EXERCISE 1

ALIGNMENT USING CLUSTAL X

1. Download the following set of sequences from in the FASTA format: `U90045, U90052, U90047, U90048, U90055, U90053, U90057, U90049, X03349, U90058, U90054, U90059, U90062, U90063, U90046, U90050, U90056, U90060, U90051, U90061, U90064`

2. Edit sequences so that it starts with the GenBank number or any identifier that you seem fit. For example:

```
>U13098 Phocoena phocoena mitochondrion 12S rRNA gene
TAAACCTAAATAGTCCTAAAACAAGACTATTCGCCAGAGTACTATCGGCAACAG
CC
```

3. Open ClustalX

4. Load the FASTA file: **file->load sequences**. The window will look like the one in Figure 1.1

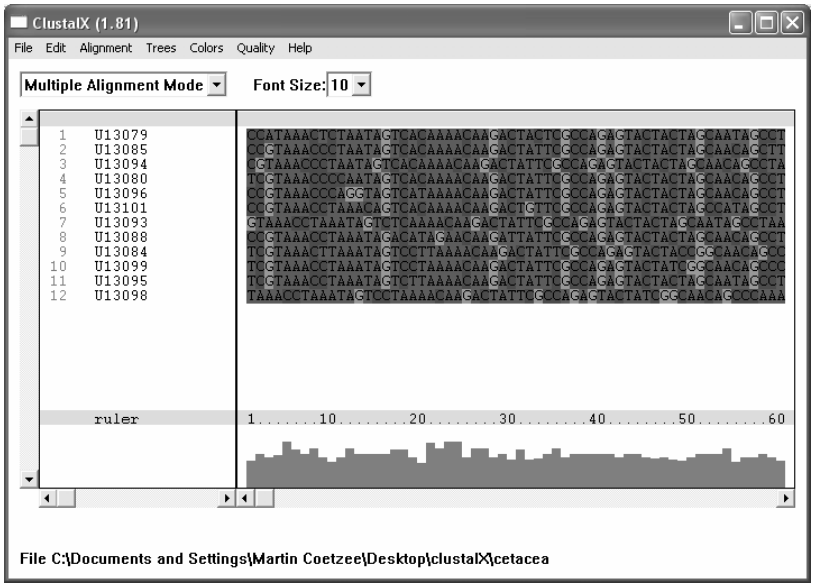


Figure 1.1

5. Set the alignment parameters (Fig. 1.2): **alignment ->alignment parameters→multiple alignment parameters**

6.

- **Gap opening** -> decreasing the values will allow more gaps, thus fewer mismatches.

- Keep the default values for the moment, you can change them and determine the differences. Figure 1.2

7. Specify the output format (Fig 1.3): **alignment -> output format options**

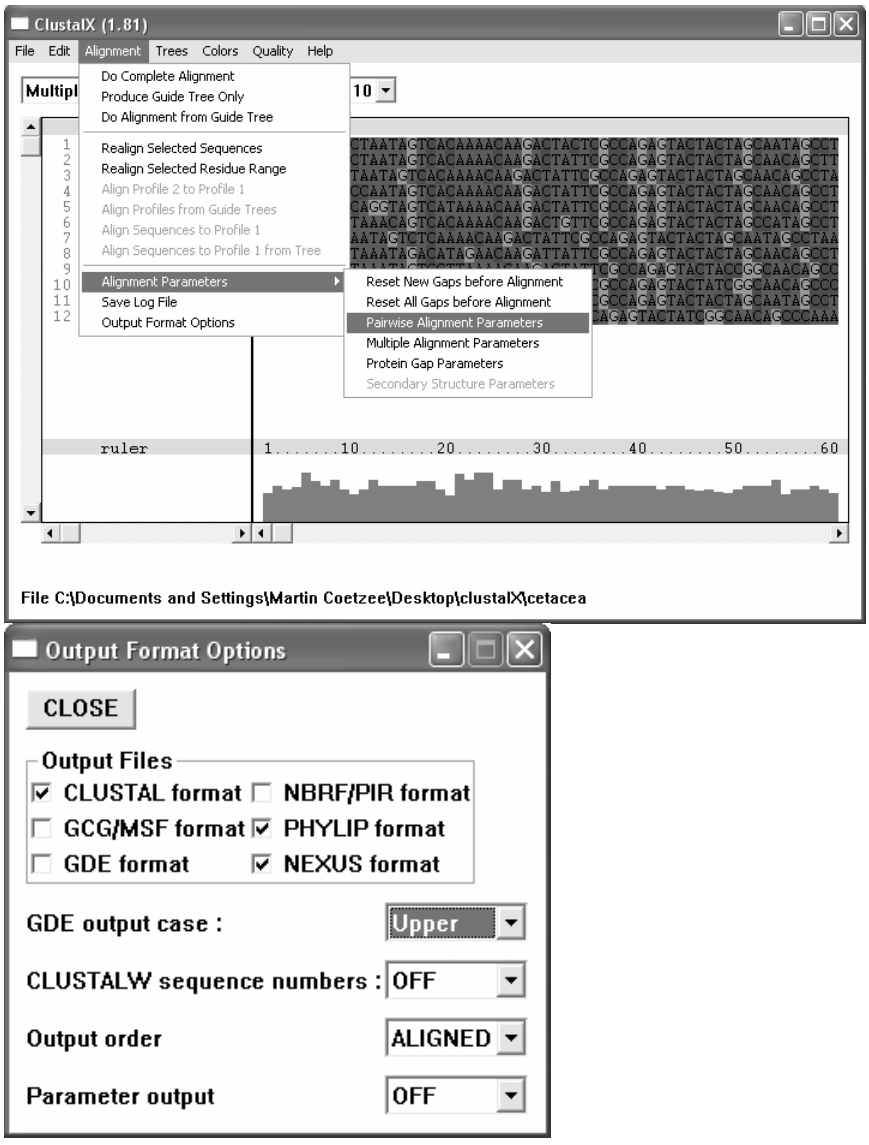


Figure 1.3

- Choose **Clustal format, Phylip format** and **Nexus format**

- Change **GDE output** case to **Upper**

5. Align the sequences: **alignment -> do complete alignment -> align**

6. You will see aligned sequences (Fig. 1.4) or your taxa upon completion. Figure 1.4

7. You can realign certain areas with apparent low similarity (e.g. blocked area in Figure 1.5)

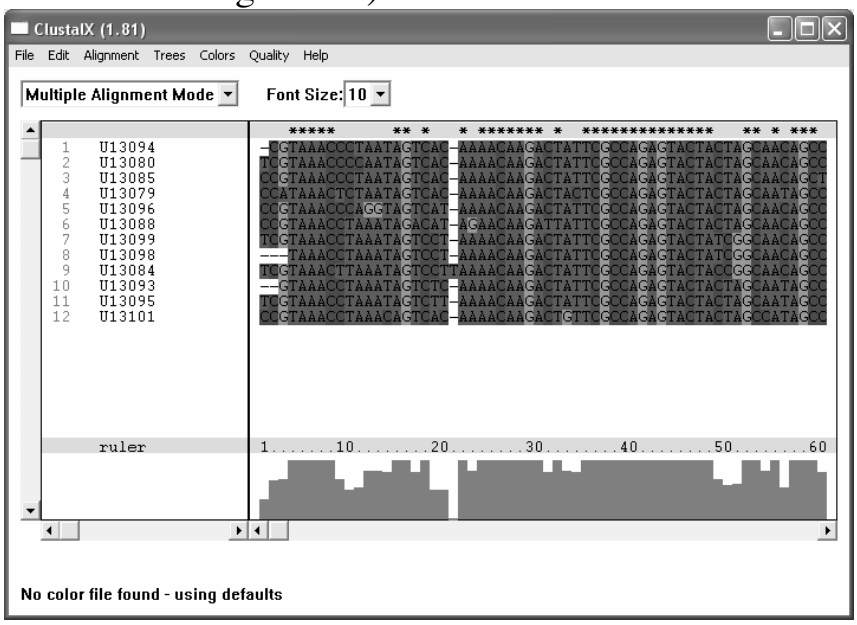


Fig. 1.4

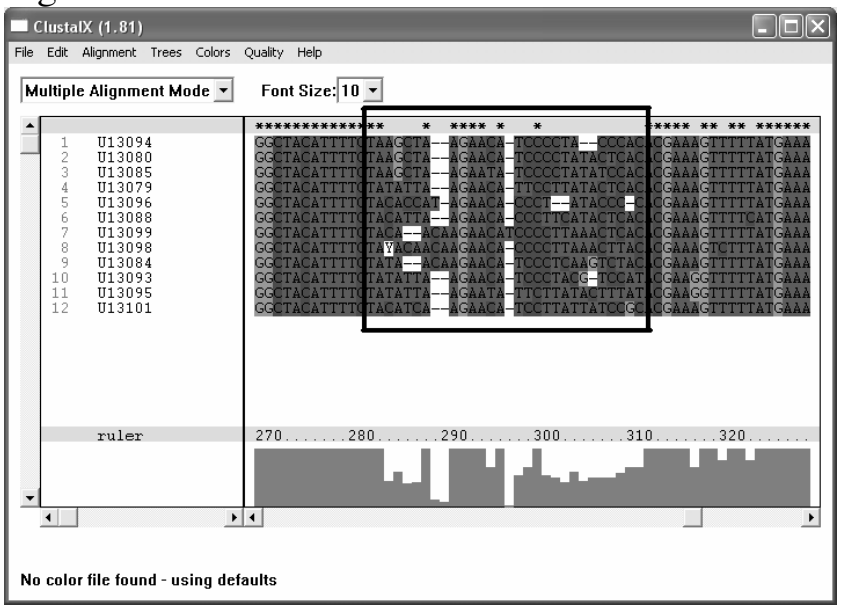


Fig. 1.5

10. Realign a sequence area:

- Place the pointer on the starting position and drag it over the area that you want to realign.

- Then choose: **alignment -> realign selected residue range -> align**.

11. Open the files in BioEdit (Fig. 1.6) and adjust miss-aligned regions.

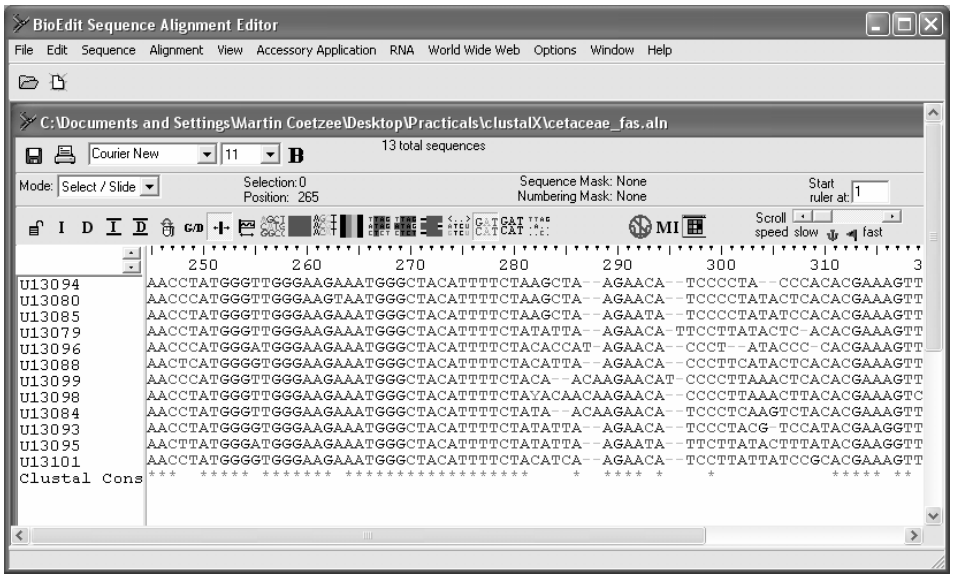


Figure 1.6

OUTPUT FILES:

8. *.aln = ClustalX alignment file

9. *.nxs = Nexus file

10. *.dnd = tree file for TreeView

NOTE: YOU CAN ALSO DO THIS IN MEGA. THE PROCEDURE WILL BE DEMONSTRATED

EXERCISE 2

ALIGNMENT USING MAFFT

11. Use the same sequence data used in Exercise 1 and re-align in MAFFT.

12. Make sure that the files are in FASTA format.

13. You can either use the web-interface or the program available on either the NBN server or other BioInformatics servers to align the sequence data.

14. **MAFT on the WWW:**

    The web-interface is available at:
    http://mafft.cbrc.jp/alignment/server/index.html

    Paste sequence or load the FASTA file.

    Choose one of the search strategies.

15. Compare the alignment in MAFFT with that obtained from CLUSTAL X.

EXERCISE 3

BASE FREQUENCY BIAS

CASE STUDY: ELONGATION FACTOR-1α (ARTHROPODS) (from Regier and Shultz 1997 MolBiolEvol 14: 902-913)

16. Obtain data from GenBank (http://www.ncbi.nlm.nih.gov) using the following accession numbers: `U90045, U90052, U90047, U90048, U90055, U90053, U90057,U90049, X03349, U90058, U90054, U90059, U90062, U90063,U90046, U90050, U90056, U90060, U90051, U90061, U90064`

17. Download the files in GenBank format.

18. Open the file in WordPad and save it as a "*.txt" file

19. Use DAMBE for the analysis:

    4.1. Load the file in DABME's memory buffer: **file-> open standard sequence file -> CDS**

    4.2. The following window will be displayed (Fig. 3.1):

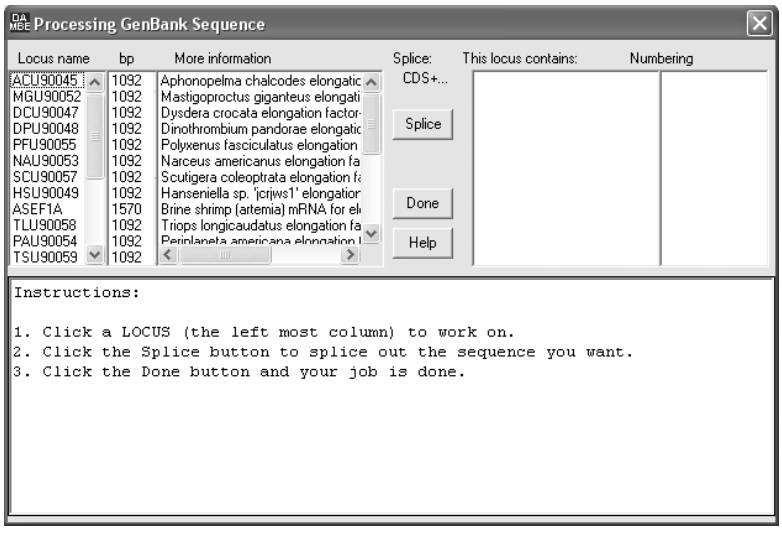


Fig. 3.1

4.3. Splice the introns out: click on the locus and the click on **Splice**.

4.4. After this, click done and you will be asked what kind of sequence it is, click on **protein-coding nucl. sequence**.

4.5. Save your sequences using the fasta format.

4.6. Translate sequences to amino acid sequences,

4.6.1. **Click sequence -> Work on amino acid sequences**.

4.7. Align amino acid sequences: **Alignment -> Align sequences using Clustalw**

4.8. Save your amino acid sequence alignment.

4.9. We are now going to align the nucleotide sequences against the aligned amino acid sequence:

4.9.1. Click **Alignment ->Align nucl sequences to aligned AA sequences in buffer.**

4.9.2. A dialog box will open asking the nucleotide sequence file, load the nucleotide file with CDS spliced out that you have saved. 4.10. Data analysis:

4.10.1. **Click Sequence | Work on codon position 3**.

4.10.2. **Click Seq. Analysis | Nucleotide Frequencies -> add all -> GO.**

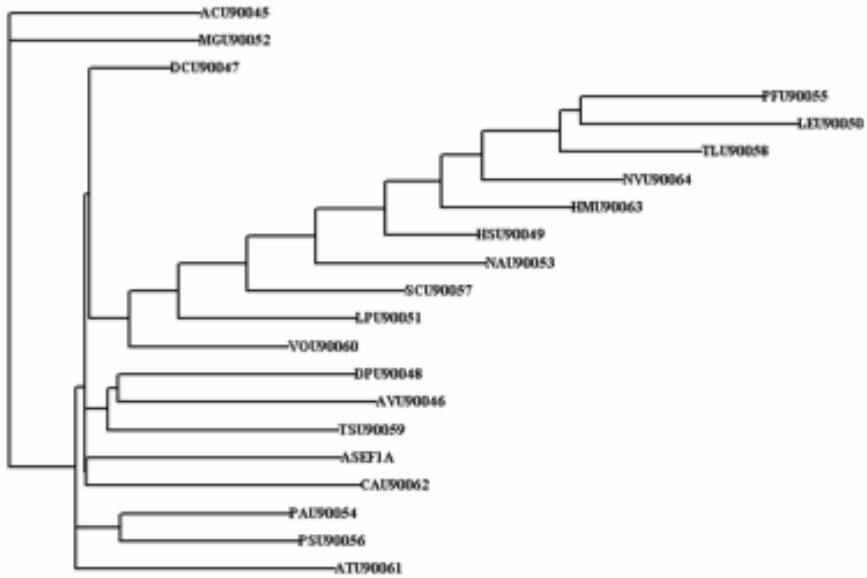
4.10.3. A simple phylogenetic analysis can be performed:

4.10.3.1. **Click Phylogenetics | Distance Methods | Nucleotide**. This is more or less what you should see:

Details of frequency distribution:

SeqName	A	C	G	T
ACU90045	0.2912	0.2170	0.1593	0.3324
TLU90058	0.0962	0.4835	0.2088	0.2115
PAU90054	0.1566	0.2308	0.2225	0.3901
TSU90059	0.1978	0.3022	0.1621	0.3379
CAU90062	0.2149	0.2479	0.2287	0.3085
HMU90063	0.1157	0.3140	0.2397	0.3306
AVU90046	0.2687	0.2604	0.1025	0.3684
LEU90050	0.0716	0.4683	0.3140	0.1460
PSU90056	0.2424	0.1736	0.2094	0.3747
VOU90060	0.1456	0.2885	0.2527	0.3132
LPU90051	0.1740	0.2155	0.2072	0.4033
MGU90052	0.2527	0.1566	0.1648	0.4258
ATU90061	0.2241	0.2661	0.1793	0.3305
NVU90064	0.1077	0.3343	0.2403	0.3177
DCU90047	0.1401	0.3407	0.2390	0.2802
DPU90048	0.2775	0.2665	0.2198	0.2363
PFU90055	0.1083	0.4389	0.2917	0.1611
NAU90053	0.1547	0.3232	0.2431	0.2790
SCU90057	0.1846	0.3003	0.2121	0.3030
HSU90049	0.1538	0.3489	0.2170	0.2802
ASEF1A	0.2078	0.2771	0.1645	0.3506

And the phylogenetic tree:



EXERCISE 4

TRANSITION / TRANSVERSION RATIO AND SATURATION

CASE STUDY: ELONGATION FACTOR-1α (ARTHROPODS) (from Regier and Shultz 1997 MolBiolEvol 14: 902-913)

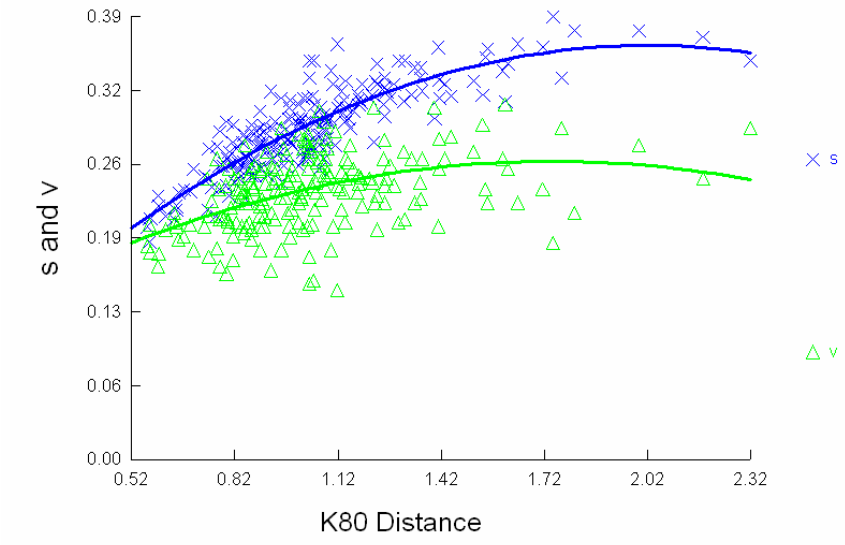
1. Load the aligned nucleotide file that you have used for the codon bias analysis.
2. We are only going to work on the third position of the codon:

   Sequences | Work on codon position 3.

3. To analysis the data:
   a) **Seq. Anlysis | Nucleotide substitution pattern | Detailed output**.
   b) **Add all -> GO**

4. Click YES if it asks you if you want a graph.

You have succeeded if you got this from the analysis:



S = transition and V = transversion

- What is the ts / tv ratio?
- Do you think there is saturation?

EXERCISE 5

DISTANCE ANALYSIS

CASE STUDY: PHYLOGENETICS OF CETACEANS(FROM MESSENGER SL AND MCGUIRE JA SYST.BIOL 47: 90- 124 1998)

Download the following 12S mt rRNA sequence from GenBank:

U13079, U13085, U13094, U13080, U13096, U13101, U13093, U13088, U13084, U13099, U13095, U13098, U13083, U13081

1. Download the files in FASTA format.

2. Edit the sequences – they must start with the species names.

3. Align your sequences in MEGA version

3. 3.1. Open MEGA

3.2. **Alignment | Alignment explorer/Clustal w -> Retrieve sequences from a file**

3.3. Open your fasta file

3.4. A window with your sequences will be displayed (Fig. 6.1)

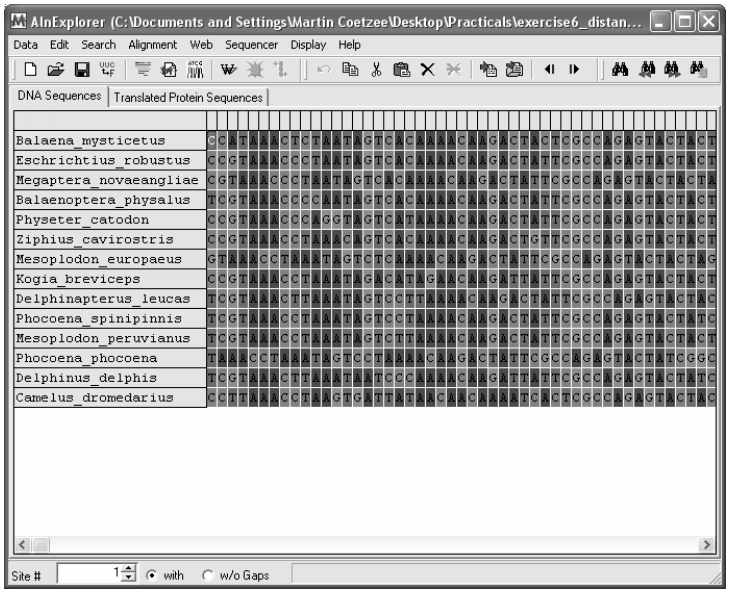


Fig. 6.1

3.5. Align | Align by ClustalW

3.6. Optimize your alignment in the editor

3.7.1. **Web | Query GenBank** -> ENTREZ will be loaded

3.7.2. Search for **`Tursiops truncates mitochondrion 12S`**

3.7.3. Select **`U13100`**

3.7.4. Display FASTA format

3.7.5. Click on **+ Add to alignment**

3.7.6. Do alignment again

4. **Save file: Data | Export | MEGA** (follow the instructions and supply answers)
   This will result in a window displaying your aligned sequences (Fig. 6.2)

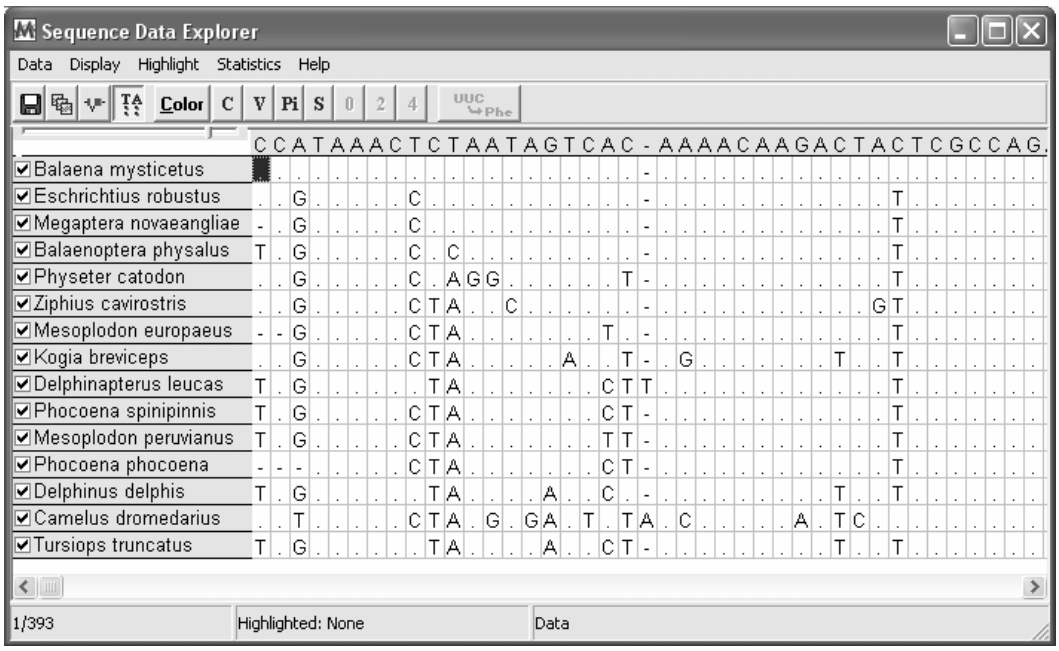


Fig. 6.2

5. You can get all kinds of statistics from the data using this window, play with it.
6. Now for the phylogenetic analyses. First do a simple NJ analysis with bootstrap (remember we do not know the correct substation

model yet):

7.1. **Phylogeny | Construct phylogeny | Neighbor-joining (NJ)**

7.2. A window with several options are displayed (Fig. 6.3)

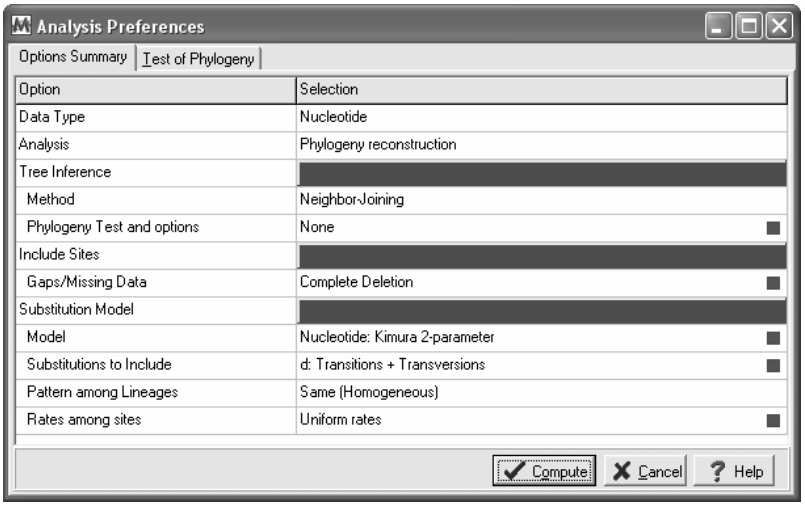


Fig. 6.3

7.3. Select a substitution model and run a bootstrap (100 replicates) analysis

7.4. A NJ tree with the bootstrap values (Fig. 6.4) will be displayed after you have clicked on **Compute**

8. I have chosen a Kimura 2-parameter substitution model. But is this the best model that explains the evolution of nucleotides in our dataset? We need to solve this question by investigating all the available models using hierarchical likelihood ratio tests or the Akaike information criterion in **ModelTest:**

8.1. Save the file: OK

EXERCISE 6

Exploring alternative approach using MEGA

Creating Multiple Sequence Alignments with Alignment Explorer.

I) *Creating multiple sequence alignment from an open text file*

1. Launch the Alignment Explorer by selecting **Alignment -> Alignment/CLUSTAL**
2. A window will appear asking you either to a) Create a new alignment, b) Open a saved alignment session, or c) Retrieve sequences from a file. Select the first option, "create a new alignment".
3. Copy and paste unaligned sequences from the text file to the Alignment Explorer.
4. In the Alignment Explorer highlight all the sequences by selecting **Edit -> Select All.**
5. Align the highlighted sequences by selecting **Alignment -> Align by ClustalW.**
6. Save the current alignment as an alignment session file by selecting **Data -> Export -> Save.** This will allow the current alignment session to be restored for future editing in a file with the extension ".mas", i.e. cox_alignment.mas
7. Save the current alignment as a MEGA file by selecting **Data -> Export -> MEGA file.** This will allow the current alignment to be analyzed by MEGA.

Estimating Evolutionary Distances from Nucleotide Sequences.

Activating a Mega file

1. Activate the MEGA file you just saved, by clicking on the link **Click me to activate a data file.**
2. Select the desired data file to activate.
3. The Sequence Data Explorer will open.

Compute the proportion of amino acid differences between each pair of sequences

1. Select **Distance -> Compute Pairwise** command to display the distance analysis preferences dialog box.
2. In the **Distance Options** tab, click on the green box in the **Models** pulldown section and then select the **Amino Acid -> p-distance** option.
3. Click "Compute" to begin the computation.

Compute distances and compare them using other methods

1. Select **Distance -> Compute Pairwise** command. Use the **Models** pulldown to select the **Amino-acid -> Poisson correction** method. Now click "OK" to begin the computation.
2. Follow the steps from 2-3 in the previous section to compute the **JTT Model** distance.
3. You now have open results windows containing the distances estimated by three different methods, which you can now compare.
4. After you've compared the results, close each one of the windows displaying the distance matrices.

Estimating the best substitution model to be incorporated in your analysis from Nucleotide/Amino acid Sequences

Determine the substitution model that best fits you data:

1. Use MEGA (MODELTEST is another option) to determine the best substitution model to be incorporated in your analysis.
2. Activate the MEGA file you just saved, by clicking on the link **Click me to activate a data file.**
3. Select **Models -> Find the best DNA/Protein models (ML)** pulldown menu. Now click "OK" to begin the computation.
4. A window with several options are displayed (Fig. 6.3). Deselect the **3rd** and **Noncoding Sites** check boxes and **click Compute.** Which is the best substitution model for your data?

   NOTE: *The model provided is not always compatible with distance analysis. Models with the lowest BIC scores (Bayesian Information Criterion) are considered to describe the substitution pattern the best. For each model, AICc value (Akaike Information Criterion, corrected), Maximum Likelihood value (lnL), and the number of parameters (including branch lengths) are also presented.*

   MEGA has a limited number of models; you can choose one from the model options that will best fit the model chosen by MODELTEST.

   Other options:
   a) Use ProtTest for amino acid sequences (http://darwin.uvigo.es/software/prottest2_server.html)
   b) Use jModelTest for DNA sequences (http://jmodeltest.org/user/dashboard)
5. Repeat Exercise 5 using the best-predicted model.

EXERCISE 7

MAXIMUM LIKELIHOOD ANALYSIS AND MOLECULAR CLOCK

CASE STUDY: HISTORICAL BIOGEOGRAPHY OF LENTINULA (SHIITAKE MUSHROOM) (FROM: HIBBETT DS JOURNAL OF BIOGEOGRAPHY 28: 231-241 2001)

The aims of this practical are to:

1) Evaluate the effect of different models on the resulting tree likelihoods using *a priori* settings and to determine the model that best fits the data using ModelTest.

2) Generate phylogenetic trees based on maximum likelihood statistics using MEGA.

Sequences, alignment and file preparation:

8. Download the following sequences from GenBank (suggest that you use MEGA): `AF287855, AF287875, AF287862, AF287890, AF287888, AF287884, AF071528, AF042595, AF261559, AF356152, AF356161, AF356164, AF356154`

9. Align the files in MEGA / ClustalX or MAFFT

10. Export the file in NEXUS format / convert it to a NEXUS file if you have used MAFFT.

11. **Save file: Data | Export | MEGA**. Save the alignment file in a format of your choice (nexus, phylip, clustal etc)

12. Now for the phylogenetic analyses. First do a simple NJ analysis with bootstrap (remember we do not know the correct substation model yet):

    - 12.1. **Phylogeny | Construct phylogeny | Neighbor-joining (NJ)**
    - 12.2. A window with several options are displayed (Fig. 6.3)

12.3. Select a substitution model.

12.4. Do a quick bootstrap analysis (Note: **we usually use 1000 reps for bootstrap; this will take some time – use only 100 for this practical**).

12.4. A NJ tree with the bootstrap values (Fig. 6.4) will be displayed after you have clicked on **Compute.**