Laboratory 3: Multiple Sequence Alignment

Key Concepts

- Appreciate how automated alignments vary when settings are changed – what it can and can't do
- Understand what aspects of multiple sequence alignment currently need manual intervention.
- Become comfortable with the manipulation of multiple alignments, including features for presentation

What you will be able to do at end of this section

- Construct an automated multiple sequence alignment
- Produce a publication-ready or presentation-ready sequence alignment
- Edit a multiple sequence alignment, in preparation for further phylogenetic analysis

Introduction

In this exercise you will construct an automated multiple alignment, and then edit it as necessary for further phylogenetic analysis. Editing is often necessary for the

- removal of "gappy bits" that tend to add "noise" to any phylogenetic analysis
- movement of some gaps that are a result of the program placing too much weight on the similarity between two residues that are more likely similar by chance

During this exercise, do spend time performing automated alignments using different settings, to gain a better understanding of how the settings affect the resulting alignment.

Worked example

Alignment of proteins: an exercise in alignment manipulation

A pathogenic bacterium, *Bioinfomaticus exerciseris*, is noted for its resistance to the antimicrobial drug imipenem, a drug used in treatment of *B. exerciseris* infections. This bacterium has been shown to be resistant due to the loss in some strains of a specific outer membrane protein, called OpdT. This outer membrane protein has been extensively studied and has been shown to also function in the uptake into the cell of the basic amino acid arginine. Notably, imipenem resembles a basic amino acid in structure. Apparently, strains not expressing OpdT are resistant to the drug because the drug is unable to be taken up into the bacterial cell through this one protein.

A family of proteins related to OpdT has also been identified in the same organism through genome analysis and other functional studies. There is interest in learning more about the function of these other proteins, but there is also a more pressing study:

Recently, a genome project was initiated for a related very pathogenic bacterium, *Bioinformaticus examinus*. Some strains of this bacterium are also resistant to imipenem, but this bacterium has not been as intensively studied because it cannot be grown (cultured) in the laboratory. However, through the genome project a number of genes related to OpdT were identified (according to BLAST analysis). It is hoped that through an analysis of the sequences of these proteins, more insight can be gained regarding the mechanism of imipenem resistance in this bacteria and what steps should be taken to combat this resistance problem. In particular, is there an ortholog of OpdT in *B. examinus*? If so, can we use the info that is known about OpdT to hypothesize new information about the function of these proteins?

In this exercise, you will create an alignment of all of the sequences related to OpdT from both *B. excerciseris* and *B. examinus*. This alignment will be subsequently used in your next laboratory exercise to construct a phylogenetic tree for further analysis. The sequences are in a file named lab_03_sequences.txt.

Automated alignment of sequences with ClustalX

The sequences you have been supplied with are in FASTA format. Start ClustalX, and "load" the sequences into this program.

Under the alignment menu, choose "Do complete alignment" to create an alignment. *Aside note: The path for your alignment file can not contain spaces (i.e. you can not save to c:\My Documents\My Alignments\First Experiment\alignment1.aln, but you can save to c:\MyAlignments\FirstExperiment\alignment1.aln).* Once complete, you will notice that many residues are now aligned, however there are a number of gaps that are not biologically relevant (residues aligned in lower similarity regions as if they are significantly similar, when they are more likely just similar by chance). Try modifying the parameters of the alignment (i.e. increase Gap cost to 14, for example) by choosing "alignment parameters" under the Alignment menu option, and then "multiple alignment parameters". Before realigning, select all the sequences by highlighting their names in the far left column of the window, and then choose "Edit" – "Remove all gaps". Do the alignment changing a few of these parameters, especially increasing the gap cost significantly (for example, try 50) to see the effect. Pick the settings that give you the most reasonable alignment (not very many gaps, but still a good alignment). Note that every time you generate an alignment you will generate a .aln file of the results. This file can be imported into Genedoc for further manipulation and better visualization of the sequence similarities.

Using Genedoc for further manipulation and analysis of an alignment

Open Genedoc and import the .aln file containing the sequence alignment you find most reasonable. You may now wish to make a few edits to the alignment. To add or remove gaps, or change the length of gaps, choose the “arrange” function and “Insert Gap Into Sequence” or “Delete Gap from Sequence”. Then click on the area you want to modify. Note that to move a gap you must insert a gap in one region and delete it in another to preserve the rest of the alignment. Try it and be careful – its easy to get confused when misaligning regions to fiddle with gaps!

Once you have an alignment that you are satisfied with, you may wish to fiddle with the display settings so that your alignment is more suitable for, say, viewing as a colourful graphic in a slide presentation. Most display settings can be found under “Project” – “Configure” and also under “Shade”. For example, to change the colours of the shading for the alignment choose “Project” – “Configure” and the “Shade” tab, and then click on the features at the bottom of the window to manipulate the colours. Try “Shade” – “Property Mode” to shade according to residue property, not by conservation.

Use the settings to generate a graphic suitable for presentation. To generate a graphic from your analysis, you will need to select all “blocks” in the sequence. Use “Edit” – “Select blocks for copy” and then “Copy selected blocks to..”.

You may also wish to view some of the other useful information available, such as “Reports” – “Statistics Report” which provides you with pairwise % similarity and % identity information.

Save your alignment as a genedoc file, and also as a Clustal .aln file for use in the next lab.

Interpretating the data

What happens when you greatly increase the gap penalty, or when you reduce it?

From a computational standpoint, how would you improve the procedure for inserting and removing gaps?

OpdT residue 431 (E or glutamic acid) has been implicated in binding to imipenem (changing this residue from its negative charge to a neutral charge leads to imipenem resistance in the mutant bacterium). Which proteins appear to share this same binding site?

How easily can you deduce which sequence is most similar to OpdT – would a phylogenetic analysis help?

Appendix

1. Resources

i) Original Papers

The ClustalX program is described in the manuscript:

Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The ClustalX windows interface: flexible strategies for multiple sequence

alignment aided by quality analysis tools. Nucleic Acids Research, 24:4876-4882.

The ClustalW program is described in the manuscript:

QU1.N92A35 - Thompson, Higgins, and Gibson (1994) Nucleic Acids Research, 22:4673-4680. (Title:CLUSTAL W: improving the sensitivity…)*The original Clustal program is described in the manuscripts:*

Higgins,D.G. and Sharp,P.M. (1989) Fast and sensitive multiple sequence alignments on a microcomputer. CABIOS 5,151-153.

Higgins,D.G. and Sharp,P.M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73,237-244.

Some tips on using Clustal W:

Higgins, D. G., Thompson, J. D. and Gibson, T. J. (1996) Using CLUSTAL for multiple sequence alignments. Methods Enzymol., 266, 383-402.

Genedoc reference (as described by author):

Nicholas, K.B., Nicholas H.B. Jr., and Deerfield, D.W. II. (1997) GeneDoc: Analysis and Visualization of Genetic Variation, EMBNEW.NEWS 4:14

ii) Software

- ClustalX and ClustalW: ftp://ftp-igbmc.u-strasbg.fr/pub/ or ftp://ftp.embl-heidelberg.de or ftp://ftp.ebi.ac.uk
- Genedoc: http://www.psc.edu/biomed/genedoc/

iv) Web Sites:

- European Bioinformatics Institute's ClustalW Online: http://www.ebi.ac.uk/clustalw/

- Baylor College of Medicine's ClustalW Online: http://dot.imgen.bcm.tmc.edu:9331/multi-align/multi-align.html
- ClustalX and ClustalW reference information: http://www-igbmc.u-strasbg.fr/BioInfo/