

Bioinformatics Lab 4

Multiple Sequence Alignment (MSA)

Assuming you have collected some protein sequences and their corresponding cDNAs (protein-coding sequences) in FASTA format. If not, try to collect >10 proteins and DNA sequences for your practice.

Option 1 (for Windown computer)

(1) Using ClustalX installed locally to do the alignment:

Note: make sure to shorten your sequence IDs. If the IDs are too long, it will be truncated by the program. I suggest to use the species name as the ID, for example, human gene: Homo_sapiens. If you have more than one sequence from the same species, you can use Homo_sapiens1, Homo_sapiens2, etc.

Do the alignment using the default parameters. You can try with different parameters if you have time.

Select your output GCG/MSF format. If you save in the CLUSTAL format, it can be opened in Word pad to see the alignment.

(2) Using GeneDoc to visualize the MSA output. There many formatting parameters you can select for your “project”, for example, you can change the colours/ fonts, etc. The output in GeneDoc can be saved (printed) into a pdf file.

GeneDoc (for Windows only)

http://proteomics.yzu.edu/courses/BIOL4800_6900/Perl/GeneDoc_Windows/

Option 2 - if you do not want to install locally or you have a MAC computer, try this new version of clustal program at <http://www.ebi.ac.uk/Tools/msa/clustalo/>

OR you can use any of the MSA programs at

<https://www.ebi.ac.uk/Tools/msa/>

Links for downloading to install on your own computer:

ClustalX: <http://www.clustal.org/clustal2/>

Mega: <http://www.megasoftware.net/> (Required for next lab)

More tools at: https://en.wikipedia.org/wiki/List_of_alignment_visualization_software