Please leave this exercise sheet on the desk and don't write on it.
It will be re-used for the other courses.

Bioinformatics 1 WiSe 2015/2016 - Phylogenetic Inference Practical

Instructor: Claudia Acquisti
Technical assistance: Anna-Lena Hallmann & Tabea Kischka

Reading a "back to the sea" story in molecular sequences: the evolution of marine mammals

General guidelines for molecular phylogenetic analysis using MEGA

- Download the three datasets in .fasta format. Make a directory and save the dataset files in .fasta format.
- Align the sequences using the two different programs available in MEGA (ClustalW and Muscle) and save the alignment in .meg format (Step A).
- Construct phylogenetic trees using two different methods: a (distance-based) Neighbor-joining approach, and a parsimony approach. Save the trees obtained as .pdf files (Steps B and C).
- Interpret and compare the results obtained using the different datasets and the different methods. Pay special attention to the different tree topologies and bootstraps values (Step D).

Datasets

Download the datasets from www.bioinformatics.uni-muenster.de/teaching/courses-2015/bioinf1/

- **Casein.fasta** K-casein exon 4 from 13 different mammalian species.
- **Haemoglobin.fasta** Concatenated protein sequence of haemoglobin -alpha and beta chains from 9 different mammalian species.
- **DNA_Concatenated.fasta** 10 Concatenated DNA sequences from six different genes from 10 different species (b-casein exon 7, K-casein exon 4, g-fibrinogen exon 2-4, g-fibrinogen introns 2-3, protamine P1 exons 1-2, protamine p1 intron1 + 5′-3' non coding regions).

These datasets have been compiled by John Gatesy and colleagues (Cladistics, 1999,15: 271-313).
Please note that **the datasets provided consist of protein and DNA sequences**. Depending on the file you are working with, **choose the MEGA software settings accordingly**.

Taxa represented in the datasets:

Artiodactyl taxa: Bovidae (sheep, cattle, bison, springbok, and antelopes), Cervidae (deer), Girafidae (giraffes), Tragulidae (chevrotains), Hippopotamidae (hippos), Camelidae (camels and llamas), Tayassuidae (peccaries), Suidae (pigs).
Cetacean taxa: Physeteridae (sperm whales), Delphinoidea (beluga whale, dolphins, and porpoises), Ziphiidae (beaked whales), Mysticeti (baleen whales).
Outgroup: (rhinos, horses, and guinea pigs).

(A) SEQUENCE ALIGNMENT

Step A.1: Start the program MEGA (Molecular Evolutionary Genetics Analysis freely available at www.megasoftware.net/, it will also be distributed on USB stick)

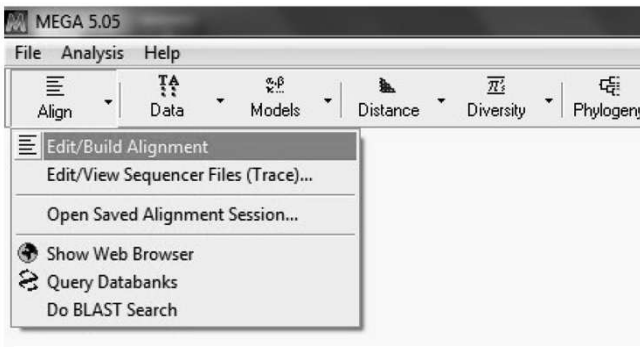

Figure 1: Step A.2

Step A.2: From the main MEGA window click on the "Align" tab, and then click the Edit/Build Alignment tab from the drop down menu, in the left hand corner of the screen.

Step A.3: A pop up window with an Alignment Editor will open up, click " Retrieve Sequence from files".

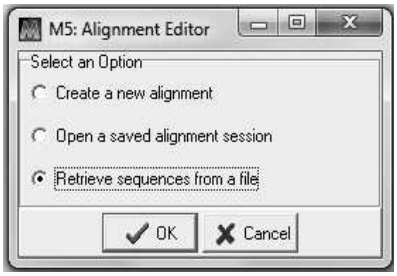

Figure 2: Step A.3

Step A.4: Choose the dataset from the directory/folder. Please be aware of the type of sequences you are working with (protein or DNA) and choose the settings accordingly.

Step A.5: *Alignment of the dataset.* Two different programs are available to perform the alignment of the data: ClustalW and Muscle. Perform both alignment types on your datasets using default parameters, and save the aligned data in .meg format. Export the alignment in .meg format (via the menu Data, Export alignment, and chose MEGA format). For each dataset you will have two different alignment files (e.g., Casein_ClustalW.meg, and Casein_Muscle.meg). For each dataset the comparison of the trees obtained from the two different alignments will help you to appreciate the relevance of the alignment process in reconstructing phylogenies.

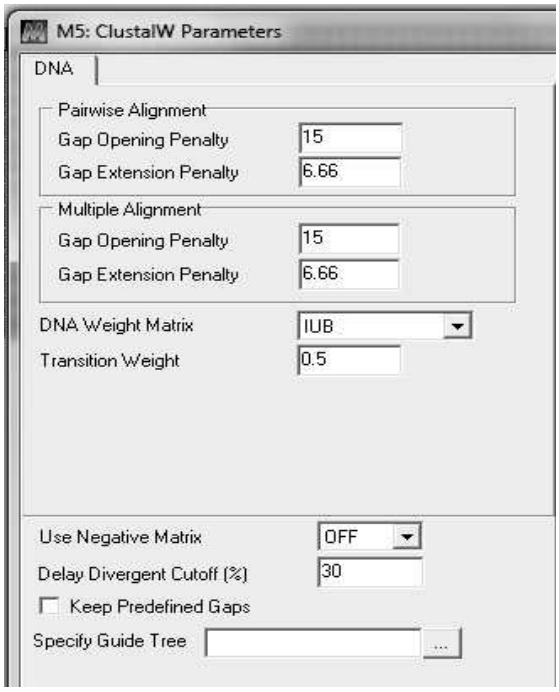

Figure 3: Step A.5, parameters for ClustalW

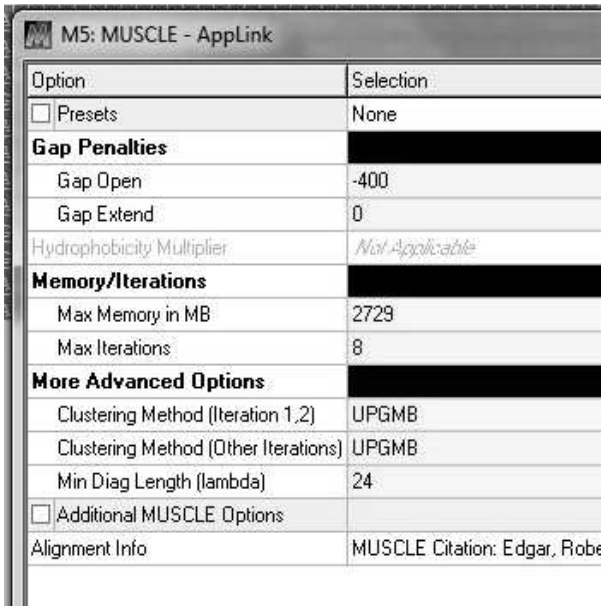

Figure 4: Step A.5, parameters for MUSCLE

(B) PHYLOGENETIC RECONSTRUCTION: NEIGHBOUR-JOINING APPROACH

Step B.1: Open the .meg files where you have stored your alignment in step A.5. Compute the pair-wise distances between each sequence pair in your dataset. Click the "Distance" tab from the main menu. Choose "Compute pairwise Distances".

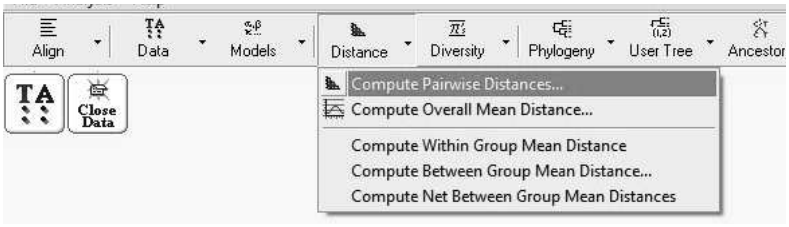

Figure 5: Step B.1

Step B.2: A window with several options for the calculation of pair-wise distances will open up. Click on the yellow tabs and choose the parameters given below. Calculate the pairwise distances using the Gamma correction. A distance matrix will be created.

Please be aware of the type of sequences you are working with (protein or DNA) and choose the settings accordingly. **A critical choice of the settings is your responsibility**. **This is a fundamentally important aspect of research!**

M5: Analysis Preferences

Options Summary

Option	Selection
Analysis	Distance Estimation
Scope	Pairs of taxa
Estimate Variance	
Variance Estimation Method	None
No. of Bootstrap Replications	*Not Applicable*
Substitution Model	
Substitutions Type	Amino acid
Genetic Code Table	Standard
Model/Method	Jones-Taylor-Thornton (JTT) model
Fixed Transition/Transversion Ratio	*Not Applicable*
Substitutions to Include	All
Rates and Patterns	
Rates among Sites	Gamma Distributed (G)
Gamma Parameter	1
Pattern among Lineages	Same (Homogeneous)
Data Subset to Use	
Gaps/Missing Data Treatment	Complete deletion
Site Coverage Cutoff (%)	*Not Applicable*
Select Codon Positions	*Not Applicable*

Figure 6: Step B.2, options for pair-wise distances

M5: Pairwise Distances (D:\Users\parijat\Desktop\dataset whippo\ali...

File Display Average Caption Help

	1	2	3	4	5	6	7	8	9
1. AHEM BOVIDAE									
2. AHEM CERVIDAE	0.10								
3. AHEM DELPHINOID	0.19	0.24							
4. AHEM PHYSETERID	0.18	0.19	0.13						
5. AHEM MYSTICETI	0.22	0.23	0.10	0.13					
6. AHEM HIPPOPOTA	0.11	0.08	0.21	0.16	0.20				
7. AHEM SUIDAE	0.15	0.13	0.25	0.20	0.26	0.12			
8. AHEM CAMELIDAE	0.16	0.18	0.22	0.20	0.24	0.14	0.15		
9. AHEM OUTGROUP	0.17	0.16	0.23	0.23	0.25	0.13	0.16	0.15	

[1,1] (AHEM BOVIDAE-AHEM BOVIDAE) / Amino: JTT matrix-based

Figure 7: Step B.2, pairwise-distances

Step B.3: *Create a distance-based Neighbor-Joining phylogenetic tree.* From the main menu click the menu "Phylogeny" and choose "Neighbor joining tree construction" from the drop down box. In the field "Test of phylogeny" set the number of bootstrap replications to 1000.

Settings for bootstrap are requested for nucleotide or proteins datasets. Think about which one is appropriate for each of your datasets and set up the substitution type accordingly (as nucleotide or amino acids).

Option	Selection
Analysis	Phylogeny Reconstruction
Scope	All Selected Taxa
Statistical Method	Neighbor-joining
Phylogeny Test	
Test of Phylogeny	Bootstrap method
No. of Bootstrap Replications	1000
Substitution Model	
Substitutions Type	Nucleotide
Genetic Code Table	*Not Applicable*
Model/Method	Maximum Composite Likelihood
Fixed Transition/Transversion Ratio	*Not Applicable*
Substitutions to Include	d: Transitions + Transversions
Rates and Patterns	
Rates among Sites	Uniform rates
Gamma Parameter	*Not Applicable*
Pattern among Lineages	Same (Homogeneous)
Data Subset to Use	
Gaps/Missing Data Treatment	Complete deletion
Site Coverage Cutoff (%)	*Not Applicable*
Select Codon Positions	☑ 1st ☑ 2nd ☑ 3rd ☑ Noncoding Sites

Figure 8: Step B.3

Step B.4 : Save the tree in PDF format by clicking on 'Save as PDF file' button in the 'Image' drop down menu of the TreeExplorer window.

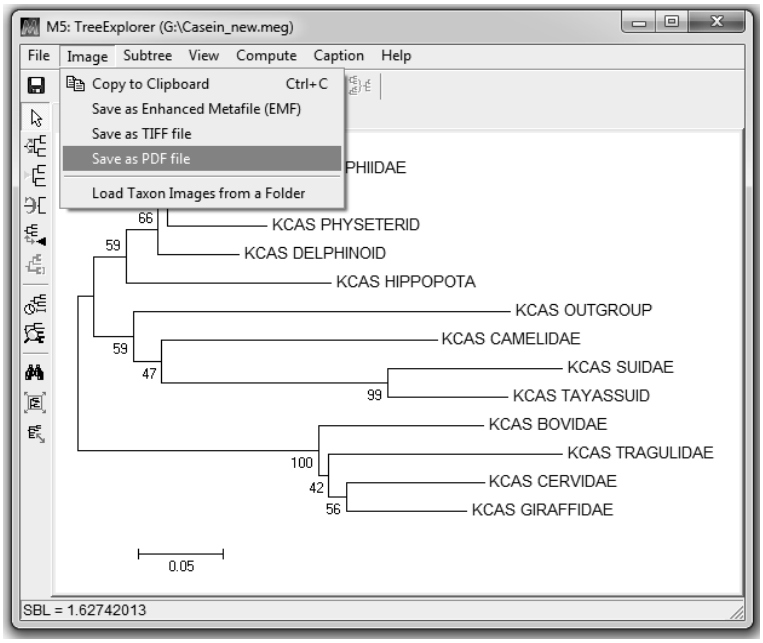

Figure 9: Step B.4

C) PHYLOGENETIC RECONSTRUCTION: PARSIMONY APPROACH

Step C.1: Construct a phylogenetic tree using a Parsimony method. Repeat step B.3 and this time choose parsimony method rather than the NJ method. Open the sequence data explorer by clicking the tab "TA" in the work space and calculate "Parsim-info-sites" in the Highlight menu. Set up bootstrap settings. Save the tree as you have done in step B.4.

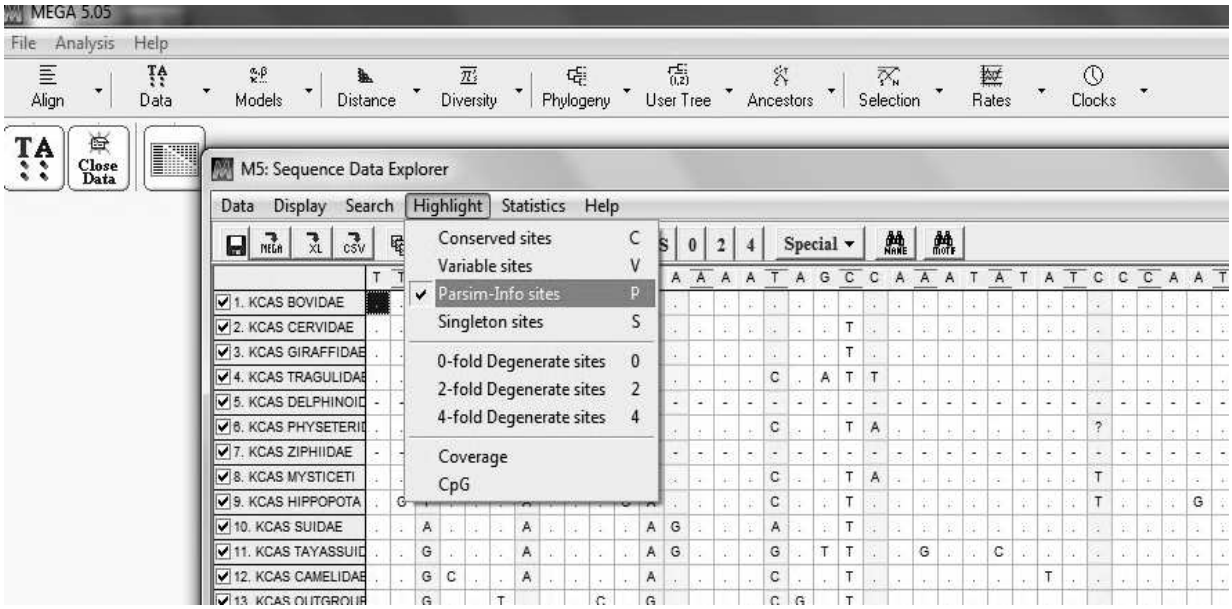

Figure 10: Step C.1, calculate "Parsim-info-sites"

BE PROUD OF THE WORK YOU HAVE DONE, AND TAKE THE TIME TO THINK CRITICALLY ABOUT THE RESULTS YOU HAVE PRODUCED!

(D) COMPARATIVE ANALYSIS OF THE RESULTS OBTAINED WITH DIFFERENT DATASETS AND METHODS

You have created 12 trees, using NJ and maximum parsimony approaches based on the three datasets, each aligned with two different alignment programs.

- Do all the datasets give the same tree topology?
- How do the bootstrap values vary between the different datasets?
- Can you see a relationship between the amount of information available in the datasets and the bootstrap values?
- Which is the terrestrial closest relative of cetaceans in the dataset?