Tutorial – Using Genetic Barcodes to Identify Earthworm Species and Assess Communities in Urban Fragments

LUIS CUNHA

Table of Contents

1. CONTEXT 4
1.1 TEACHING OBJECTIVES 4
1.2 CYCLE SEQUENCING (ALSO KNOWN AS SANGER SEQUENCING) 6
2. VISUALISING AN ELECTROPHEROGRAM AND CURATING FOR SEQUENCES 7
2.1 SOFTWARE TOOLS 7
2.2 FILE TYPE REQUIRED AND EXAMPLES 7
2.3 MENUS AND FUNCTIONS 7
2.4 BASIC VISUALISATION 7
2.5 REMOVE SECTION/S OF "BAD" SEQUENCE 9
2.6 EXPORTING THE RESULTING SEQUENCE FILE 9
3. FASTA SEQUENCE FILES (UNDERSTANDING THE FILE FORMAT) 10
3.1 SOFTWARE TOOLS 10
3.2 FILE TYPE REQUIRED AND EXAMPLES 10
3.3 MENUS AND FUNCTIONS 10
3.4 BASIC STRUCTURE 10
4. ADD SEQUENCES TO A UNIQUE FASTA FILE FOR YOUR SET OF SPECIES 11
4.1 OPEN SEQUENCES IN 'NOTEPAD++' 11
4.2 APPENDED OTHER SEQUENCES 11
5. DETERMINING SEQUENCE HOMOLOGY BY BLAST (N, X OR P) 11
5.1 SOFTWARE TOOLS 11
5.2 FILE TYPE REQUIRED AND EXAMPLES 11
5.3 WHAT IS BLAST 11
5.4 TYPES OF BLAST 12
5.5 SUBJECT DATABASES 12
5.6 SEQUENCE INPUT 12
5.7 PARAMETERS TO ADJUST 12
5.8 EXECUTING A SEARCH 13
5.9 OUTPUT 13
5.12 DOWNLOAD CLOSE HOMOLOGS FROM YOUR BLAST SEARCH REPRESENTING RELEVANT SPECIES 14
5.13 EDIT SEQUENCE NAMES AS APPROPRIATE 16
6. CLUSTALW ALIGNMENT IN MEGA 16
6.1 IMPORT FILE INTO MEGA ALIGNMENT 16
6.2 ALIGN SEQUENCES 17
6.3 TRIM SEQUENCES 17
7. PHYLOGENETIC ANALYSIS WITH MEGA (USE MEGA X) 18

7.1 USING YOUR ALIGNMENT FOR ANALYSIS 18
7.2 DATA ANALYSIS 19
7.3 PHYLOGENETIC HYPOTHESIS 21

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 (Earthworms) for your urban fragment;
2. Clean/process the sequence file (with FinchTv);
3. Create a multiple 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 and add putative species name;
6. Pick a “root” sequence for your phylogeny;
7. Data analysis (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:

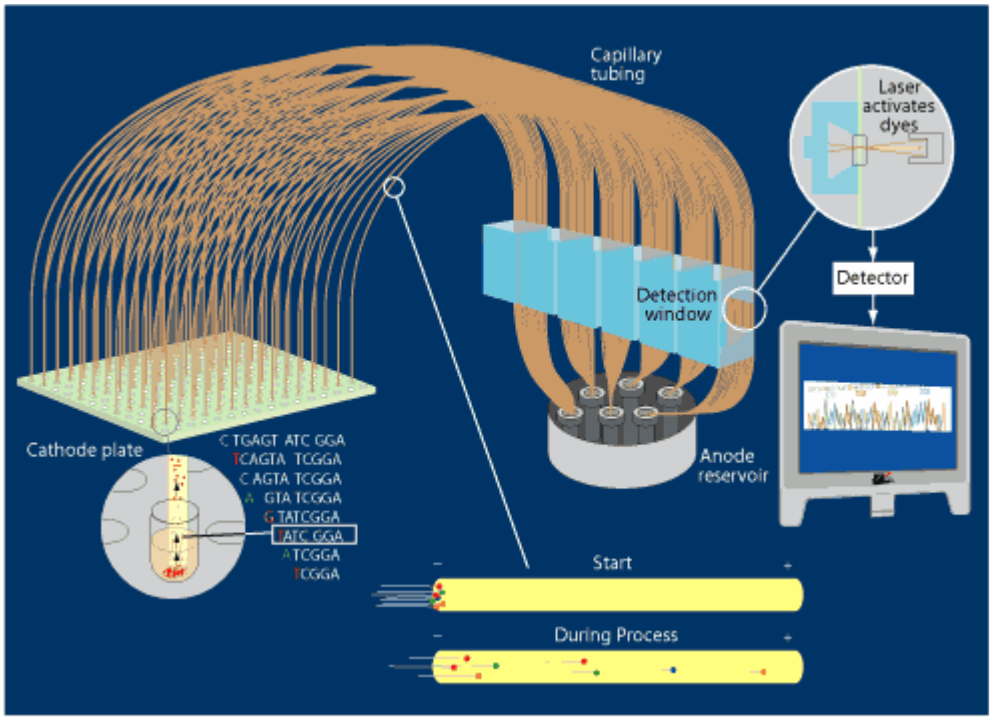
3. Evaluate pairwise distances matrix (genetic distances) between your sequence and appropriate sequences downloaded from Genebank.
4. A further challenge can be done by creating a phylogenetic tree for your sequence, include your sequences, references sequences and appropriate sequences downloaded from 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

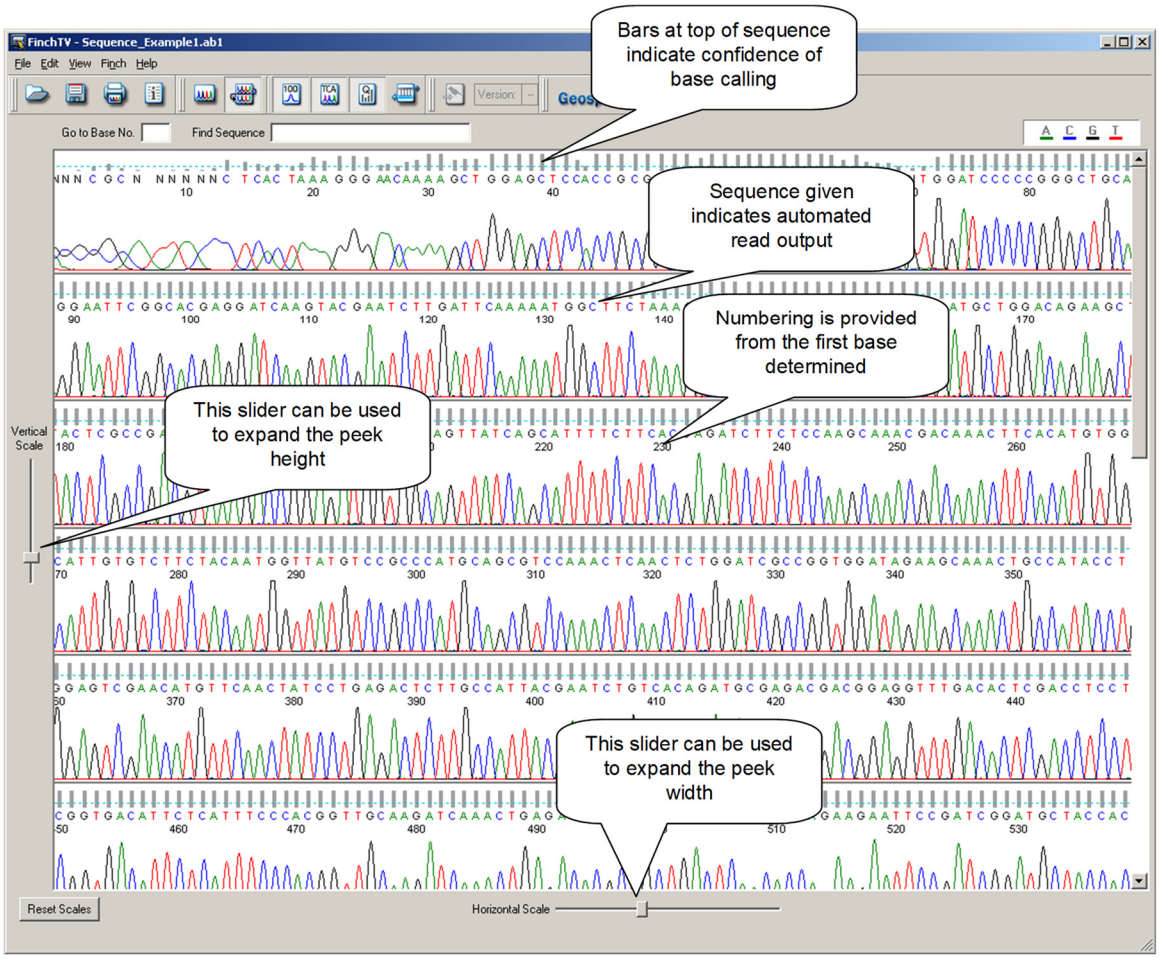
2.4 Basic Visualisation

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

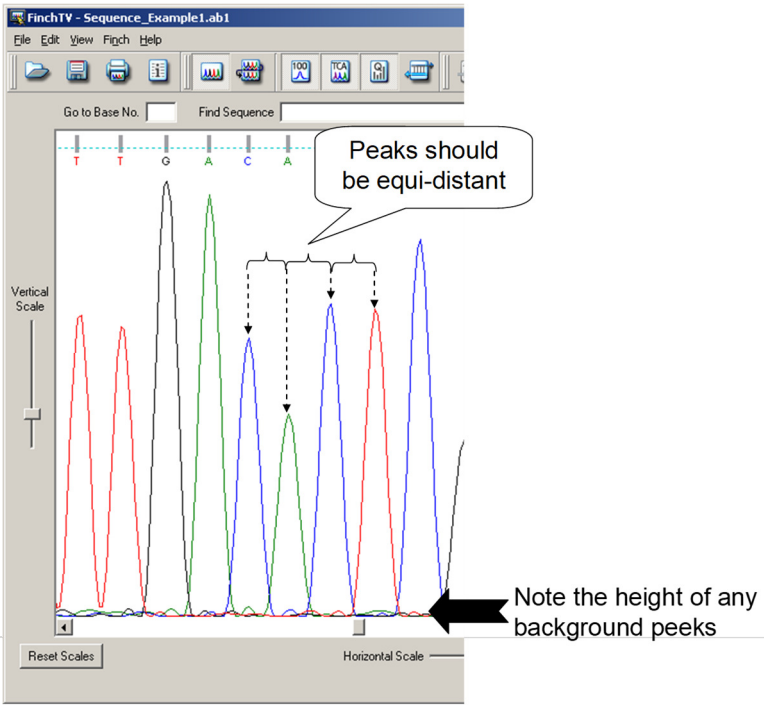
2. Choose **File > Open** and select the Sequence File

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), reject and 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, *.fas, *.fasta, *.seq or *.txt
X.seq is provided on UC Teacher (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 set of 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 and save the file as "multiple_sequences.fasta".

5. Determining Sequence Homology by Blast (N, X or P).

5.1 Software Tools

Recommended: *Blast at NCBI*
Source: http://www.ncbi.nlm.nih.gov/BLAST/

Alternatives:

EBI - http://www.ebi.ac.uk/blastall/index.html

5.2 File Type Required and Examples

FASTA text file: *.fa, *.seq or *.txt
Sequence_Example1.seq is provided on Blackboard

5.3 What is Blast

BLAST (Basic Local Alignment Search Tool) is a method to ascertain sequence similarity. The program takes a query sequence and searches it against the database selected by user. It aligns a query sequence against the every subject sequence in the database. The results are reported in a form of a ranked list followed by a series of individual sequence alignments, plus various statistics and scores. Every hit in that list is assigned with a similarity score S. Further, that score is analysed how likely it is to arise by chance. For that purpose so called E-value is calculated for every hit. E-value for the score S tells the expected number of hits of the score S or higher in the database.

5.4 Types of Blast

There are five main implementations of the blast algorithm, which can be distinguished by the type of the query sequence (DNA or protein) and the type of the subject database:

BLASTP compares an amino acid query sequence against a protein sequence database;

BLASTN compares a nucleotide query sequence against a nucleotide sequence database;

BLASTX compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database;

TBLASTN compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands).

TBLASTX compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

5.5 Subject Databases

There are many databases to use as subject databases. One of the most commonly used is nr database: collection of "non-redundant" sequences from GenBank and other sequence databanks.

5.6 Sequence input

BLAST accept the sequence in FASTA format (see Section 4) or Accession Number (GI number).

5.7 Parameters to adjust

EXPECT value: The statistical significance threshold for reporting matches against database sequences; the default value is 10, such that 10 matches are expected to be found merely by chance. If the statistical significance ascribed to a match is greater than the EXPECT threshold, the match will not be reported. Increasing the EXPECT value forces the program to report less isgnificant matches.

FILTER (Low-complexity): Mask off segments of the query sequence that have low compositional complexity (i.e. regions of biased composition, such as short-period repeats)

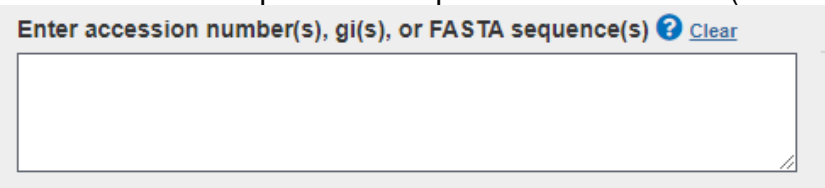
TARGET DATABASES: this can be refined to specific groups of organisms or by a NCBI Taxonomic group code.

5.8 Executing a search

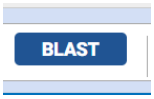
1. Open FASTA file in notepad++. Select contents of file & copy (**Edit>Copy** or **Ctrl-C**)

2. Open *Blast Page* http://www.ncbi.nlm.nih.gov/BLAST/ and navigate to the appropriate Blast page – try BlastN (nucleotide blast).

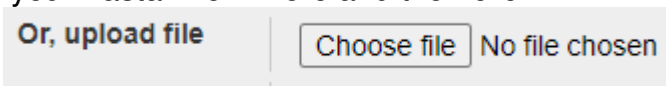
3. Click into input box and paste in FASTA File (**Edit>PASTE** or **Ctrl-V**).



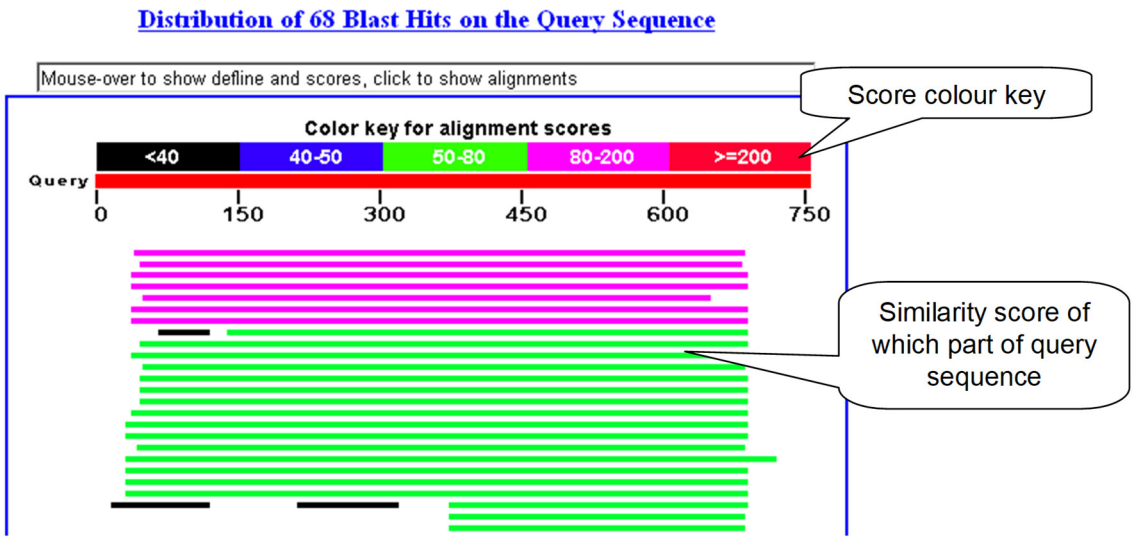
4. Keep the standard parameters and press BLAST at the bottom.



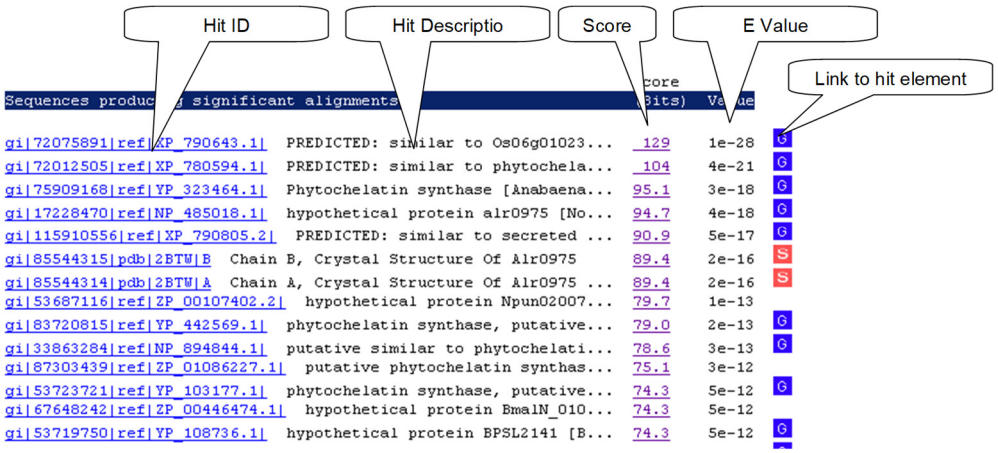
5. Alternatively, you can upload your "fasta" file in here and then click



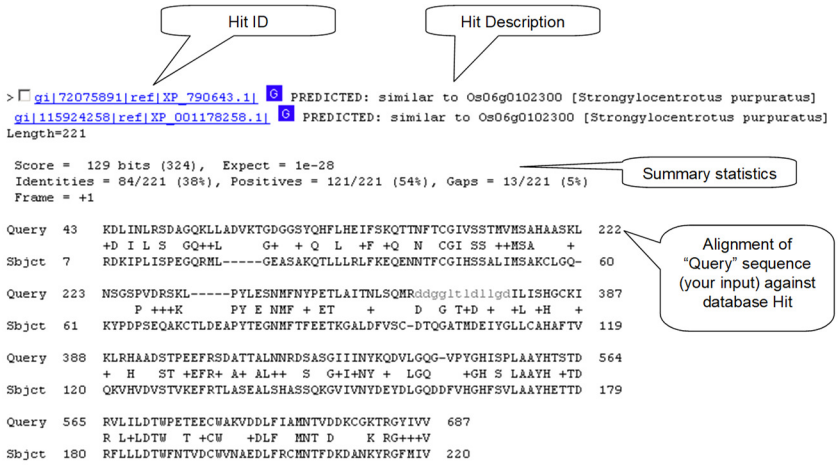
5.9 Output



Hit Table



Alignments



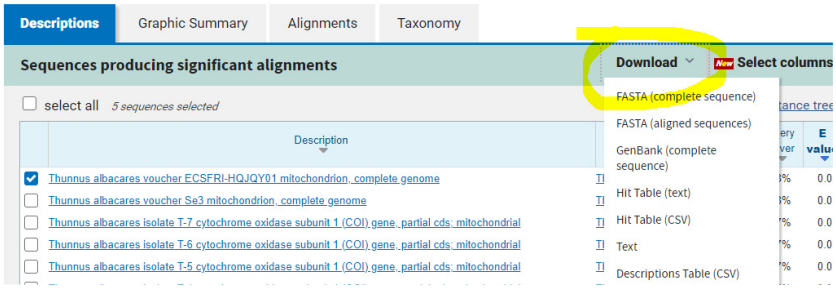
At this stage, if you have failed to retrieve a good sequence or your blast hit shows contamination (bacterial?). Record it and consider the result for your report. However, to conduct the following tasks you should go back to your dataset and pick an alternative sequence. Keep in mind, that you can use the full dataset for your report.

5.12 Download close homologs from your Blast search representing relevant species

Do this once. Select 5 (COI) sequences representing (up to five) different species, including at least one representative of your species of interest (highest hit):

Lumbricus rubellus L2 voucher INA132 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	580	580	100%	2e-162	96%	JQ909114.1
Lumbricus rubellus L2 voucher INA149 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	580	580	100%	2e-162	96%	JQ909110.1
Lumbricus rubellus voucher 1413 cytochrome c oxidase subunit I (COI) gene, partial cds; mitochondrial	580	580	99%	2e-162	96%	JN869925.1
Lumbricus rubellus voucher 1412 cytochrome c oxidase subunit I (COI) gene, partial cds; mitochondrial	580	580	99%	2e-162	96%	JN869924.1
Lumbricus rubellus strain Lr1 cytochrome c oxidase subunit I (COI) gene, partial cds; mitochondrial	580	580	99%	2e-162	96%	FJ214213.1
Lumbricus rubellus L2 voucher INA122 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	575	575	100%	1e-160	96%	JQ909127.1
Lumbricus rubellus voucher 1420 cytochrome c oxidase subunit I (COI) gene, partial cds; mitochondrial	575	575	100%	1e-160	96%	JN869932.1
Lumbricus castaneus isolate 1 cytochrome oxidase subunit I (COI) gene, partial cds; mitochondrial	575	575	100%	1e-160	96%	DQ092905.1
Lumbricus rubellus isolate 15 cytochrome oxidase subunit I (COI) gene, partial cds; mitochondrial	569	569	100%	5e-159	96%	DQ092903.1
Lumbricus rubellus isolate 12 cytochrome oxidase subunit I (COI) gene, partial cds; mitochondrial	569	569	100%	5e-159	96%	DQ092901.1
Lumbricus rubellus voucher 1419 cytochrome c oxidase subunit I (COI) gene, partial cds; mitochondrial	470	470	98%	5e-129	91%	JN869931.1
Lumbricidae sp. DPEW54597 voucher EW-SJ-650 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	342	342	98%	1e-90	84%	GU014047.1
Hoplochaetella stuarti voucher EW-HS-02-1-KOLLI_HILLS cytochrome oxidase subunit I (COI) gene, partial cds; mitochondrial	331	331	98%	2e-87	84%	JN793519.1
Hoplochaetella stuarti isolate SR-04-E2A cytochrome C oxidase subunit I (COX1) gene, partial cds; mitochondrial	331	331	98%	2e-87	84%	JN887890.1
Hoplochaetella stuarti isolate SR-04-E2B cytochrome C oxidase subunit I (COX1) gene, partial cds; mitochondrial	331	331	98%	2e-87	84%	JN887891.1
Aporrectodea nocturna isolate AnBRE1 cytochrome oxidase subunit I (COI) gene, partial cds; mitochondrial	326	326	98%	1e-85	84%	JQ763493.1
Aporrectodea nocturna isolate AnBRE2 cytochrome oxidase subunit I (COI) gene, partial cds; mitochondrial	326	326	98%	1e-85	84%	JQ763494.1
Eisenia nordenskioldi aff. nordenskioldi 3 SVS-2012 isolate E0246 cytochrome oxidase subunit 1 (cox1) gene, partial cds; mitochondrial	318	318	97%	2e-83	83%	JX531552.1
Eisenia nordenskioldi aff. nordenskioldi 3 SVS-2012 isolate E0244 cytochrome oxidase subunit 1 (cox1) gene, partial cds; mitochondrial	313	313	97%	9e-82	83%	JX531553.1
Eisenia nordenskioldi aff. nordenskioldi 3 SVS-2012 isolate E0056 cytochrome oxidase subunit 1 (cox1) gene, partial cds; mitochondrial	313	313	97%	9e-82	83%	JX531551.1
Metaphire soulensis mitochondrial COI gene for cytochrome oxidase subunit 1, partial cds, isolate: Msou-2196	292	292	94%	1e-75	82%	AB542667.1
Metaphire soulensis mitochondrial COI gene for cytochrome oxidase subunit 1, partial cds, isolate: Msou-1848 >dbjAB542668.1| Metaphire soulensis mitochondrial COI gene	292	292	94%	1e-75	82%	AB542665.1
Metaphire soulensis mitochondrial COI gene for cytochrome oxidase subunit 1, partial cds, isolate: Msou-1652	292	292	94%	1e-75	82%	AB542664.1
Metaphire soulensis mitochondrial COI gene for cytochrome oxidase subunit 1, partial cds, isolate: Msou-1718 >dbjAB542666.1| Metaphire soulensis mitochondrial COI gene	292	292	94%	1e-75	82%	AB542663.1
Metaphire soulensis mitochondrial COI gene for cytochrome oxidase subunit 1, partial cds, isolate: Msou-1814	292	292	94%	1e-75	82%	AB542662.1
Amynthas purpuratus mitochondrial COI gene for cytochrome oxidase subunit 1, partial cds, isolate: Apur-1851	292	292	93%	1e-75	83%	AB542524.1
Amynthas purpuratus mitochondrial COI gene for cytochrome oxidase subunit 1, partial cds, isolate: Apur-1852	287	287	93%	5e-74	82%	AB542525.1
Megascolecidae sp. DPEW51284 voucher EW-SJ-416 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	287	287	99%	5e-74	81%	GU014031.1
Lumbricidae sp. DPEW54242 voucher EW-SJ-408 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	285	285	98%	2e-73	81%	GU014046.1
Amynthas purpuratus mitochondrial COI gene for cytochrome oxidase subunit 1, partial cds, isolate: Apur-1662	281	281	97%	2e-72	81%	AB542523.1

Now return to the menu at the top to the significant hit list and select Download>FASTA (complete sequences)>continue. Save under an appropriate name and in a appropriate local.



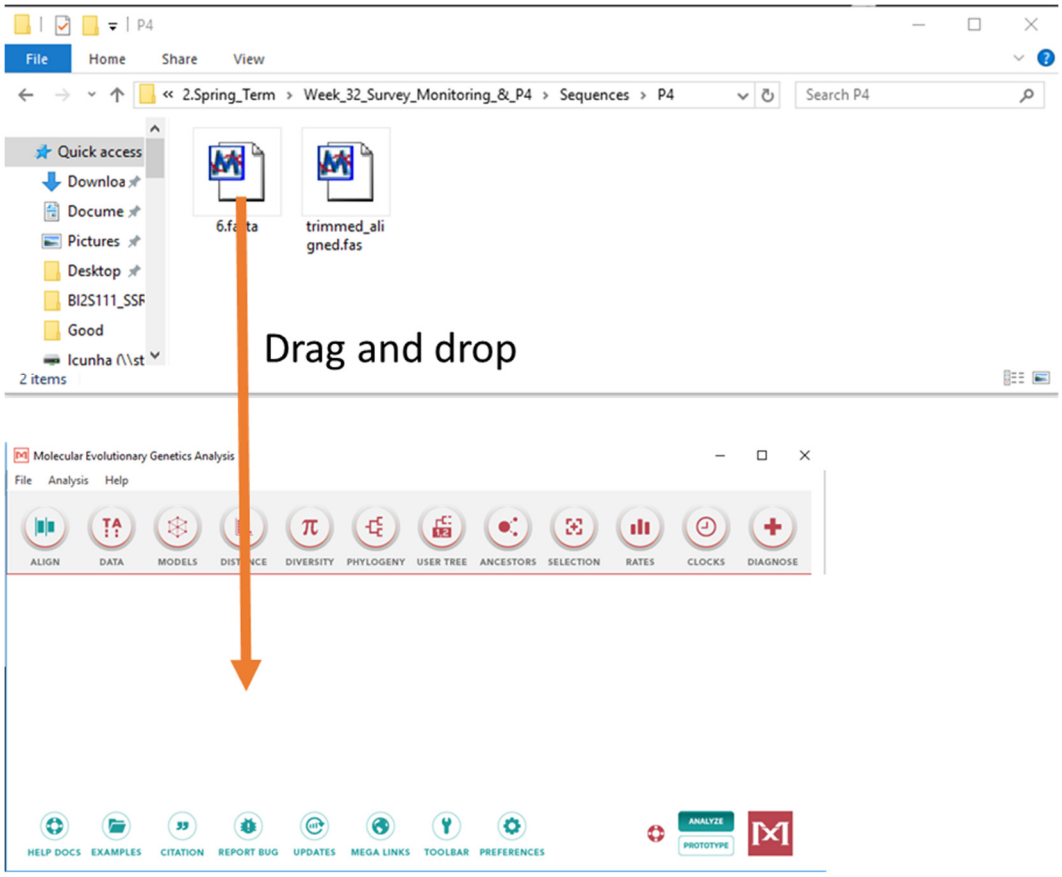
5.13 Edit sequence names as appropriate

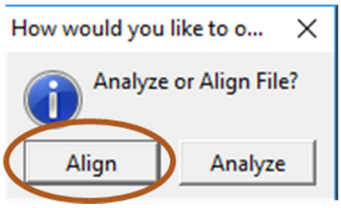
Using the best hit (reflected by the highest identity and lowest e-score) from blast search (homology search), rename the name of each sequence with its putative specie name found.

6. ClustalW alignment in Mega

6.1 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"





6.2 Align Sequences

Select all sequences

Edit>Select All

Align Sequences

Alignment > Align by ClustalW

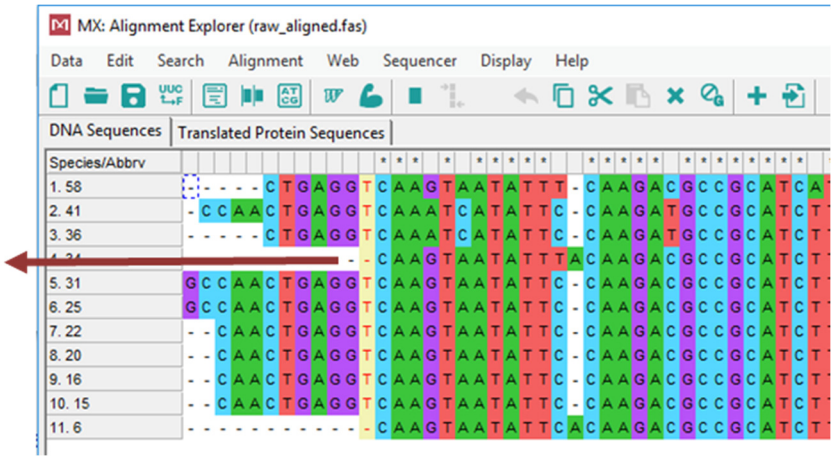
Keep default alignment parameters

>OK

6.3 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

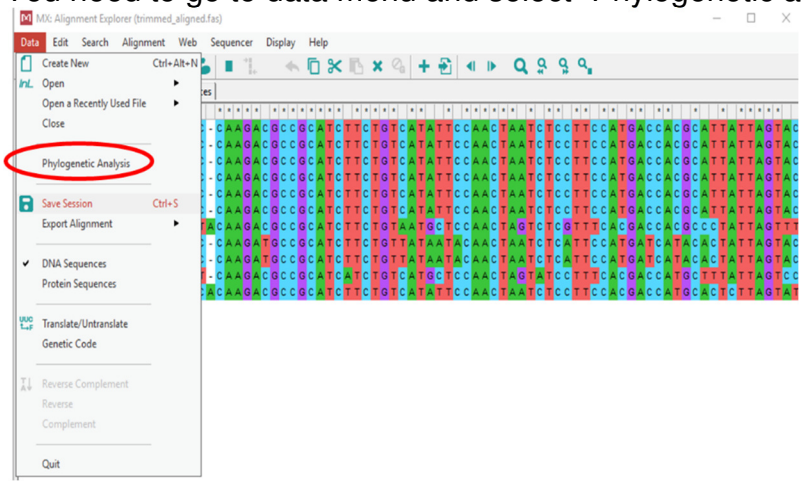
>Data>Export>MEGA format (or fasta format)

Name the sequence and save.

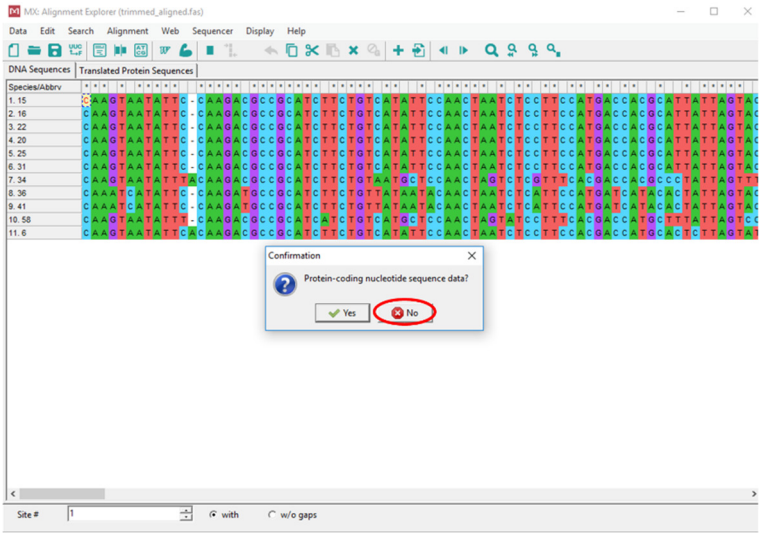
7. Phylogenetic analysis with MEGA (Use MEGA X)

7.1 Using your alignment for analysis

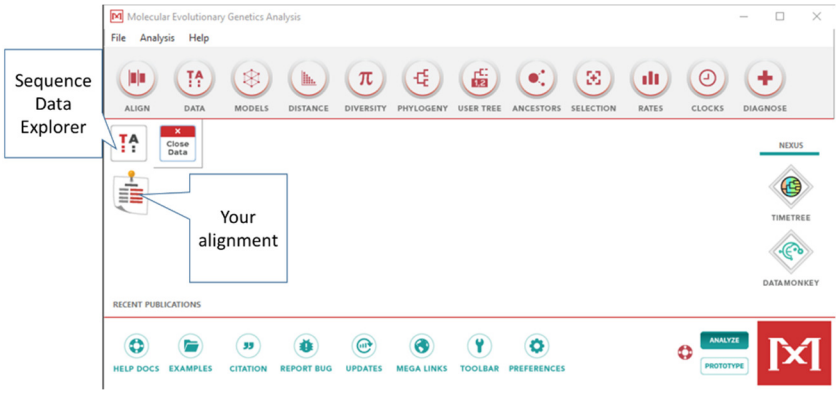
You need to go to data menu and select “Phylogenetic analysis”



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



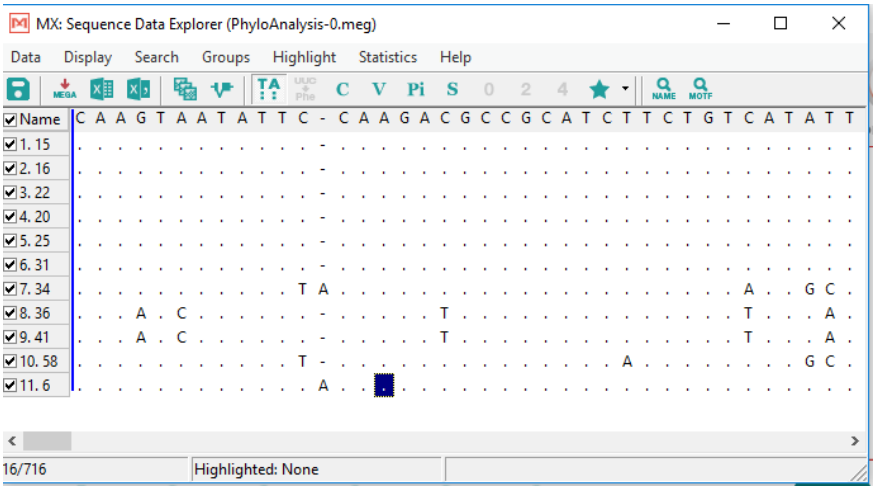
Now, go back to the Mega-X main window and you should see this:



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

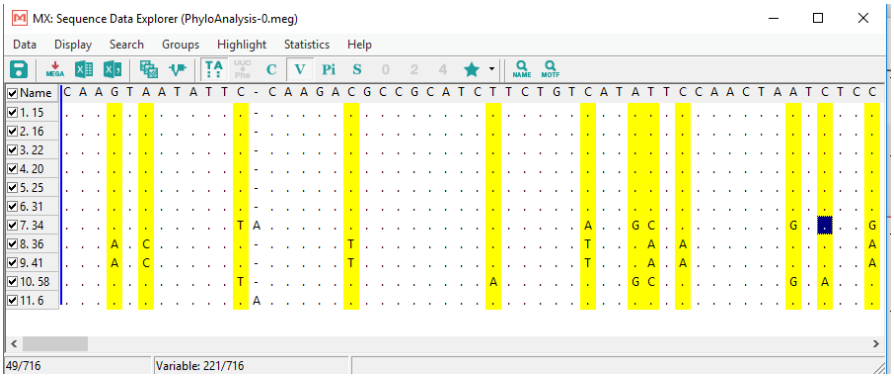
7.2 Data Analysis

View the **Sequence Data Explorer** window:

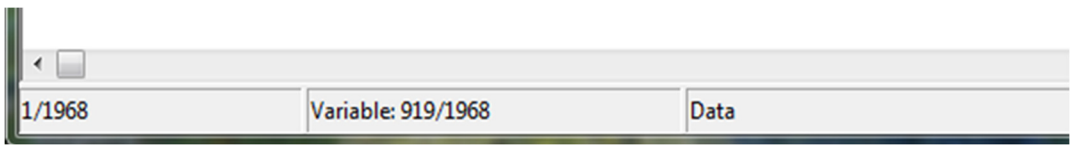


You can use the  button on the menu bar to toggle between showing all nucleotides and polymorphic positions only. A colour option is also available.

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



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

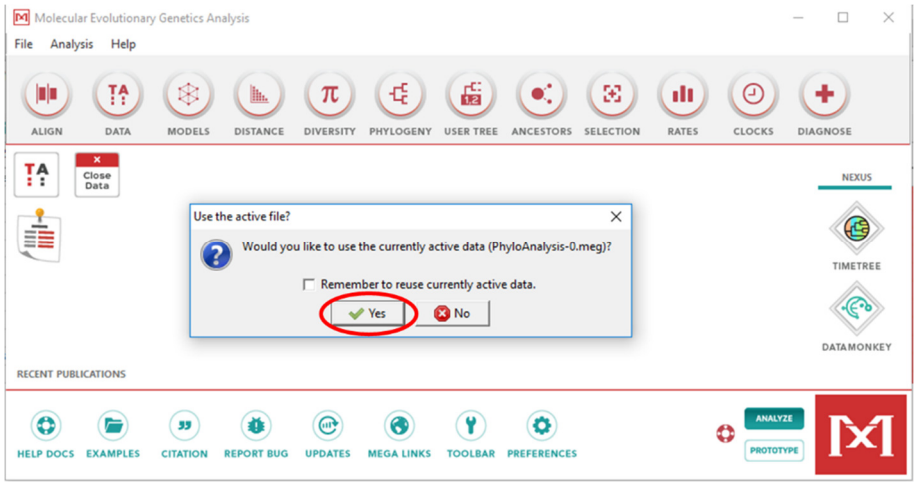


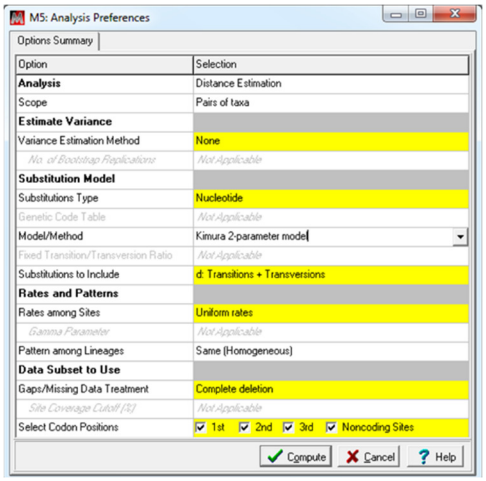
Click the translate menu button to view amino acid translations:

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…:**

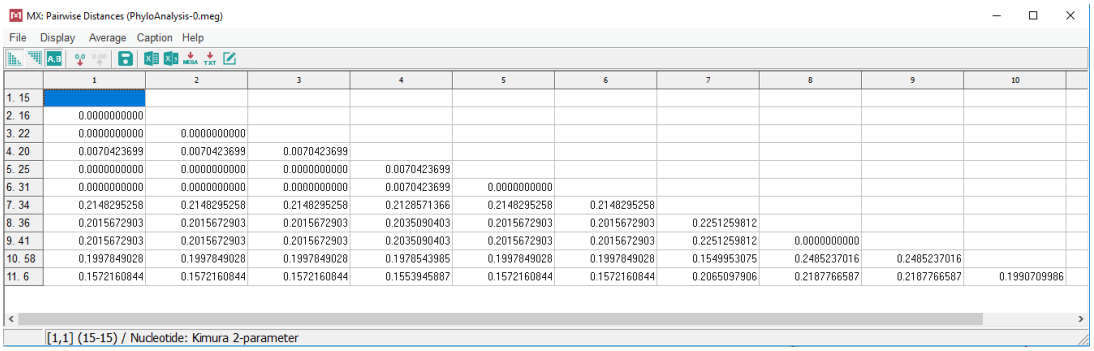
Use currently active data:





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.

7.3 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