

Exercise 2: Creating a phylogenetic tree using the maximum likelihood method

The first step in creating a phylogenetic tree is to collect sequence data for analysis. For whole genomes, DNA genomic sequences are obtained from databases like GenBank, collated together, and then used to create trees. For individual genes, either the DNA sequences or the amino acid sequences are obtained and then used to produce phylogenetic trees. For example, the cytochrome c protein sequence, a protein found in prokaryotes and eukaryotes and involved in the electron transport chain can be used to determine evolutionary relationships between these organisms. [1]

Fasta sequences contain the linear sequence of all nucleotides (DNA) or amino acids (proteins) and are the basic input used to produce phylogenetic trees. In this exercise, you will follow the steps below to collect fasta sequences and to create a tree using the program phylML. Be sure to read each step carefully and ensure that the fasta sequences are formatted correctly.

Exercise 2A: Building a Fasta file for Cytochrome C protein sequences

Collect the cytochrome C protein sequences in a text file using the Notepad program on a PC or the TextEdit program for Macs. You must use these programs as they will produce files that are “clean” and do not have additional formatting often added by other programs.

Step 1: Go to the National Institutes of Health National Center for Biotechnology site (NCBI) to search the protein sequence bank for the sequences in Table 1 (see step 2)

<https://www.ncbi.nlm.nih.gov/protein/>

Table 1. Cytochrome C protein accession numbers for a diverse set of eukaryotic organisms.

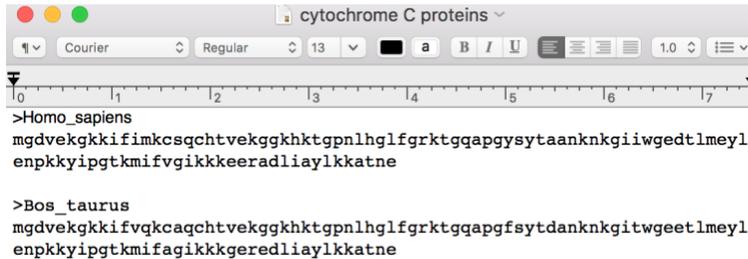
Species	Accession number
Homo sapiens	CAG46972
Bos taurus	NP_001039526
Equus caballus	NP_001157486
Orcinus orca	XP_004265696
Apis mellifera	NP_001170961
Salmo trutta	XP_029596713
Gorilla gorilla	XP_018886566
Drosophila melanogaster	AAA28437

Step 2: Paste the accession number from Table 1 into the search line and click search. A page with information regarding the sequence will appear and the amino acid sequence will be at the bottom.

Step 3: Copy the sequence into your Notepad or TextEdit file. Be sure to label each sequence correctly- see the example below.

First line has >name

Second line starts with amino acid sequence. Must have no spaces, numbers or hard returns in the sequence.



```
>Homo_sapiens
mgdvekqkkifimkcsqchtvekggkxhktgpnlhglfgrktggapgsysyaanknkgiiwgedtlmeyl
enpkkiyigtkmifvgikkkeeradliaylkkatne

>Bos_taurus
mgdvekqkkifvqkcaqchtvekggkxhktgpnlhglfgrktggapgsysydanknkgitwgeetlmeyl
enpkkiyigtkmifagikkkeredliaylkkatne
```

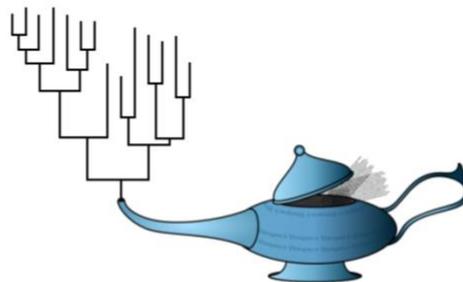
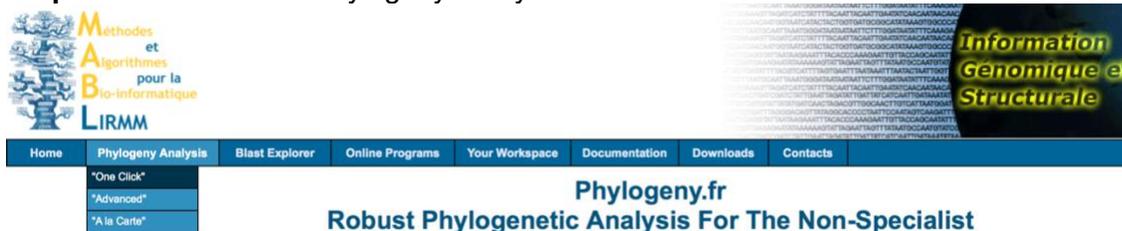
Step 4: Be sure to save this fasta file on your local computer or in a cloud storage environment for use in Exercise 3.

Step 5: Submit your fasta file to your instructor or copy and paste it into your worksheet

Exercise 2B: Use your Fasta file to build a Cytochrome C protein tree

To build a tree using the maximum likelihood method go to www.phylogeny.fr and follow the steps below

Step 1: Under the tab 'Phylogeny Analysis' select 'One Click'



Phylogeny.fr is a free, simple to use web service dedicated to reconstructing and analysing phylogenetic relationships between molecular sequences.

Step 2: Paste your fasta file sequences (Step 4 in Exercise 2A above) from the Notepad or TextEdit program into the box

Step 3: Uncheck the box for 'Use the Gblock program...'

Step 4: Click the submit button

Home Phylogeny Analysis Blast Explorer Online Programs Your Workspace Documentation

"One Click" Mode

Alignment MUSCLE ⇌ Phylogeny PhyML ⇌ Tree Rendering TreeDyn

1. Overview 2. Data & Settings

Name of the analysis (optional):

Upload your set of sequences in FASTA, EMBL or NEXUS format from a file:

Choose File no file selected

Or paste it here (load example of sequences)

Clear

Maximum number of sequences is 200 for proteins and 200 for nucleic acids.
Maximum length of sequences is 2000 for proteins and 6000 for nucleic acids.

Use the Gblocks program to eliminate poorly aligned positions and divergent regions

To receive the results by e-mail, enter your address(es):

Submit

A tree will be created that includes bootstrapping values.

After creating the cytochrome C protein tree in phylogeny.fr use the resulting information to answer the following questions.

Question Set 4:

- 2.1 Submit a screen shot of your tree by pasting it into your worksheet.
- 2.2 Which branch are you most confident in? Justify your response.
- 2.3 Which branch are you least confident in? Justify your response.
- 2.4 Did you produce a rooted or unrooted tree? How do you know?
- 2.5 After the gorilla, which organism does your tree predict is most closely related to humans?
- 2.6 There are many types of phylogenetic trees available on the internet and in other scientific resources. What distinguishes them is the information used to construct them (genomic data, protein sequence data, phenotypic data, etc.). Compare and contrast your tree to the one found [here](#) and discuss similarities and differences in the two trees.

References:

1. Fitch WM. The molecular evolution of cytochrome c in eukaryotes. *J Mol Evol.* 1976;8(1):13-40. doi:10.1007/BF01738880