Project DNA5 Comparison of sequences using CLUSTAL

Investigators often wish to compare DNA, RNA or amino acid sequences. They could be looking at different alleles of the same gene in a eukaryotic organism or comparing homologous sequences from different organisms. The first step in this process is aligning the sequences to be compared. Once aligned the sequence can examined for identical residues, conservative changes or divergent regions. The aligned sequences can also be examined using statistical analysis to measure similarity.

There are a number of programs that align multiple DNA, RNA or amino acid sequences available. Many are written for mainframe computers but Mac and PC (Windows) variants are available. We will use CLUSTAL, the most common PC-based alignment program (free) at <http://www.clustal.org/>. Clustal Omega is an online alignment program with a very strong engine (<https://www.ebi.ac.uk/Tools/msa/clustalo/>). It can do a lot of sequences, hundreds, but you have to format the sequences and use a command line interface. If you remember DOS you can do this. We will use Clustal X2, which is a much simpler (<ftp://ftp.ebi.ac.uk/pub/software/clustalw2/2.1/>). Many DNA analysis suites, such as MEGA contain Clustal and other alignment programs. It often tends to be simpler to use program in the suite. We will.

CLUSTALW/X

**Exercise D5: Aligning DNA sequences from herpesvirus glycoprotein B Genes**

In this exercise, you will align a number of provided DNA sequences. The goal is to open sequences and align them using the CLUSTALX2 program.

The sequences are in Project DNA5 Alignment Data Zip File. We’ll use MEGA to create a single file and ClustalW to align the sequences.

Procedure:

1. Import the sequences into a single Mega Project by starting Mega and selecting Align > Edit/Build Alignment > Create a new alignment

*-we need to get all the gB sequence files together in single file whose format Clustal can read.*

1. It will ask if you are building a DNA or protein alignment. Click DNA.
2. Click Edit > insert sequence from file (or use Control I) and import all files at a time once by browsing to the Project DNA5 gB Alignment Data, selecting all sequences and hitting open
3. The sequences will be open in the main Mega Alignment explorer window.
4. Name and save the project as a Mega file (Glycoprotein\_B raw.mas).
5. The files are not yet aligned but show a lot of similarity.
6. You can rearrange the order of sequences by highlighting the sequence name in the lefthand window and dragging it to the top.
7. Select all the files then click Alignment and choose Align by Clustal W.
8. In the new alignment file being displayed, identical residues are marked by an askerisk.
9. Save the aligned file as Glycoprotein\_B aligned.mas
10. There are a lot of things you can do with a set of aligned sequences. We will in another project.
11. Close Mega

These sequences have been trimmed for you!! When you align a sequence with others, they need to be nearly the same length or the alignment program will put in a large number of gaps, degrading the alignment data. The usual process is to align various sequences (obtained from GenBank or collaborators) and compare them to your sequence then trim off sequence from the ends. Then you re-align them. The gB data sequences you are using are <1300 bp yet the gB gene is ~2700 bp so I trimmed them to match the DrCMV sequence.

You will have to align and trim sequences in Project D7

Data to be produced:

Computer files Acrobat documents

Glycoprotein\_B raw.mas

Glycoprotein\_B aligned.mas