BNS2002 Evolution and Genetics

Practical 2. Creating a DNA sequence based molecular identification tool: Aligning and translating sequences, treebuilding and generating distance matrices.

In this practical, we will align all of our quality checked sequences from last time using ClustalW embedded in the software program MEGA11, followed by exporting the fasta aligned sequence file into a MEGA formatted file. From there, we will translate all nucleotide sequence data into amino acids to check for the existence of stop codons, prior to tree building, interpretation of sequence identity (with reference to known species and unknown sequences) and calculating intra- and interspecific genetic distances. The latter will provide information about how closely related the taxa from which the sequences were derived are from each other. Details of how MEGA files are formatted are given towards the end of the handout.

Creating a phylogenetic framework for sequence/species identification

1. To begin, collate all of your sequence data into a single fasta file, to include your quality checked sequences and the Tetragnathidae_allforclass.fas file from the blackboard website (Si Creer DNA Data) (make sure that you use notepad or wordpad to copy/paste the files together and save as a text file).

Aligning sequence data and creating a MEGA input file – details of how the MEGA windows are organized are presented at the end of the practical handout.

2. Make sure that you have a version of MEGA11 on your machine. The most recent version is MEGA11.

https://www.megasoftware.net/

If you cannot access MEGA versions yourself, do not worry and I will provide working examples of data files as we move through the practicals.

Then, using MEGA11 itself (search in the Windows tool bar), import your sequence dataset using the **align** button and selecting **edit/build alignment** and selecting **retrieve sequences from a file**. Then, the alignment Explorer window will open, showing your sequences. To align, navigate to **alignment**, **align by ClustalW** and using the default parameters. From there, you need to export your sequence alignment as a MEGA formatted file, by navigating to **data, export alignment** and selecting the **MEGA format**, giving it a title and specifying Protein-coding nucleotide sequence data. ***From here, shut the alignment explorer window, go to main MEGA shell and open your MEGA formatted file thus: File, Open a file/session, navigate to your MEGA file, open.***

Translating DNA sequences into amino acid sequences, to check for stop codons

3. Now, you can check that your sequences translate into amino acids, without any errors or stop codons. There should be no stop codons, or gaps (insertion/deletion events) in our dataset in AT LEAST ONE OF THE THREE POSSIBLE READING FRAMES

(stop codons are indicated by *) in our putative COI genes (see Lecture 3!). Using the Sequence Data Explorer window in MEGA (TA), use the **data** drop-down menu and **Select Code Table**, to use the invertebrate mitochondrial DNA coding translation. Then, click on **the translate button (UUCàPhe).**

If you're lucky, you will have translated the nucleotide sequences at the correct starting base, but since translation can occur in three reading frames, you may have to interact with the program to perform your translation using the first, second and third starting codon positions before seeing the translation of your nucleotides that do not contain any stop codons. The latter can be achieved by using the **Data/Select & Edit Genes and Domains**, you can translate from all three reading frames by editing Codon Start.

If at this stage, you cannot visualise a protein alignment that is free from stop codons, you will either have badly edited sequence data in your alignment, or potentially the presence of non-homologous genes. If you suspect the former, now is the time to double check your sequences and edit appropriately. If you suspect the latter, then which sequences (and why?) do you think should be removed from the alignment?

In order to proceed you will have to continue only with sequences that have:

a. BLASTed well (e.g. length, match and identity) as spider (Araneae) COI DNA.
b. been aligned correctly and can be translated into amino acids without the presence of any stop codons.

Any other sequences that do not meet the above criteria can be discarded from future spider analyses. Once you have identified if you do have non-homologous sequences in your dataset, you can simply deselect them within MEGA (tick box) and ensure that your dataset contains only homologous genes.

To continue, with any other permutation of data will be scientifically flawed and will be likely to yield incorrect results due to poor data quality and/or flawed assumptions of homology in the alignment.

Treebuilding and assigning bootstrap support

4. Okay, let's build a tree! In the main menu of MEGAX, navigate to the **phylogeny** button and **construct neighbor-joining tree.** Apart from the default settings, please select **phylogeny test, bootstrap method, 1000 replications, substitutions type nucleotide, model/method P-distance** (this is just the number of nucleotides differing between two taxa represented as a percentage, ie 3 differences / 100 base pairs = 0.03, or 3%). For gaps/missing data treatment, please select pairwise deletion, since this uses the maximum amount of data in your alignment. Leave all other options as they are and then press **compute**!

Sequence/taxon identification

5. In this tree, you will be able to see a phylogenetic tree based upon your aligned sequence data and a pre-written figure legend. Hopefully, you will be able to see a number of highly supported clades/monophyletic groups that contain an assemblage of sequences of known origin and those that we are trying to assign the identity of. For the purpose of this practical we will use a bootstrap support of 100 to interpret as strongly supported clades. Save this tree and legend in a suitable format that you will be able to reproduce in a Word document as it will form part of your assessment. **Via interpreting the structure of your tree and referring to high bootstrap values at the base of monophyletic clades, can you now infer what species your unknown sequences came from/are? Can you make a note of your interpretation of the identity of your unknown taxa according to the analyses from both practical 1 and 2, as it will form part of your assessment.** Annotating this tree is relatively easy to perform using PowerPoint (and you can italicize latin names) and then copying into a Word document.

6. Once you are happy with what your unknown taxa are, an interesting thing to find out is how the levels of genetic variation compare between individuals of the same species (intraspecific variation) and between species (interspecific variation). You can compare these by assigning taxa to groups using the Data, Select & Edit Taxa/Groups (in the Sequence Data Explorer window) and adding a new group called, for example *M menardi* for all the *Meta menardis* and then highlighting in the right hand column and pushing them across using the arrows with the target group highlighted. Do this for all the species that have more than one representative (*P. listeri* is very rare and so we only have one sequence). Once you have sorted out all your species groups, you can use the Distances tab (Main MEGA window) to compute a range of statistics, but we will calculate Within Group Means and Between Group Means. The within group will give you the intraspecific divergence within a species and the between will give you the interspecific distance. **Either save these tables, or make a note of the values as these will form part of your assessment. For your assessment, it will be better to represent the distances currently presented as proportions out of 100, to percentages, by multiplying by 100 – this will give you a percentage sequence divergence.**

Are all the intraspecific distances in the same sort of range, or are some higher than the others? If so, what do you think could have caused such a pattern, ie why some taxa have elevated levels of divergence compared to others?

The assessment will be available on Blackboard and discussed during the practical in addition to our final lecture synthesis session. Please do not miss the synthesis session – it is very important to complete the teaching cycle.

Notes:

A MEGA formatted file must adhere to the following structure:

#MEGA
Title Something that tells you what the data is

```
#TaxonA         ATCAGATCAGATACAGCAGCAGCAGCTACAGACGACTACGA
#TaxonB         ATCAGATCAGATACAGCAGCAGCAGCTACAGACGACTACGA
#TaxonC         ATCAGATCAGATACAGCAGCAGCAGCTACAGACGACTACGA
```

Notes on MEGA format are available here:

Key Words

Every data file must contain the key words #MEGA and TITLE. These key words can be written in any combination of lower- and upper-case letters.

#MEGA This key word indicates that the data file is prepared for analysis using MEGA. It must be present on the very first line in the data file.

TITLE The word TITLE must be written on the second line. It may be followed by some description of data on the same line. This description is written in all the output files containing results. If the specified description exceeds 128 characters in length, the additional characters are ignored. After the MEGA format identifier (#MEGA) and the title (TITLE), the data should follow. Comments may be written on one or more lines right after the TITLE line and before the data (see examples in sections 2.2 and 2.3).

OTU Labels

Distance matrices as well as sequence data may come from species, populations, or individuals. These evolutionary entities are designated as OTUs (Operational Taxonomic Units). Each OTU must have an identification tag, i.e., an OTU Iabel. In the input files prepared for use in MEGA, these labels should be written according to the following conventions.

'#' Sign Every OTU Iabel must be written on a new line, and a '#' sign must proceed the label. OTU Iabels cannot be longer than 40 characters; extra characters are disregarded. OTU Iabels are not required to be unique, but identical labels may result in ambiguities.

Forbidden
Characters The '#' sign, blanks, and tabs cannot be a part of an OTU Iabel. For multiple word labels, an underscore can be used to represent a blank space. All underscores are converted into blank spaces, and subsequent displays of the OTU label show this change. For example, E._coli becomes E. coli.

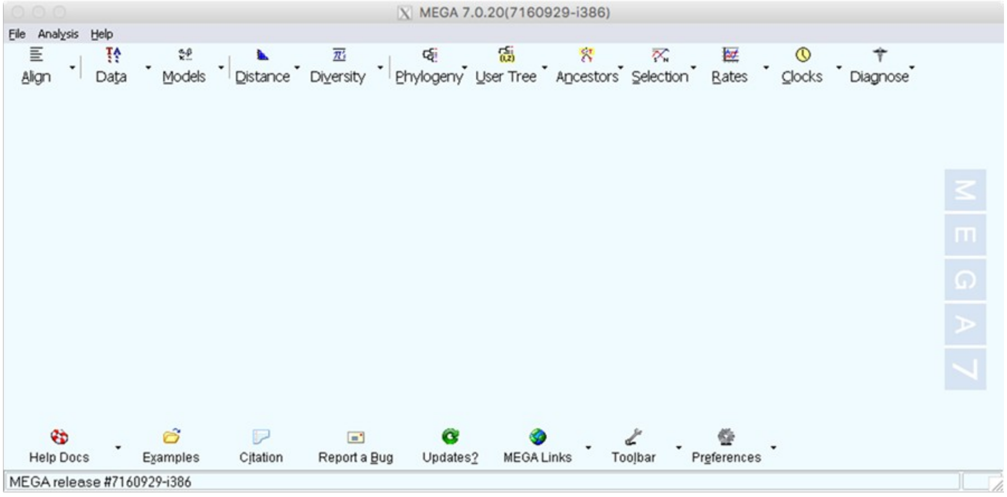
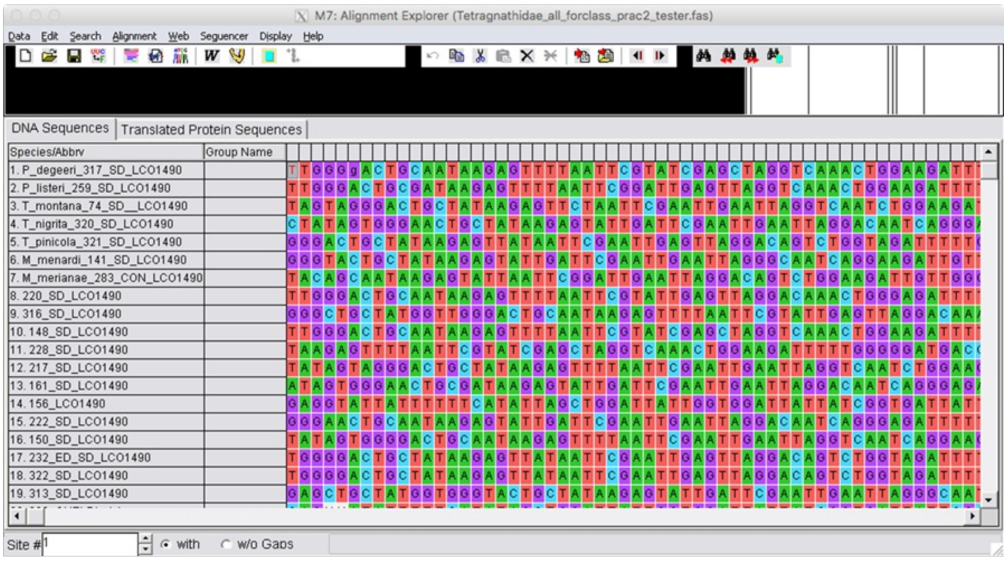
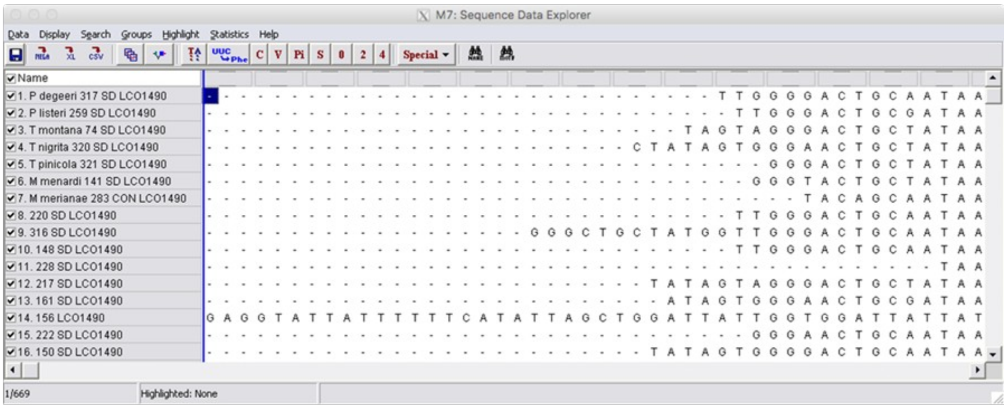


Figure 1. Normal MEGA analysis window, offering main functionality.


M7: Alignment Explorer (Tetragnathidae_all_forclass_prac2_tester.fas)
Data
Edit
Search
Alignment
Web
Sequencer
Display
Help
DNA Sequences
Translated Protein Sequences
Site #
with
w/o Gaps

```
Species/Abbrv                    Group Name
1. P_degeeri_317_SD_LCO1490                 TTGGgACTGCAATAAGAGTTTTAATTCGTATCGAGCTAGGTCAAACTGGAAGATTT
2. P_listeri_259_SD_LCO1490                 TTGGGACTGCGATAAGAGTTTTAATTCGGATTGAGTTAGGTCAAACTGGAAGATTT
3. T_montana_74_SD__LCO1490                 TAGTAGGGACTGCTATAAGAGTTCTAATTCGAATTGAATTAGGTCAATCTGGAAGA
4. T_nigrita_320_SD_LCO1490                 CTATAGTGGGAACTGCTATAAGAGTATTGATTCGAATTGAATTAGGACAATCAGGG
5. T_pinicola_321_SD_LCO1490                GGGACTGCTATAAGAGTTATAATTCGAATTGAGTTAGGACAGTCTGGTAGATTTTT
6. M_menardi_141_SD_LCO1490                 GGGTACTGCTATAAGAGTATTGATTCGAATTGAATTAGGGCAATCAGGAAGATTGT
7. M_merianae_283_CON_LCO1490               TACAGCAATAAGAGTATTAATTCGGATTGAATTAGGACAGTCTGGAAGATTGTTGG
8. 220_SD_LCO1490                           TTGGGACTGCAATAAGAGTTTTAATTCGTATTGAGTTAGGACAAACTGGGAGATTT
9. 316_SD_LCO1490                           GGGCTGCTATGGTTGGGACTGCAATAAGAGTTTTAATTCGTATTGAGTTAGGACAA
10. 148_SD_LCO1490                          TTGGGACTGCAATAAGAGTTTTAATTCGTATCGAGCTAGGTCAAACTGGAAGATTT
11. 228_SD_LCO1490                          TAAGAGTTTTAATTCGTATCGAGCTAGGTCAAACTGGAAGATTTTTGGGGGATGAC
12. 217_SD_LCO1490                          TATAGTAGGGACTGCTATAAGAGTTTTAATTCGAATTGAATTAGGTCAATCTGGAA
13. 161_SD_LCO1490                          ATAGTGGGAACTGCGATAAGAGTATTGATTCGAATTGAATTAGGACAATCAGGGAG
14. 156_LCO1490                             GAGGTATTATTTTTTCATATTAGCTGGATTATTGGTGGATTATTATCGGTGATTAT
15. 222_SD_LCO1490                          GGGAACTGCAATAAGAGTATTGATTCGAATTGAATTAGGACAATCAGGGAGATTTT
16. 150_SD_LCO1490                          TATAGTGGGGACTGCAATAAGAGTTTTAATTCGAATTGAATTAGGTCAATCAGGAA
17. 232_ED_SD_LCO1490                       TGGGGACTGCTATAAGAGTTATAATTCGAATTGAGTTAGGACAGTCTGGTAGATTT
18. 322_SD_LCO1490                          TGGGGACTGCTATAAGAGTTATAATTCGAATTGAGTTAGGACAGTCTGGTAGATTT
19. 313_SD_LCO1490                          GAGCTGCTATGGTGGGTACTGCTATAAGAGTATTGATTCGAATTGAATTAGGGCAA
```

Figure 2. MEGA alignment explorer window, enabling ClustalW alignments.



```
☑ Name
☑ 1. P degeeri 317 SD LCO1490         - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - T T G G G A C T G C A A T A A
☑ 2. P listeri 259 SD LCO1490         - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - T T G G G A C T G C G A T A A
☑ 3. T montana 74 SD LCO1490          - - - - - - - - - - - - - - - - - - - - - - - - - - - - T A G T A G G G A C T G C T A T A A
☑ 4. T nigrita 320 SD LCO1490         - - - - - - - - - - - - - - - - - - - - - - - - - C T A T A G T G G G A A C T G C T A T A A
☑ 5. T pinicola 321 SD LCO1490        - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - G G G A C T G C T A T A A
☑ 6. M menardi 141 SD LCO1490         - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - G G G T A C T G C T A T A A
☑ 7. M merianae 283 CON LCO1490       - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - T A C A G C A A T A A
☑ 8. 220 SD LCO1490                   - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - T T G G G A C T G C A A T A A
☑ 9. 316 SD LCO1490                   - - - - - - - - - - - - - - - - - - - G G G C T G C T A T G G T T G G G A C T G C A A T A A
☑ 10. 148 SD LCO1490                  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - T T G G G A C T G C A A T A A
☑ 11. 228 SD LCO1490                  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - T A A
☑ 12. 217 SD LCO1490                  - - - - - - - - - - - - - - - - - - - - - - - - - - T A T A G T A G G G A C T G C T A T A A
☑ 13. 161 SD LCO1490                  - - - - - - - - - - - - - - - - - - - - - - - - - - - A T A G T G G G A A C T G C G A T A A
☑ 14. 156 LCO1490                     G A G G T A T T A T T T T T T C A T A T T A G C T G G A T T A T T G G T G G A T T A T T A T
☑ 15. 222 SD LCO1490                  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - G G G A A C T G C A A T A A
☑ 16. 150 SD LCO1490                  - - - - - - - - - - - - - - - - - - - - - - - - - - T A T A G T G G G G A C T G C A A T A A
1/669        Highlighted: None
```

Figure 3. MEGA sequence data explorer window, enabling translation to amino acids.