**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.

More information on MUSCLE:

<http://www.ebi.ac.uk/Tools/msa/muscle/help/index.html>

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

Sequence A and B are identical in the regions corresponding to the first and last 30%. The region in the middle is not identical or may be missing in one of the sequences

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.

100 bp

100 bp

1,1

**Sequence A**

**Sequence B**

**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, Click on the Run BLAST option

([http://blast.ncbi.nlm.nih.gov/Blast.cgi)](http://blast.ncbi.nlm.nih.gov/Blast.cgi%29)



Click on **Run Blast**, and check if the accession number (the WP number) is entered in the search box. The default setting for the Search Set is the nr (non-redundant) 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?

The sequence homologs are present in both the bacteria as well as the phages

3. Next we’ll format the sequences downloaded from the BLAST website. Open the Text Editor on your computer and copy the sequences in FASTA format. (This can be done in Word but you need to watch for hidden formatting – returns, etc.)

\*One note: oftentimes problems arise due to incorrect formatting. A few things to check to make sure the sequences are in the correct format:

* Make sure the carat symbol (>) is there
* Make sure that the header line contains **no** spaces. If you want you can type in underscores between the words. Remove any parenthesis that may be present in the header line.
* Make sure there is an “enter” or “return” after the header line
* Numbers in the sequence don’t usually matter but it may be necessary to take them out
* Make sure there are no stray symbols at the end of each sequence
* Try and use a shorter name in the header file. This name should be unique to the sequence and easily identifiable, e.g.

>C\_pneumoniae\_LPCoLN\_PRIP

mrelnafelt qpeeyrnrwv lmpclkcrfc rtqhakvwsy rcvheaslye kncfltltyd dkhlpqygsl vklhlqlflk rlrdrisphk iryfgcgeyg tklqrphyhl lifnydslldg

\*Note that this is FASTA format – just the header and the sequence, no numbers or other symbols

4. 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?

Asterisks indicate amino acids (in case of protein alignments) or nucleotides (in case of DNA or RNA alignments) that are identical between the sequences that are compared.

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

The dashes in the middle indicate gaps. Remember the gaps are introduced in a multiple sequence alignment to align as many identical characters as possible

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

<http://outreach.gtldna.com/origin/proc_man/Clustal/Clustal_tutorial.html>

**Reference:**

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