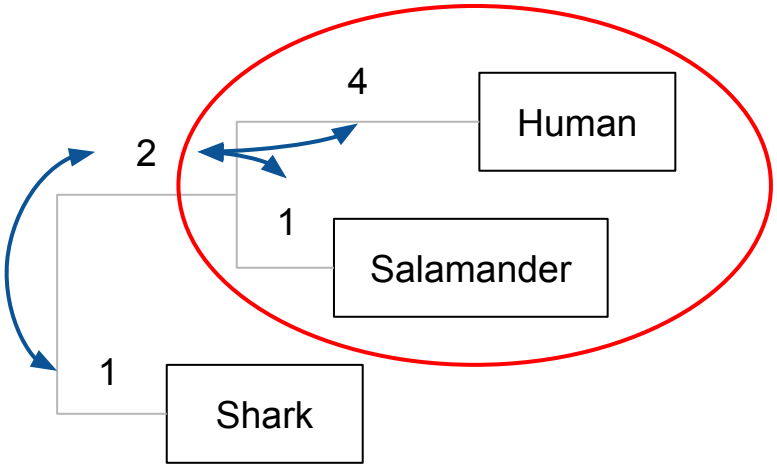
The result is a distance matrix

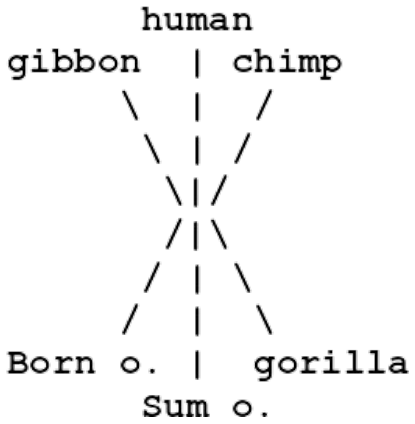
	Human	Chimp	Gorilla	Sumatran orangutan	Borneo orangutan	Gibbon
Human	0					
Chimp	5	0				
Gorilla	4	7	0			
Sumatran orangutan	7	10	7	0		
Borneo orangutan	6	9	6	5	0	
Gibbon	8	11	8	9	8	0

Building the neighbour joining tree

	H	C	G	S	B	G
H	0					
C	5	0				
G	4	7	0			
S	7	10	7	0		
B	6	9	6	5	0	
G	8	11	8	9	8	0

The starting shape is a star, where we start to build the branches by iteratively finding the pairs that are TOPOLOGICALLY (not by distance!!!) the closest to each other





To achieve this, one have to create a modified matrix to search for the first pair to group together

	H	C	G	S	B	G
H	0					
C	5	0				
G	4	7	0			
S	7	10	7	0		
B	6	9	6	5	0	
G	8	11	8	9	8	0

Step 1:

Sum of distances (r(i)):

Human: 0 + 5 + 4 + 7 + 6 + 8 = 30

r(human) = 30

Chimp: 5 + 0 + 7 + 10 + 9 + 11 = 42

r(chimp) = 42

etc.

To achieve this, one have to create a modified matrix to search for the first pair to group together

	H	C	G	S	B	G
H	0					
C	5	0				
G	4	7	0			
S	7	10	7	0		
B	6	9	6	5	0	
G	8	11	8	9	8	0

Step 2:

Use the formula to find the lowest modified value (M) in the matrix, that will be the first two OTU to pair (here pair i and j, N is the number of OTUs)

$$M(i,j) = d(i,j) - [r(i) + r(j)] / (N - 2)$$

e.g.:

$$M(\text{human, chimp}) = 5 - [30 + 42] / (6 - 2) = -13$$

To achieve this, one have to create a modified matrix to search for the first pair to group together

	H	C	G	S	B	G
H	0					
C	-13	0				
G	-11,5	-11,5	0			
S	-10	-10	-10,5	0		
B	-10	-10	-10,5	-13	0	
G	-10,5	-10,5	-11	-11,5	-11,5	0

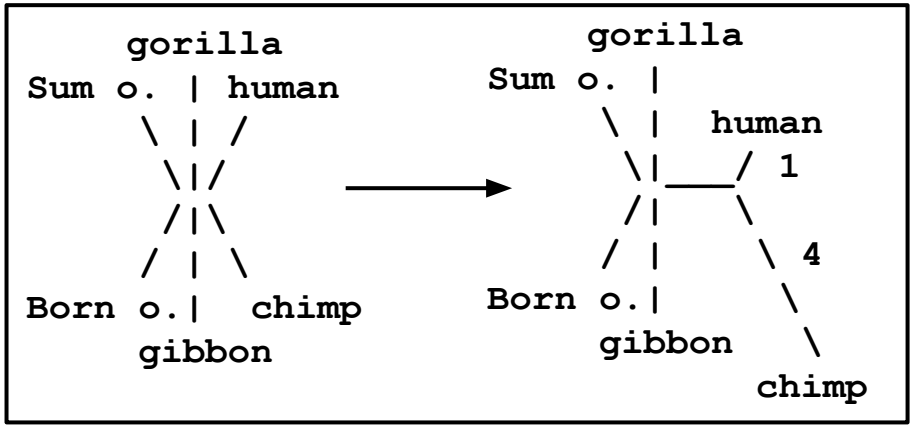
Now we can see that there is two lowest value in the matrix, we can start with any of them. Creating a new node (U) by grouping together say the chimp and human first.

To get the branch lengths (S), we have this formula:

S(human, U) = d(human, chimp) / 2 + [r(human) - r(chimp)] / 2(N-2) = 1

S(chimp, U) = d(human, chimp) - S(human, U) = 4

Now we have a new tree and a new matrix by recalculating the distances



	U(H,C)	G	S	B	G
U(H,C)	0				
G	3	0			
S	6	7	0		
B	5	6	5	0	
G	7	8	9	8	0

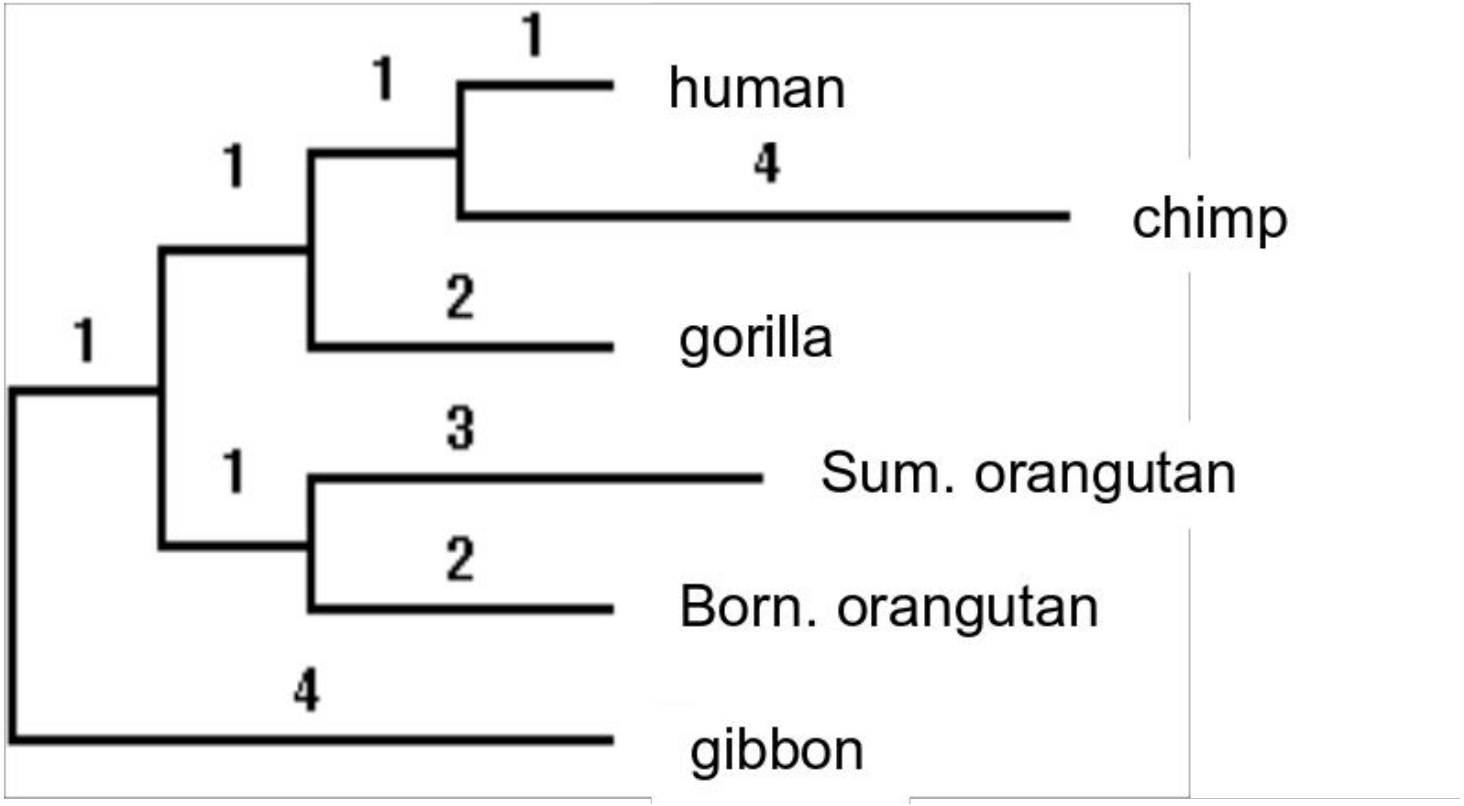
d(i,U(Human,Chimp)) = [d(i,Human) + d(i,Chimp) - d(Human,Chimp)] / 2

e.g.: d(gorilla, U(Human,Chimp)) = [d(gorilla, Human) + d(gorilla, Chimp) - d(Human, Chimp] / 2

d(gorilla, U(Human,Chimp) = [4 + 7 - 5] / 2 = 3

N now becomes N - 1 = 5; reiterate the whole process until everything got grouped

The result tree



Pros and cons of a NJ tree

Cons:

- very basic, only a rough estimation of a phylogenetic tree
- a lot of complicated evolutionary changes remain hidden

Pros:

- easy to use and to understand
- fast (still in use for large datasets)
- in most of the times it provides the correct tree topology anyway

Step 3: Choosing the phylogenetic method II. - character based methods

Maximum parsimony (MP) is one of the character based methods

Parsimony principle: based on the theory of **Occam's razor**: the simplest hypothesis that explains the data should be selected

In terms of phylogeny: the tree we are choosing assumes the minimum character changes (=substitutions) during evolution of the sequences

Basically this means that we are going to have a huge number of possible trees, and the method is going to select the one that has the less amount of necessary changes (the tree that assumes the less evolution that possible), minimizing the number of homoplasies (convergent evolution, back mutations, parallel changes, etc.).

Maximum parsimony working method

Get aligned sequences (four at least)

Get a tree topology drawn from the data, calculate the less amount of character changes required for that tree

Repeat this for all possible tree topologies for our data

Scoring the trees, considering the best tree as the one with the less amount of substitutions

Definitions for character based methods

Branch length: length of the tree branch that reflects the number of substitutions (=evolutionary changes)

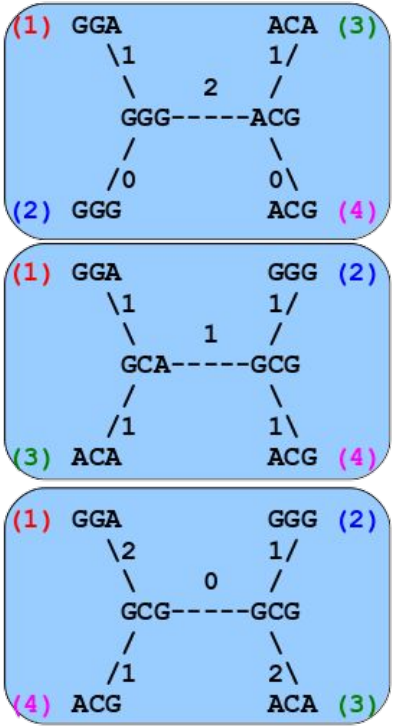
Total branch length: sum of all branch lengths of the tree

Uninformative sites: characters in the alignment that have no information (missing data, singletons (mutation that occurs in only one OTU), unchanged sites)

Informative sites: where at least two type of character state appears, and each appears at least in two OTUs

Calculation

	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	G	C	A
2	A	G	A	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G
					*		*		*



In terms of calculation it is quite simple, however, by increasing the number of taxa the number of possible tree topolgies are exponentially increasing

4 taxa - 3 possible unrooted trees, 3*5 possible rooted trees

5 taxa - 105 possible trees

10 taxa - 34,5 million possible trees

20 taxa - $8{,}2*10^{21}$ possible trees

Computationally exhaustive and in most of the time there is more than 1 best trees exist

Various heuristic methods are available to decrease the computational time, but that is beyond the boundaries of this class

Pros and cons of maximum parsimony

Cons:

- only calculates the correct topology when no or minimal amount of invisible mutations exist in the data
- often needs a lot of informative characters
- highly sensitive to different evolutionary rates between taxa (=branch lengths vary a lot), tends to group together branches with similar lengths (long branch attraction)

Pros:

- could be effective for modeling the evolution of complex traits (e.g.: morphological evolution)

Other character based methods

Maximum likelihood:

A heuristic tree searching method, it finds the best tree by using the likelihood of tree topologies under the given paramteres (alignment). The output is the most likely tree in the tree search area.

Bayesian method:

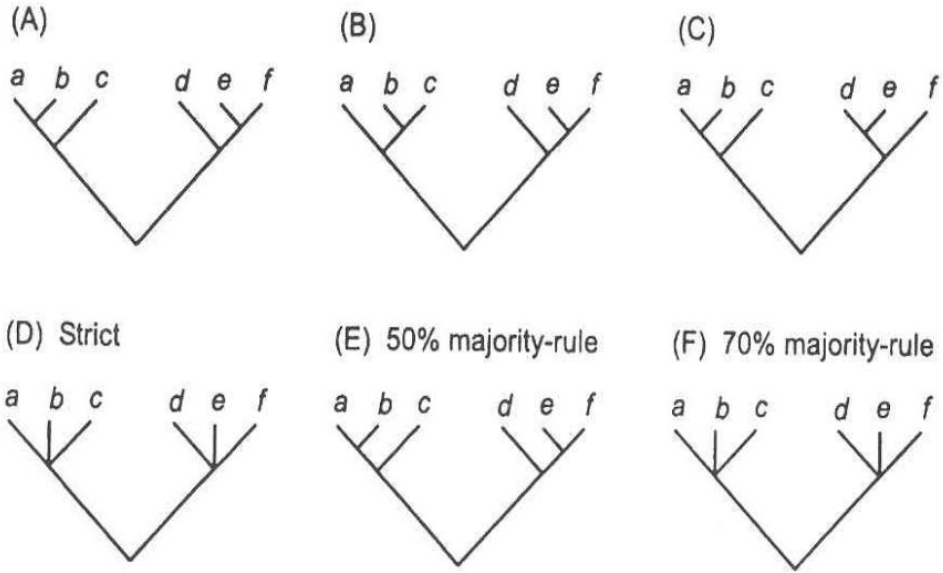
A heuristic tree searching method, it provides the best tree by merging all the most likely tree topologies.

Both are used frequently, and these are the most sophisticated methods available nowadays. It is highly recommended to use one of these methods.

Bayes and Maxium Parsimony: more than one best tree?

These two methods are providing more than one possible best tree topologies, which then can be merged to create the final tree topology. This output tree is called the Consensus tree.

This is as simple as it looks: by using the majority rule, the branching that appears the most considered to be the best. 50% majority rule means that we consider 50% of the trees with the most similar branching, 100% (strict) means that we consider all trees, and in case of discrepancy, the sub-branch considered to be unresolved.



Bootstrap analysis: statistical validation of the tree

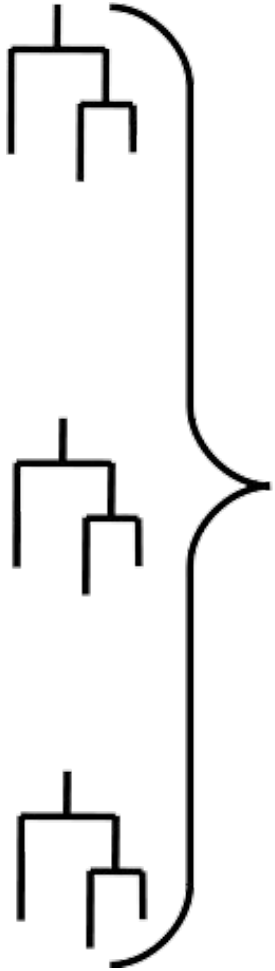
To assess whether we got the best branching, we have to use a statistical approach called bootstrapping. The method in general is quite simple

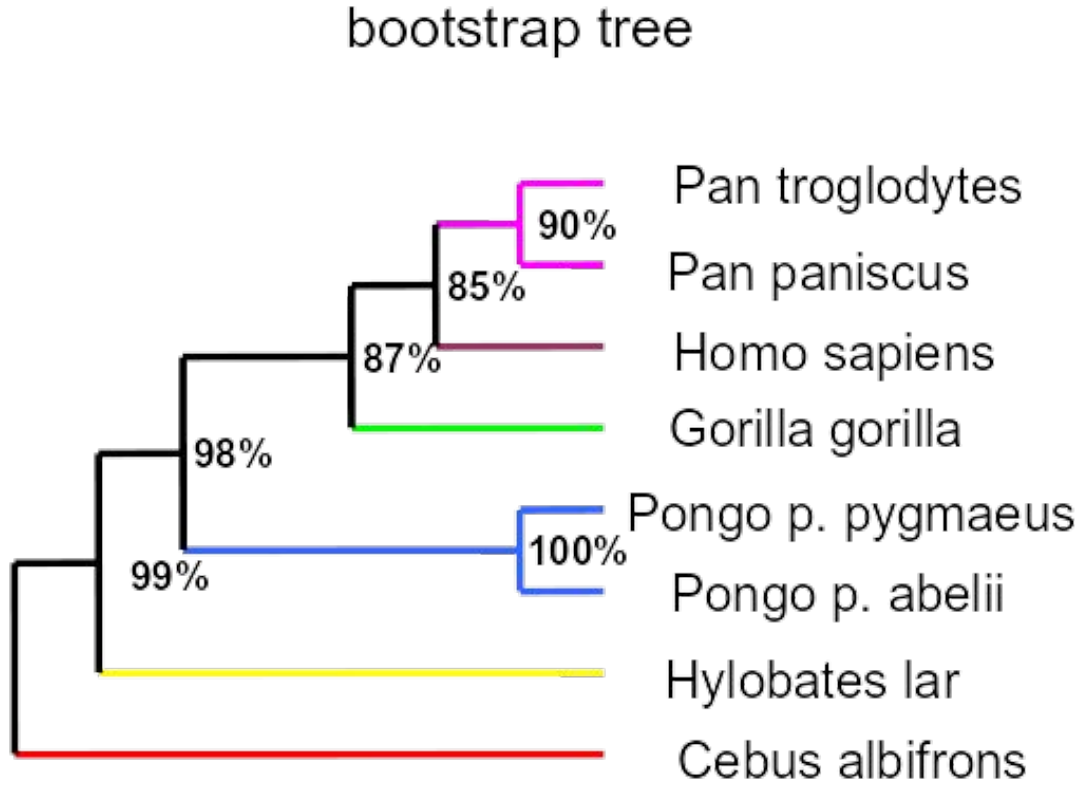
1) generating a new alignment by randomly subsampling the original one, this creates noise and/or hides informative sites
2) calculate the tree for the new dataset
3) repeat the process (the repeat number is optional, it is used to be 100 to 1 million)
4) create a consensus tree from it, check the differences between this and the original one OR calculate the percent probability of a certain node and OTU grouping according to how many times it appears in the bootstrap series

```
           0000000001111
           1234567890123
Cebu_a_Ala GAGGAGTTGAAAG
Homo_s_Ala AAGGGGCTGAGGA
Pan_tr_Ala AAGGGGCTGAGGA
Pan_pa_Ala AAGGGGCTGAGGA
Gorill_Ala AAGGGGCTGAGGA
Pong_p_Ala GAGGGGCTGAGAA
Pong_a_Ala GAGGGGCTGAGGA
Hylo_l_Ala AAGGGACTGGGAA

           0000000010111
           1343388415333
Cebu_a_Ala GGGGGTTGAAGGG
Homo_s_Ala AGGGGTTGGGAAA
Pan_tr_Ala AGGGGTTGGGAAA
Pan_pa_Ala AGGGGTTGGGAAA
Gorill_Ala AGGGGTTGGGAAA
Pong_p_Ala GGGGGTTGGAAAA
Pong_a_Ala GGGGGTTGGGAAA
Hylo_l_Ala AGGGGTTGGAAAA

           0000011000100
           1222411664055
Cebu_a_Ala GAAAGAAGGGAAA
Homo_s_Ala AAAAGGGGGGAGG
Pan_tr_Ala AAAAGGGGGGAGG
Pan_pa_Ala AAAAGGGGGGAGG
Gorill_Ala AAAAGGGGGGAGG
Pong_p_Ala GAAAGGGGGGAGG
Pong_a_Ala GAAAGGGGGGAGG
Hylo_l_Ala AAAAGGGAAGGGG
```





Literature

Felsenstein, Joseph (2004). Inferring Phylogenies. Sinauer Associates, Inc., Sunderland, MA, USA.

Hall, Berry G. (2004). Phylogenetic Trees Made Easy; A How-To Manual, Second edition, Sinauer Associates, Inc., Sunderland, MA, USA.

Podani János (2000). Introduction to the exploration of multivariate biological data, Backhuys Publishers

Salemi, Marco & Vandamme, Anne-Mieke (2003). The phylogenetic handbook : A practical approach to DNA and protein phylogeny. Cambridge University Press, Cambridge, U.K. ; New York.

Barton N.H. et al: Evolution, Chapter 27: Phylogenetic Reconstruction
http://evolution-textbook.org/content/free/contents/ch27.html

Joe Felsenstein honlapja: PHYLIP is a free package of programs for inferring phylogenies
http://evolution.genetics.washington.edu/phylip.html

MrBayes: Bayesian Inference of Phylogeny
http://mrbayes.csit.fsu.edu/

Thank you for your attention

We are looking for master thesis writers!

Laboratory of Archaeogenetics
https://ri.btk.mta.hu/archaeogenetika/kutatas.html

Topics:

- Recent human population genetics in the Carpathian Basin
- Ancient human population genetics in Central-Eastern Europe and Northern Asia

Further details in person

Find us:
Dániel Gerber (gerberd1990[at]gmail.com)
Anna Szécsényi-Nagy (Szecsenyi-Nagy.Anna[at]btk.mta.hu)