Investigating Sequence Similarity Exercise 3

# Instructor Preface

This document includes:

1. A student handout / worksheet containing a compilation of the questions integrated throughout the student handout for Exercise 3 that students can complete and turn in for a post-lab assignment
2. The instructor solutions key for Exercise 3

Instructors should distribute the student handout / worksheet to students, while retaining the instructor solutions key.

Investigating Sequence Similarity

EXERCISE 3

**Phylogenetic Analysis of Homologous Sequences**

**Objectives**

After completing this exercise, you should be able to:

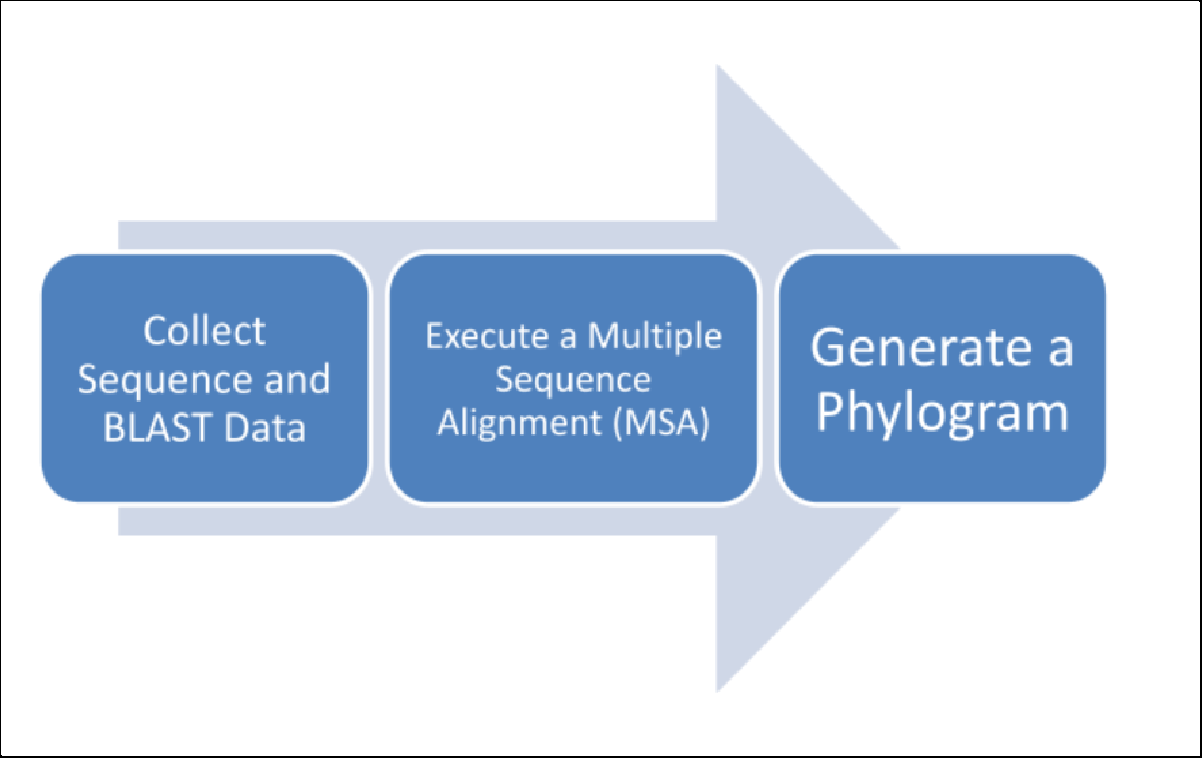
1. Create a distance matrix and multiple sequence alignment
2. Create a phylogram based on similarity of amino acid sequences.
3. Distinguish between a rooted and unrooted phylogenetic tree.

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 closer the relation, the more similarities should be expected. Similarly, the more distantly related, the more dissimilar the sequences will be.

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(1).

Constructing phylogenetic trees based on molecular data can provide additional insight into evolutionary relationships between species. It can even, in some cases, change historical phylogeny that had previously been 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(2) - coupled with use of a distance matrix(3) to create a phylogenetic tree - can be useful for validating phylogenetic relationships or assigning phylogeny to a newly identified organism.



**Overview of the module.**

# Collect Sequence and BLAST Data

#### Computational Procedure:

The first step toward constructing a molecular phylogenetic tree is to target a conserved protein(4) to use as a molecular clock(5).

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 a web 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 toward constructing a molecular phylogenetic tree is to carry out a series of pairwise sequence alignments, with a program such as BLAST(6), to find homologs(7) 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. First, obtain the FASTA formatted amino acid sequence.

“FASTA” is a method of formatting amino acids in a way that many online programs can read/compute with.

1. To do this, find the NP\_accession number associated with *Homo sapiens* chromosome c and find the “FASTA” link at the top of the page. Copy and paste this into a separate document, NotePad file, or other such recording platforms. You will need this for later, so make sure to hang onto this file!
2. Next, 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. Click on the “FASTA” link toward the top of the page and copy & paste the sequence along with the identifiers (Figure 3) into that separate file with the *Homo sapien* data.

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

**Figure 3** Example FASTA formatted amino acid sequence with identifiers. Note that there should be no text break of any kind between the lines, however they are condensed here for format sake. Having line/paragraph breaks may cause an error later, when computing the piecewise function.

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 (Including *Zea mays*) :

* 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

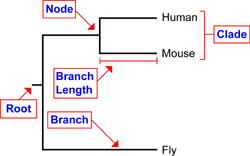
The third step toward constructing a molecular phylogenetic tree is to execute a **multiple sequence alignment** (MSA)(8). A MSA directly aligns three or more sequences that have similarity. In addition to the number of sequences being compared MSA’s differ from pairwise sequence alignment algorithms, 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.

A **cladogram** represents an evolutionary branching pattern in which the branch length *does not* represent evolutionary time.

**A phylogram** is a distance based tree that measures pairwise differences among sequences, and generates 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)*.* Here, this molecular phylogram shows that Humans, Mice, and Flies share a similar root (a common ancestor) but because the Mice and the Humans share a clade, their molecular sequences diverged later in evolution while the Fly’s sequence diverged earlier in evolution.

The simplest method for making a phylogram is the **distance method**(9). In this method, a **distance matrix** is set up by individually comparing similarities in each sequence with the other sequences, and then 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

When the amino acid in both sequences is identical, it is represented by a “|” in the space between. If they are different, they do not get marked with a “|” and should be counted for later use. For example:

### Seq A MGDVEKGKKIFVMKCSQCHTVEKGG

| ||||| |||||||||||||| || Seq B MVDVEKGMKIFVMKCSQCHTVEAGG

Between Seq A and Seq B, there are three (3) instances where the amino acids of the two sequences are not the same. We use this number to find the fraction of amino acid differences in the sequences. The lower the fraction, the more similar the sequence is. This fraction (number of differences / total amino acids) is the “distance” value.

For this example, the distance value is 3/25 = 0.12. 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 |  |  |  |  | - |

**Directions**: Calculate the following alignment identity values and place them into the Distance Matrix above.

Seq A MGDVEKGKKIFVMKCSQCHTVEKGG Seq C MVDVEKGMLIFVMKCSQCHTVEAGG

Alignment Value

Seq A MGDVEKGKKIFVMKCSQCHTVEKGG Seq D MGDIEKGKLIFVMSCSQCHTVYAGY

Alignment Value

Seq A MGDVEKGKKIFVMKCSQCHTVEKGG Seq E MIDTEKGYIIFVMTCSQCHTVLMGT

Alignment Value

Seq B MVDVEKGMKIFVMKCSQCHTVEAGG Seq C MVDVEKGMLIFVMKCSQCHTVEAGG

Alignment Value

Seq B MVDVEKGMKIFVMKCSQCHTVEAGG Seq D MGDIEKGKLIFVMSCSQCHTVYAGY

Alignment Value

Seq B MVDVEKGMKIFVMKCSQCHTVEAGG Seq E MIDTEKGYIIFVMTCSQCHTVLMGT

Alignment Value

Seq C MVDVEKGMLIFVMKCSQCHTVEAGG Seq D MGDIEKGKLIFVMSCSQCHTVYAGY

Alignment Value

Seq C MVDVEKGMLIFVMKCSQCHTVEAGG Seq E MIDTEKGYIIFVMTCSQCHTVLMGT

Alignment Value

Seq E MIDTEKGYIIFVMTCSQCHTVLMGT Seq D MGDIEKGKLIFVMSCSQCHTVYAGY

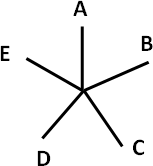
Alignment Value

**Make sure** all of your calculations are in the Distance Matrix above before

**continuing!**

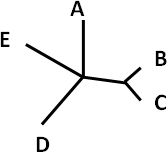
# Neighbor Joining and Building Your Phylogram

Now that you have calculated the distances in the distance matrix, you can make a phylogram by joining the most similar sequences. The smaller the value in the distance matrix, the closer and more related the two sequences are. This specific type of distance method is called “**neighbor-joining**”(10).

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.

This is an unrooted(11) phylogram tree. All species are present, but no branch lengths or other such relationships have been proven yet. This is why it looks like a star - don’t worry, we’ll fix this soon.

Next, the most similar sequences (“neighbors”) are “joined” onto a single branch. This is where we use our distance matrix! The initial distances serve to help us find what sequences are most alike - aka which sequences are “neighbors”.



Now that we have a more related structure, we are going to recalculate our distance matrix. Seq B and Seq C are very similar - they share a node. So, now we need to figure out how long the branch between the node and the closest value (in this case, Seq D) is. We do this by averaging the distances relative to sequences B and C.

The distance between Seq B and Seq D is 0.28. 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 = 0.26

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Seq A | **Seq B/C** | Seq D | Seq E |
| Seq A | - | - | - | - |
| Seq B | 0.12 | - | - | - |
| Seq C | 0.16 | - |  | - |
| **Seq D** | 0.2 | **0.26** | - | - |
| Seq E | 0.32 | 0.32 | 0.32 | - |

#### Now, do the same for Seq B/C and Seq A.

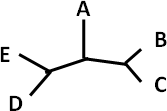
The distance between Seq B and Seq A is . The distance between Seq C and Seq A is .

Thus, the distance between Seq B/C and Seq A is the average distance between Seq A and Seqs B and C

( + / 2 = ).

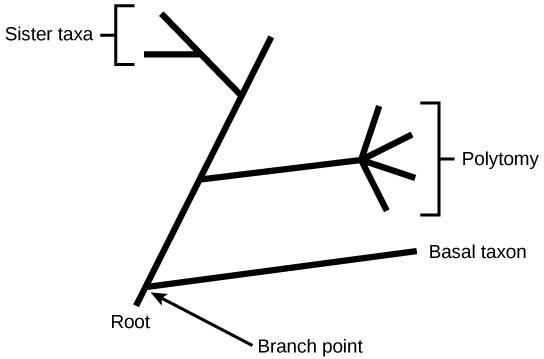
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Seq A** | Seq B/C | Seq D | Seq E |
| Seq A | - | - | - | - |
| Seq B/C |  | - | - | - |
| Seq D | 0.2 | 0.26 | - | - |
| Seq E | 0.32 | 0.32 | 0.32 | - |

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 has 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, we could use this species as an outgroup(12) to root the tree.

In the figure below, a rooted tree is represented,

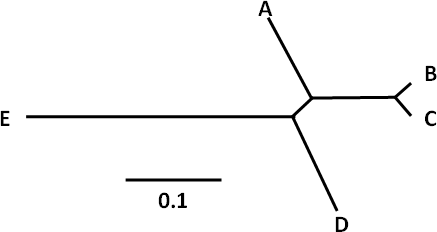
**Figure 4.** Image from OpenStax Biology 2e*.*

*What are the similarities and differences between an unrooted tree and a rooted tree?*

*Looking back on the list of organisms we created before (Homo sapiens, Zea mays, etc) what species could you utilize as an “outgroup” if you wanted to build a rooted tree?*

Now that tree topology(13) and branch order(14) 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).



The figure shows the difference in branch lengths between the sequences of the example above. For this, 0.1 is given a length - think of it as a scale in a map. If it is easier for you, feel free to make each 0.1 equal to an inch or a centimeter - that may make it easier to map out.

For example, Seq E has a distance value of 0.32 when looking at its relation to Seq D. If we use centimeters, 0.1 would equal 1 centimeter. Then, when drawing the tree, we would measure 3 ⅕ or 3.2 centimeters between Seq E and Seq D.

Fill in the chart below using this idea. If you have trouble, refer to the explanation above. A few have been done for you:

#### 0.1 = 1 cm

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 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 | - |

|  |  |  |
| --- | --- | --- |
|  | Distance Value | Centimeters |
| Seq A - Seq B/C | 0.14 | 1.4 cm |
| Seq A - Seq D |  |  |
| Seq A - Seq E |  |  |
| Seq B/C - Seq D |  |  |
| Seq B/C - Seq E |  |  |
| Seq D - Seq E | 0.32 | 3.2 cm |

Now that you have been introduced to the neighbor joining method, let's apply it to the real world data we collected via BLAST (i.e., human, Rhesus monkey, cattle, and chicken). In your list of collected sequences for cytochrome *c*, use *Drosophila melanogaster* as an outgroup.

**Fill in the following tables, which are focused on the species we chose earlier (with the exception of *Zea mays*). Use the *D. melanogaster* as an outgroup.**

<http://www.ncbi.nlm.nih.gov/homologene>

1. Go to Homologene and type the human **cytochrome C** accession number (NP\_061820.1) into the text box toward the top of the page and click “Search”.
2. Scroll down the page to the “Protein Alignments” section and click on “Show Pairwise Alignment Scores”.
3. Find the Percent Protein Identity **for each pair in the table below** and

**convert it to a dissimilarity value**(15) **(% non-identity divided by 100)**.

1. Part of the table has been pre-filled for you to speed up the process.

**Distance Matrix**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ------------ | *H. sapiens* | *M. mulatta* | *B. taurus* | *G. gallus* | *D. melanogaster* |
| *H. sapiens* | - | - | - | - | - |
| *M. mulatta* |  | - | - | - | - |
| *B. taurus* |  | 0.095 | - | - | - |
| *G. gallus* |  | 0.124 | 0.086 | - | - |
| *D.*  *melanogaster* |  | 0.170 | 0.157 | 0.167 | - |

**Distance Matrix – First Round of Neighbor Joining**

(Remember - Neighbor joining just refers to finding the average distances between two sequences. We did this earlier with Seq B and Seq C in the example.)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| -------- | *H. sapiens*  */*  *M. mulatta* | *B. taurus* |  |  |
| *H. sapiens*  */*  *M. mulatta* | - | - | - | - |
| *B. taurus* | 0.095 | - | - | - |
|  |  |  | - | - |
|  |  |  |  | - |

**Distance Matrix – Second Round of Neighbor Joining**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | - | - | - |
|  |  | - | - |
|  |  |  | - |

**Distance Matrix – Third Round of Neighbor Joining**

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | - | - |
|  |  | - |

#### Make sure all available squares have been filled before continuing!

*Below, briefly sketch the resulting neighbor joining phylogram (provide a scale for branch length, using whatever measurement is easier for you - it doesn’t have to be perfect, just try your best!)*

Additionally, statistical tests may also be applied to phylograms to compute the confidence levels(14) 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). This website is a one-stop web-based MSA and phylogenetic tree generator (Sievers et al., 2011).
2. We will use the default settings in this application for simplicity, but the program does allow for customization.
3. In the text box, paste the entire 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

4. 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**.

|  |  |  |
| --- | --- | --- |
| Bases | Color | Shared Properties |
| A, V, F, P, M, I, L, W | Red | Small, usually hydrophobic  ( sometimes includes aromatic base Y) |
| D, E | Blue | Acidic |
| R, H, K | Magenta | Basic |
| S, T, Y, H, C, N, G, Q | Green | Hydroxyl +Amine + Basic + Q |

The MSA also includes an additional row under each alignment that denotes the degree of conservation(16) at each amino acid position using consensus symbols.

**Consensus Symbols:**

**\* (asterisk)** 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 a new Word file, print it out, and* ***attach to the end of the packet.***

*Highlight the amino acid residues that are not conserved between the organisms in the MSA.*

|  |  |
| --- | --- |
| 5. Next, click on the “Results Summary” tab. Under the “Result files” heading you will find a link to a pairwise **percent identity matrix**. (17) | |
|  | 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, dissimilarity 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?*

5.

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:

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=fl](https://www.viprbrc.org/brc/tree.spg?method=ShowCleanInputPage&decorator=flavi_zika) [avi\_zika](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”
4. Paste your sequences into the text box and select “unaligned FASTA”
5. Click “Build Tree”.
6. On the following screen, choose “Archaeoptryx-js” as the tree viewer mode and click “View Tree.”
7. A new window will open with a visual representation of the MSA sequence percent identity in the form of a phylogenetic tree.
8. *Take a screenshot of the tree, print this image, and* ***attach*** *to the back of the packet.*

*Based on your knowledge of eukaryotic organisms, does your phylogram represent how you would predict the relative evolutionary relationships between the organisms to be? Explain.*

**References**

Ruiz-Trillo, I., Riutort, M., Fourcade, H. M., Baguñà, J., & Boore, J. L. (2004). Mitochondrial genome data support the basal position of Acoelomorpha and the polyphyly of the Platyhelminthes. *Molecular Phylogenetics and Evolution*, 33(2), 321-332. doi: 10.1016/j.ympev.2004.06.002

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., ... & Thompson, J. D. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(1), 539.112(2), 458-463. doi: 10.1038/msb.2011.75.

## Terms and Examples

1. **Phylogenetic Tree** - A phylogenetic tree (also called a phylogeny or evolutionary tree) is a branching diagram that shows the evolutionary relation between several species/entities - based on their biological, physical, or genetic similarities. The closer two branches are, the evolutionary time separating them.
2. **Sequence Alignment** - In bioinformatics, a sequence alignment is a way of lining up DNA, RNA, or proteins in order to identify regions of similarity. Regions of similarity (or difference) can show evolutionary patterns or connections that can help biologists infer about ancestry, evolution, or even evolutionary pressures.
3. **Distance Matrix** - A distance matrix is a table that shows the distance between two objects, where the object’s own distance from itself is 0. The distance between the two objects is put in at the intersection in the chart where they overlap. Below is an example:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | A | B | C |  | Distances: |
| A | 0 | 16 | 24 |  | A:B 16  A:C 24  B:C 12 |
| B | 16 | 0 | 12 |  |
| C | 24 | 12 | 0 |  |

See how the distance between A:A, B:B, and C:C is 0? That’s because they are the same - they can’t be distant from themselves! But the distance between two different objects, like A to B, has a numerical value greater than zero; in this case, A:B or B:A is 16.

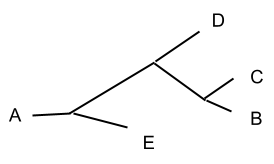
1. **Conserved Protein** - A conserved protein refers to when identical sequences of nucleotides (in DNA or RNA) is present in many species of related origin. The conservation of this sequence indicates that this protein, genomic code, or taxa has been maintained through natural selection.
2. **Molecular Clock** - A figurative term for a technique called the Molecular Clock Hypothesis, which posits that the mutation rate of biomolecules (like DNA, RNA, proteins, etc.) can be used to deduce the time in which two or more organisms diverged in an evolutionary tree.
3. **BLAST** - The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity by comparing the nucleotides or proteins of two or more sequences.
4. **Homologs** - a homologous thing; homology is defined in biology as similarity due to shared ancestry between a pair of structures or genes in different taxa. This often results in shared genetic or morphological structures.
5. **Multiple Sequence Alignment (MSA)** - The process or result of placing multiple biological sequences (nucleotides or proteins) of the same length into a tool that will measure their similarities to each other sequence. This often results in a chart
6. **Distance Method** - The distance method is a form of sequence alignment where two sequences are compared at a time and a distance matrix is built. The number of changes or differences between the bases of the two sequences is then presented as a proportion of the overall sequence length.
7. **Neighbor Joining** - The term neighbor joining refers to the “bottom up”

clustering method for the creation of phylogenetic trees. Neighbor joining is based on the genetic distances, and clumps the most genetically alike sequences into nodes on the tree.

1. **Unrooted / Rooted Trees** - The main difference between a rooted and unrooted phylogenetic tree is whether or not ancestry is shown. Rooted trees show the ancestry of the organisms, often referring back to a single common ancestor. Unrooted trees merely show the relationships between the organisms on the tree. Below is an example:

|  |  |
| --- | --- |
| Unrooted Tree | Rooted Tree |
|  |  |

1. **Outgroup** - The outgroup of a phylogenetic tree is the root; it is used in phylogenetics as the root of the tree and thus is often outside the clade that follows. Below, A is an example of an outgroup:



1. **Tree Topology** - Tree topography summarizes the relatedness of a group of species *independent* of branch length.
2. **Branch Order** - The overall relation between branches of a phylogenetic tree.
3. **Dissimilarity Value** - A numerical value that is meant to show the degree in which two objects differ; The more similar the objects, the smaller the dissimilarity value and vice versa. Example below:

Seq A: ATBOTRGSERGH

| | | | | | | | | |

Seq B: ATLOTRGSERYH

How many differences out of 12: 2/12 Percent Dissimilarity: ~ 17% Dissimilarity Value: 0.17

1. **Degree of Conservation** - The amount or proportion of similarities between two sequences that indicates the conservation/passing of specific genetic code through evolution.
2. **Percent Identity Matrix** - A matrix composed of the values of similarity between several sequences. Instead of measuring the differences between sequences like a Distance Matrix, the Percent Identity Matrix measures the similarities between sequences.

**Instructor Solution Key**

## Exercise 3 – Phylogenetic Analysis of Homologous Sequences

#### Neighbor Joining Exercise

Now, do the same for Seq B/C and Seq A.

The distance between Seq B and Seq A is \_ 0.12 . The distance between Seq C and Seq A is \_ 0.16 .

Thus, the distance between Seq B/C and Seq A is the average distance between Seq A and Seqs B and C

( 0.12 + 0.16 / 2 = 0.14 ).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **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 | - |

Fill in the chart below using this idea. If you have trouble, refer to the explanation above. A few have been done for you:

#### 0.1 = 1 cm

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 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 | - |

|  |  |  |
| --- | --- | --- |
|  | Distance Value | Centimeters |
| Seq A - Seq B/C | 0.14 | 1.4 cm |
| Seq A - Seq D | 0.2 | 2 cm |
| Seq A - Seq E | 0.32 | 3.2 cm |
| Seq B/C - Seq D | 0.26 | 2.6 cm |
| Seq B/C - Seq E | 0.32 | 3.2 |
| Seq D - Seq E | 0.32 | 3.2 cm |

1. Calculate 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.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Seq A | Seq B | Seq C | Seq D | Seq E |
| Seq A | - | - | - | - | - |
| Seq B | 0.12 | - | - | - | - |
| Seq C | 0.16 | 0.04 | - | - | - |
| Seq D | 0.2 | 0.28 | 0.24 | - | - |
| Seq E | 0.32 | 0.32 | 0.32 | 0.32 | - |

*Students would align each combination of pairwise sequences and determine the fraction of differing residues for each combination and place it into the distance matrix above. The smaller the fraction the more similar two sequences are.*

1. What species in your list of sequences derived from homologene could you utilize as an “outgroup” if you wanted to build a rooted tree?

*S. cerevisiae would be a good choice since we know that this simple single celled yeast is distantly related to the other more complex multicellular eukaryotes on the list.*

*Cytochrome C protein sequences students should obtain*:

*>Homo\_sapiens MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGI IWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE*

*>Macaca\_mulatta MGDVEKGKKIFVMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSNTAANKNKG ITWGEDTLMEYLENPKKYIPGTKMIFVGIKKREERADLIAYLKKATNE*

*>Bos\_taurus MGDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFSYTDANKNKG ITWGEETLMEYLENPKKYIPGTKMIFAGIKKKGEREDLIAYLKKATNE*

*>Gallus\_gallus MGDIEKGKKIFVQKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAEGFSYTDANKNKGI TWGEDTLMEYLENPKKYIPGTKMIFAGIKKKSERVDLIAYLKDATSK*

*>Xenopus\_(Silurana)\_tropicalis MGDAEKGKKIFVQKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAEGFSYTDANKNKG IVWDEGTLLEYLENPKKYIPGTKMIFAGIKKKGERQDLIAYLKQSTSS*

*>Drosophila\_melanogaster MGVPAGDVEKGKKLFVQRCAQCHTVEAGGKHKVGPNLHGLIGRKTGQAAGFAYTDAN KAKGITWNEDTLFEYLENPKKYIPGTKMIFAGLKKPNERGDLIAYLKSATK*

*>Zea\_mays MASFSEAPPGNPKAGEKIFKTKCAQCHTVDKGAGHKQGPNLNGLFGRQSGTTAGYSY SAGNKNKAVVWEEDTLYEYLLNPKKYIPGTKMVFPGLKKPQERADLIAYLKEATA*

*>Saccharomyces\_cerevisiae\_S288c MTEFKAGSAKKGATLFKTRCLQCHTVEKGGPHKVGPNLHGIFGRHSGQAEGYSYTDA NIKKNVLWDENNMSEYLTNPKKYIPGTKMAFGGLKKEKDRNDLITYLKKACE*

1. Using the matrix, which two organisms have the least evolutionary time separating them based on the molecular data analyzed?

*Homo sapiens and the Rhesus monkey. They have 96.19% identity according to the matrix. In other words, 3.81% of the amino acids in cytochrome C are different.*

1. Briefly sketch the resulting neighbor joining phylogram (estimate branch length) and **attach** the handout to your laboratory report

*Students should attach a filled-out Cytochrome C neighbor joining handout (suggested solutions below) along with a sketch that resembles the tree below.*

**Distance Matrix**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *H. sapiens* | *M. mulatta* | *B. taurus* | *G. gallus* | *D. melanogaster* |
| *H. sapiens* | - | - | - | - | - |
| *M. mulatta* | 0.038 | - | - | - | - |
| *B. taurus* | 0.095 | 0.095 | - | - | - |
| *G. gallus* | 0.124 | 0.124 | 0.086 | - | - |
| *D.*  *melanogaster* | 0.216 | 0.170 | 0.157 | 0.167 | - |

**Distance Matrix – First Round of Neighbor Joining**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *H. sapiens*  */M. mulatta* | *B. taurus* | *G. gallus* | *D. melanogaster* |
| *H. sapiens /M. mulatta* | - | - | - | - |
| *B. taurus* | 0.095 | - | - | - |
| *G. gallus* | 0.124 | 0.086 | - | - |
| *D. melanogaster* | 0.193 | 0.157 | 0.167 | - |

**Distance Matrix – Second Round of Neighbor Joining**

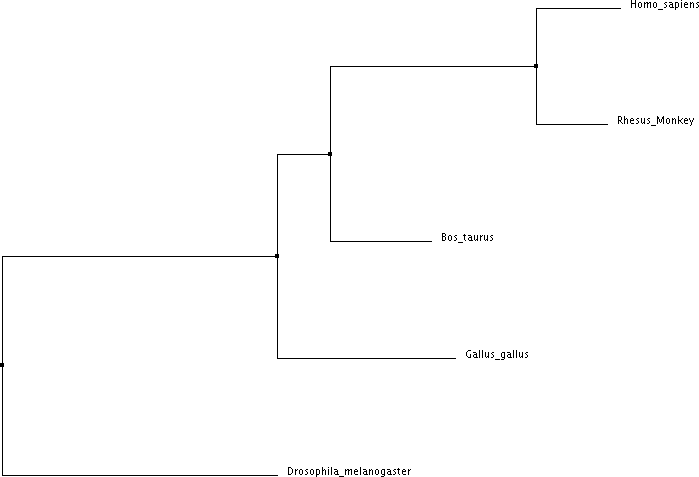
|  |  |  |  |
| --- | --- | --- | --- |
|  | *H. sapiens /M. mulatta* | *B. taurus /G. gallus* | *D. melanogaster* |
| *H. sapiens /M. mulatta* | - | - | - |
| *B. taurus /G. gallus* | 0.110 | - | - |
| *D. melanogaster* | 0.193 | 0.162 | - |

**Distance Matrix – Third Round of Neighbor Joining**

|  |  |  |
| --- | --- | --- |
|  | *H. sapiens /M. mulatta /B. taurus /G. gallus* | *D. melanogaster* |
| *H. sapiens /M. mulatta /*  *B. taurus /G. gallus* | - | - |
| *D. melanogaster* | 0.193 | - |

**Draw the resulting phylogram (estimate branch length)**

*The tree that students draw based on the series of neighbor joining steps using the distance matrices above should look something like the following phylogram.*



1. Go back to the “Alignments” tab and take a screenshot of the MSA, place it in a new Word file, print it out, and attach to your lab report. Highlight the amino acid residues that are not conserved between the organisms in the MSA.



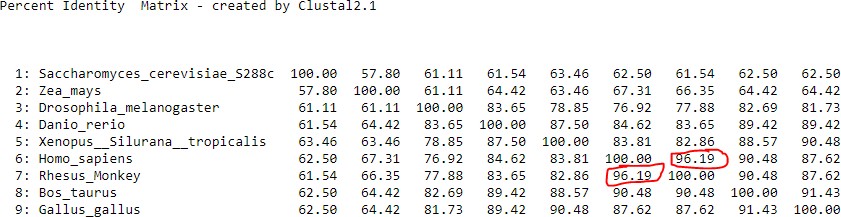
[*https://www.ebi.ac.uk/Tools/msa/clustalo/*](https://www.ebi.ac.uk/Tools/msa/clustalo/)

*Students should see that most of the amino acid residues across evolutionary time are either perfect identities, conserved, or semi-conserved.*

1. Using the percent identity matrix, which two organisms have the least evolutionary time separating them based on the molecular data analyzed?

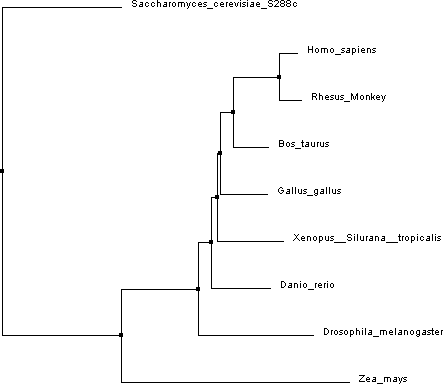
*Homo sapiens and Macaca mulatta (Rhesus Macaque) align with 96.19% identity as indicated by the Clustal Omega percent identity matrix, which is the highest percent identity between the sequences compared, thus humans and Rhesus Macaques have the least evolutionary time separating them.*

***(please note that Danio rerio is present in the two charts below but was taken out of the exercise)***



[*https://www.ebi.ac.uk/Tools/msa/clustalo/*](https://www.ebi.ac.uk/Tools/msa/clustalo/)

A new window will open with a visual representation of the MSA sequence identity in the form of a phylogenetic tree. Take a screenshot of the tree, place it in a Word document, print this image, and attach to your laboratory report.



[*https://www.ebi.ac.uk/Tools/msa/clustalo/*](https://www.ebi.ac.uk/Tools/msa/clustalo/)

1. Based on your knowledge of eukaryotic organisms, does your phylogram represent how you would predict the relative evolutionary relationships between the organisms to be? Explain.

*The phylogram generated from the cytochrome c taxonomic marker gene protein sequences is similar to what we could expect in terms of relative proximity between the whole organisms and a common ancestor. Cytochrome c is highly conserved between organisms that rely on oxygen as a final electron acceptor. Due to its importance in metabolism it is under purifying selection, but changes at a rate that is useful for broad taxonomic identification.*

*The fungus Saccharomyces cerevisiae is used as an outgroup with plants (Zea mays) separating from bilateral metazoans. Protostomes (Drosophila melanogaster) separate from deuterostomes and in this case also chordates and vertebrates, which further subdivides into tetrapods, amniotes, and then mammals.*