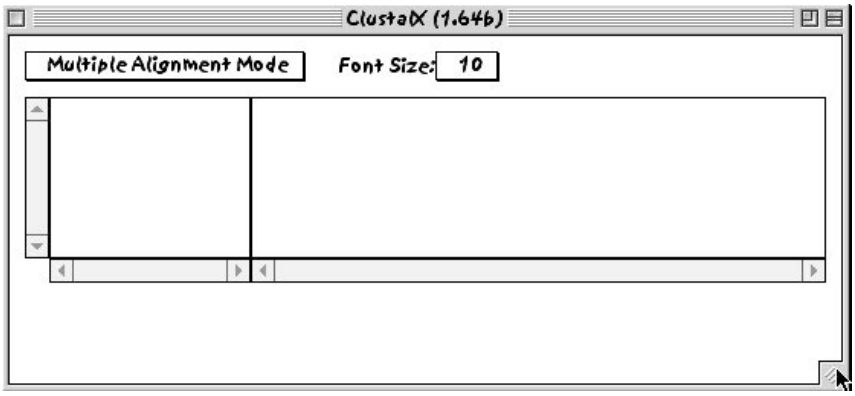
ClustalX Practical

Introduction:

The purpose of this practical is to become familiar with the use of the ClustalX program. This program is written with the assistance of the NCBI toolbox for software development. This means that the program runs on all operating systems with a common graphical user interface (GUI) on each system.

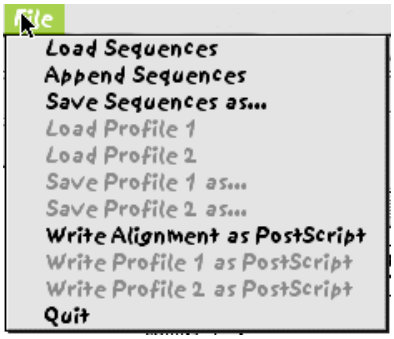
Depending on the operating system, you will start the program either by double-clicking the ClustalX icon or by typing 'clustalx' (no parentheses) at the command line (UNIX). The first screen that you will see will look like this:



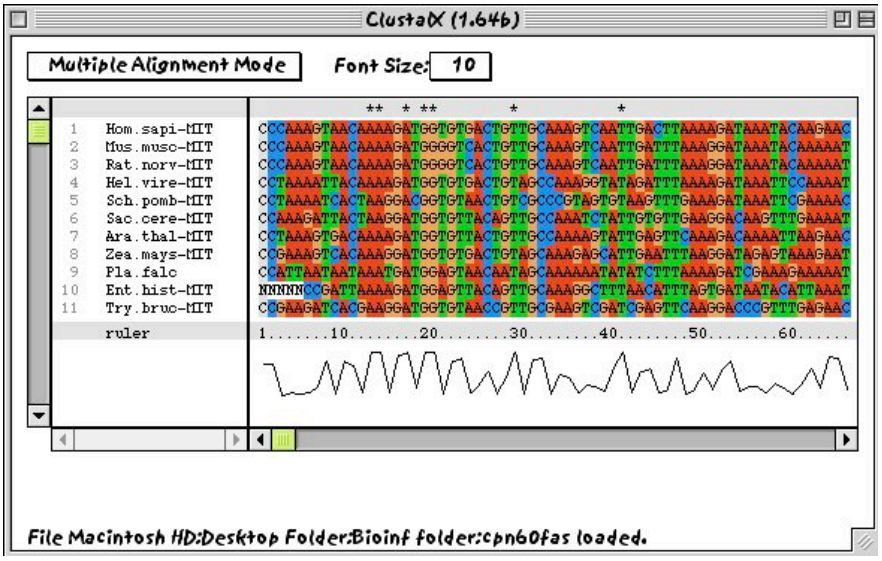
Note: This is a screen shot from the macintosh version. Other versions might look slightly different, but will have the same general appearance. My apologies for the unusual font, this is the one I use on my macintosh.

You will notice at the top, left hand side of the screen is a pull-down menu. Clicking on this menu allows you to change between the two major modes of alignment of ClustalX:

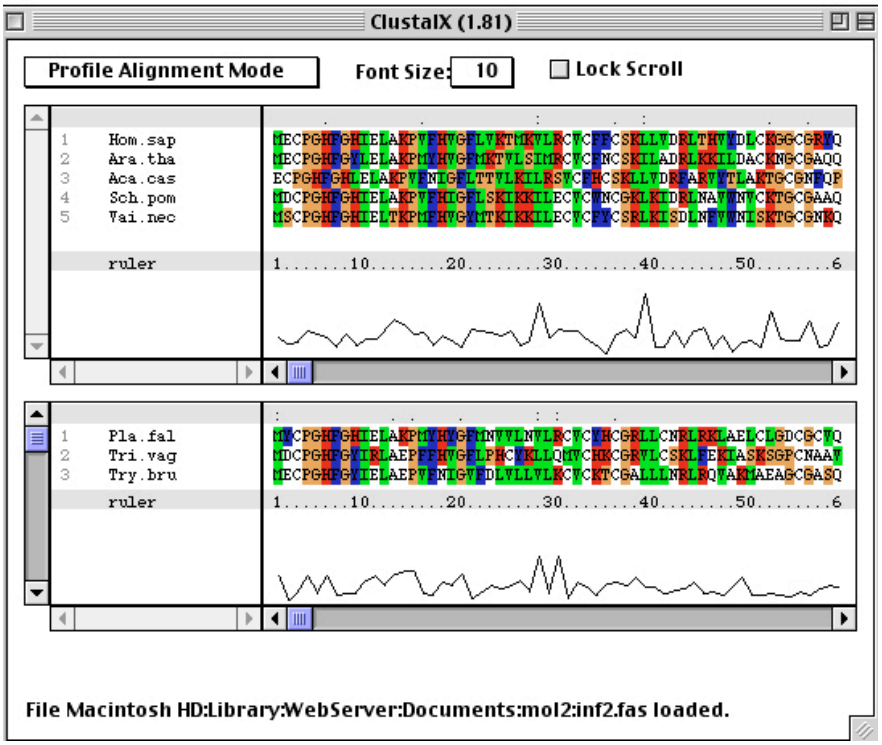
In order to load some sequences into memory, you must select the file menu with the mouse:



If you select the option to load sequences, they will be displayed in the main window. ClustalX accepts 7 formats at present.



If, on the other hand, you wish to align two existing alignments together or to align a single sequence to a pre-existing alignment, then you can switch to profile alignment mode. In this case, the screen is split. When you have two profiles loaded, the screen will look something like this:

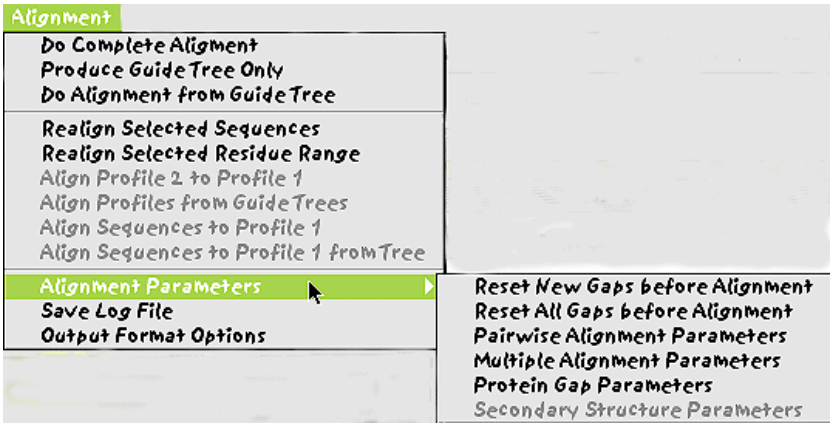


If you wish to manipulate sequences, for example, re-arranging the order of the sequences in the window, removing gaps, removing columns of data that contain only gaps (this can sometimes happen when you remove sequences, the remaining sequences have columns with gaps only), you can do so using the commands available under the edit menu.

Perhaps the most important menu is the one dealing with alignments. Under this menu, it is possible to change the alignment parameters, both for pairwise alignment and for the multiple alignment stages. If a profile alignment is being carried out, or if sequences with gap characters are being used in the alignment process, then it is possible to remove the gap characters prior to alignment.

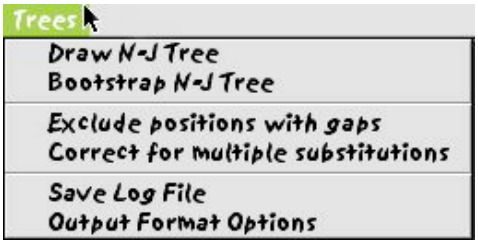
A complete alignment may be carried out, or alternatively, the first stage of the alignment process (producing the guide tree from the

initial pairwise alignments) may be carried out, or even an old guide tree may be used as the starting point for the alignment process, thereby eliminating the need for the pairwise alignments.



ClustalX may also be used in order to infer the relationships between sequences. This part of the program should be treated with extreme caution. The only alignments that should be used are those where positional homology is guaranteed. Unfortunately, automated alignment can be prone to error and uncertainty.

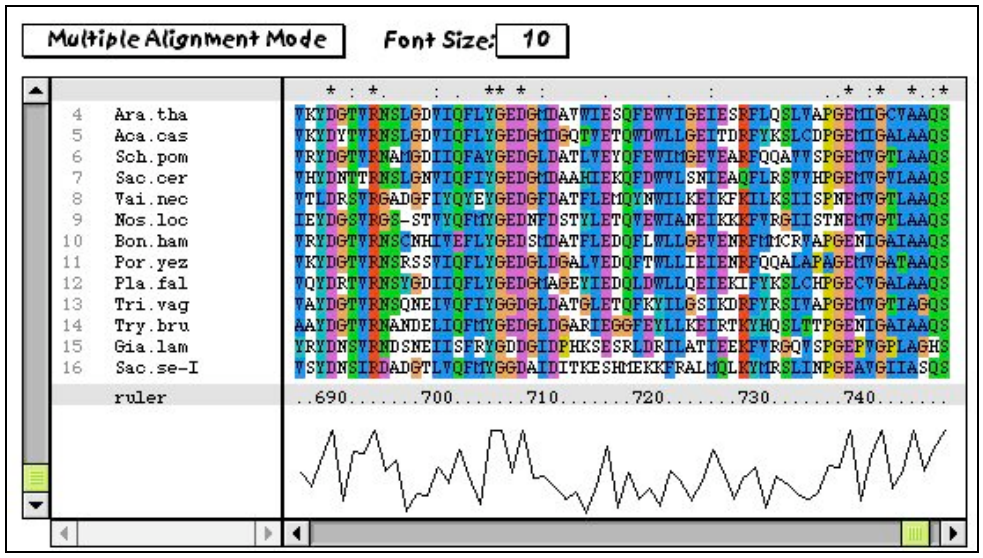
However, given a reliable alignment, where every position is aligned with absolute certainty to its homologs in the other sequences, ClustalX can be very useful for drawing phylogenetic trees. There are some basic corrections for multiple substitutions and the reliability of the resulting relationships can be estimated by bootstrapping. However, the most important features of ClustalX are not in its tree-drawing capabilities and that generating phylogenetic hypotheses is best carried out using other products.



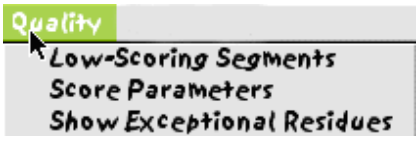
There is also a series of options for changing the colour of the residues in the alignment. There is a default colour parameter file, but it is possible to change it. It might be interesting to do this if you wish to group certain residues in different ways. Grouping of protein residues will be dealt with later in the course.

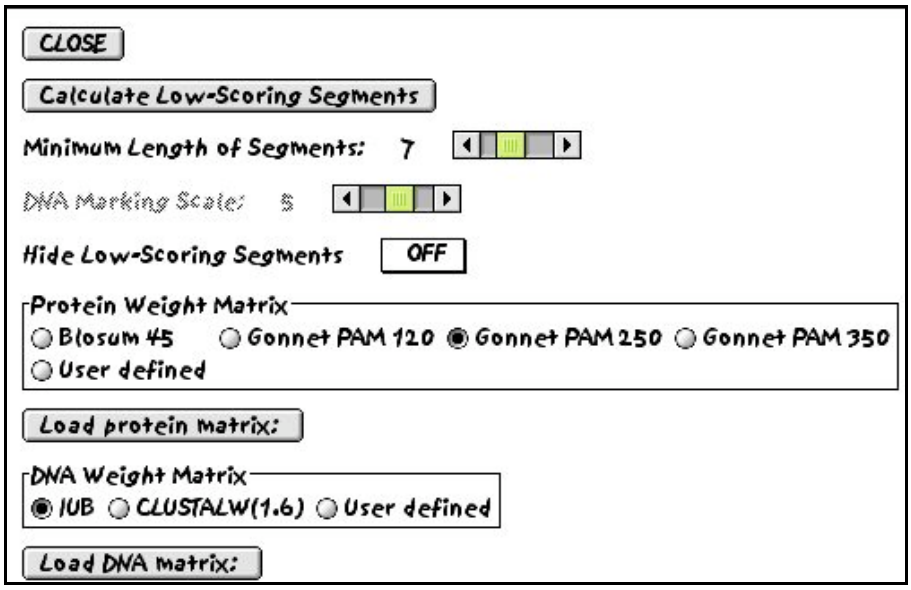
One of the interesting advantages of using ClustalX over ClustalW is the ability to visually evaluate the quality of the alignment. In particular, the ability to highlight areas where the alignment is poor is a significant advantage. It is possible that one sequence has a short region that shows low levels of residue similarity to the rest of the alignment. This can be due to evolutionary processes, or quite frequently, due to errors in sequencing. A frameshift error that is corrected by a frameshift downstream might go unnoticed and become submitted to GenBank. ClustalX is capable of detecting these kinds of errors. An example of the utility of this method is shown:

Here is a typical alignment:

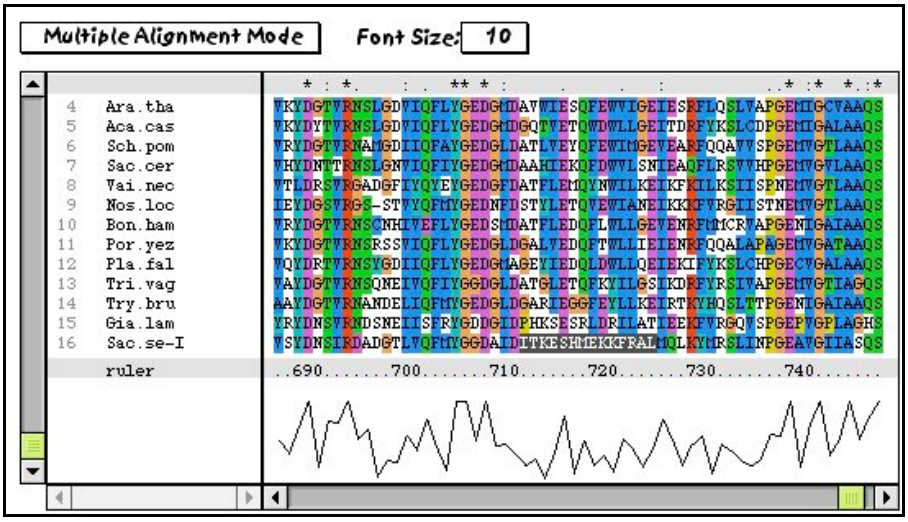


However, choosing a segment length of 7 residues and using the Gonnet 250 matrix, the program can choose regions of the alignment that are showing exceptionally low scores.





The bottom sequence has a region of 14 amino acid residues that show a surprising amount of *dissimilarity* compared to the others (note the darkly-shaded residues). It is possible that either this region of exceptional divergence is due to misalignment, relaxed selective pressures on this region, or due to sequencing errors.



In addition to all of the other features, ClustalX offers a short description of the available commands, using the help menu. You can select any of these options for a brief description of their usage.

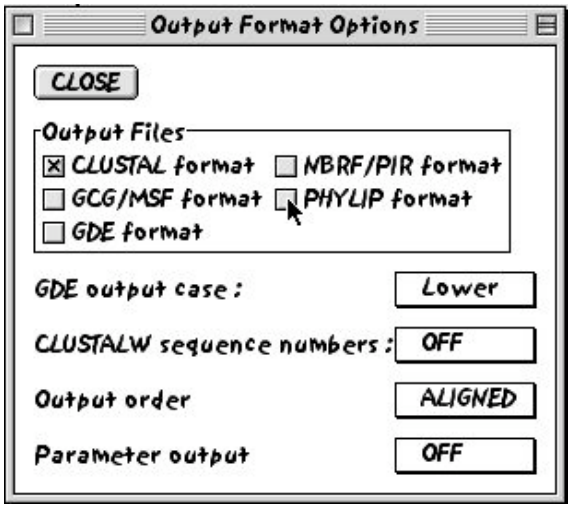
This brief description is not intended to replace the ClustalX information page. Please visit this page and read its contents.

Practical Exercise

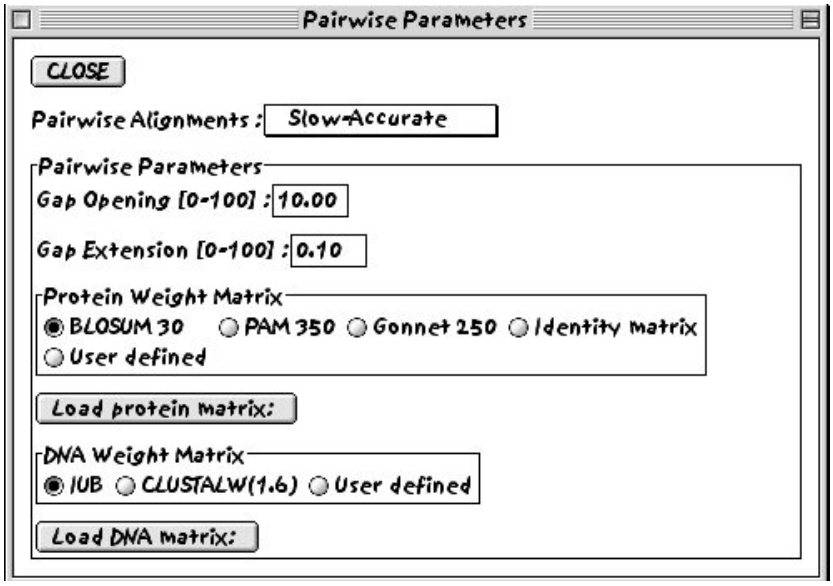
This is the practical exercise for ClustalX. In this practical you will become familiar with performing alignments, changing parameters, performing profile alignments and assessing the quality of the alignments.

1. Load the sequences in the file "inf1.fas" into memory. Save the file to your home directory, or an appropriate location. There are 5 amino acid sequences in this file.

2. Under the alignment menu, choose the output format options and select the PHYLIP output format (we might wish to manipulate the alignment manually later, so this format is compatible with the SEAVIEW manual alignment software).

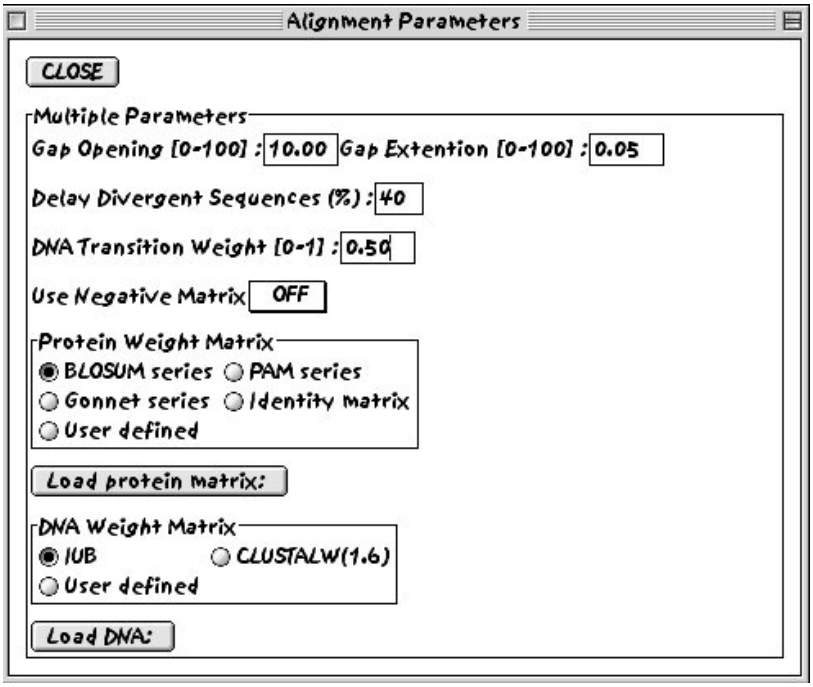


3. Look at the pairwise alignment parameters settings.



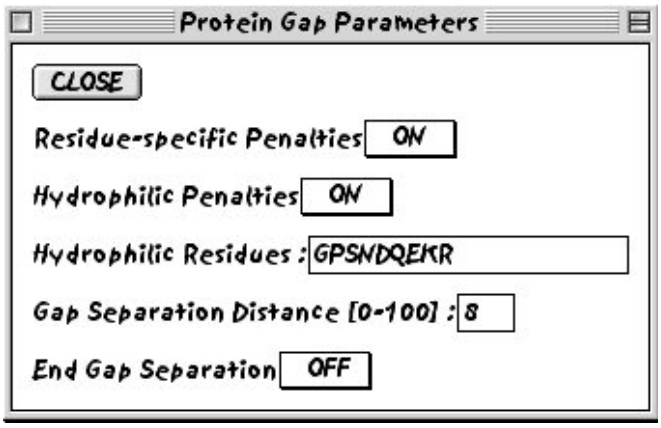
You can feel free to change these settings (Gap Opening Penalty, Gap Extension Penalty, Protein Weight Matrix). The values that have been entered are only the suggested values and are used as the default simply because they are thought to be broadly appropriate for many alignments. You can switch from slow-accurate alignment to fast-approximate alignment. Naturally, the first option is superior, however with large datasets, it may only be possible to use the latter. If we were dealing with DNA sequences, we would be able to change the DNA weight matrix parameters.

4. Look at the multiple alignment parameters settings.



Again, feel free to change the protein weight matrices, the percentage divergence cutoff for delaying sequence addition to the growing alignment and so on. If we were dealing with a nucleotide sequence alignment, we could change the parameters for DNA sequences.

5. You can also decide to change the protein-specific gap parameters.

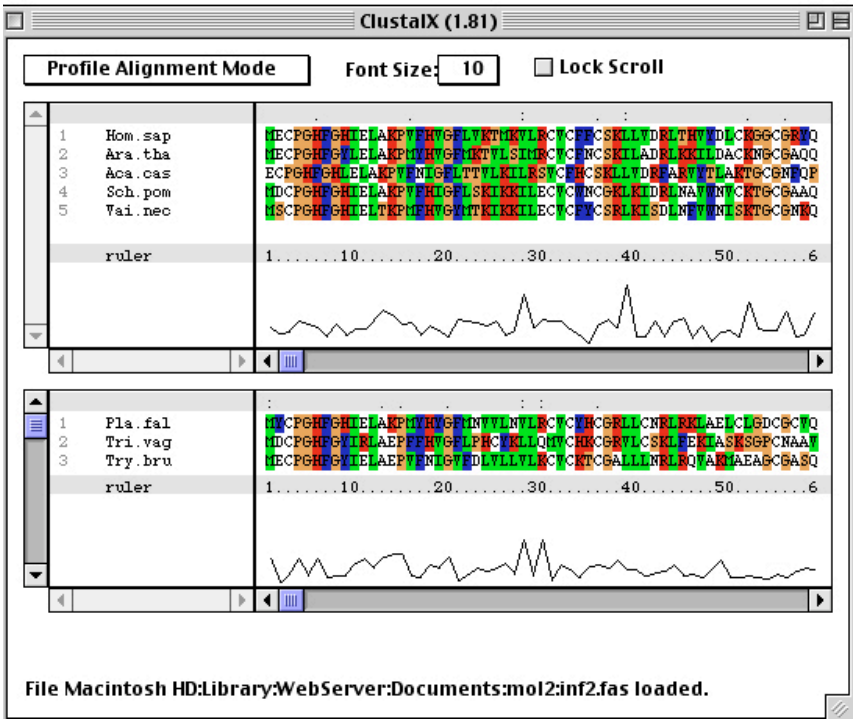


6. When you are satisfied with the alignment parameters, you should carry out a complete alignment.

7. Remove the gaps and realign with different penalties (either very small or very large values).

8. Remove the gaps and realign with different penalties (the opposite of whatever you chose at step 7).

9. Now we are going to perform a profile alignment. You must change the mode of operation of ClustalX from Multiple to profile alignment.



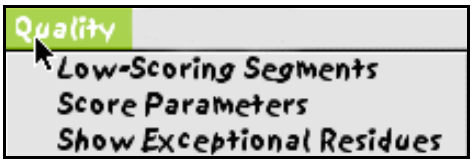
10. When you switch to profile alignment mode, the screen will split and the bottom half of the screen is reserved to the second profile (note: you could have started the program in profile alignment mode and input 2 profiles from disk).

11. The file for performing the profile alignment is "inf2.fas". Save this file to your home directory.

12. Although there is the choice of aligning the second profile to the first, it is not a sensible option in most situations. In this case, the second profile is unaligned so this option makes very little sense. It is much more profitable to align the sequences in the second profile to the first profile. Do this now.

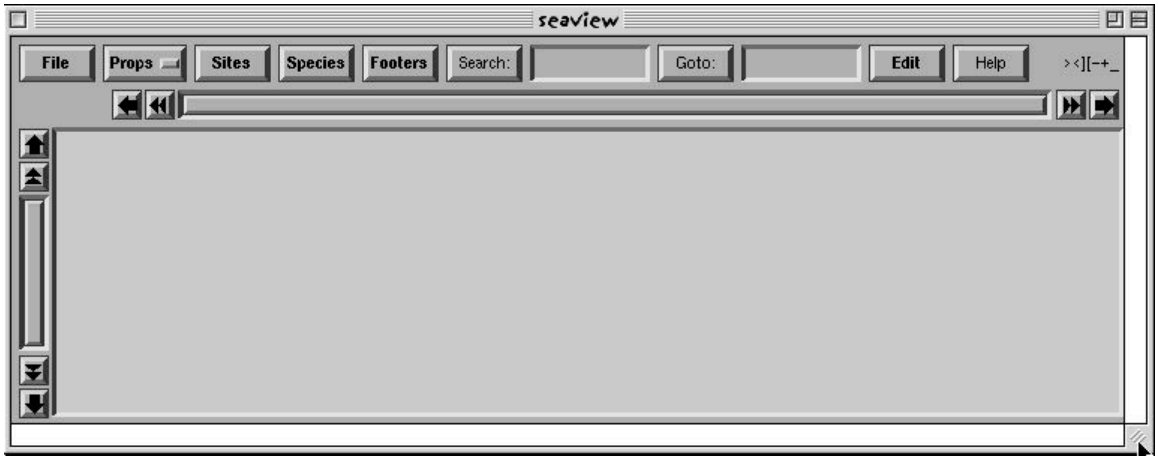
13. You can lock the scroll bars together and look along the length of the alignment. If there are regions to be realigned, perhaps using a different scoring scheme then you can select those regions by switching back to multiple alignment mode, selecting the badly-aligned region and choosing the appropriate option in the alignments menu. Note: There might be a computer 'bug' in this section of the code.

14. You should now choose to ask the program to evaluate the alignment for regions of poor alignment. This is done under the Quality menu. I would suggest looking for low-scoring segments, initially looking for long stretches (say 7 or more residues).

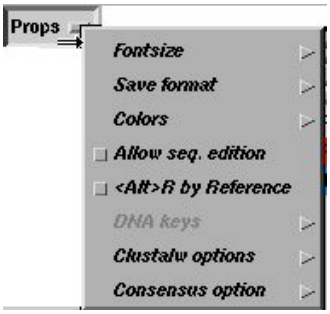


15. When you are satisfied that the alignment is not going to be improved by automated methods it might be appropriate to move to a manual alignment method in an effort to improve the alignment. You can now execute 'seaview' and load the PHYLIP-formatted output file from ClustalX.

16. Seaview is a manual alignment program, again using the Vibrant library (which is part of the NCBI toolkit). It is available for most computer platforms and can be retrieved from http://pbil.univ-lyon1.fr/. The initial screen looks like this:



17. You can alter the properties of the program by choosing the props menu. One of the important things to change is the save format. The default is MASE, but the PHYLIP format is more useful.



If the program is configured to use ClustalW (text-version of ClustalX), it is possible to do some automated alignment. However, most of the features of automated alignment in seaview simply proceed by calling the Clustal program and it is probably a better idea to do any of these things directly in ClustalX. It is also possible to generate a consensus sequence.

18. You can define sets of sites and this will allow you to exclude regions of poor alignment. You can do this by creating a new set (remember to give this set a sensible name). Then you can select/deselect sites using the mouse.

19. When you are happy that *only* positions of unquestionable homology remain in the alignment, then you can save all of the sequences to

disk. Note: it is also possible to just work on a subset of the sequences using the 'species' menu.

20. Finally, save the finished datamatrix in PHYLIP-format. Most other computer programs can read PHYLIP-formatted files. You have two choices for saving the alignment. You can save the complete alignment (this is done using the save/save as... options under the file menu). Alternatively, if you are using a sequence set and a species set, you can save a subset of the alignment. Remember, even though the program automatically prompts with a name, you should name the data matrix in a way that reflects its format (in other words, the usual extension for PHYLIP formatted files is .phy, for FastA-format it is .fas).