**Worksheet 3 – Learning How to Create a Multiple Sequence Alignment**

In this section you will be generating an alignment between all of your sequences. This is called a multiple sequence alignment (MSA). The alignment can provide indications of what sequence residues are well-conserved and what sequence residues are variable, or can delineate species.

**Learning Objectives**

*At the end of this worksheet, students will be able to*...

* create and interpret a multiple sequence alignment using CLUSTALOmega

The next morning you wake up to find a telegram slid under your door. Dr. Shipton will be sending additional samples back to you for analysis. She informs you that she will also be contacting you for an update on the sequencing and artifact identification that you have been doing. While you have the sequences identified, you decide that you need to present a more cohesive analysis of the samples that you have analyzed. You remember something from your bioinformatics course about presenting a number of sequences together in a MSA (maybe it was many sequence at-once). You decide that you need to spend part of the morning figuring out how to create an MSA for your meeting. While the internet is working at the field station, you find what you think might be a good document to guide you. Excerpts from that document are shown below. While this is not directly related to your research, the examples look like they could help you understand the analysis process better.

The first thing that you realize when looking at the document is that MSA stands for **multiple sequence alignment**, and you make note to explain this correctly to Dr. Shipton. On a separate page of the internet document that you found, it explains that you can go to the NCBI webportal(http:/[www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)) and type in the accession number to find the sequence and related information.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Internet document

****

**Question 1.** Look at the multiple sequence alignment in Figure 1. What can you infer from the results of the human (NC\_012920) sequence aligned with the other sequences? (*Hint: Look at the columns that lack an asterisk (\*) at the bottom.*

**Question 2**. If you look closely at the six non-human sequences, you should be able to determine that these sequences (99 nucleotides) are 100% identical. What might you be able to state about these results?

**Question 3.** Using the Genbank identification in Figure 1 (left side), fill out the table below to indicate the animal species that are represented in the figure.

|  |  |  |
| --- | --- | --- |
| **Genbank accession number** | **Genus and species** **(use proper scientific format)** | **Common name** |
| MF579937 |  |  |
| KY616981 |  |  |
| KY616977 |  |  |
| KY364233 |  |  |
| KY99558 |  |  |
| AJ428946 |  |  |

**Question 4.**  Based on your identifications, are the non-human animal 12S rDNA sequences aligned in Figure 1 representing identical species? If not, are they related? Are any of the species extinct?

After reviewing the information on interpreting an MSA (and taking a break for lunch), you patiently wait for the internet to re-connect so that you can perform your MSA. As you wait, you prepare a document outlining your procedure so that others in the lab can follow what you have done. As you finish typing, the internet connects and you proceed to perform an MSA on your six Yeti sequences.

1. A multiple sequence alignment (MSA) of the sequences allows for the comparison of all of the sequences being investigated at the same time. This approach aligns all of the sequences that you input into the program and creates a stacked alignment similar to the pairwise alignments seen in the BLAST searches.
	1. There are different programs available to do MSA. One common bioinformatics tool is the CLUSTAL package. The algorithms that are employed with CLUSTAL have changed over time (the changes are not pertinent to this assignment). You will be using a newer version of CLUSTAL called CLUSTALOmega.
	2. Go to the website <https://www.ebi.ac.uk/Tools/msa/clustalo/>.



Figure 2. CLUSTAL Omega window.

* 1. Scroll down to Step 1 and in the box below that (which should say PROTEIN), select DNA. This changes the input format from amino acid sequences to nucleotide sequences.
	2. Paste your FASTA file (from previous assignment) into the window titled “sequences supported in any format” shown in Figure 2.
	3. Scroll down to Step 2 and select CLUSTALW as the output format.
	4. Scroll down and click on the ‘Submit’ button. You do not need to change any of the defaults.



Figure 3. CLUSTAL Omega Submit button.

* 1. When the alignments are completed, copy and paste the MSA into the box below (Question 5). If the formatting does not look aligned, change the text to Courier font (a uniform font width) and set the size of the font to 9 or smaller.

**Question 5:**

paste your sequence alignment here

* 1. The asterisks below the aligned sequences indicate 100% consensus (all of the sequences have that base found in that position).

|  |
| --- |
| **Question 6**: With the information that you have now generated in the MSA, compare the alignments. Are your purported Yeti sequences (1-6 from the previous worksheets) * Identical
* Similar
* Totally unrelated
 |

|  |
| --- |
| **Question 7**: Are there regions in this MSA that might be good indicators of where a variable region might be to distinguish sample identity? (*Keep in mind that sometimes gaps shift the alignment and make a region look more variable than it is. Look at the sequence, not just the \*.)* |
|  |

|  |
| --- |
| **Question 8**: What can you tell Dr. Shipton when she calls about the data that you have so far collected? |
|  |

|  |
| --- |
| **Question 9**: Since Mr. Norgay provided you with 30 total specimens, how might you explain to Dr. Shipton about what you might be looking for in the remaining samples and those that she is sending to you? |
|  |