**Bioinformatics:**

**Investigating Sequence Similarity**

EXERCISE 3

**Phylogenetic Analysis of Homologous Sequences**

**Objectives**

After completing this exercise, you should be able to:

1. Create a dissimilarity matrix and multiple sequence alignment
2. Create a phylogram based on similarity of amino acid sequences.
3. Distinguish between a rooted and unrooted phylogenetic tree.

Humans and chimpanzees are very closely related, so similarity in amino acid sequences is expected. The degree of similarity in amino acid sequences should reflect the evolutionary distance between two species; the more distantly related, the more dissimilar the sequences. In this exercise, we will use some of the same techniques, in addition to a few new ones, to compare the **cytochrome C** amino acid sequence among several eukaryotes. Then, using the similarity information we will build a phylogenetic tree.

Constructing phylogenetic trees based on molecular data can provide additional insight into evolutionary relationships and even in some cases change historical phylogeny based on morphological characteristics. For example, molecular data has led to the creation of a new subphylum known as *Acoelomorpha*, for a group of soft bodied flatworms that had traditionally been classified as platyhelminths (Ruiz-Trillo et al., 2004). Thus, sequence alignment coupled with use of a distance matrix to create a phylogenetic tree can be useful for validating phylogenetic relationships or assigning phylogeny to a newly identified organism.

**Collect Sequence and BLAST Data**

**Computational Procedure:**

The first step in constructing a molecular phylogenetic tree is to target a conserved protein to use as a molecular clock. In this case, we will use the protein cytochrome C, a small protein which functions as an electron carrier in the electron transport chain, within the inner mitochondrial membrane of eukaryotes. Find the amino acid sequence for human cytochrome C using the steps below.

1. Using Firefox as your browser, return to the NCBI home page (ncbi.nlm.nih.gov).
2. Type “**NP\_061820.1**” in the textbox toward the top of the page. This is the NCBI accession number associated with Human cytochrome C protein. Click “search” in order to search for the record associated with this accession number.

The second step in constructing a molecular phylogenetic tree is to carry out a series of pairwise sequence alignments, with a program such as BLAST, to find homologs in a diverse group of species.

1. On the protein record page for Human cytochrome C protein, find the “Run BLAST” link under the “Analyze this Sequence” heading in the column toward the right-hand side of the page.
2. On the BLASTp page, you will align the Human cytochrome C protein against a database of your choice. For example, you could look for cytochrome C homologs in *Zea mays* (corn). To do this change the **Database** to “refseq\_protein” and type “*Zea mays*” in the text box next to **Organism** and click “BLAST”.
3. On the BLASTp results page find the *Zea mays* cytochrome C homolog and obtain the FASTA formatted amino acid sequence. To do this, click on the NP\_ accession number associated with *Zea mays* cytochrome C to obtain the protein record. Next, click on the “FASTA” link toward the top of the page and copy & paste the sequence along with the identifiers (Figure 3) into a new Notepad (TextEdit for Mac) file.
4. Repeat this process for the cytochrome C homolog found in *Danio rerio*, a popular fish used as a developmental research model organism, and copy & paste the respective amino acid sequence into the same Notepad file that currently holds the *Zea mays* amino acid sequence.

>gi|293335855|ref|NP\_001170028.1| cytochrome c [Zea mays] MASFSEAPPGNPKAGEKIFKTKCAQCHTVDKGAGHKQGPNLNGLFGRQSGTTAGYSYSAGNKNKAVVWEE DTLYEYLLNPKKYIPGTKMVFPGLKKPQERADLIAYLKEATA

**Figure 3** Example FASTA formatted amino acid sequence with identifiers

Conceivably, a scientist could continue this process with a collection of target species to be used for phylogenetic analysis. To expedite this process, we will obtain cytochrome C amino acid sequences from a collection of species using the HomoloGene database at NCBI. HomoloGene is a database that can be utilized to detect homologs in 20 sequenced eukaryotic genomes.

**Navigate to HomoloGene at:** [http://www.ncbi.nlm.nih.gov/HomoloGene](http://www.ncbi.nlm.nih.gov/homologene)

1. Type the Human cytochrome C accession number (NP\_061820.1) into the text box toward the top of the page and click “Search”.
2. The resulting page will bring up a number of genes and their predicted proteins with the same name found in various species. In the protein column, one may click on the NP\_ or XP\_ prefixed accession number to obtain FASTA formatted amino acid sequence.

Repeat this process for the following species and copy & paste the sequence into your expanding Notepad file:

* + 1. *Homo sapiens* – human
		2. *Macaca mulatta* - Rhesus monkey
		3. *Bos taurus* – cattle
		4. *Gallus gallus* – chicken
		5. *Xenopus tropicalis* – western clawed frog
		6. *Drosophila melanogaster* – fruit fly
		7. *Saccharomyces cerevisiae* – budding yeast

**Execute a Multiple Sequence Alignment and Generate a Phylogram**

The third step in constructing a molecular phylogenetic tree is to execute a **multiple sequence alignment** (MSA). A MSA directly aligns three or more sequences that have similarity. In addition to the number of sequences being compared MSA’s differ from pairwise sequence alignment algorithms, such as BLAST, in that they do not cut the sequences up into short segments, thus the sequences must be of similar length.

Once the MSA is executed, a number of computational methods can be applied to generate a phylogenetic tree such as a phylogram or cladogram from the alignment. While a **cladogram** represents an evolutionary branching pattern, the branch length does not represent evolutionary time as it does in a phylogram. **Phylograms** are distance based and measure pairwise differences among sequences and generate a tree from a matrix made with these values. A molecular phylogram is scaled with the branch length (**Figure 4**) representing the amount of evolutionary divergence between sequences.



**Figure 4.** Image from NCBI's[*Science Primer Phylogenetics Factsheet*](http://www.ncbi.nlm.nih.gov/About/primer/phylo.html)

The simplest method for making a phylogram is the **distance method**. In this method, a **distance matrix** is set up by individually comparing similarity in each sequence with the other sequences and calculating the fraction of identical bases.

**Example MSA:**

Seq A MGDVEKGKKIFVMKCSQCHTVEKGG

Seq B MVDVEKGMKIFVMKCSQCHTVEAGG

Seq C MVDVEKGMLIFVMKCSQCHTVEAGG

Seq D MGDIEKGKLIFVMSCSQCHTVYAGY

Seq E MIDTEKGYIIFVMTCSQCHTVLMGT

From the aligned sequences above we can individually calculate the fraction of differing amino acid residues between the 25 amino acids that make up Seq A and B (identical residues represented by “|”):

Seq A MGDVEKGKKIFVMKCSQCHTVEKGG

 | ||||| |||||||||||||| ||

Seq B MVDVEKGMKIFVMKCSQCHTVEAGG

This value (3/25 = 0.12) is then entered in the Distance Matrix below:

**Distance Matrix**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Seq A | Seq B | Seq C | Seq D | Seq E |
| Seq A | - | - | - | - | - |
| Seq B | 0.12 | - | - | - | - |
| Seq C |  |  | - | - | - |
| Seq D |  |  |  | - | - |
| Seq E |  |  |  |  | - |

Using the MSA Calculation Template on the Exercise 3 Worksheet determine the fraction of identical bases between the other possible combinations of amino acid sequences and fill in the blanks in the table. Show your calculations on the MSA calculation template.

The distances from the matrix can now be used to make a phylogram by joining the most similar sequences. This specific type of distance method is called “**neighbor-joining**”. The basic structure of the phylogram is first inferred and the branch lengths will later be determined. The phylogram starts out with a single node as shown below.



The most similar sequences (“neighbors”) are “joined” onto a single branch. The distance matrix is then recalculated by averaging the distances relative to sequences B and C (bold values in table below). For example, the distance between Seq B and Seq D is 0.28, while the distance between Seq C and Seq D is 0.24. Thus, the distance between Seq B/C and Seq D is the average distance between Seq D and Seqs B and C (0.28+0.24/2).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Seq A | **Seq B/C** | Seq D | Seq E |
| Seq A | - | - | - | - |
| **Seq B/C** | **0.14** | - | - | - |
| Seq D | 0.2 | **0.26** | - | - |
| Seq E | 0.32 | **0.32** | 0.32 | - |



In the new distance matrix, sequences B/C and A are most similar, thus they are joined below in the new iteration of the tree.



At this point the phylogram is completely resolved with three branches extending from each node. This leaves us with an unrooted tree since these species have an unknown common ancestor. If we had a known organism that is distantly related to the other organisms being analyzed, we could use this species as an outgroup to root the tree.

What species in your list of sequences derived from HomoloGene could you utilize as an “outgroup” if you wanted to build a rooted tree? **(Worksheet Question 1)**

Now that tree topology and branch order has been obtained calculations can be made to determine the relative length of the branches to represent evolutionary time. Generally, evolutionary time in a molecular tree is represented as the number of average changes per amino acid residue (see figure with a scale below).



Now that you have been introduced to the neighbor joining method, lets apply it to the amniotes in your list of collected sequences for cytochrome C, using *Drosophila melanogaster* as an outgroup.

Additionally, statistical tests may also be applied to phylograms to compute the confidence levels for tree branch support. In the exercise below, a simple web-based computational algorithm will be utilized without statistical analysis due to the high computational cost (time) needed to carry out such an analysis.

Besides the neighbor joining distance method outlined above, other more complex computational approaches may be used by biologists to generate a phylogram. These computational algorithms are beyond the scope of this introductory activity, thus we will next focus on using a web-based, neighbor-joining algorithm to create a phylogenetic tree.

Follow the protocol below to execute a MSA and generate a phylogram using the **neighbor-joining method** using a computational tool**.**

**Computational Procedure:**

1. Go to the Clustal Omega home page at [www.ebi.ac.uk/Tools/msa/clustalo/](http://www.ebi.ac.uk/Tools/msa/clustalo/) (it is suggested to use Firefox for a web browser for functionality of Jalview later in the activity). This website is a one stop web-based MSA and phylogenetic tree generator (Sievers et al., 2011). We will use the default settings in this application for simplicity, but the program does allow for customization.
2. In the text box, paste your collection of FASTA formatted amino acid sequences you collected for homologs of cytochrome C. Modify the identifiers for each sequence in such a way in which only the species name remains and click “Submit” (**Figure 5**).

>Zea mays MASFSEAPPGNPKAGEKIFKTKCAQCHTVDKGAGHKQGPNLNGLFGRQSGTTAGYSYSAGNKNKAVVWEE DTLYEYLLNPKKYIPGTKMVFPGLKKPQERADLIAYLKEATA

**Figure 5.** Example FASTA formatted amino acid sequence with modified identifiers

1. The results page consists of a series of tabs, with the “Alignments” tab view set as the default. Click on the “Show Colors” button under the tabs.

Please note how amino acids single letter designations are colored by shared properties.

AVFPMILW-**Red**: Small (small + hydrophobic [includes aromatic –Y])
DE-**Blue**: Acidic
RHK-**Magenta**: Basic
STYHCNGQ -**Green**: Hydroxyl +Amine+ Basic + Q

The MSA also includes an additional row under each alignment that denotes the degree of conservation at each amino acid position through the use of consensus symbols.

**Consensus Symbols:**

**\* (asterix)** means that the residues or nucleotides in that column are identical in all sequences in the alignment.

**: (colon)** means that conserved substitutions have been observed, according to the color table above

**. (period)** means that semi-conserved substitutions are observed.

Go back to the “Alignments” tab and take a screenshot of the MSA, place it in the worksheet under **Cytochrome C MSA**.

1. Next, click on the “Results Summary” tab. Under the “Result files” heading you will find a link to a pairwise **percent identity matrix**. The resulting matrix was created by converting the distance matrix, which was generated using a very similar strategy as we did manually earlier in this exercise, to an identity matrix. To do this, dissimilarly matrix values were subtracted from 1 and then converted into a percentage.

Using the percent identity matrix, which two organisms have the least evolutionary time separating them based on the molecular data analyzed? **(Worksheet Question 2)**

1. Now construct a phylogenetic tree of the MSA. Within the “Phylogenetic Tree” tab, find the stereo button to switch between Cladogram and Real (phylogram) to view relationships and relative evolutionary distance respectively. To obtain a customizable and interactive phylogenetic tree for easier interpretation, you may:

**(A)** navigate to the “Result Summary” tab, find the box labeled “Jalview” and click on the “Start Jalview” button if Java is enabled in your web-browser and continue with the following steps:

1. A new browser window will open with a more interactive and dynamic MSA visual. Within this window click on the “Calculate” menu and navigate through the following submenus: Calculate>Calculate Tree>Neighbor Joining Using % ID
2. A new window will open with a visual representation of the MSA sequence percent identity in the form of a phylogenetic tree. Take a screenshot of the tree, place it in the worksheet under Cytochrome C Phylogenetic Tree

**In my experience, Jalview does not display on my browser when I use my Mac. I use the procedure below to complete the construction of phylogenetic tress.**

**Or** if Java is not enabled in your web-browser, you may **(B)** go ahead and use an alternative web-based Phylogenetic tree generator program hosted by the Virus Pathogen Database and Analysis Resource (VIPR): <https://www.viprbrc.org/brc/tree.spg?method=ShowCleanInputPage&decorator=flavi_zika>

1. Input a name for your phylogenetic tree
2. Choose Quick Tree followed by the sequence type being analyzed.
3. Choose “Paste sequence in FASTA” and cut and paste your sequences into the text box and select “unaligned FASTA” prior to clicking “Build Tree”.
4. On the following screen choose “Archaeoptryx-js as the tree viewer mode and click “View Tree”.
5. A new window will open with a visual representation of the MSA sequence percent identity in the form of a phylogenetic tree. Take a screenshot of the tree, place it in the worksheet under Cytochrome C Phylogenetic Tree.