LEARN PYTHON & R FOR BIOINFORMATICS

Introduction:

Multiple Sequence Alignment:

Multiple Sequence Alignment (MSA), as the name suggests, requires more than 2 sequences (Nucleotide or Protein) and outputs the results for the analysis of similarity and identity among the query sequences.

Basically, MSA analysis is done to:

- Find the phylogenetic relationship between more than 2 sequences.
- When and where those sequences diverged from each other.
- MSA helps in finding the mutated region in the (sequence of) same protein/nucleotide found in different species.
- Similarly, we can assign a putative function to a novel sequence by comparing it with other sequences (if the sequence is known).

There are several tools available for MSA analysis, but the most commonly used tool for MSA analysis in Bioinformatics is **Clustal Omega** hosted by EBI (European Bioinformatics Institute).

Clustal Omega:

Clustal Omega is the most commonly used tool that is utilized for MSA in Bioinformatics. Clustal Omega inputs hundreds to thousands (upto 4000) sequences and provides results in a few minutes to hours depending upon the number of your sequences. Clustal Omega is an online as well as stand-alone program. If you're dealing with a hundred or more sequences, you should install Clustal Omega on your PC having Linux or Mac OS for the optimal and efficient output. Otherwise, it will take a lot of time for your sequence alignment and the results might not be optimal.

Steps:

- Click on the link given below to go to the webserver page of Clustal Omega:
  https://www.ebi.ac.uk/Tools/msa/clustalo/
  [It will open the web-server page of Clustal Omega].
- **Step 1** is to 'Enter your query sequences'. Here, it will show you a drop-down window from where you have to select the correct option according to your query sequence, i.e., whether it is a protein sequence, DNA sequence or RNA sequence.

  [Remember all the sequences you are comparing must be from the same category, i.e., either Protein or DNA or RNA. Do not use mixed sequences of these three categories or it will result in an error].

- In the next box, you need to paste the sequences in FASTA format.
  [Although you can paste your query sequence yet you must create a FASTAfile of those sequences, first].

Note: There would be garbage information in your file like the one-line definition, which is not required for your MSA analysis in Clustal Omega. So, you should remove that garbage information except the '>' symbol and the accession numbers of the sequences.

- So, you can either paste your sequences in the box present on the web-server page of Clustal Omega or you can upload the FASTA file by clicking on the 'Choose File' option present just below the box.
- **Step 2** is 'Set your parameters' where you need to choose the specific output format considering which format would be more favourable for your further analysis, or you can just leave it to the **Default** option.
  [For better understanding of different file formats, kindly watch our video on File Formats].
- **Step 3** is 'Submit your job' where you just need to click on the 'Submit' option in order to get the results of your query.
- Now it will display your results, i.e., the alignment of your multiple sequences. On top of the alignment portion, it gives multiple options to make efficient analysis of your sequences. For example, by changing colour, it will show you different colours for each residue that corresponds to the physicochemical properties of amino acids (in case of proteins).
- In the alignment portion of the result page, each row belongs to a separate sequence against which the accession number for each corresponding sequence is given on the left side.

- **Interpretation of resulting alignment of Clustal Omega:**

1. No similarity among sequences	_
2. Gaps among different sequences	_
3. Highly converged area (with	*

identical residues)	
4. Highly converged area (with similar residues, i.e., same physicochemical properties)	:
5. Somewhat converged area (one or a few having different physicochemical properties)	.
6. May or may not be converged area	**(single space)**

Note: The highly conserved regions, where the sequence alignment shows maximum identity between all the sequences might represent the Domain of proteins (if you are dealing with amino acid sequences).

- To download the MSA file click on ‘Download Alignment File’ option on the top of the page and then press Ctrl+S to save it on your PC.

Note: There are options provided by Clustal Omega for building phylogenetic trees based on your query sequences, but we recommend not to rely upon them, instead use other tools like MEGA.
[Kindly watch our tutorial on MEGA to get a better understanding of it].

Summary:

In this video, we came to know why MSA analysis is important and how we can align multiple sequences using Clustal Omega. We also learned how to interpret the output of the final alignment results after the query submission in Clustal Omega. Besides, we also discussed the process of downloading multiple sequences in a single FASTA file.