Tutorial – Using Genetic Barcodes to Evaluate Background Genetic Diversity in Lab Cultures

LUIS CUNHA

Table of Contents

1. Context

1.1 Teaching Objectives

The purpose of this practical is to introduce you to the very basic techniques of analysis of a DNA sequence generated using a cycle sequence reaction followed by capillary separation of the products (see Section 1.3). This process generates an electropherogram, which may be used to derive the sequence of the template molecule. The steps in the process are given below:

1. Pick a sequence set (electropherograms) from the batch provided (Earthworm and Springtail);
2. Clean/process the sequence file (with FinchTv);
3. Create a fasta sequence file (using a text editor, notepad++);
4. Submit the sequence to the blastn tool;
5. Evaluate the identity/similarity to the hit in the database;
6. Pick a reference sequence for your species (for comparison)
7. Do some comparisons (phylogeny, pairwise distance table);
8. Discuss.

Particular attention should be paid to the following points:

[1] The manual is designed to provide the individual steps required in the process. These steps may be used in a variety of ways to aid analysis of the sequence under investigation.

[2] A majority of the tools used come from web sites as this is the most convenient way of performing each task. These sites may go offline, if so, there are alternative resources available on the internet.

[3] There are commercial software options (expensive) for performing these operations, these can be used if you have access.

The first part you will be instructed on how to use the various components of the sequence analysis pipeline. In the subsequent session you will then be expected to analysis the sequence/s provide to you derived from sequencing reactions performed on a COI barcode gene. You will be expected to discuss your findings.

You may wish to consider some of the following elements as part of your discussion:

1. If your sequencing reactions have not yielded a sequence trace use traces other group providing the appropriate acknowledgement. You can use other or even all the sequences provided for this exercise.
2. A short discussion of the homology search of your sequences against relevant database sequences using Blastn (e.g Pubmed) will be done after your finish.

Then:

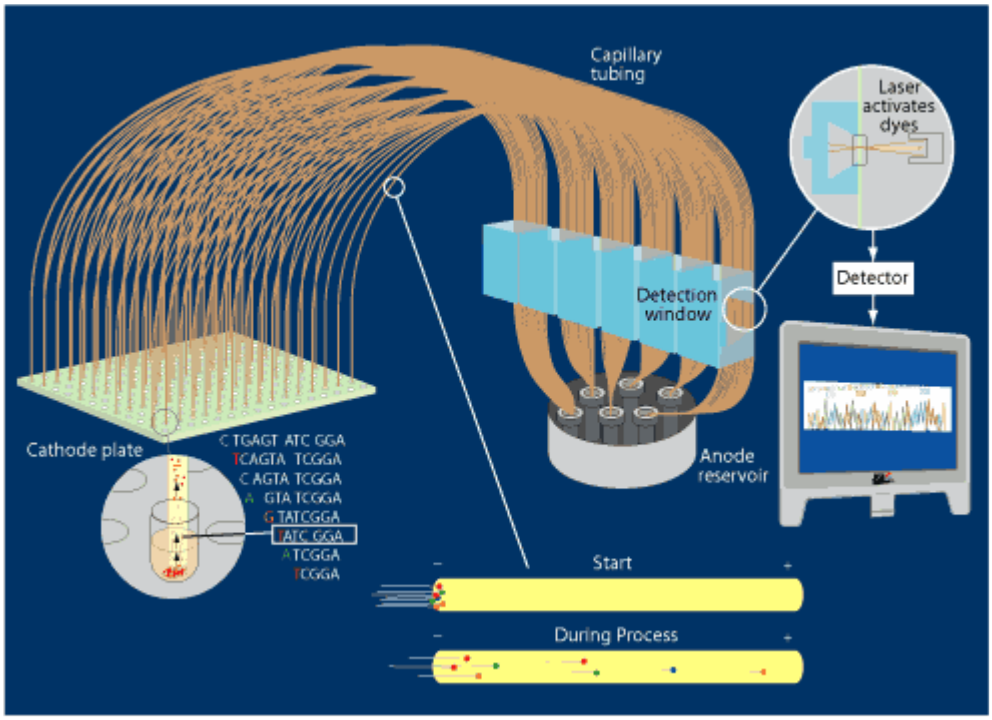
3. Evaluate pairwise distances matrix (genetic distances) between your sequence and appropriate sequences downloaded from Genebank.
4. A further challenge can be done by creating a phylogenetic tree for your sequence, include your sequences, references sequences and appropriate sequences downloaded from Genebank. An ideal result will be appropriately rooted and contain bootstrap values.

1.2 Cycle sequencing (also known as Sanger sequencing)

An excellent animation describing cycle sequencing is given at;

http://www.dnalc.org/ddnalc/resources/cycseq.html

In this example the fluorescent products from the cycle sequencing reaction are separated by size fractionation using an Acrylamide gel. Recently this separation technique has been replaced by capillary electrophoretic separation since there is no need to physically poor a gel and the loading of the capillaries can be automated. The fragments, which have a negative charge, move through the capillaries of the sequencing machine toward the positively charged pole. Shorter fragments move faster than longer ones. At the detection window, a laser excites the dyes. A detector "reads" the colours, one at a time, to determine the sequence of the tagged As, Ts, Gs, and Cs.



2. Visualising an electropherogram and curating for sequences

2.1 Software Tools

Recommended: *FinchTV*
Source: Geospiza (https://digitalworldbiology.com/FinchTV)
Application object available under departmental applications.

Alternatives:

- Chromas Lite 2.0 (Free) (Technelysium Pty Ltd)
  Windows specific chromatogram viewer.
  http://www.technelysium.com.au/chromas_lite.html

- Edit View (ABI)
  MAC specific sequence viewer
  http://www.appliedbiosystems.com/support/software/...

- Sequence scanner v1.0 (Applied Biosystems)
  The free Sequence Scanner Software enables you to view, edit, print and export sequence data generated using the Applied Biosystems Genetic Analyzers. The software generates graphically expressive reports on results.

2.2 File Type Required and Examples

*.ab1 contain electrophoretogram information
X.ab1 is provided on Unilearn (where X is your sample number)

2.3 Menus and Functions

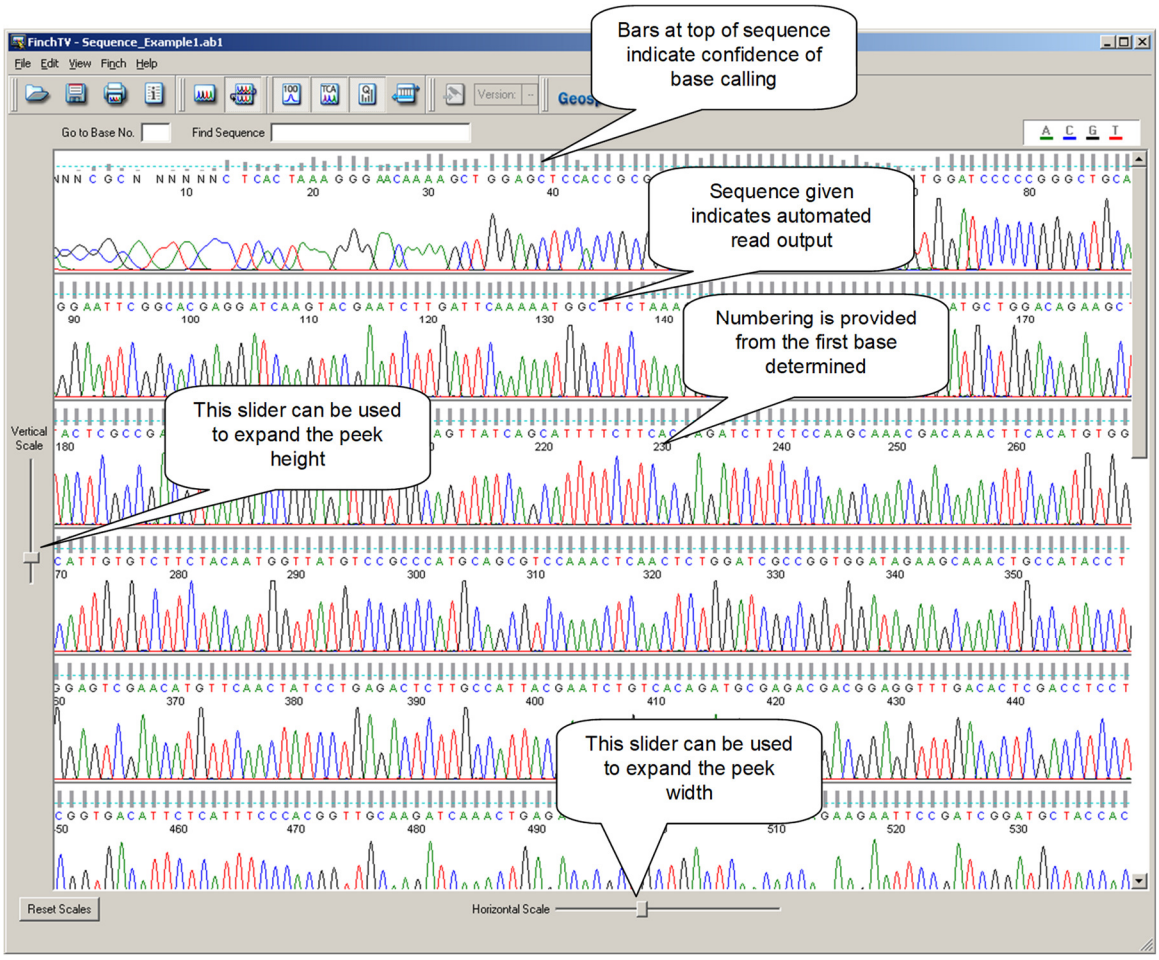
2.4 Basic Visualisation

1. Download file from blackboard, Open software **Network Applications>Departmental Applications>Finch TV**.

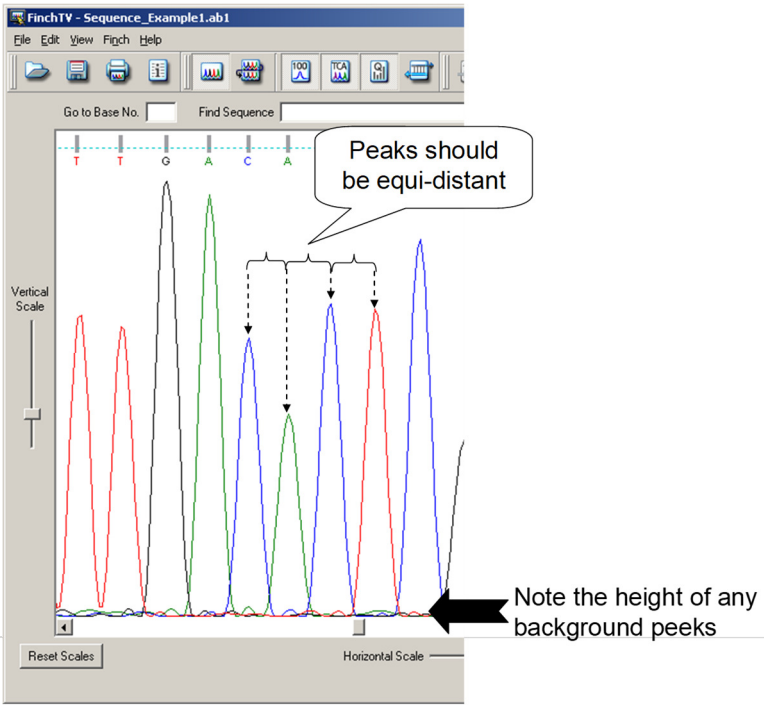
2. Choose **File > Open** and select the Sequence File

3. Click on Wrapped Sequence View 

4. Overall Sequence View



5. Detailed View



2.5 Remove Section/s of "Bad" Sequence

At the front and towards the end of the sequence the you can observe the bars that indicate the quality of base calling are small and the program show Ns in the sequence output. This indicates that this section of sequence is "bad" or cannot be determined with any degree of confidence. Before saving the sequence remove these sections by highlight the areas by holding the left-hand mouse button down and dragging across the sequence to be deleted. When the sequence to be removed is highlighted press the backspace or delete key.

2.6 Exporting the resulting sequence file

To export your edited sequence selected **File>Export>DNA Sequence:FASTA**
I recommend you DO NOT Save any edited AB1 files since this will affect the original raw data that you may wish to review in the future.

NOTE: If you sequence is not useful for downstream tasks (e.g. too "dirty" with double peaks), pick another one from the dataset provided.

3. FASTA Sequence Files (understanding the file format)

3.1 Software Tools

Recommended: *Notepad++*
Source: https://notepad-plus-plus.org/

Alternatives:

Any word processing / text editor. If you use alternatives remember always to save the FASTA as a "text only" file.

3.2 File Type Required and Examples

FASTA text file: *.fa, *.seq or *.txt
X.seq is provided on Unilearn (where X is the sample number/filename)

3.3 Menus and Functions

Normal text processor functions

3.4 Basic Structure

1. Open software **Notepad++**

2. Choose **File > Open** and select Files of Type (*.*) followed by the appropriate file. Open the file from previous step (curated sequence)

Fasta Format:

- The description line starts with a greater than symbol (">").
- The word following the greater than symbol (">") immediately is the "ID" (name) of the sequence, the rest of the line is the description.
- The "ID" and the description are optional.
- All lines of text should be shorter than 80 characters.
- The sequence ends if there is another greater than symbol (">") symbol at the beginning of a line and another sequence begins.

e.g.

>Sequence_Example1
GATCCCCCGGGCTGCAGGAATTCGGCACGAGGATCAAGTACGAATCTTGATTCAAAAATGGCTTCTA
AAGATCTGATCAACCTACGTTCGGATGCTGGACAGAAGCTACTCGCCGACGTAAAGACCGGAGATG
GTGGCAGTTATCAGCATTTTCTTCACGAGATCTTCTCCAAGCAAACGACAAACTTCACATGTGGCATT
GTGTCTTCTACAATGGTTATGTCCGCCCATGCAGCGTCCAAACTCAACTCTGGATCGCCGGTGGATA
GAAGCAAACTGCCATACCTGGAGTCGAACATGTTCAACTATCCTGAGACTCTTGCCATTACGAATCT
GTCACAGATGCGAGACGACGGAGGTTTGACACTCGACCTCCTCGGTGACATTCTCATTTCCCACGGT
TGCAAGATCAAACTGAGACACGCTGCTGACTCAACACCAGAAGAATTCCGATCGGATGCTACCACTG
CACTGAATAACCGTGACTCTGCCTCTGGTATCATCATCAACTACAAGCAGGATGTGTTGGGTCAGGG
TGTACCATACGGGCATATCAGCCCTCTCGCGGCGTACCATACTTCCACTGATCGCGTTCTGATTTTG
GATACTTGGCCCGAAACGGAGGAATGCTGGGCCAAAGTGGAT

4. Add sequences to a unique fasta file for your culture box species

4.1 Open sequences in ‘notepad++’

As before open sequences in ‘notepad++’ by selecting file with right hand click of the mouse and selecting open with ‘notepad++’.

4.2 Appended other sequences

Cut and paste sequence from your other curated fasta files (remember to include your reference sequence for the species of interest.

4.3 Edit sequence names as appropriate

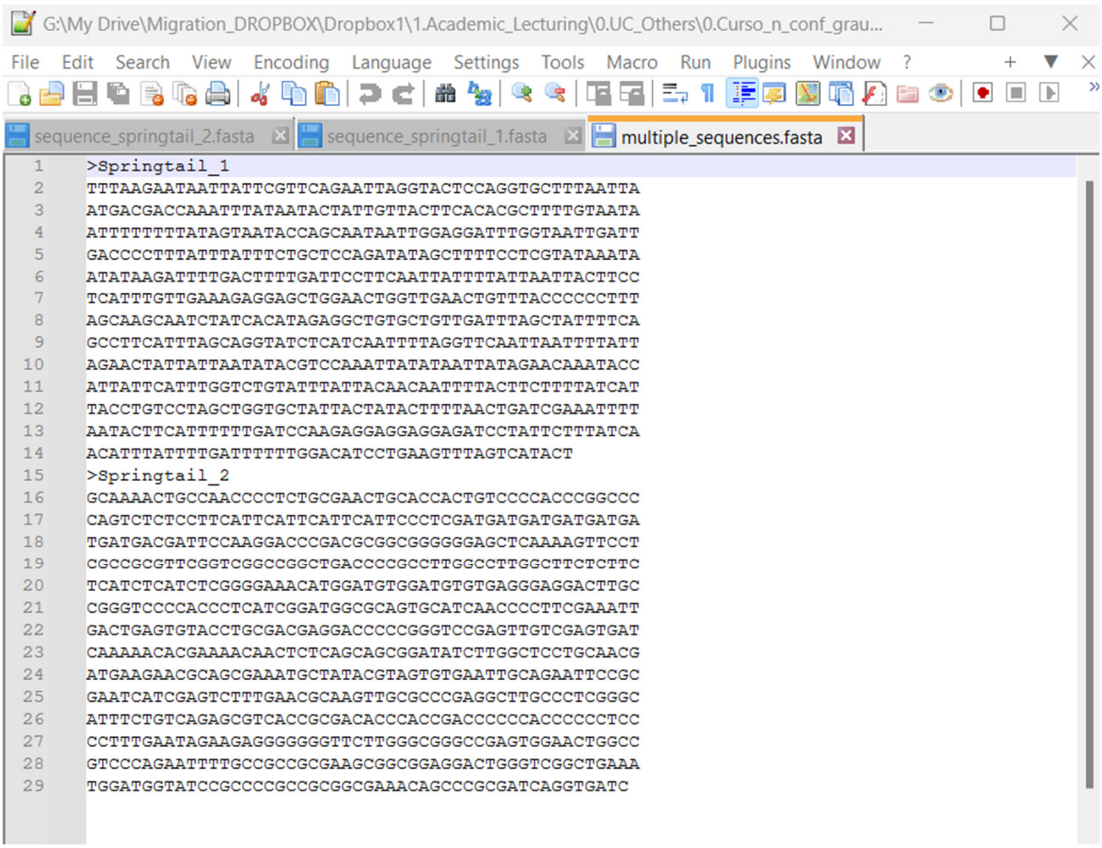
In notepad++ rearrange the name of the file so the species name (blast hit) is given first and the description is truncated – keep GenBank ID for future reference.

e.g.

>EHX191_290919121912

would become

>Springtail_1

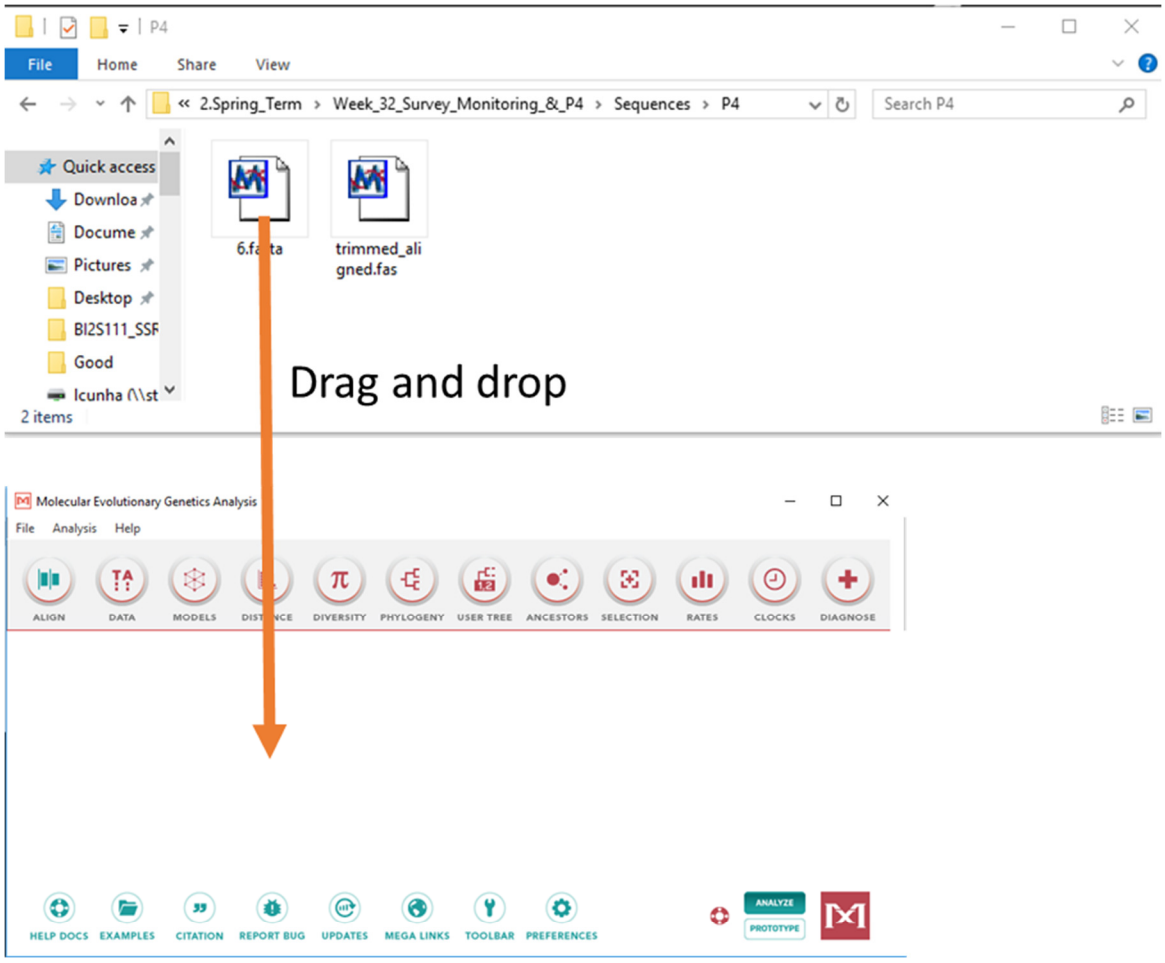


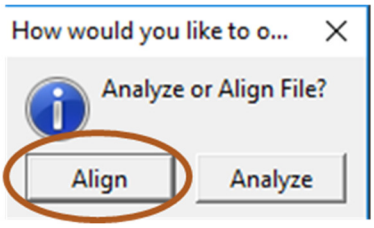
Now in file menu ‘save as’ select Text Document but ensure the file has a “.fasta” suffix.

5. ClustalW alignment in Mega

5.1 Import file into mega alignment

Open mega and open a windows explore window containing your ".fasta" file.
Now drag and drop you fasta file into the mega window….it will now give you the choice of "analyse" or "align", select "align"





5.2 Align Sequences

Select all sequences

Edit>Select All

Align Sequences

Alignment > Align by ClustalW

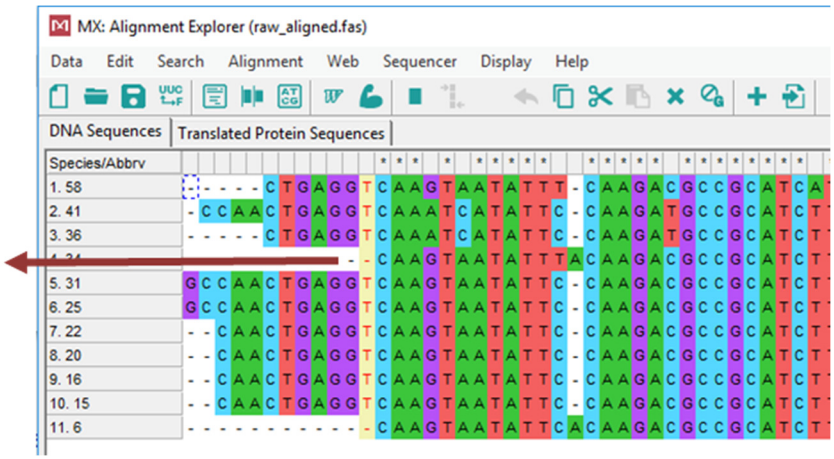
Keep default alignment parameters

>OK

5.3 Trim sequences

Trim away non-aligned sequences from the 5’ and 3’ of the alignment.

Identify the core alignment and select the column representing the first base prior to the alignment. Now whilst holding the ’shift’ key select ‘home’ key and then delete.



Navigate to the end of the core alignment and select the column representing the first base after to the alignment. Now whilst holding the ‘shift’ key select ‘end’ key and then delete.

Now Export the alignment

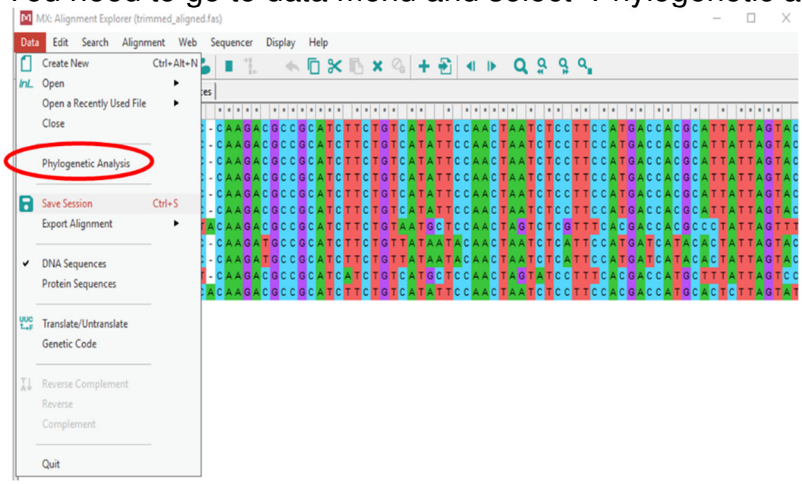
>Data>Export>MEGA format (or fasta format)

Name the sequence and save.

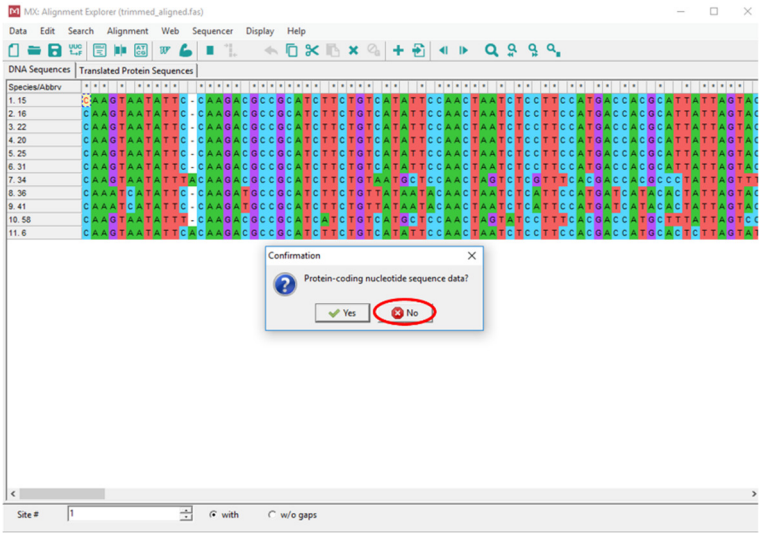
6. Phylogenetic analysis with MEGA (Use MEGA X)

6.1 Using your alignment for analysis

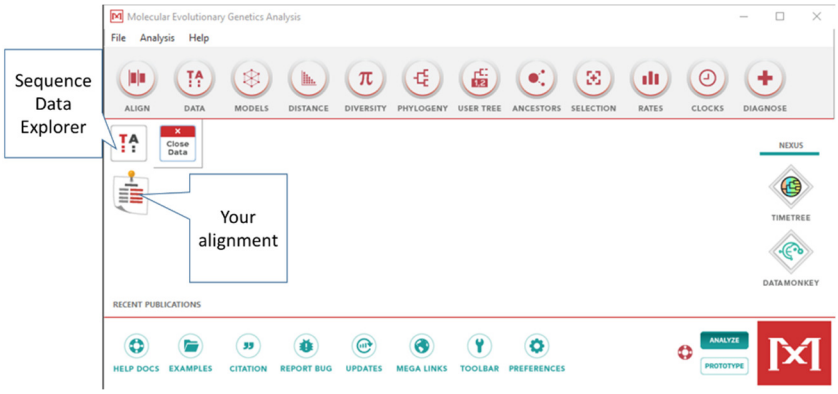
You need to go to data menu and select “Phylogenetic analysis”



Then select nucleotide sequence data by refusing to convert to protein-coding data



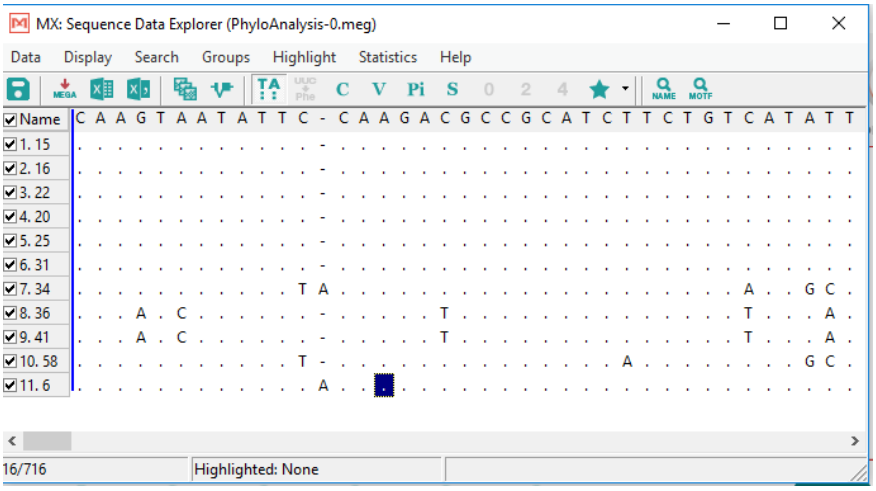
Now, go back to the Mega-X main window and you should see this:



Click on the **Sequence Data Explorer** icon.

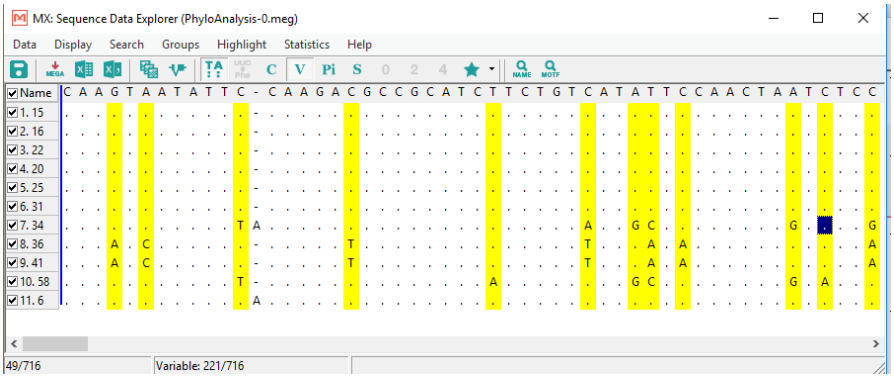
6.2 Data Analysis

View the **Sequence Data Explorer** window:

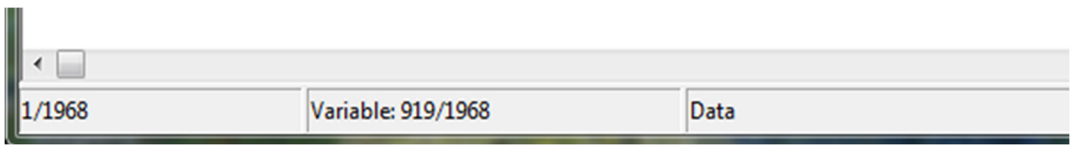


You can use the  button on the menu bar to toggle between showing all nucleotides and polymorphic positions only. A colour option is also available.

Click **[V]** to show **v**ariable (polymorphic) nucleotides:



Observe the number of polymorphic nucleotides at the bottom of the window

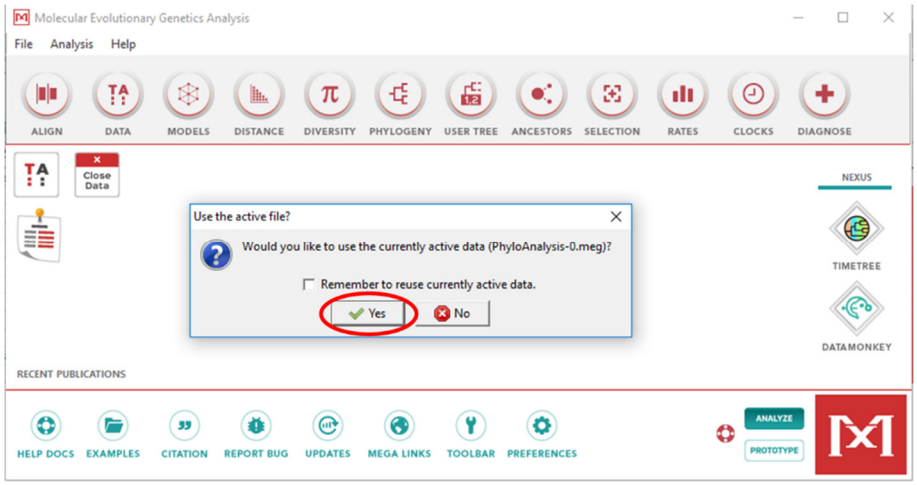


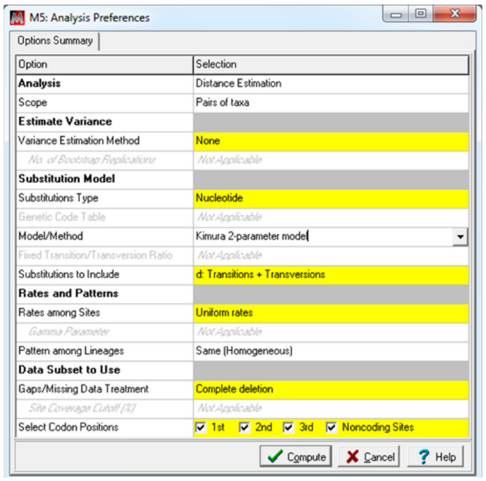
Click the translate menu button to view amino acid translations:

Close the Sequence Data Explorer window to return to the MEGA main/home panel

To compute distances between sequences on click **Distances>Compute Pairwise Distances…:**

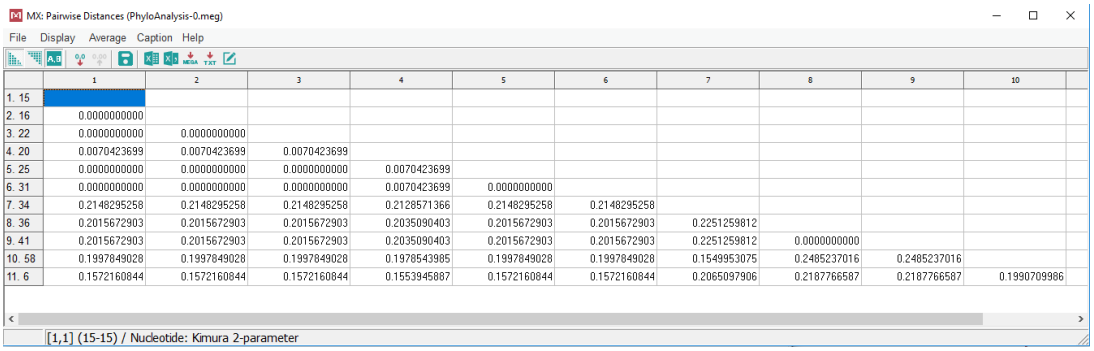
Use currently active data:





Click on the appropriate parameters from the dropdown menus, selecting the appropriate distance measure for your data. We will choose one basic genetic distance estimate – **Kimura's 2-parameter model**, which assumes that transitions and transversions have a different mutation probability. We will compute distances using **transitions plus transversions (d)**. You can leave the rest of the parameters as it is.

Observe Pairwise Distances (genetic distances) between sequences, represented in a similarity matrix:



	1	2	3	4	5	6	7	8	9	10
1. 15										
2. 16	0.0000000000									
3. 22	0.0000000000	0.0000000000								
4. 20	0.0070423699	0.0070423699	0.0070423699							
5. 25	0.0000000000	0.0000000000	0.0000000000	0.0070423699						
6. 31	0.0000000000	0.0000000000	0.0000000000	0.0070423699	0.0000000000					
7. 34	0.2148295258	0.2148295258	0.2148295258	0.2128571366	0.2148295258	0.2148295258				
8. 36	0.2015672903	0.2015672903	0.2015672903	0.2035090403	0.2015672903	0.2015672903	0.2251259812			
9. 41	0.2015672903	0.2015672903	0.2015672903	0.2035090403	0.2015672903	0.2015672903	0.2251259812	0.0000000000		
10. 58	0.1997849028	0.1997849028	0.1997849028	0.1978543985	0.1997849028	0.1997849028	0.1549953075	0.2485237016	0.2485237016	
11. 6	0.1572160844	0.1572160844	0.1572160844	0.1553945887	0.1572160844	0.1572160844	0.2065097906	0.2187766587	0.2187766587	0.1990709986

You can save this file to record the genetic distances between each sequence.

6.3 Phylogenetic hypothesis

You will now investigate the **phylogenetic relationship** of your sequences by selecting from the **Phylogeny>Construct/Test…** sub-menu on the **MEGA main menu**.

You can choose from several methods of phylogenetic reconstruction; we will use both **Neighbor Joining (NJ)** and **Maximum Likelihood (ML)**.

Under **NJ**, construct trees under **d** (transitions + transversions) and **v** (transversions only). To assess the statistical significance of the resultant phylogeny, select **500 bootstrap replicates** in the "**Test of Phylogeny**" section of the box.

For ML, we are going to use the "general time reversible" model that usually is appropriated for this mitochondrial gene. Then click **Compute**. This should take a couple of minutes. The numbers at the branching points of the trees are the number times (out of 500) that the same branch is created by the method, if you re-sample from a subset of the sequences. **Values of 400 (i.e 80%) or higher, usually indicate good statistical support for this grouping.** See the **MEGA help file** for a further explanation.

You may edit the resulting tree using the tools provided. To save the tree, Click **Image>Save as Enhanced Metafile (EMF)**.

NJ uses the genetic distance estimates between each tree as a basis for a simple clustering procedure that groups those sequences that have the lowest genetic distance together first, then searches for the next most similar sequences to that group of sequences, and so on until the tree is complete.

ML uses standard statistical techniques to assign probabilities to particular possible phylogenetic relation. The method requires a substitution model to assess the probability of particular mutations, in this case we are using GTR; roughly, a tree requiring more mutations at interior nodes to support the observed phylogeny will be assessed as having a lower probability. Examining all possible topologies in any method is a very time consuming process. This algorithm reduces the searching time by first producing a temporary tree, (e.g., an NJ tree), and then examining all topologies that are different from this temporary tree. If this is repeated many times, avoiding all the topologies previously examined, one can usually obtain the tree being sought.

Examine the trees you have produced. Is NJ concordant with ML?

I hope you have enjoyed, and would love to hear your feedback

Luis