**Comparative Genomics Exercise**

The purpose of this exercise is to become familiar with pairwise alignments and multiple sequence alignment. We will be using Muscle algorithm 1 to do the multiple sequence alignment.

Learning Goals:

* Become familiar with the utility of dot plots in pairwise alignments
* Become familiar with sequence alignments
* Become familiar with multiple sequence alignment and be able to use it for making phylogenies
* Become familiar with the alignment output and be able to interpret the results

**Part 1 - Dot plots:**

Consider two **identical** 100 bp DNA sequences A and B. A dot plot comparing these two sequences would look like:

100 bp

**Sequence A**

1,1

**Sequence B**

100 bp

1. Consider the following dot plot: How will you explain this alignment?

**Sequence A**

**Sequence B**

100 bp

100 bp

1,1

2. Draw a dot plot for two 100 bps sequences A and B that are identical except that B has a 40 bp inversion relative to A in the center.

**Part 2 – Multiple sequence alignment:**

Learning Goals

* Become familiar with some of the online tools available for sequence analysis
* Investigate the relationships among related sequences
* Develop hypotheses for the evolution of these sequences

1. On NCBI (<http://ncbi.nlm.nih.gov>) - find the protein sequence that corresponds to WP\_014517873.1

(*Hint:* Select ‘Protein’ from the drop-down menu next to the search bar*)*

2. Using this sequence, BLAST against the nr (non-redundant) database.

(<http://blast.ncbi.nlm.nih.gov/Blast.cgi>).

Click on **Protein Blast**, then on the next page either enter the sequence in FASTA format\* or enter the accession number (the WP number). The default setting for the Search Set is the nr database.

Capture the top 7 “hits” from the Blast return. These accession numbers and sequences should go into your Word or Text document.

What is unusual about the sequences you’ve retrieved?

3. Go to <http://www.ebi.ac.uk/Tools/msa/muscle/> and upload the file to the server.

Use Clustal-W or Clustal-W (strict) as the output format and submit your job.

Note: You can either submit your email ID for notification of your results or wait for the window to refresh.

Multiple Sequence Alignment Results:

The simplest output of a multiple sequence alignment will look like this:



What do the asterisks at the bottom indicate?

What do the dashes in the middle indicate in terms of the sequence alignment?

For more information on interpreting multiple sequence alignment results and for information on parameters, refer to the following:

<https://www.american.edu/cas/hpc/upload/Muscle_User_Guide.pdf>

**Reference:**

Edgar RC. (2004). *Nucleic Acids Res.* **32:** 1792

Edgar, R.C. (2010). Nucleic Acids Res., 1–9