Bioinformatics 1 WiSe 2012/2013 - Phylogenetic Inference Practical
Instructor: Claudia Acquisti; Technical assistance: Stefanie Henze and Parijat Tripathi

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 (Clustal W 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 http://www.compgen.uni-muenster.de/teaching/courses-2012/bioinf1

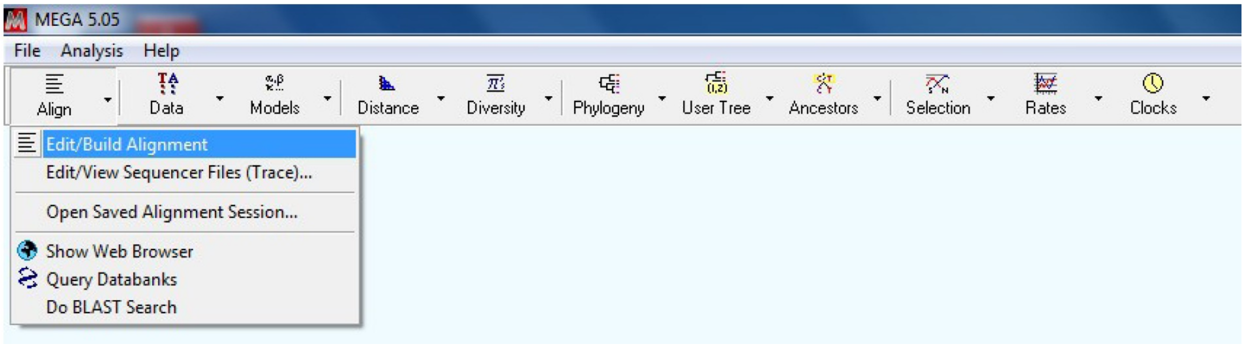
1. **Casein.fasta** K-casein exon 4 from 13 different mammalian species.
2. **Haemoglobin.fasta** Concatenated protein sequence of haemoglobin -alpha and beta chains from 9 different mammalian species.
3. **DNA_Concatenated.fasta** 10 Concatenated DNA sequences from six different genes from 10 different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

These datasets have been compiled by John Gatesy and colleagues (Cladistics, 1999,15: 271-313).
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).

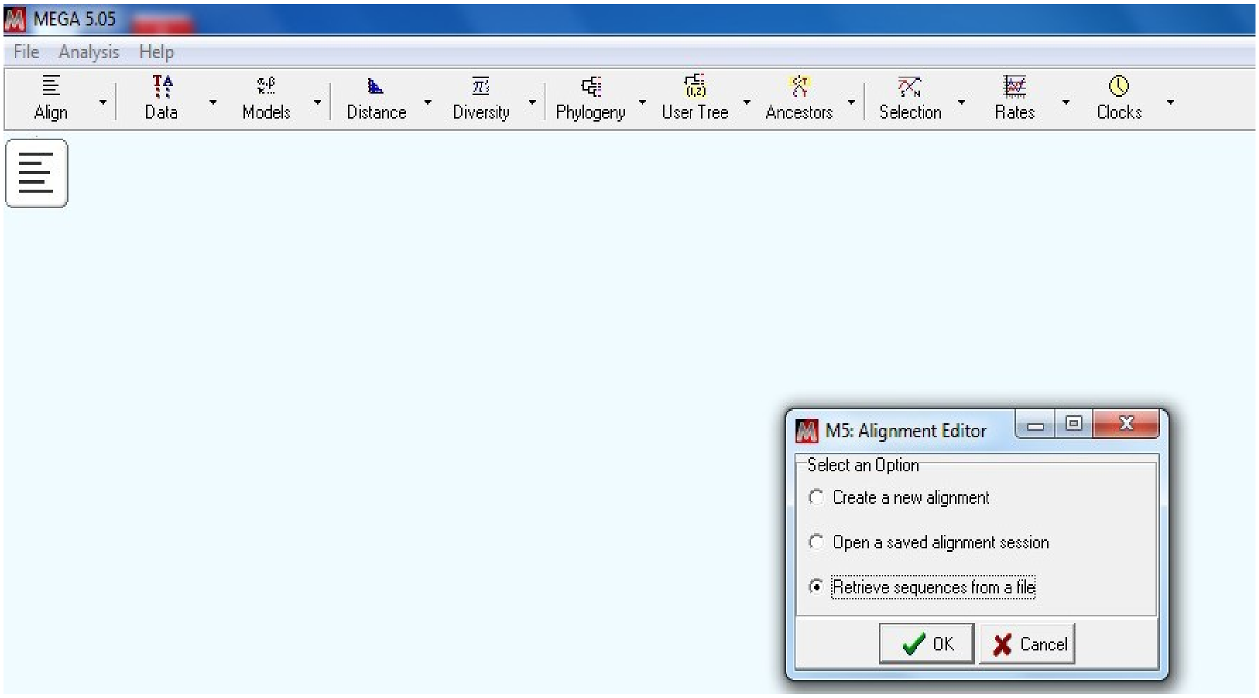
(A) SEQUENCE ALIGNMENT

Step A.1: Start the program MEGA (Molecular Evolutionary Genetics Analysis freely available at http://www.megasoftware.net/)

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



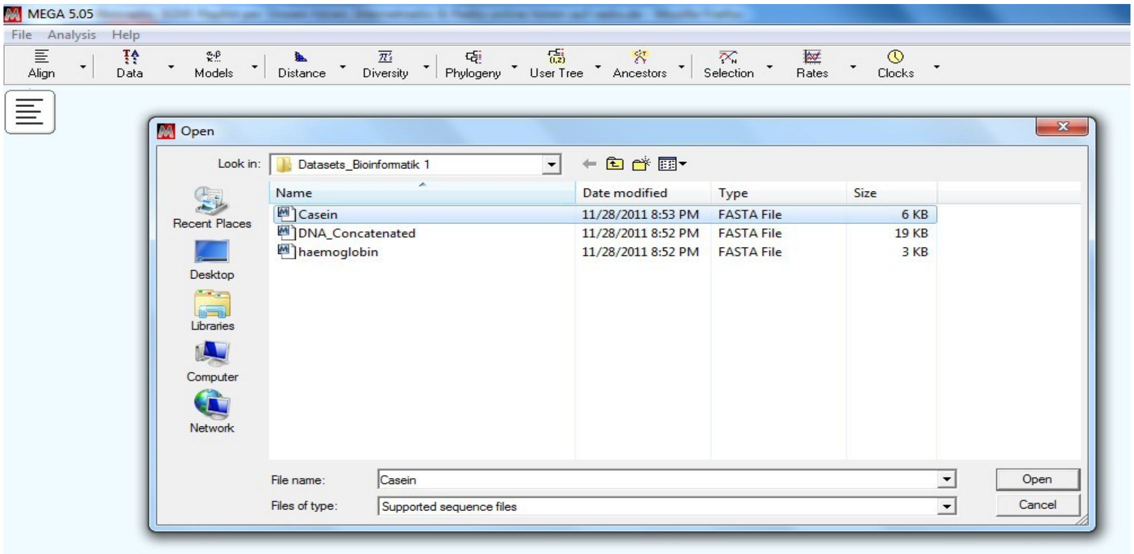
Step A.3: Once you have carried out the first step, a pop up window with an Alignment Editor will open up, click " Retrieve Sequence from files".



Select an Option
○ Create a new alignment
○ Open a saved alignment session
● Retrieve sequences from a file

\

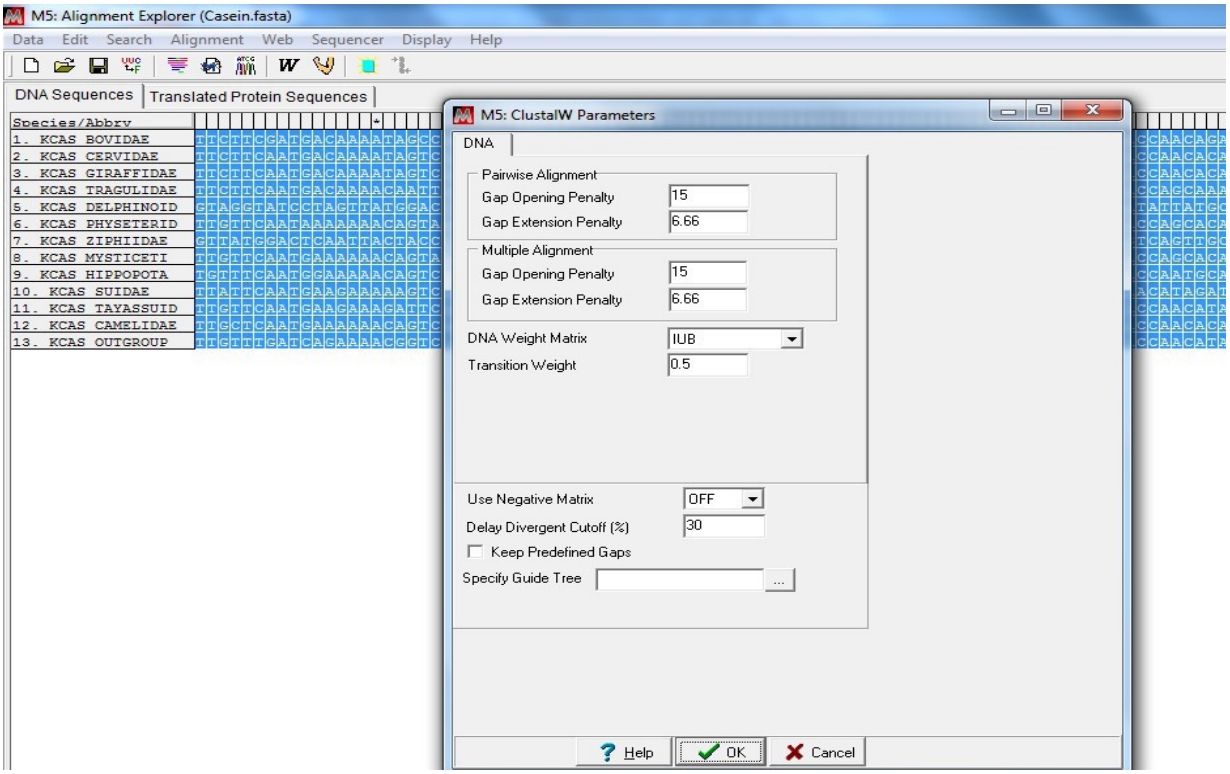
Step A.4: Choose the dataset of sequences from the directory/folder.



Name	Date modified	Type	Size
Casein	11/28/2011 8:53 PM	FASTA File	6 KB
DNA_Concatenated	11/28/2011 8:52 PM	FASTA File	19 KB
haemoglobin	11/28/2011 8:52 PM	FASTA File	3 KB

File name: Casein
Files of type: Supported sequence files

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 dataset using default parameters, and save the aligned data in .meg format. Export the alignment in .meg format (via the tab 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 later to understand the relevance of the alignment process in reconstructing phylogenies.



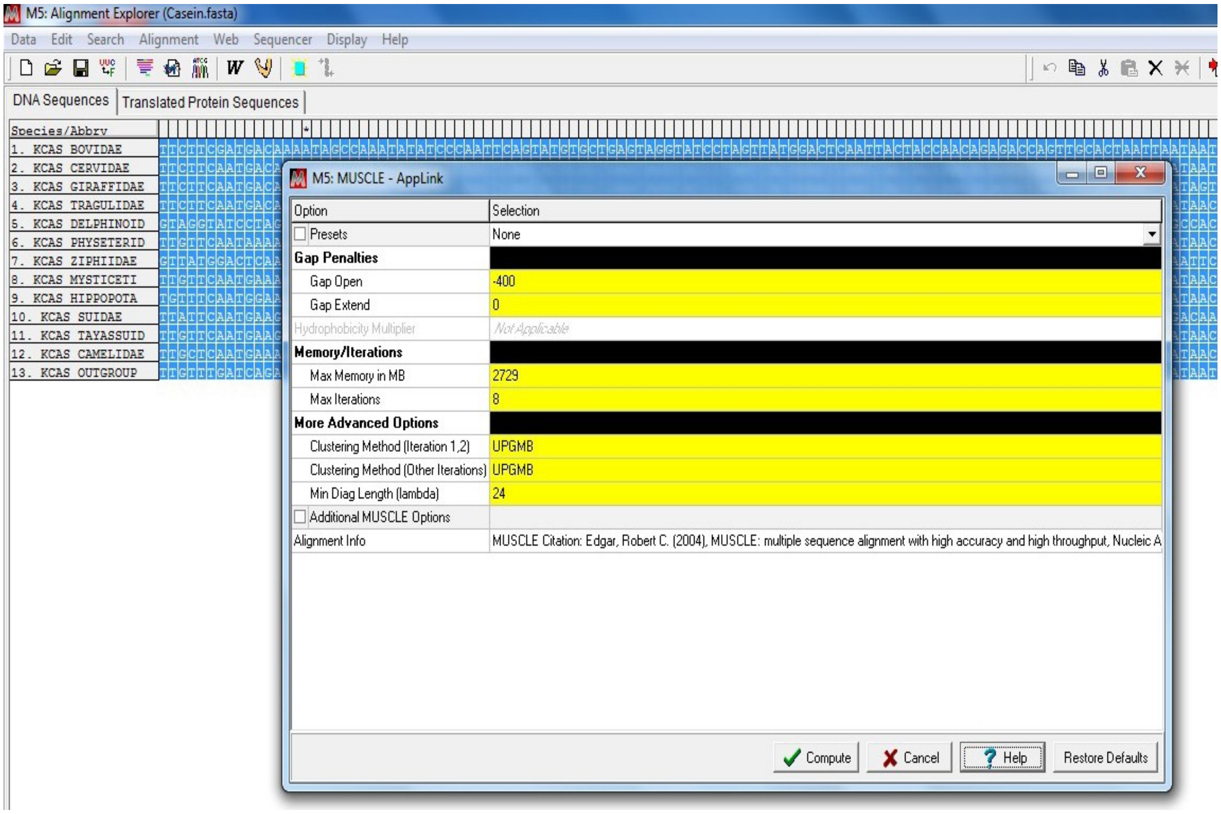
Species/Abbrv	
1. KCAS BOVIDAE	TTCTTCGATGACAAAATAGCC
2. KCAS CERVIDAE	TTCTTCAATGACAAAATAGTC
3. KCAS GIRAFFIDAE	TTCTTCAATGACAAAATAGTC
4. KCAS TRAGULIDAE	TTCTTCAATGACAAAACAATT
5. KCAS DELPHINOID	GTAGGTATCCTAGTTATGGAC
6. KCAS PHYSETERID	TTGTTCAATAAAAAAACAGTA
7. KCAS ZIPHIIDAE	GTTATGGACTCAATTACTACC
8. KCAS MYSTICETI	TTGTTCAATGAAAAAACAGTA
9. KCAS HIPPOPOTA	TGTTTCAATGGAAAAACAGTC
10. KCAS SUIDAE	TTATTCAATGAAGAAAAAGTC
11. KCAS TAYASSUID	TTGTTCAATGAAGAAAGATTC
12. KCAS CAMELIDAE	TTGCTCAATGAAAAAACAGTC
13. KCAS OUTGROUP	TTGTTTGATCAGAAAACGGTC

Pairwise Alignment
Gap Opening Penalty: 15
Gap Extension Penalty: 6.66

Multiple Alignment
Gap Opening Penalty: 15
Gap Extension Penalty: 6.66

DNA Weight Matrix: IUB
Transition Weight: 0.5

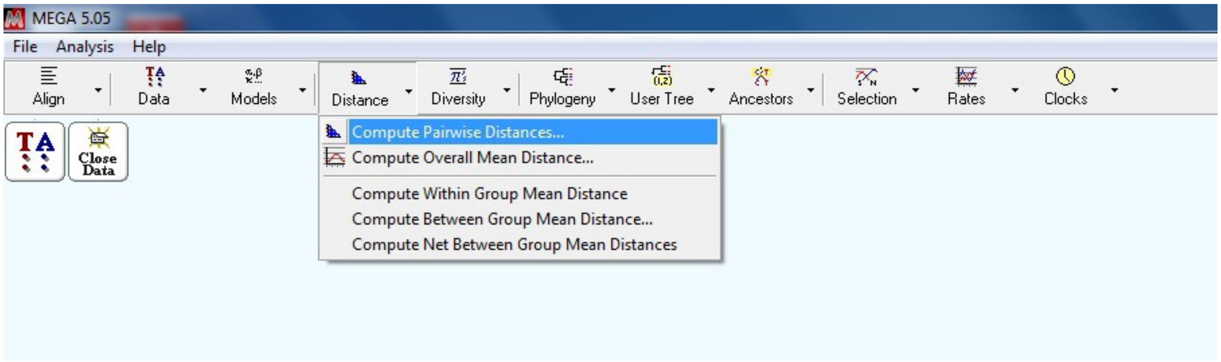
Use Negative Matrix: OFF
Delay Divergent Cutoff (%): 30
☐ Keep Predefined Gaps
Specify Guide Tree:



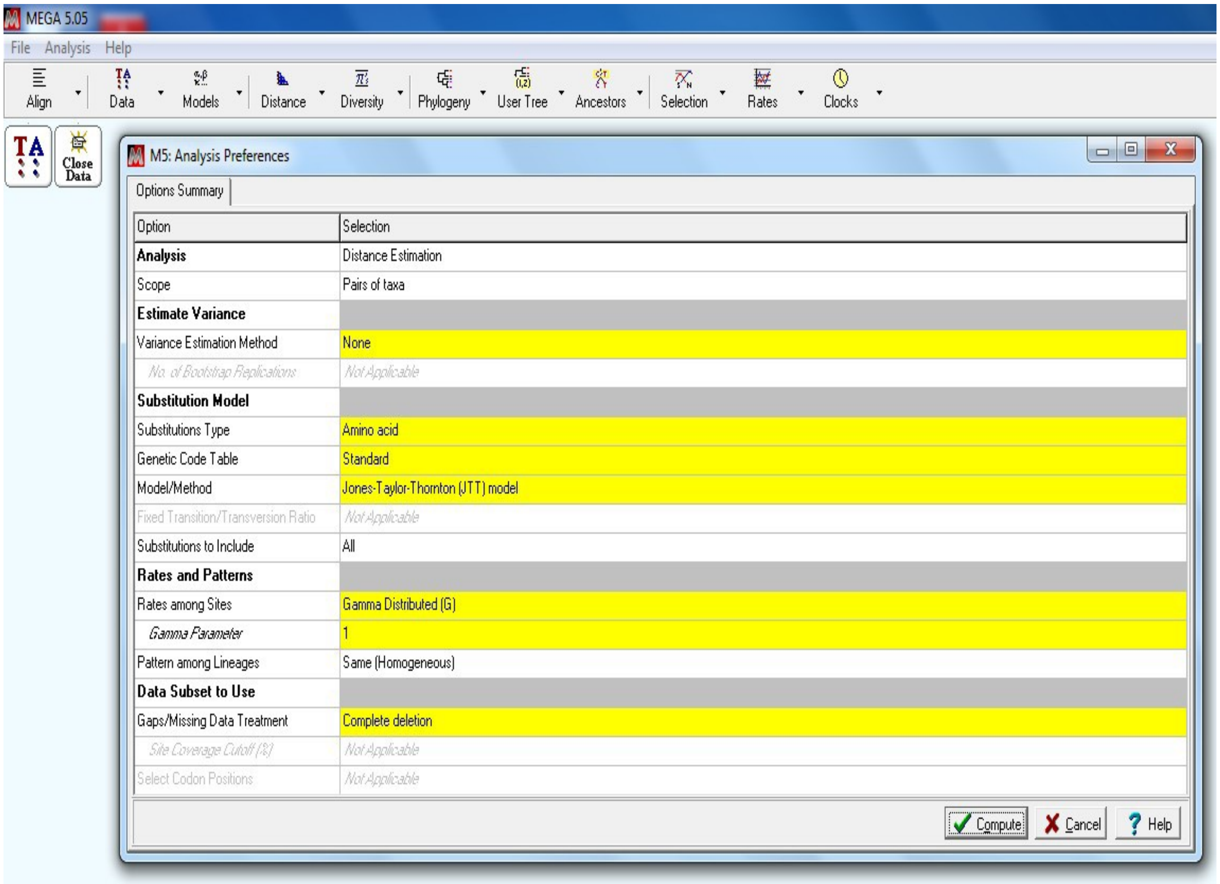
Option	Selection
☐ Presets	None
Gap Penalties	
Gap Open	-400
Gap Extend	0
Hydrophobicity Multiplier	Not Applicable
Memory/Iterations	
Max Memory in MB	2729
Max Iterations	8
More Advanced Options	
Clustering Method (Iteration 1,2)	UPGMB
Clustering Method (Other Iterations)	UPGMB
Min Diag Length (lambda)	24
☐ Additional MUSCLE Options	
Alignment Info	MUSCLE Citation: Edgar, Robert C. (2004), MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic A

(B) PHYLOGENETIC RECONSTRUCTION: NEIGHBOUR-JOINING APPROACH

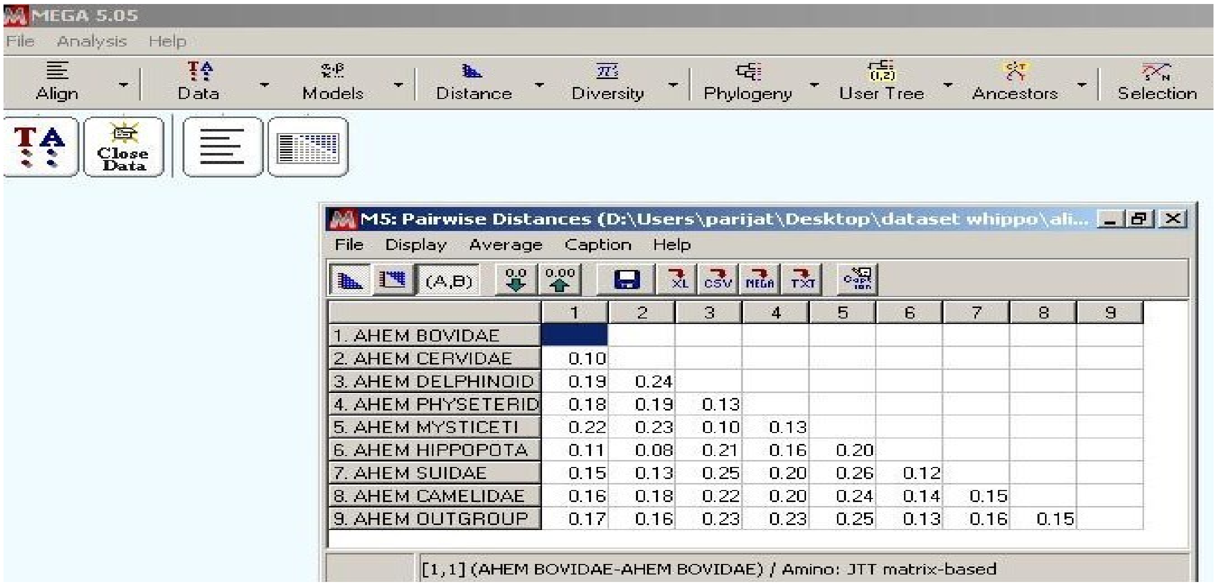
Step B.1: Open the .meg files where you have stored your alignment in step A.5. Compute the pair-wise distances between all the sequences in your dataset. Click the "**Distance**" tab from the main menu. Choose "Compute pairwise Distances".



Step B.2: A window with analysis preference will open up for the calculation of pair-wise distances. Click on the yellow color tab and choose parameters given below. Calculate the pairwise using the gamma correction. Once the distance matrix is created, save it.



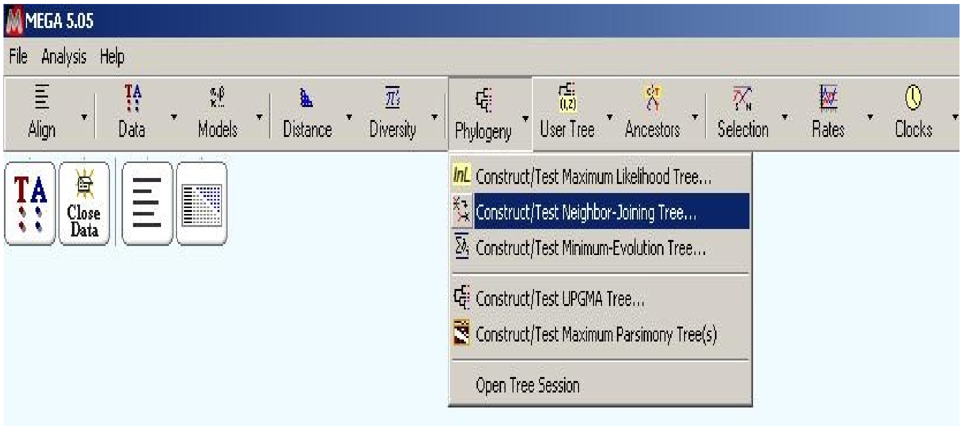
Option	Selection
Analysis	Distance Estimation
Scope	Pairs of taxa
Estimate Variance	
Variance Estimation Method	None
No. of Bootstrap Replications	*Not Applicable*
Substitution Model	
Substitutions Type	Amino acid
Genetic Code Table	Standard
Model/Method	Jones-Taylor-Thornton (JTT) model
Fixed Transition/Transversion 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*

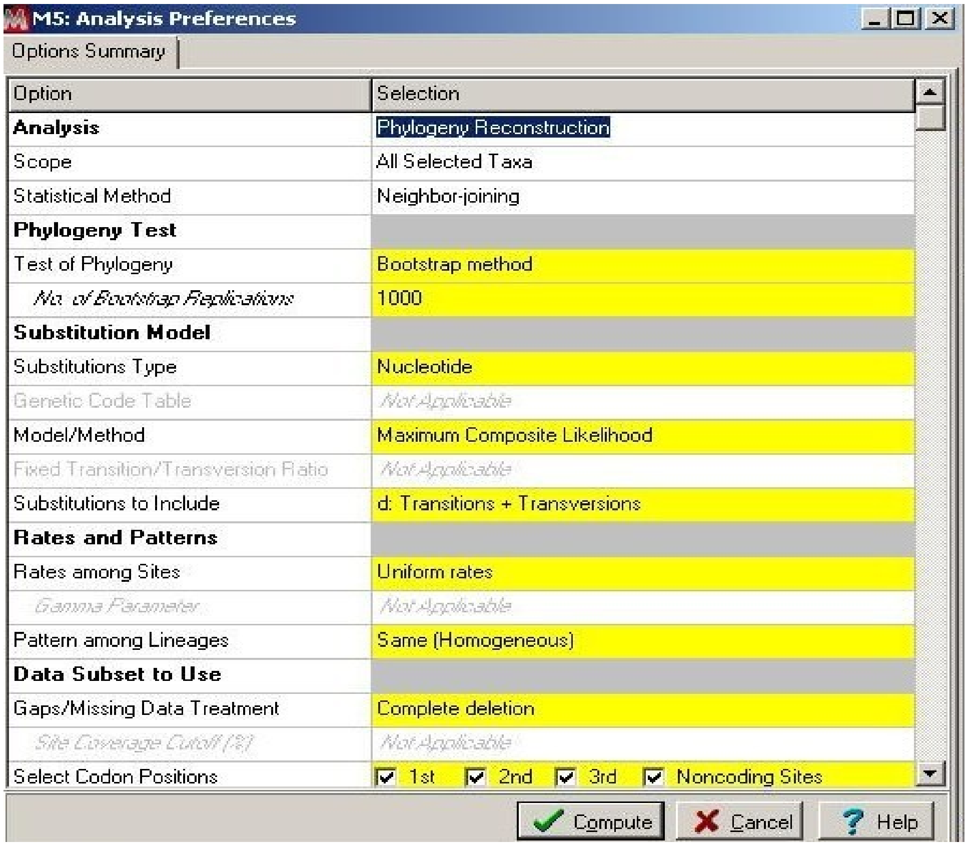


	1	2	3	4	5	6	7	8	9
1. AHEM BOVIDAE									
2. AHEM CERVIDAE	0.10								
3. AHEM DELPHINOID	0.19	0.24							
4. AHEM PHYSETERID	0.18	0.19	0.13						
5. AHEM MYSTICETI	0.22	0.23	0.10	0.13					
6. AHEM HIPPOPOTA	0.11	0.08	0.21	0.16	0.20				
7. AHEM SUIDAE	0.15	0.13	0.25	0.20	0.26	0.12			
8. AHEM CAMELIDAE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

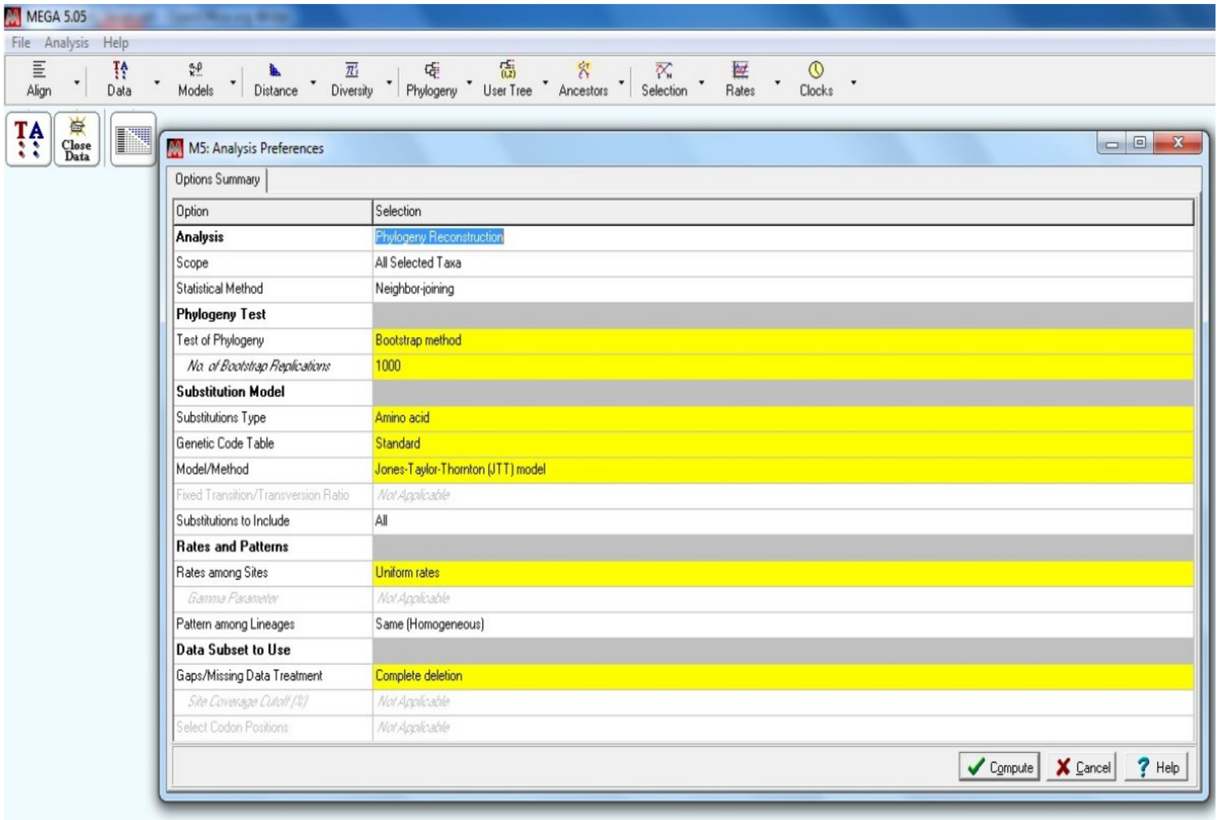
[1,1] (AHEM BOVIDAE-AHEM BOVIDAE) / Amino: JTT matrix-based

Step B.3: Create a distance-based Neighbor-Joining phylogenetic tree: Click the tab **Phylogeny** from the main menu 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.



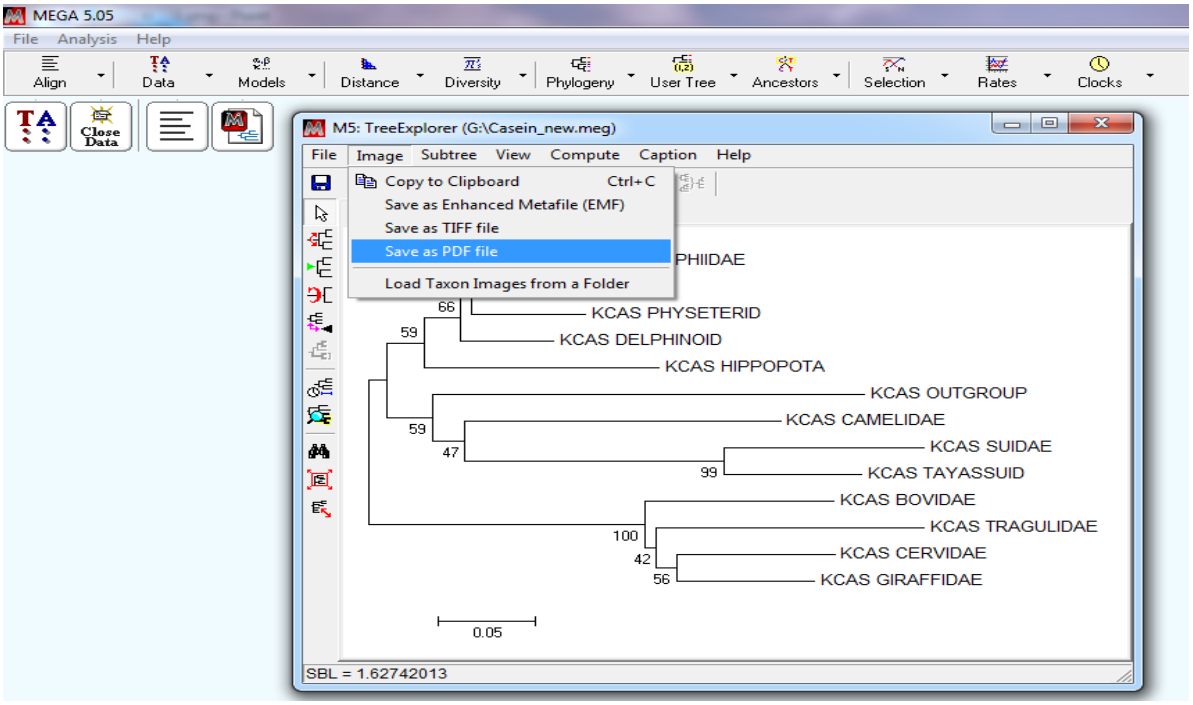


Option	Selection
Analysis	Phylogeny Reconstruction
Scope	All Selected Taxa
Statistical Method	Neighbor-joining
Phylogeny Test	
Test of Phylogeny	Bootstrap method
No. of Bootstrap Replications	1000
Substitution Model	
Substitutions Type	Nucleotide
Genetic Code Table	*Not Applicable*
Model/Method	Maximum 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



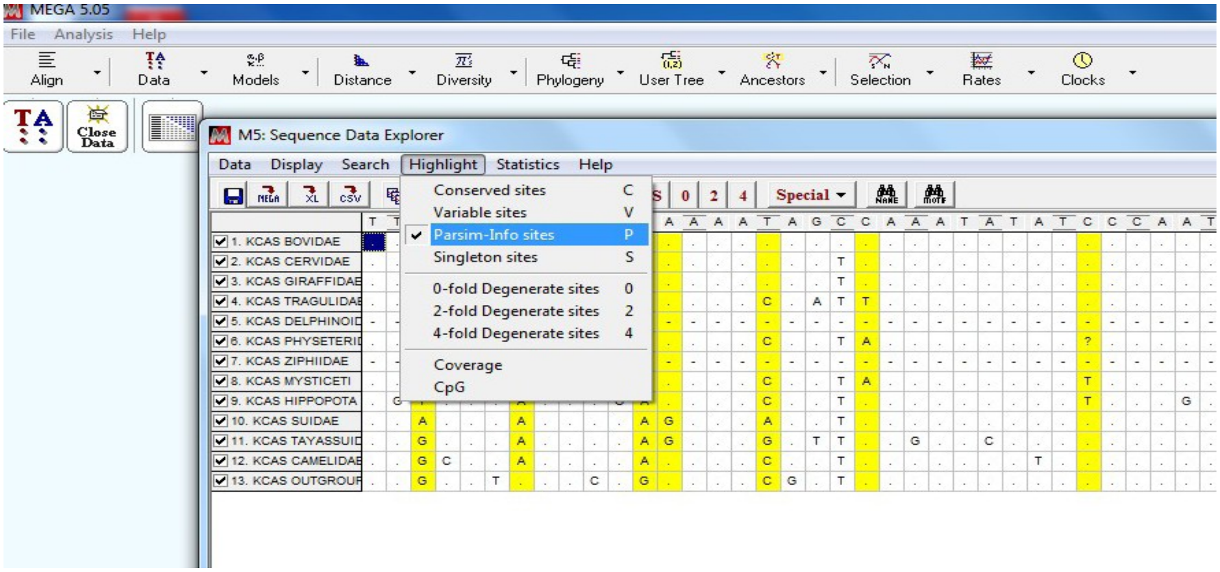
Option	Selection
Analysis	Phylogeny Reconstruction
Scope	All Selected Taxa
Statistical Method	Neighbor-joining
Phylogeny Test	
Test of Phylogeny	Bootstrap method
No. of Bootstrap Replications	1000
Substitution Model	
Substitutions Type	Amino acid
Genetic Code Table	Standard
Model/Method	Jones-Taylor-Thornton (JTT) model
Fixed Transition/Transversion Ratio	*Not Applicable*
Substitutions to Include	All
Rates and Patterns	
Rates among Sites	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	*Not Applicable*

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



C) PHYLOGENETIC RECONSTRUCTION: PARSIMONY APPROACH

Step C.1: Construct a phylogenetic tree using a **parsimony method. Repeat step B.3** This time choose parsimony method for construction. Open the Sequence data explorer just by clicking the tab " TA" in the work space and calculate " Parsim-info-sites", by clicking the **highlight** section. Save the tree as you have done in step B.4 .



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

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?