16.3.09

Practical analysis

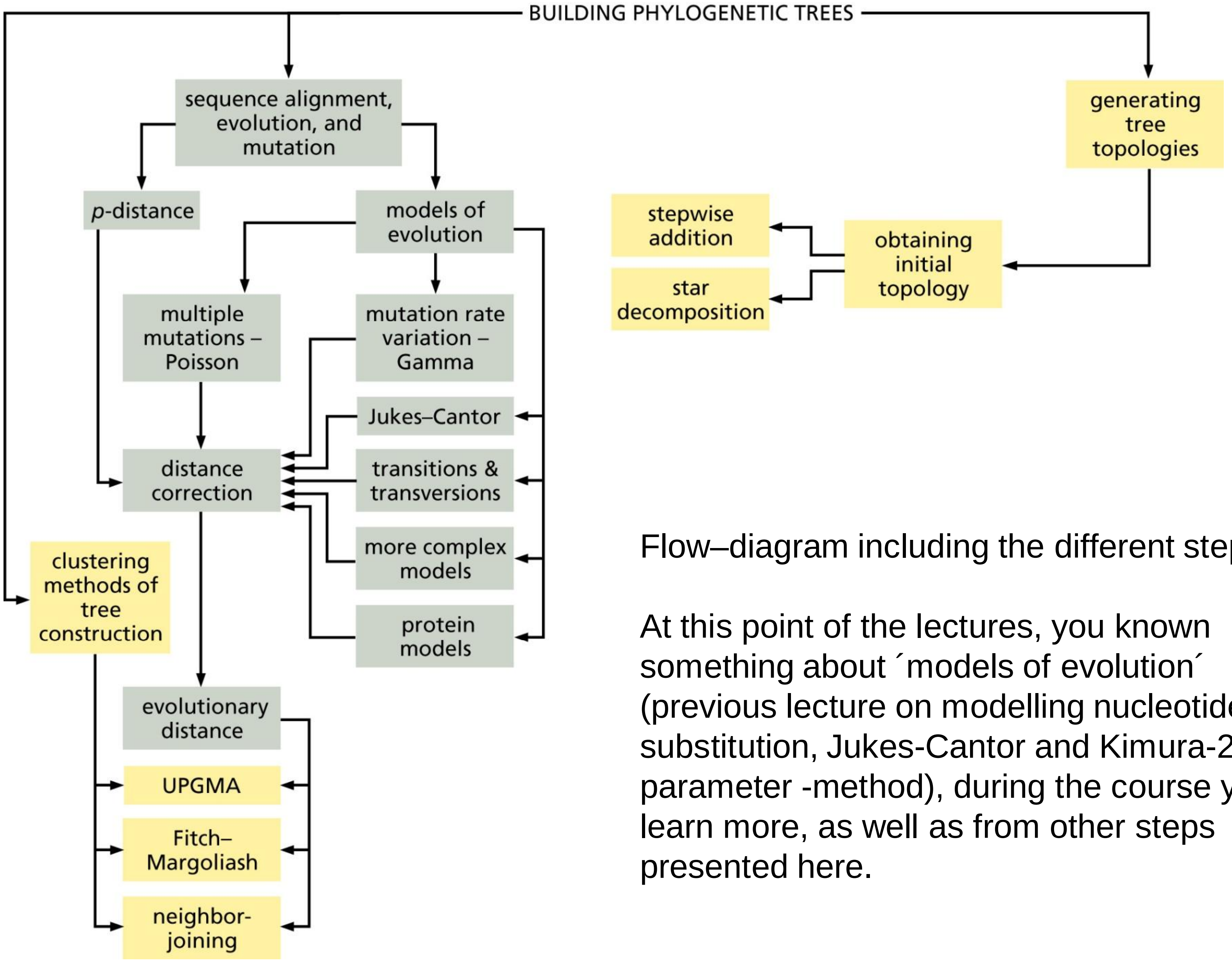
Neighbor-joining phylogeny by MEGA-software

This lecture:

- Introduction to getting started with phylogenies in practice
- Distance matrix methods, the simple ones
- MEGA-software, the easy-to-use software for distance matrix methods

Distance matrix methods

- UPGMA
- Fitch-Margoliash
- Neighbor-joining
- Because these (simple) methods (at least UPGMA and neighbor-joining) are included in several other lectures (for example ´*Introduction to bioinformatics*´), here we just briefly go through their basics, and proceed to their usage in practice

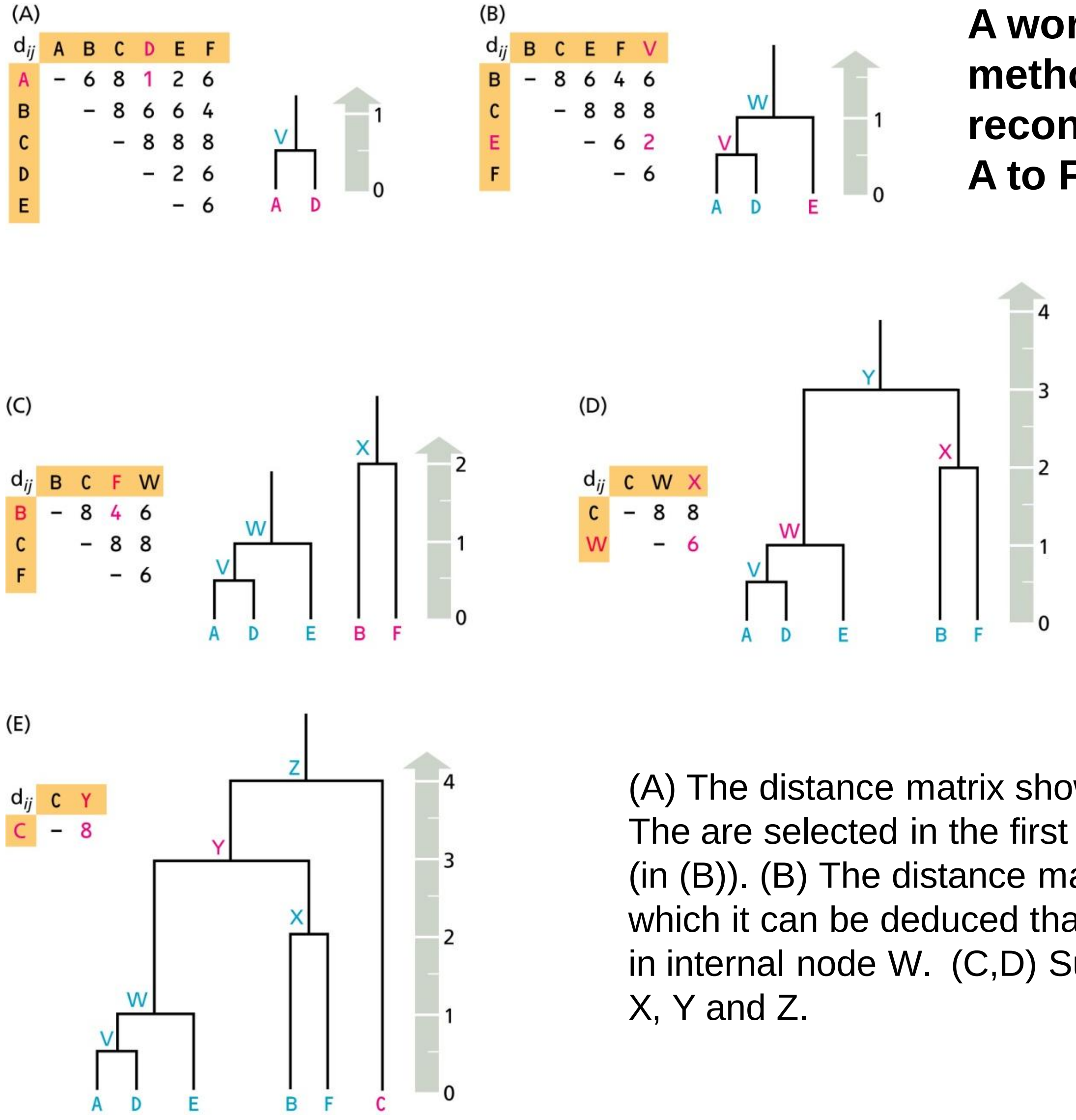

BUILDING PHYLOGENETIC TREES
sequence alignment, evolution, and mutation
p-distance
models of evolution
multiple mutations – Poisson
mutation rate variation – Gamma
Jukes–Cantor
transitions & transversions
more complex models
protein models
distance correction
clustering methods of tree construction
evolutionary distance
UPGMA
Fitch–Margoliash
neighbor-joining
generating tree topologies
obtaining initial topology
stepwise addition
star decomposition

Flow–diagram including the different steps .

At this point of the lectures, you known something about ´models of evolution´ (previous lecture on modelling nucleotide substitution, Jukes-Cantor and Kimura-2-parameter -method), during the course you will learn more, as well as from other steps presented here.

This picture is from : Zvelebid&Baum, Understanding Bioinformatics, 2008, Garland Science, Page 277.

A worked example of the UPGMA method of phylogenetic tree reconstruction for six sequences, A to F.



(A)

d_{ij}	A	B	C	D	E	F
A	–	6	8	1	2	6
B		–	8	6	6	4
C			–	8	8	8
D				–	2	6
E					–	6

(B)

d_{ij}	B	C	E	F	V
B	–	8	6	4	6
C		–	8	8	8
E			–	6	2
F				–	6

(C)

d_{ij}	B	C	F	W
B	–	8	4	6
C		–	8	8
F			–	6

(D)

d_{ij}	C	W	X
C	–	8	8
W		–	6

(E)

d_{ij}	C	Y
C	–	8

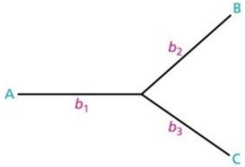
(A) The distance matrix showing that A and D are closest. The are selected in the first step to produce internal node V (in (B)). (B) The distance matrix including node V from which it can be deduced that V and E are closest, resulting in internal node W. (C,D) Subsequent steps defining nodes X, Y and Z.

This picture is from : Zvelebid&Baum, Understanding Bioinformatics, 2008, Garland Science, Page 279.

A worked example of the Fitch-Margoliash method.

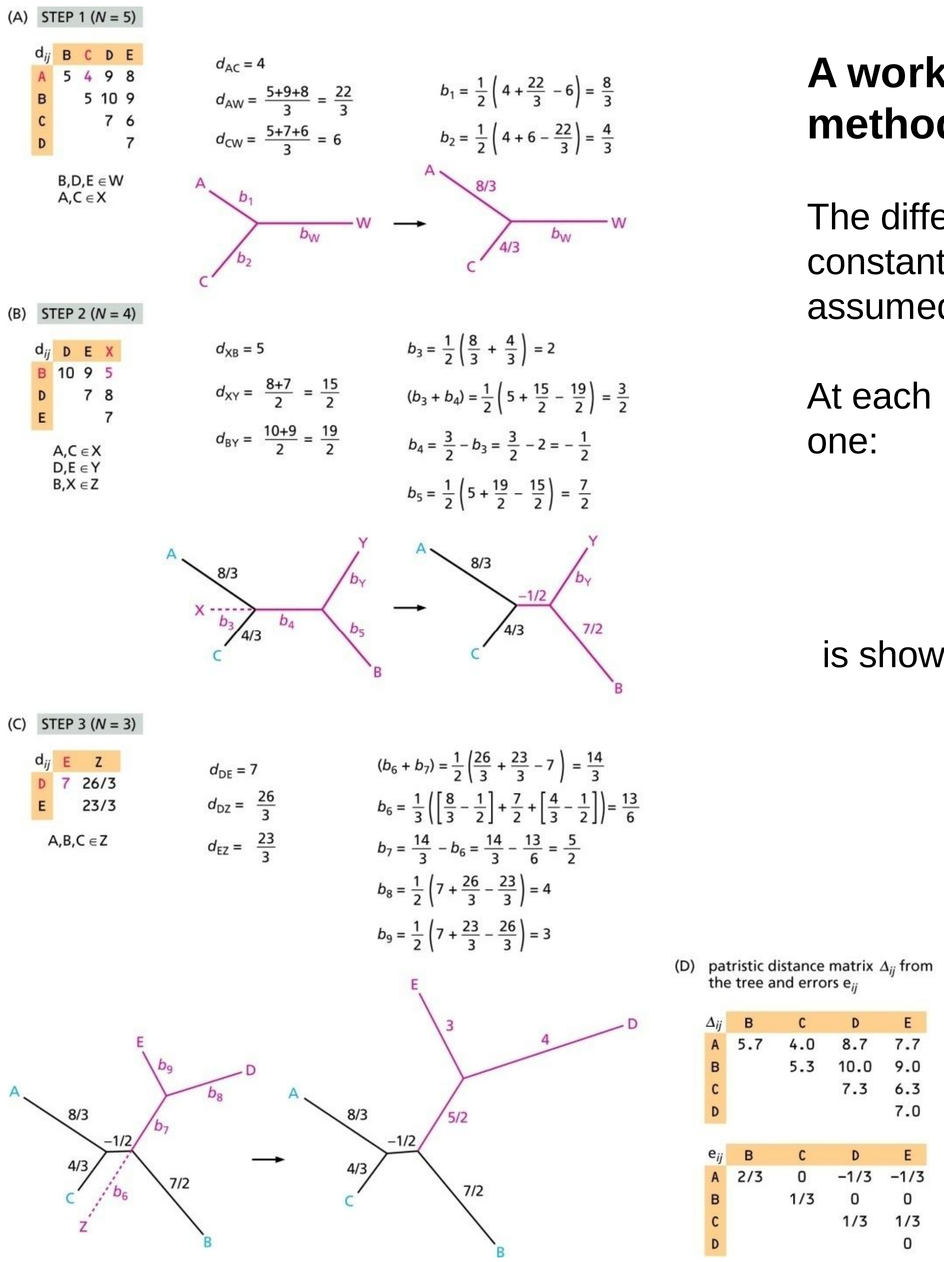
The difference to UPGMA is that the assumption of constant mutation rate is not made. Distances are assumed to be additive.

At each step the three-leaf tree that is equivalent to this one:



is shown in red on the left-hand tree.

In the first step the shortest distance is used to identify two clusters which are combined tp create the next internal node. A temporary cluster W is defined as all clusters except these two, and the distances are calculated from W to both A and C.



(A) STEP 1 (N = 5)

d_{ij}	B	C	D	E
A	5	4	9	8
B		5	10	9
C			7	6
D				7

B,D,E $\in$ W
A,C $\in$ X

$$d_{AC} = 4$$
$$d_{AW} = \frac{5+9+8}{3} = \frac{22}{3}$$
$$d_{CW} = \frac{5+7+6}{3} = 6$$
$$b_1 = \frac{1}{2}\left(4 + \frac{22}{3} - 6\right) = \frac{8}{3}$$
$$b_2 = \frac{1}{2}\left(4 + 6 - \frac{22}{3}\right) = \frac{4}{3}$$

(B) STEP 2 (N = 4)

d_{ij}	D	E	X
B	10	9	5
D		7	8
E			7

A,C $\in$ X
D,E $\in$ Y
B,X $\in$ Z

$$d_{XB} = 5$$
$$d_{XY} = \frac{8+7}{2} = \frac{15}{2}$$
$$d_{BY} = \frac{10+9}{2} = \frac{19}{2}$$
$$b_3 = \frac{1}{2}\left(\frac{8}{3} + \frac{4}{3}\right) = 2$$
$$(b_3 + b_4) = \frac{1}{2}\left(5 + \frac{15}{2} - \frac{19}{2}\right) = \frac{3}{2}$$
$$b_4 = \frac{3}{2} - b_3 = \frac{3}{2} - 2 = -\frac{1}{2}$$
$$b_5 = \frac{1}{2}\left(5 + \frac{19}{2} - \frac{15}{2}\right) = \frac{7}{2}$$

(C) STEP 3 (N = 3)

d_{ij}	E	Z
D	7	26/3
E		23/3

A,B,C $\in$ Z

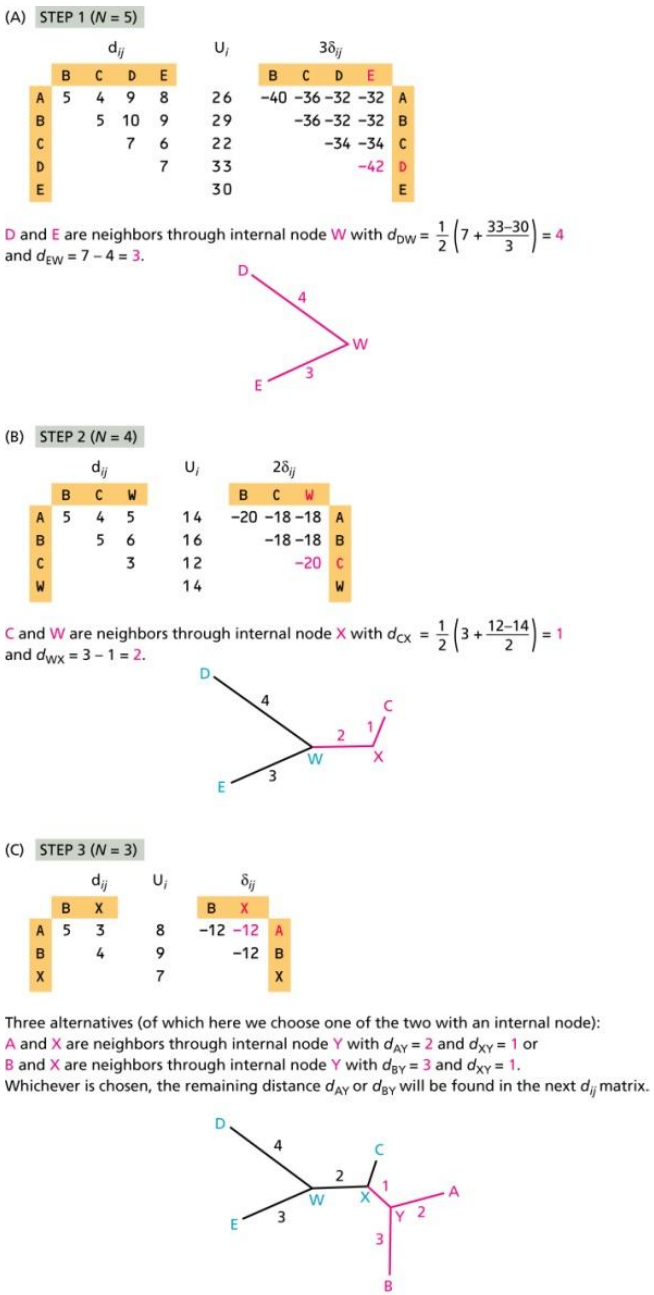
$$d_{DE} = 7$$
$$d_{DZ} = \frac{26}{3}$$
$$d_{EZ} = \frac{23}{3}$$
$$(b_6 + b_7) = \frac{1}{2}\left(\frac{26}{3} + \frac{23}{3} - 7\right) = \frac{14}{3}$$
$$b_6 = \frac{1}{3}\left(\left[\frac{8}{3} - \frac{1}{2}\right] + \frac{7}{2} + \left[\frac{4}{3} - \frac{1}{2}\right]\right) = \frac{13}{6}$$
$$b_7 = \frac{14}{3} - b_6 = \frac{14}{3} - \frac{13}{6} = \frac{5}{2}$$
$$b_8 = \frac{1}{2}\left(7 + \frac{26}{3} - \frac{23}{3}\right) = 4$$
$$b_9 = \frac{1}{2}\left(7 + \frac{23}{3} - \frac{26}{3}\right) = 3$$

(D) patristic distance matrix Δ_{ij} from the tree and errors e_{ij}

Δ_{ij}	B	C	D	E
A	5.7	4.0	8.7	7.7
B		5.3	10.0	9.0
C			7.3	6.3
D				7.0

e_{ij}	B	C	D	E
A	2/3	0	−1/3	−1/3
B		1/3	0	0
C			1/3	1/3
D				0

This picture is from : 277. Zvelebid&Baum, Understanding Bioinformatics, 2008, Garland Science, Page 281.



(A) STEP 1 ($N = 5$)

	d_{ij} B	C	D	E	U_i	$3\delta_{ij}$ B	C	D	E	
A	5	4	9	8	26	−40	−36	−32	−32	A
B		5	10	9	29		−36	−32	−32	B
C			7	6	22			−34	−34	C
D				7	33				−42	D
E					30					E

D and E are neighbors through internal node W with $d_{DW} = \frac{1}{2}\left(7 + \frac{33-30}{3}\right) = 4$ and $d_{EW} = 7 - 4 = 3$.

(B) STEP 2 ($N = 4$)

	d_{ij} B	C	W	U_i	$2\delta_{ij}$ B	C	W	
A	5	4	5	14	−20	−18	−18	A
B		5	6	16		−18	−18	B
C			3	12			−20	C
W				14				W

C and W are neighbors through internal node X with $d_{CX} = \frac{1}{2}\left(3 + \frac{12-14}{2}\right) = 1$ and $d_{WX} = 3 - 1 = 2$.

(C) STEP 3 ($N = 3$)

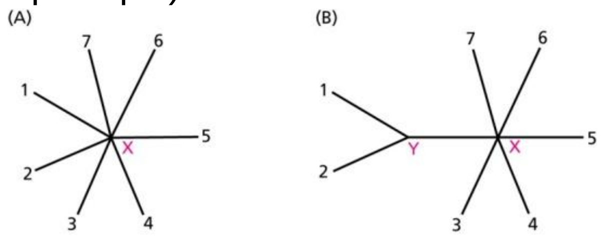
	d_{ij} B	X	U_i	δ_{ij} B	X	
A	5	3	8	−12	−12	A
B		4	9		−12	B
X			7			X

Three alternatives (of which here we choose one of the two with an internal node):
A and X are neighbors through internal node Y with $d_{AY} = 2$ and $d_{XY} = 1$ or
B and X are neighbors through internal node Y with $d_{BY} = 3$ and $d_{XY} = 1$.
Whichever is chosen, the remaining distance d_{AY} or d_{BY} will be found in the next d_{ij} matrix.

A worked example of the neighbor-joining method.

The distance matrix is the same as in the Fitch-Margoliash example.
At each step the distances are converted by using the algorithm which minimizes the total tree distance (the minimum evolution principle).

The first step:



(A) Star-tree in which all sequences are joined directly to a single internal node X with no internal branches.
(B) After sequences 1 and 2 have been identified as the first pair of nearest-neighbors, they are separated from node X by and internal node Y. The method calculates the brabch lengths from sequences 1 and 2 to node Y to complete the step.

This picture is from : Zvelebid&Baum, Understanding Bioinformatics, 2008, Garland Science, Page 284.

New Features

Excel and CSV Output

Update Notification

Interface Improvements

http://www.megasoftware.net/

Home-exercise 2

Download MEGA-software from http://www.megasoftware.net/

Perform an exercise with one of the example data-sets:

- D-loop_Vigilant
- these sequences are from human mitochondrial D-loop, from different populations and continents
- do neighbor-joining and UPGMA
  - the sequences are not protein coding, you have to notice this when the program asks!
  - familiarize yourself with different tree styles etc.

You are not supposed to submit this part of the exercise, submit the exercise given in the next page.

Although you are not supposed to submit answers to exercises until April 17., which is the deadline for all 5 home-works, it is advisable to perform the exercise (previous page) rather soon, because it facilitates following the lectures.

And leave this part (below) for which you have to submit answer, later:

Select 5 sequences from the D-loop_Vigilant dataset, for example one from Europe, one Asian, one New Guinean, two from Africa.
Now you have a ”similar” (N=5 => easy to handle...) dataset than those which were used in the examples for neighbor-joining, Fitch-Margoliash and UPGMA . Calculate and draw these phylogenies by hand, so you familiarize with the procedures (what is the computer program doing, actually....)