Multiple Sequence Alignment Using ClustalW and ClustalX

The Clustal programs are widely used for carrying out automatic multiple alignment of sets of nucleotide or amino acid sequences. The most familiar version is ClustalW (Thompson et al., 1994), which uses a simple text menu system that is portable to more or less all computer systems. ClustalX (Thompson et al., 1997) features a graphical user interface and some powerful graphical utilities for aiding the interpretation of alignments, and is the preferred version for interactive usage. ClustalW and ClustalX are developed in parallel, and the same version-numbering system is used for both in order to synchronize changes (e.g., bug fixes, improvements, and additions). In January 2002, the latest version for both programs was 1.81. The programs can both be run interactively, but the protocols below give instructions on how to do this using ClustalX. Alternatively, ClustalW supports a full command-line interface which allows it to be used automatically as part of larger analyses (e.g., it can be run from scripts). In the simplest usage (see Basic Protocol), the programs are employed to take a set of homologous sequences (all DNA/RNA or all protein) and to produce a single multiple alignment. This covers the vast majority of Clustal usage and will be sufficient for most cases. Nonetheless, Clustal also has extensive facilities for adding sequences to existing alignments, merging existing alignments (so-called profile alignment as described in the Alternate Protocol), realignment of sections of alignment, detecting and fixing alignment errors, and basic phylogenetic analysis. Users may run Clustal remotely from several sites using the Web, or the programs may be downloaded to be run locally on PCs, Macintosh, or Unix computers (Support Protocol).

USING CLUSTALW AND CLUSTALX TO DO MULTIPLE ALIGNMENTS

BASIC PROTOCOL

The programs ClustalW and ClustalX provide alternative user interfaces to the Clustal multiple alignment software. The alignments produced by the two programs are exactly the same; the only difference between ClustalW and ClustalX is the way in which the user interacts with the program. ClustalW is now mainly used as a command-line program by Web servers and automatic batch systems, although the program does provide text menus which can be used to input sequences and perform multiple alignments. Most users who run Clustal interactively now use the graphical interface provided by ClustalX. This protocol therefore uses ClustalX (here on a Silicon Graphics Unix workstation) to illustrate the basic multiple alignment procedure. Although the example given here uses protein sequences, the same protocol can be performed with nucleic acid sequences.

Necessary Resources

Hardware

Unix (including Linux) workstation (e.g., Sun, Alpha, Silicon Graphics, PC), PC with MS Windows, or Power Macintosh

Software

ClustalW or ClustalX program (see Support Protocol)

Files

Sequences can be input to both ClustalW and ClustalX in one of seven file formats. All sequences must be in the same file. The formats that are automatically recognized are: NBRF/PIR, EMBL/Swiss-Prot, Pearson (FASTA; *APPENDIX 1B*), Clustal, GCG/MSF, GCG9/RSF, and GDE flat file. The sequences

Contributed by Julie D. Thompson, Toby. J. Gibson, and Des G. Higgins
Current Protocols in Bioinformatics (2003) 2.3.1-2.3.22
Copyright © 2003 by John Wiley & Sons, Inc.

must be all nucleotide or all amino acid, and the program will attempt to guess which by the composition of the letters. Upper- or lowercase can be used and most symbols and numbers will be ignored (removed); unrecognized residues will be counted as X or N.

If using a word processor to prepare the input file, save the file as plain text with line breaks—i.e., as a simple ASCII file. ClustalX cannot deal with native word processor formats.

1. Download and install ClustalX on your local machine (see Support Protocol).

Construct an initial alignment with the default parameters

2. *Start a ClustalX session.* On PC and Macintosh computers, click on the ClustalX icon. On Unix systems, at the prompt type `clustalx &`.

   The ClustalX window will appear, as shown in Figure 2.3.1. The window on Unix or PC systems has a series of menu items across the top. For Macintosh users, the menu items are displayed at the top of the screen, separate from the ClustalX window itself. Options can be selected by moving the mouse cursor to one of the menu items and clicking the left mouse button to display the list of menu options under that item, then moving the cursor to the appropriate option and clicking the mouse button again.

3. *Load sequences in ClustalX.* Select Load Sequences from the File menu in the ClustalX window.

   A new window will appear (Fig. 2.3.2) that displays the user's subdirectories and files.

4. *Select a file containing the unaligned sequences.* Use the mouse cursor to highlight the filename in the file selection window, then click the OK button at the bottom of the window.

   If the selected file contains more than one sequence and these are in one of the seven recognized file formats, then the unaligned sequences will be displayed in the ClustalX window (Fig. 2.3.3) with the sequence names on the left-hand side. Figure 2.3.3 shows the sequences of five immunoglobulin superfamily domains for which the three-dimensional structures have been resolved. The sequence alignment is for display only; it cannot be edited here. A ruler is displayed below the sequences, starting at 1 for the first residue position (residue numbers in the sequence input file are ignored). The line above the alignment is used to mark strongly conserved positions. Sequence residues are colored to highlight conserved features in a multiple alignment. At this stage, as the sequences are not yet aligned, the residue coloring will not be informative. ClustalX also provides an indication of the quality of an alignment by plotting a "conservation score" below the alignment.

5. By default, the output file of the program is produced in Clustal format, which can be read by many other sequence-analysis packages. To change this, select the output format using Output Format Options window, selected from the Alignment menu (Fig. 2.3.4). The user can save the final multiple alignment in one (or more than one) of six file formats: Clustal, NBRF/PIR, GCG/MSF, PHYLIP, NEXUS or GDE. Select the output file options and close the Output Format Options window by clicking the Close button.

   The different output file formats are provided for compatibility with a wide range of multiple alignment analysis programs. Users can also change the default case of the residues from lowercase to uppercase for GDE output by clicking the appropriate button in this window. Residues are not normally numbered in the output, but users can choose to use numbers here. The order of the sequences is changed to reflect the order of alignment. Crudely, this puts similar sequences beside each other in the output. This can be changed by setting the output order to be the same as the input order. Finally, the values of the parameters (e.g., gap penalties, amino acid weight matrix) can be printed out in the output file by changing the Parameter output option in this window to On.

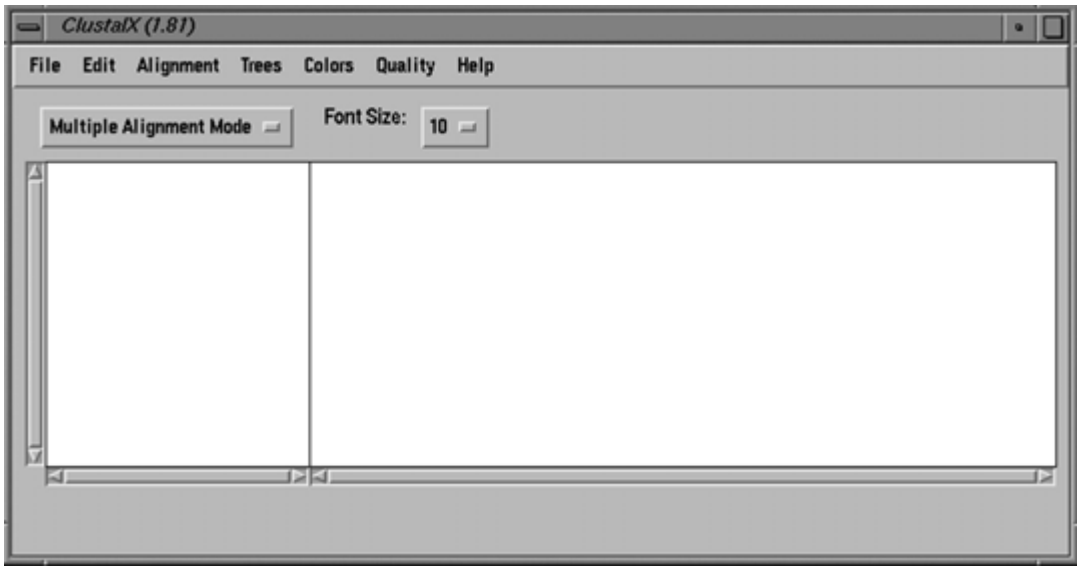

Figure 2.3.1 The ClustalX window on a Unix workstation before any sequences are loaded.

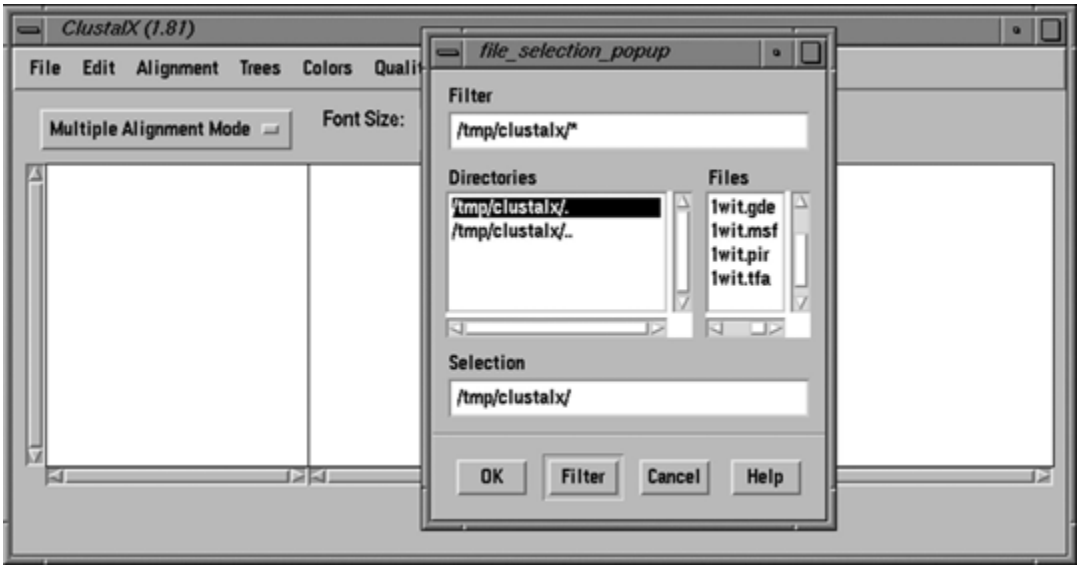

Figure 2.3.2 The input file selection window for ClustalX.

The output files are produced as plain text or ASCII. Use a fixed-space font such as Courier to view these using a word-processing package. This ensures that the aligned residues from the different sequences will be placed neatly in columns.

6. Construct a multiple alignment of the sequences by selecting the Do Complete Alignment option from the Alignment menu. A new window will appear (Fig. 2.3.5) that displays the default filenames for the output guide tree file and the output alignment file. If required, these filenames may be edited, before clicking on the Align button.

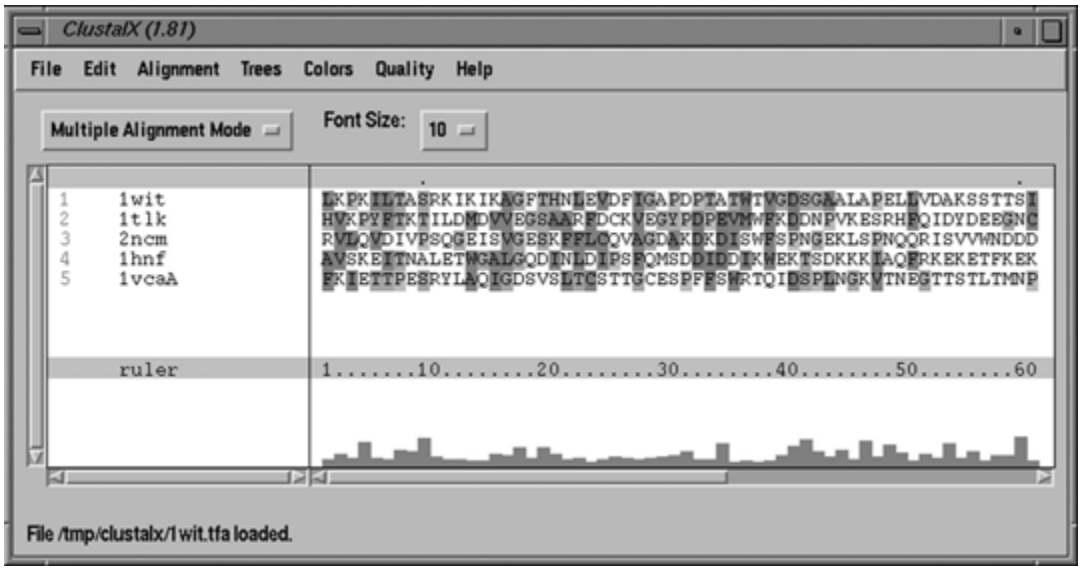

Figure 2.3.3 ClustalX with five loaded but unaligned sequences.

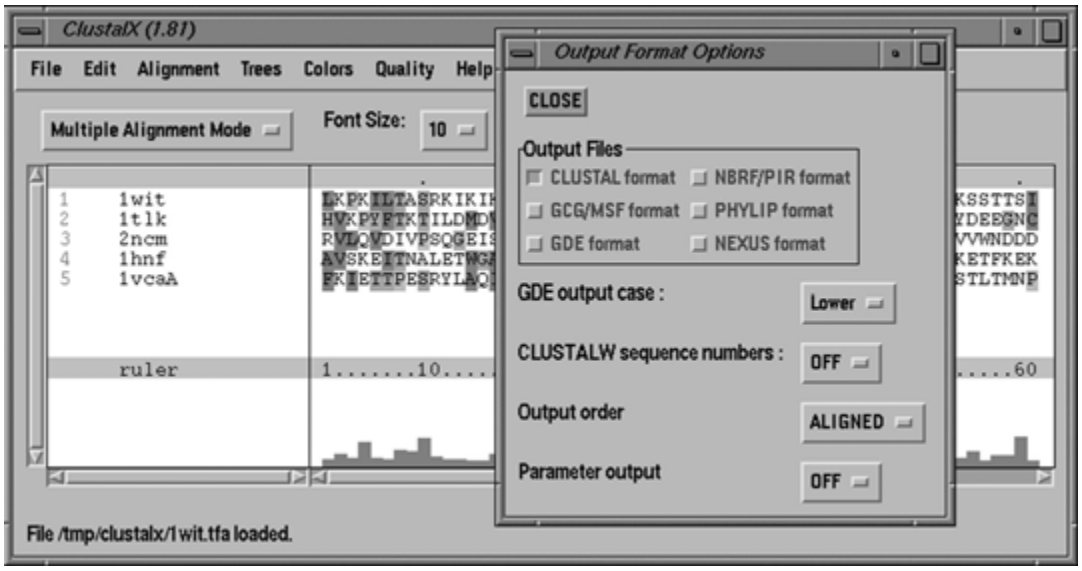

Figure 2.3.4 Changing the format of the multiple alignment output in ClustalX. Clustal format is the default.

ClustalX will perform the complete multiple alignment of the sequences shown in the window. The alignment consists of three steps: first, all the sequences are compared to each other in a pairwise fashion; next, a guide tree is created from the pairwise sequence distances and written to a file; finally, the multiple alignment is built up following the order given by the guide tree (see Background Information). The current status of the alignment process is continuously updated in the message area at the bottom of the ClustalX window. When the alignment is complete, the window display is updated to show the aligned sequences with gaps represented by "-" characters (Fig. 2.3.6).

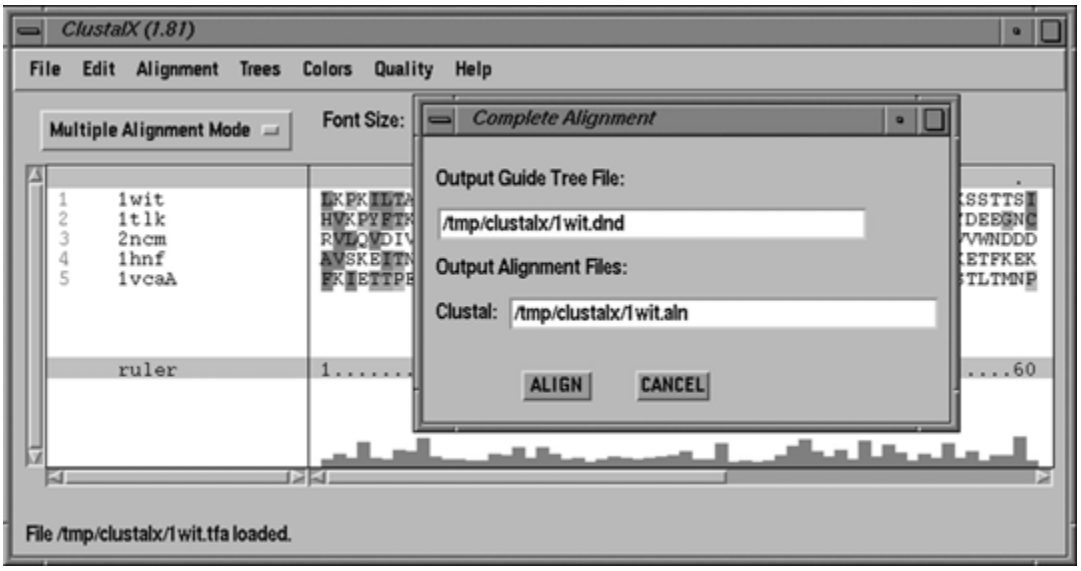

Figure 2.3.5 Selecting the names for the output files for the dendrogram (`1wit.dnd` is offered as the default) and the multiple alignment (`1wit.aln` is the default) for an input file called 1wit.

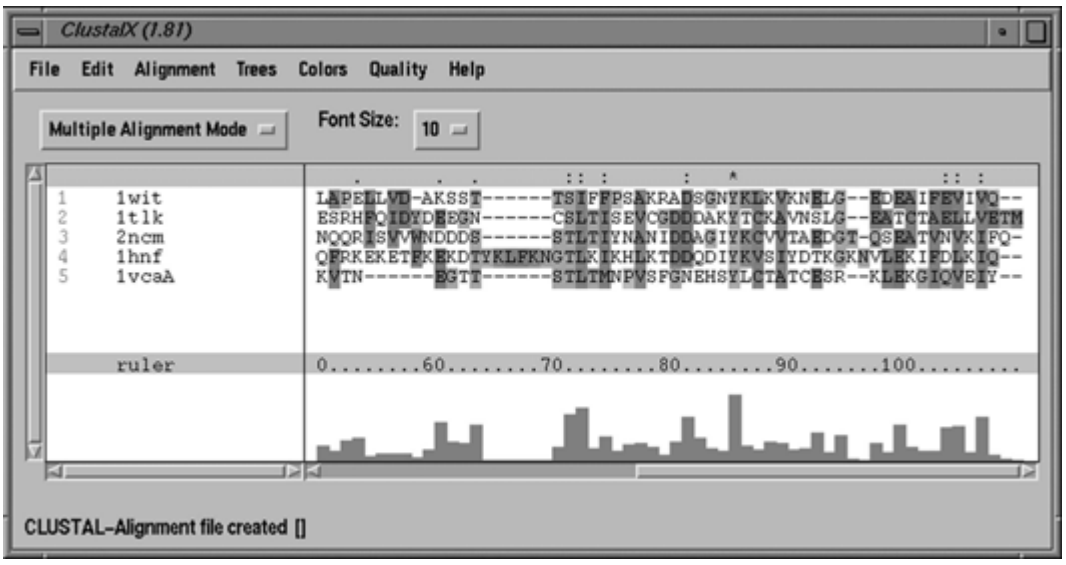

Figure 2.3.6 ClustalX after a multiple alignment has been carried out on the five sequences. The alignment has been written to a text file which can be used for further analysis. The user can also choose to analyse this alignment further within ClustalX (e.g., to calculate a phylogenetic tree).

Evaluate and realign if necessary

7. *Examine the multiple alignment in the ClustalX window.* The ClustalX graphical interface offers several methods of analyzing the multiple alignment (see Guidelines for Understanding Results).

   First, strongly conserved positions are indicated on the line above the alignment. The "" character indicates positions which have a single, fully conserved residue. e.g., the conserved tyrosine in column 85. The ":" and "." characters indicate that the column is "strongly" or "weakly" conserved, respectively. The definitions of strong and weak conservation are described in detail in the ClustalX documentation. These depend on the amino acid scoring system being used and can be changed by the user (see step 8). These symbols ("*",":" and ".") are also included in the output text file when Clustal format is used.*

   Second, the sequence residues are colored either by assigning a color to specific residues (default), or on the basis of an alignment consensus. In the latter case, the alignment consensus is calculated automatically, and the residues in each column are colored according to the consensus character assigned to that column. In this way, the user can choose to highlight, for example, conserved hydrophilic or hydrophobic positions in the alignment. More details about the ClustalX color scheme and how to customize it are given in the documentation and in the online help. These colored alignments cannot be seen in the normal alignment output files. To print these out using the colors, produce a PostScript file (see step 12) and print it with a PostScript-capable printer.

   Third, the quality curve displayed below the alignment plots a "conservation" score for each column in the alignment. A high score indicates a well conserved column; a low score indicates low conservation. The algorithm used to calculate the quality scores is described in detail in Thompson et al. (1997).

   Finally, there are extensive facilities for directly highlighting sections of sequences or blocks of alignment that appear to be very unreliable or poorly aligned, or where the alignment is very ambiguous. These facilities are found under the Quality item of the main menu at the top of the ClustalX window. This is invaluable where one suspects that a sequence is not homologous to the rest of the sequences in a data set, or has sequencing errors or where one wishes to select reliably aligned regions of an alignment for further analysis.

8. *Change the alignment parameters.* If the alignment that is obtained using default settings is not optimal, i.e., if the alignment shows no clearly conserved blocks separated by gapped regions, or if conserved residues or motifs have been misaligned in some sequences (see Guidelines for Understanding Results), the user can modify a large number of alignment parameters. Pairwise alignment parameters will mainly affect the speed/sensitivity of the initial alignments that are used to construct the guide tree, but will not normally have a great effect on the final multiple alignment. In contrast, the multiple alignment parameters control exactly how the final multiple alignments are carried out. To modify the alignment parameters, select the Alignment Parameters option from the Alignment menu, then select either Pairwise Alignment Parameters or Multiple Alignment Parameters. Figure 2.3.7 displays the default settings.

   Under Pairwise Parameters, the most important choice is that between Slow-Accurate and Fast-Approximate pairwise alignments. The Accurate alignments are carried out using a dynamic programming method (Myers and Miller, 1988; UNIT 3.1) to align every pair of sequences. This may be too slow for large numbers (e.g., >100) of long (e.g., >1000 residue) sequences. In this case, the Fast/Approximate alignments using the method of Wilbur and Lipman (1983) may be more suitable. These are several orders of magnitude faster to construct than the former and allow huge data sets to be aligned. The effects on the accuracy of the final alignments are minor except in cases where the alignment is especially difficult.

Under Multiple Parameters, each step in the final multiple alignment consists of aligning two alignments or sequences. This is done progressively, following the branching order in the guide tree. The multiple alignment parameters window allows the user to change the scoring matrices (UNIT 3.5) and the penalties for opening and extending gaps in the sequences. Gap penalties usually need to be altered for aligning nucleic acids, e.g., they are likely to require reduction if divergent sequences are present in the set. In this case, a gap-opening penalty of 7.5 and a gap extension penalty of 3.33 may be more appropriate. For proteins, this is not so often the case, as there is a (hidden) scaling for divergence built into the algorithm.

The Delay Divergent Sequences option delays the alignment of the most distantly related sequences. These sequences are usually the most difficult to align correctly, and it is generally better to delay their incorporation into the alignment until the more easily aligned sequences are aligned. By default, sequences sharing less than 30% residue identity with all other sequences are delayed. If this option is set to 0, the alignment will follow the guide tree exactly. For alignments containing a large number of sequences (e.g., more than 100), it may be useful to reduce the Delay option to 20% or even 10% residue identity.

Invoking the Use Negative Matrix option ensures that the best matching subregion of the alignment will be found. This is a useful precaution when the sequences may be related only over a small part of their full lengths, as often occurs when a sequence set is taken directly from a database search output. However, for sequences that are related over their entire lengths, the default gives slightly (but clearly) better alignments.

For nucleic acid sequences, the Transition Weight option gives transitions (A↔G or C↔T, i.e., purine-purine or pyrimidine-pyrimidine substitutions) a weight between 0 and 1; a weight of 0 means that the transitions are scored as mismatches, while a weight of 1 gives the transitions the match score. For distantly related DNA sequences, the weight should be near zero; for closely related sequences it can be useful to assign a higher score.

The ClustalX alignment options are described more fully in the documentation and in the online help that is available by selecting the Help menu in the ClustalX window.

9. *Rebuild the multiple alignment.* If the pairwise parameters have been changed, it will be necessary to rebuild the complete multiple alignment, as described in step 6, in order to make a new alignment. If only the multiple alignment parameters have been changed, the first stages (pairwise alignments, guide tree) can be reused by using the Do Alignment from Guide Tree option, selected from the File menu.

   In the latter case, a window appears with the default filenames of the input guide tree (written during the multiple alignment process in step 6), and the output alignment file (Fig. 2.3.8). If the user changes the file names in step 6, a similar change should be made when running the alignment from an existing tree guide. ClustalX will perform only the final multiple alignment of the sequences shown in the window. When the alignment is complete, the window display is updated to reflect the new multiple alignment.

10. *Perform alignment quality control.* To highlight sections of sequences or blocks of alignment that are unreliable or badly aligned in the ClustalX window, select the Show Low Scoring Segments option from the Quality menu.

    Sequence segments which obtain low quality scores are displayed with white characters on a black background (Fig. 2.3.9). These segments may be due to one of various reasons—e.g., (i) partial or total misalignments caused by a failure in the alignment algorithm, (ii) partial or total misalignments because at least one of the sequences in the given set is partly or completely unrelated to the other sequences, or (iii) frameshift translation errors in a protein sequence causing local mismatched regions to be heavily highlighted. The calculation of the ClustalX alignment quality scores is described in the documentation and in the online help.

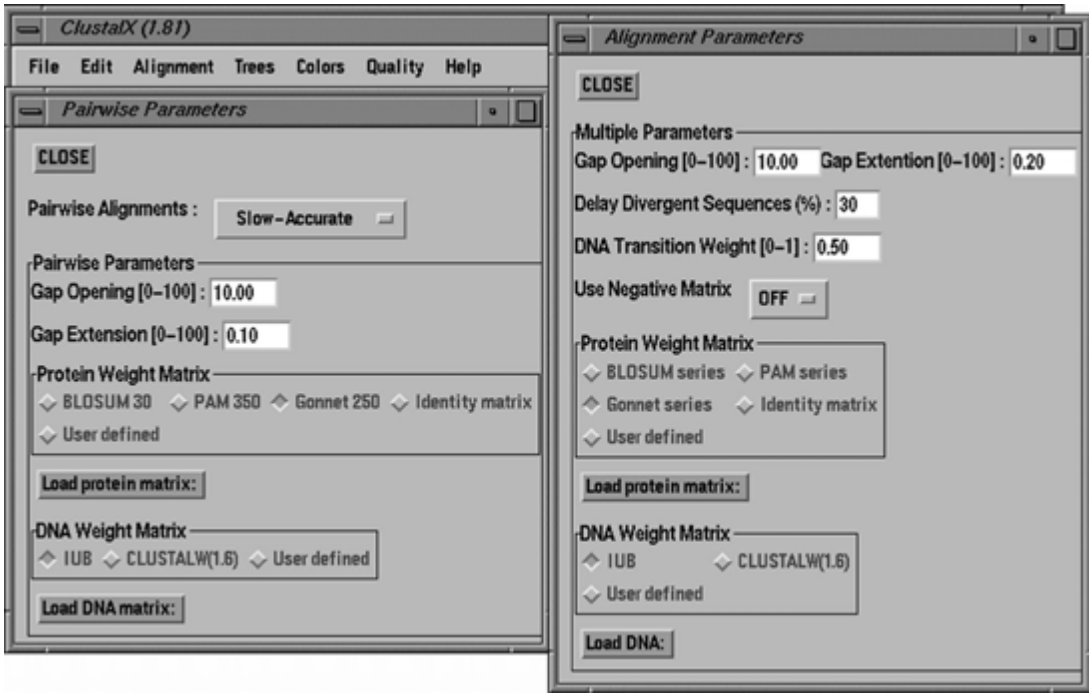

Figure 2.3.7 The windows containing the buttons and (default) settings for the pairwise alignment parameters (left) and the multiple alignment parameters (right).

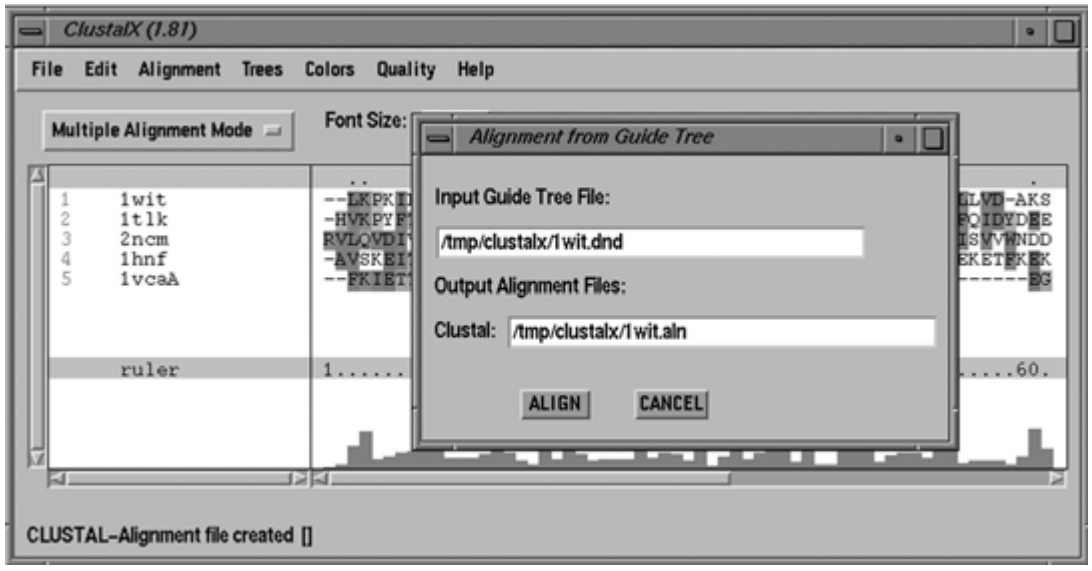

Figure 2.3.8 Producing a new multiple alignment (`1wit.aln`) using an old guide tree file (`1wit.dnd`).

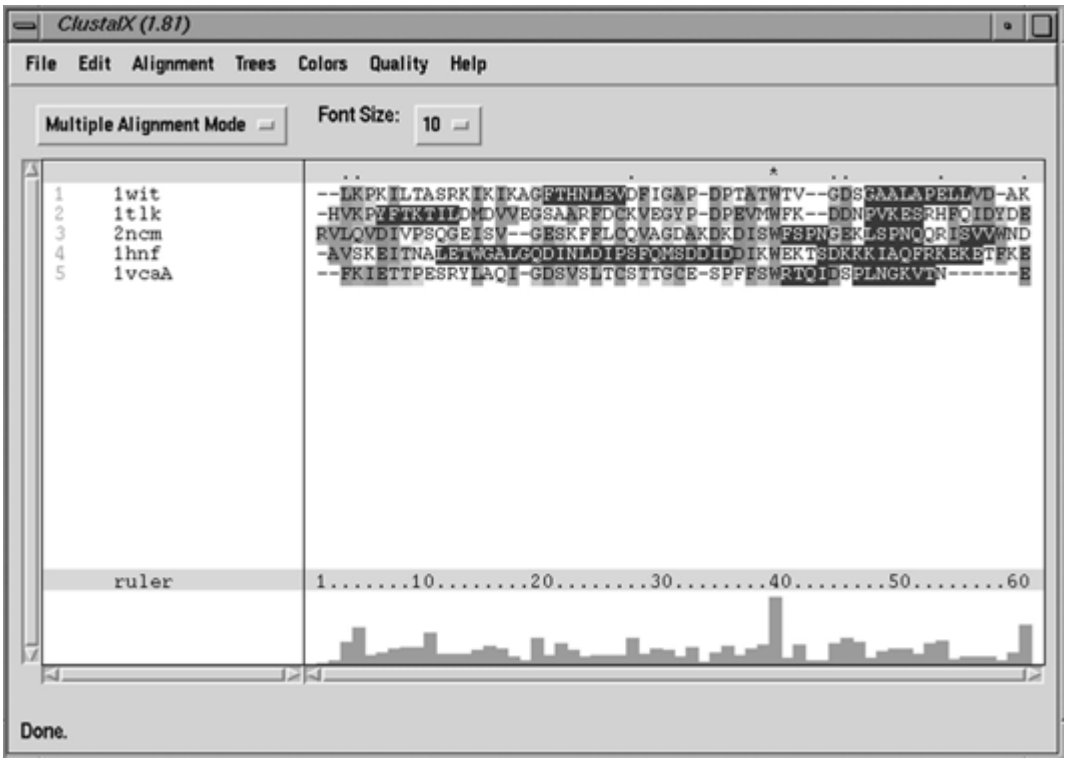

Figure 2.3.9 Window displayed upon selecting the Show Low Scoring Segments option from the Quality menu.

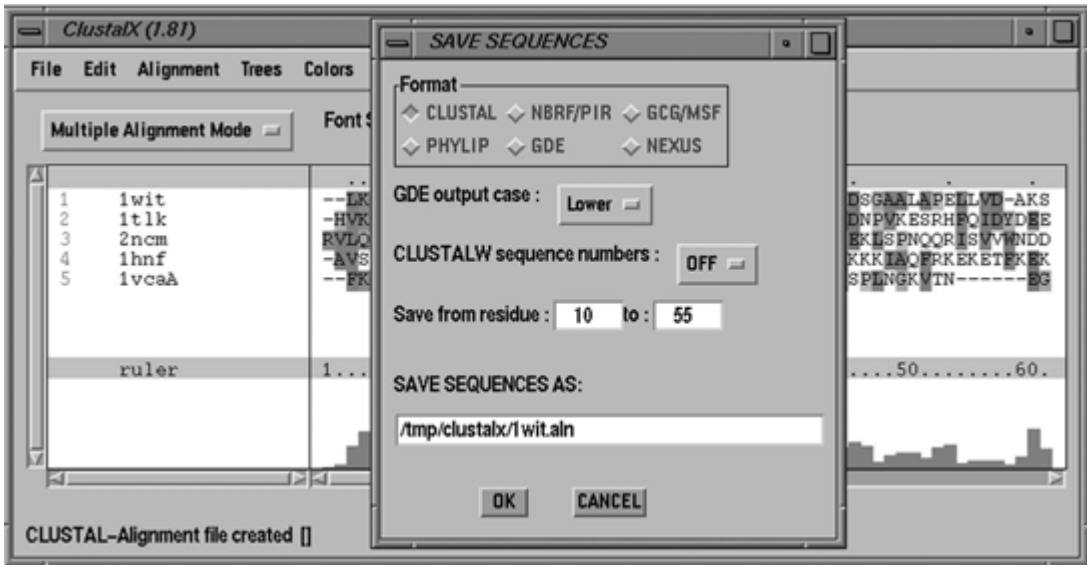

Figure 2.3.10 The Save As menu from ClustalX which is used to save an alignment after it is produced. Alignments are written to output files by default anyway, but this option allows users to save the output afterwards, perhaps in a different format. The full alignment is saved by default; here the user has chosen to save residues 10 to 55.

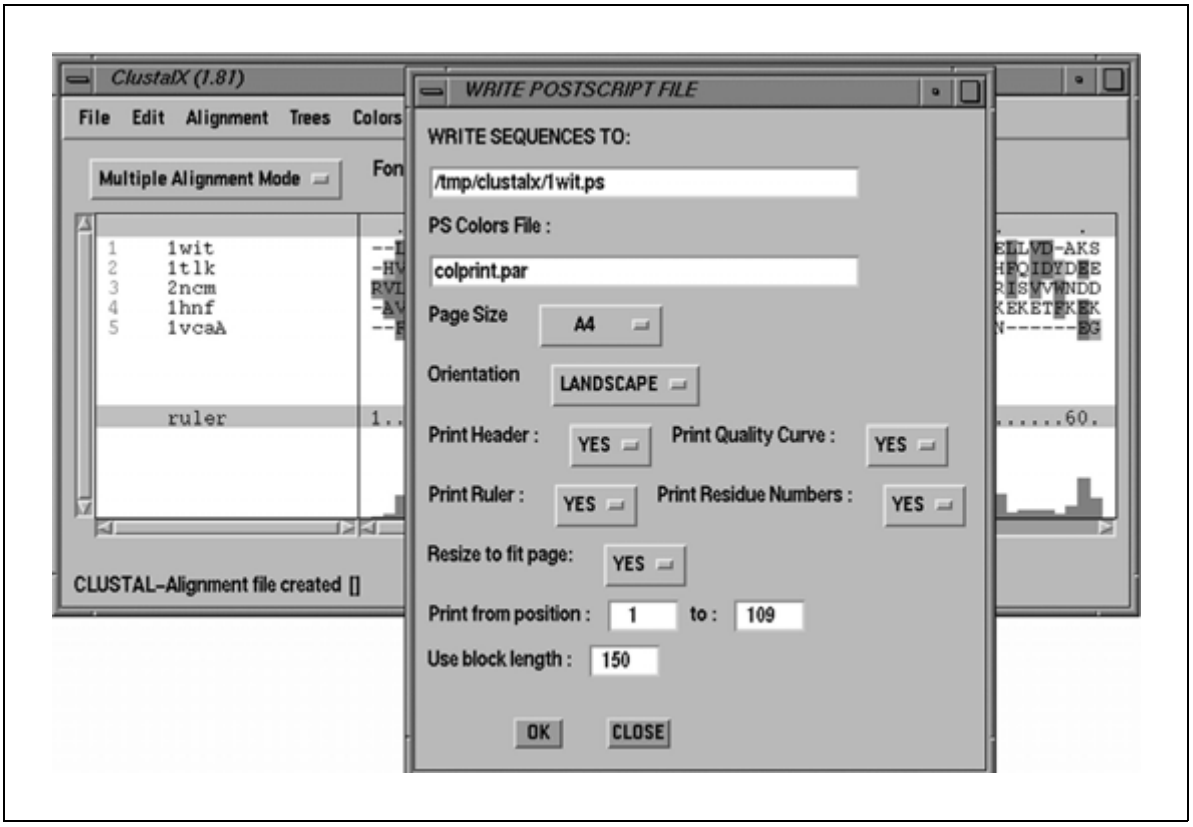

Figure 2.3.11 The PostScript output menu from ClustalX. This is used to save the colored alignment with or without some of the ornamentation in the window.

11. *Save the alignment.* During the alignment process, the final multiple alignment is automatically written to the output file. This file may be specified by the user or the default may be used (the name and the format type are normally chosen by default; see step 6). In addition, after the multiple alignment is completed, the user has the option of changing the output file format or saving only a selected part of the whole alignment and getting the output alignment written out to a file again. Select the Save Sequences As option from the File menu.

    A window will appear (Fig. 2.3.10) offering the user a choice of one of the six output formats (see step 5). Options are also available to switch between Upper/Lower case for GDE files, to output Sequence Numbering for Clustal files, and to save a range of the alignment. In addition, the output filename may be specified by the user. Clicking on the OK button will save the sequence alignment to the selected file.

12. *Create a PostScript image of the alignment.* The ClustalX alignment display can be saved in a PostScript file, which can then be either sent directly to a printer or loaded into a graphics-editing program. This is done by selecting the Write alignment as PostScript option from the File menu.

    A window will appear with a number of options for customizing the PostScript output (Fig. 2.3.11). The options are explained in detail in the ClustalX documentation and online help. The file will automatically include the colored sequences, and the consensus and ruler lines. The Alignment Quality curve can be optionally included in the output file.

USING CLUSTALW AND CLUSTALX FOR PROFILE ALIGNMENTS

ALTERNATE PROTOCOL

ClustalW and ClustalX allow the user to reuse an old alignment and add new sequences to it, or even merge two alignments together. This is known as profile alignment (the term profile analysis was first used by Gribskov et al., 1987). This is useful in any ongoing project where new sequences are being generated and alignments need updating. Adding new sequences to an old alignment has some advantages. First, it is much faster than redoing the alignment from scratch each time. Second, the original sequence alignment is kept intact, which is especially useful if the alignment had been hand-edited. A profile is simply an alignment of two or more sequences (e.g., an alignment output file from Clustal). One or both sets of input sequences may include secondary structure assignments or gap penalty masks to guide the alignment. Profile alignment allows the user to read in an old alignment (in any of the allowed input formats) and align one or more new sequences to it.

Necessary Resources

Hardware

Unix (including Linux) workstation (e.g., Sun, Alpha, Silicon Graphics, PC), PC with MS Windows, or Power Macintosh

Software

ClustalW or ClustalX program (see Support Protocol)

Files

Sequences and existing alignments can be input to both ClustalW and ClustalX in one of seven file formats. All sequences must be in the same file. The formats that are automatically recognized are: NBRF/PIR, EMBL/Swiss-Prot, Pearson (FASTA; *APPENDIX 1B*), Clustal, GCG/MSF, GCG9/RSF, and GDE flat file. In the examples here, unaligned sequences are in FASTA format and existing alignments are in Clustal and GCG/MSF formats.

Merge two existing alignments

1. Download and install ClustalX on a local machine (see Support Protocol).
2. Start a ClustalX session (see Basic Protocol, step 2) and switch to Profile Alignment Mode by clicking on the Multiple Alignment Mode toggle button just above the sequence display area.

   The single sequence display area will be replaced by two display areas (Fig. 2.3.12). Initially, both areas are empty.

3. Load the first profile by selecting the Load Profile 1 option from the File menu. A file selection window will appear, allowing the user to select a file. The procedure is similar to that used for loading unaligned sequences (see Basic Protocol, steps 3 to 4). Profile 1 should contain a single sequence or an existing alignment of two or more sequences, e.g., an alignment file that was produced by ClustalX at an earlier stage (these file names have the extension `.aln`).

   The selected alignment will be displayed in the top half of the ClustalX window (Fig. 2.3.13). See Basic Protocol, step 4, for a description of the alignment display. In Figure 2.3.13, the alignment consists of immunoglobulin superfamily domain sequences, generated with default parameters.

4. Load the second profile by selecting the Load Profile 2 option from the File menu. The procedure is the same as that used for loading the first profile. Profile 2 should contain a single sequence or several aligned sequences.

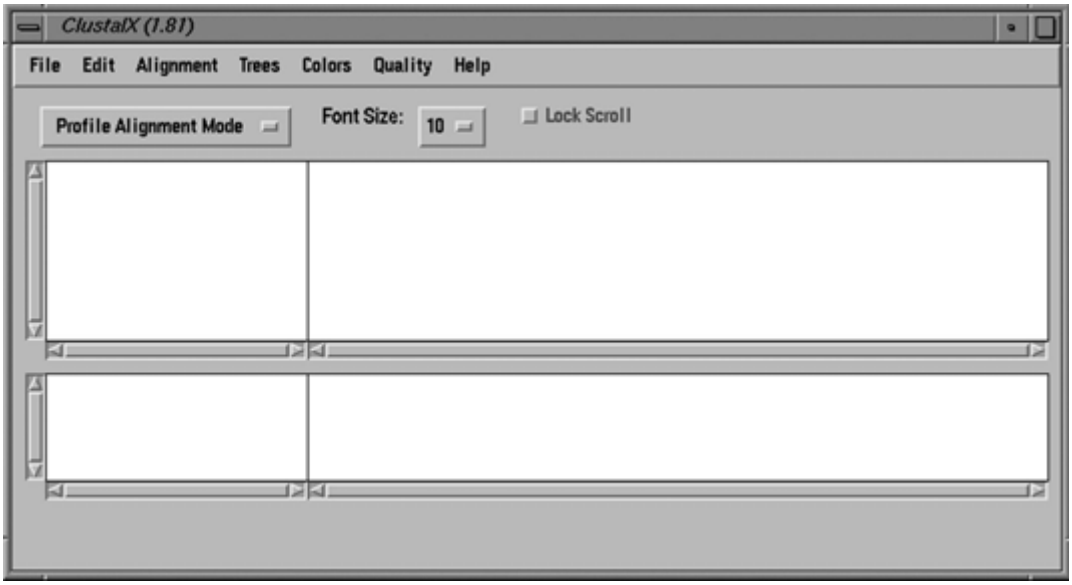

Figure 2.3.12 ClustalX in profile alignment mode before any sequences or profiles are loaded. The two empty windows will hold the two profiles (existing alignments) or groups of sequences.

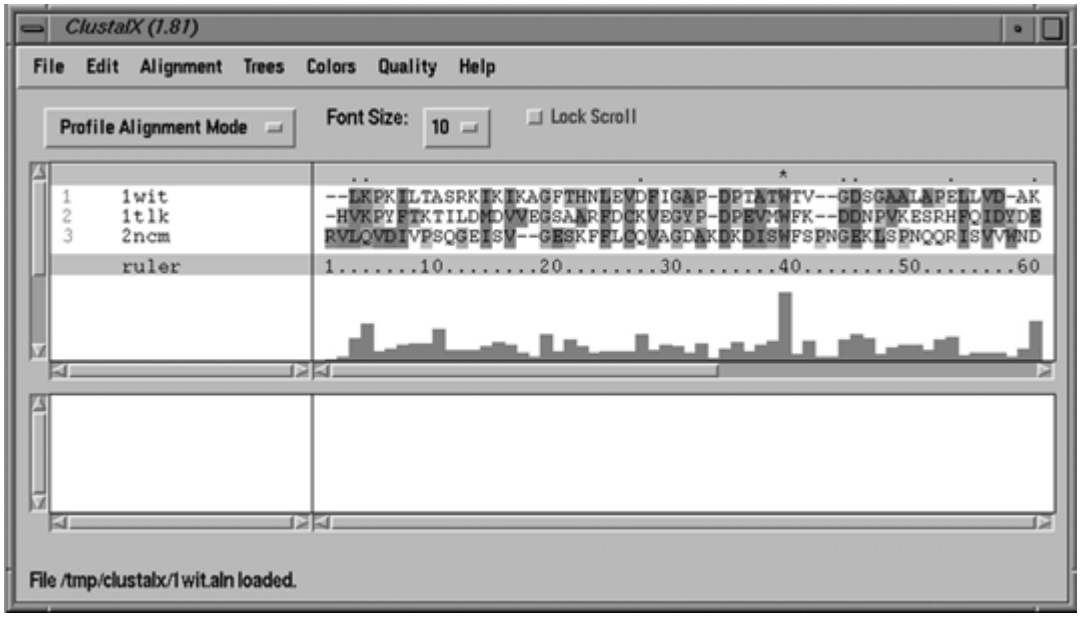

Figure 2.3.13 ClustalX in profile alignment mode after the first profile (a five-sequence alignment) has been loaded (only three are visible in scrollable window).

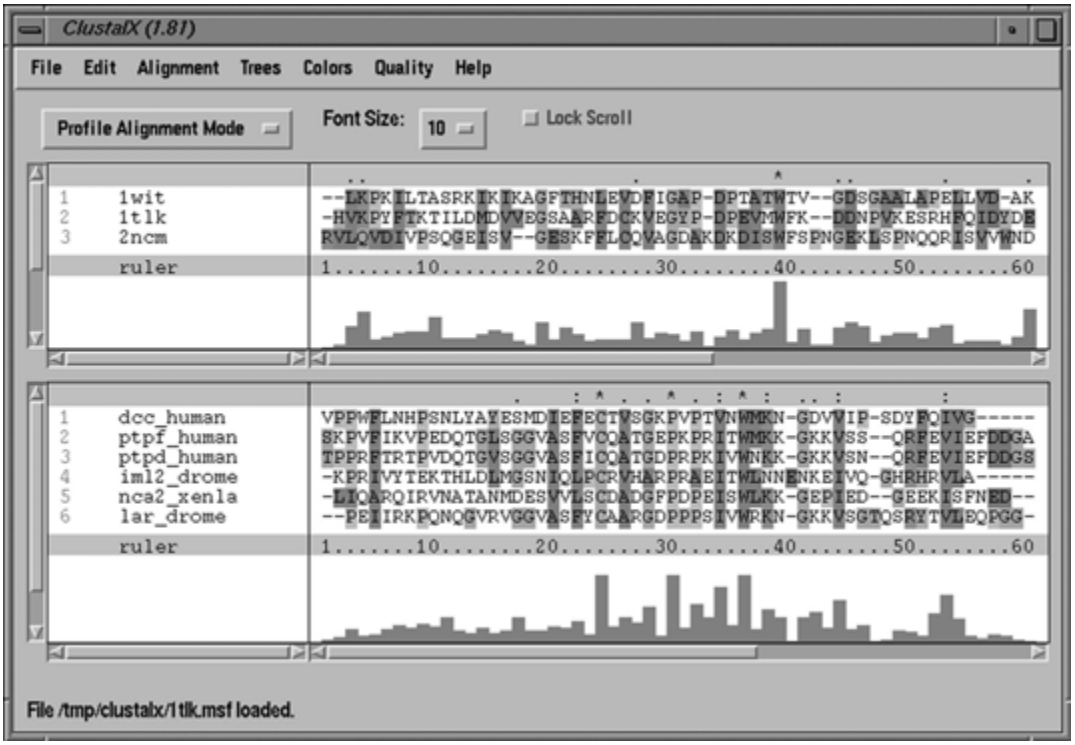

Figure 2.3.14 ClustalX in profile alignment mode with both profiles loaded. Alignment was based on secondary structure superposition and manually adjusted.

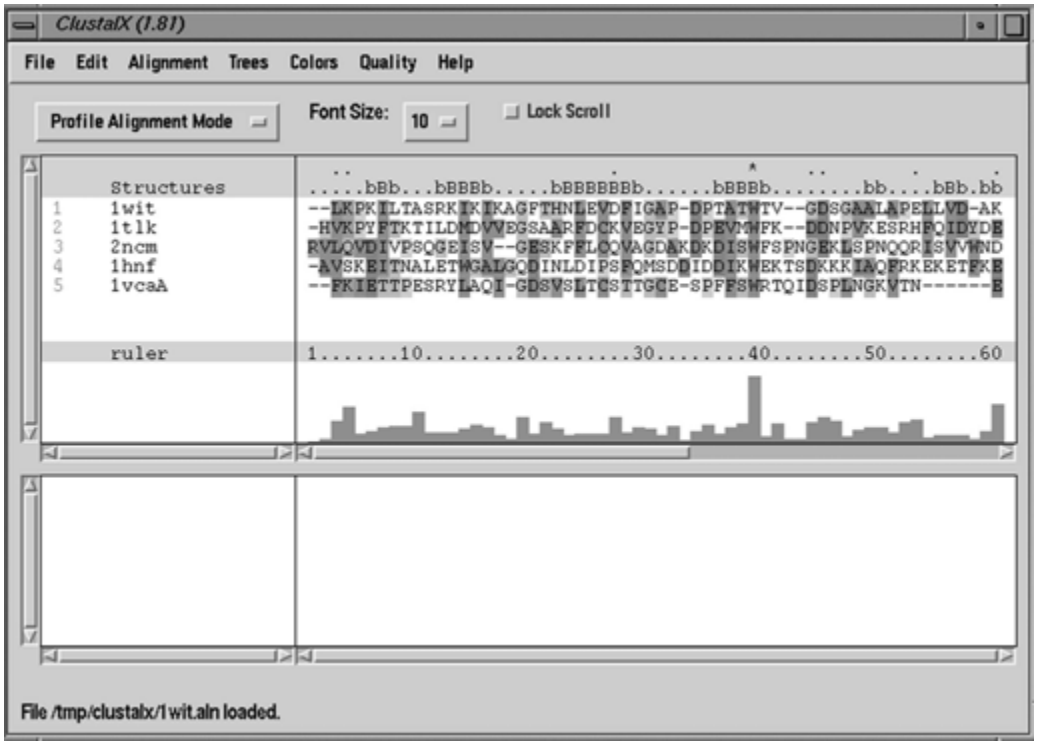

Figure 2.3.15 Window displayed upon loading a profile with a structure mask in Profile Alignment Mode.

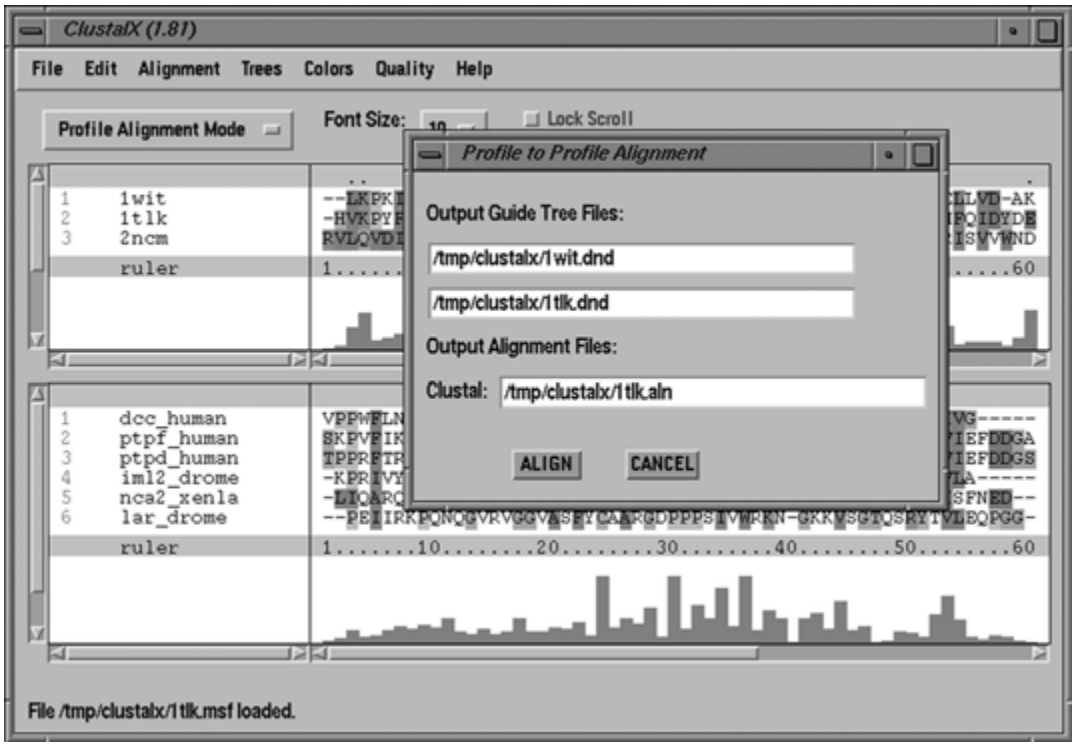

Figure 2.3.16 The default file names for the output files from the profile alignment.

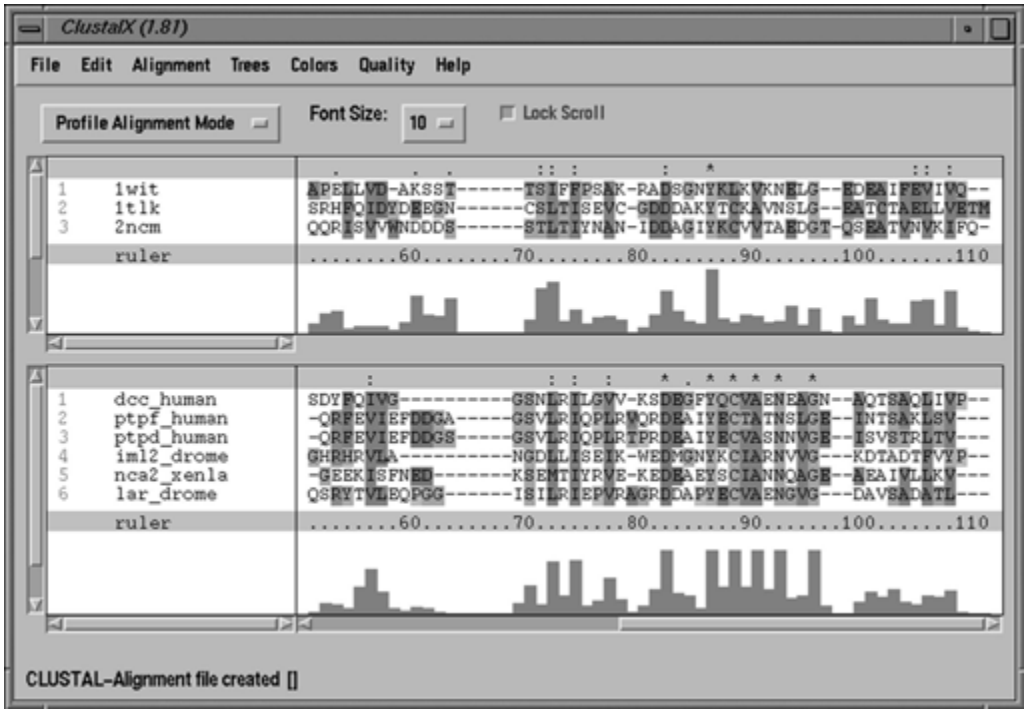

Figure 2.3.17 The two profiles after they have been aligned together. They are still in separate windows but have been locked together by pressing the Lock Scroll button. They are moved together by the single scroll bar at the bottom of the screen.

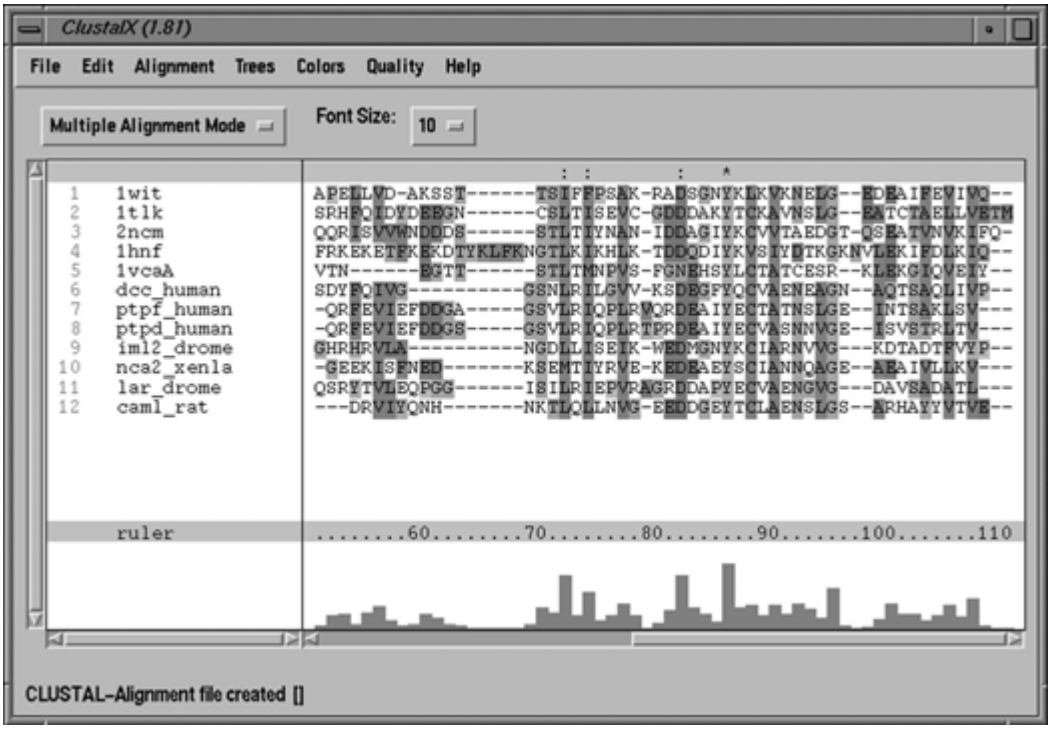

Figure 2.3.18 The final profile alignment can be viewed in a single window by reverting back to Multiple Alignment Mode (from Profile Alignment Mode).

The selected alignment will be displayed in the bottom half of the ClustalX window (Fig. 2.3.14). The example alignment shown here contains sequences belonging to the C-2-type subfamily of the immunoglobulins.

5. *Optional:* Supply secondary structure and/or gap penalty masks with the input sequences used during profile alignment (note that the secondary structure information is not used during multiple sequence alignment).

   The secondary structure elements can be read from Swiss-Prot, Clustal, or GDE format input files. For many 3-D protein structures, secondary structure information is recorded in the feature tables of Swiss-Prot database entries and ClustalX recognizes Swiss-Prot HELIX and STRAND assignments. Alternatively, the Clustal or GDE files can be edited manually. The format for the masks is described in the documentation and in the online help.

   ClustalX reads the structure or gap penalty masks automatically when a profile is loaded in Profile Alignment Mode and displays the information in the ClustalX window above the alignment display (Fig. 2.3.15). The masks work by raising gap penalties in specified regions (typically secondary structure elements) so that gaps are preferentially opened in the less well conserved regions (typically surface loops). The values for raising the gap penalty at particular secondary structure elements may be modified using the Alignment Parameters, Secondary Structure Parameters options from the Alignment menu.

6. Align the two profiles by selecting Align Profile 2 to Profile 1 from the Alignment menu. A window will appear (Fig. 2.3.16) that displays the default filenames for the output guide tree files and the output alignment file. If required, these filenames may be edited by the user before clicking on the Align button.

   ClustalX will align the two profiles together to form a single multiple alignment. The original alignments are not altered. The two profiles are simply aligned together by

introducing complete columns of gaps into one or both of the profiles. The current status of the alignment process is continuously updated in the message area at the bottom of the ClustalX window. When the alignment is complete, the window display areas are updated to show the aligned profiles. Clicking on the Lock Scroll button just above the top display area will remove the horizontal scroll bar from the top display area (Fig. 2.3.17). The single remaining scroll bar at the bottom of the window will then allow both profile display areas to be scrolled together.

A second option is to align the sequences from the second profile, one at a time, to the first profile. This is useful for incorporating a set of new sequences (not aligned) into an older alignment. The procedure to follow is very similar to that used above to merge two existing alignments. In this case, however, the second profile should contain one or more unaligned sequences. Each sequence is aligned individually with the existing alignment, starting with the most closely related. In step 6 above, the sequences can be aligned to profile 1, by selecting the Align Sequences to Profile 1 option from the Alignment menu.

7. Merge the two profiles by switching back to multiple alignment mode using the toggle button just above the top sequence display area.

   The sequences from both profiles are merged into a single alignment (Fig. 2.3.18).

SUPPORT PROTOCOL

OBTAINING THE CLUSTALW AND CLUSTALX PROGRAMS

The Clustal series of programs are available by anonymous FTP from *ftp-igbmc.u-strasbg.fr* or *ftp.ebi.ac.uk*. ClustalW is written in ANSI standard C and has been tested on a number of Unix platforms, including DEC, SGI, and Sun, as well as Macintosh and PC systems. However, it can be compiled on any platform which supports a C compiler. Executable programs are supplied for Power Macintosh computers and for PCs running either the Windows or DOS operating systems. ClustalX uses the Vibrant multiplatform user interface development library, developed by the National Center for Biotechnology Information (Bldg. 38A, NIH 8600 Rockville Pike, Bethesda, MD 20894) as part of their NCBI Software Development Toolkit. As executable programs are supplied for most major platforms, it is not usually necessary to download the Vibrant toolkit in order to use ClustalX. To compile ClustalX on an unsupported platform, the toolkit should be obtained by anonymous FTP from *ftp://ncbi.nlm.nih.gov*.

Necessary Resources

Hardware

Unix (including Linux) workstation (Sun, Alpha, Silicon Graphics, PC), PC with either MS-DOS or MS Windows, Power Macintosh, or any other computer supporting a C compiler

1a. To obtain the latest ClustalW software, run an FTP session as follows:

```
%ftp ftp-igbmc.u-strasbg.fr
Name: anonymous
Password: [your internet address]
ftp> cd pub/ClustalW
ftp> binary
ftp> get clustalw1.81.DOS.zip (for PC computers)
ftp> get clustalw1.81.PPC.sea.Hqx (for Macintosh)
ftp> get clustalw1.81.UNIX.tar.gz (for Unix systems)
ftp> quit
```

1b. Similarly, to obtain the latest ClustalX software:

```
%ftp ftp-igbmc.u-strasbg.fr
Name: anonymous
```

```
Password: [your internet address]
ftp> cd pub/ClustalX
ftp> binary
ftp> get clustalx1.81.msw.zip (for MS Windows)
ftp> get clustalx1.81.PPC.sea.Hqx (for Macintosh)
ftp> get clustalx1.81.sgi.tar.gz (for Silicon Graphics)
ftp> get clustalx1.81.sun.tar.gz (for Sun Solaris)
ftp> get clustalx1.81.alpha.tar.gz (for Alpha)
ftp> get clustalx1.81.linux.tar.gz (for Linux ELF for x86 PCs)
ftp> quit
```

2. Complete instructions for compilation and installation are available in the `README` files included in the ClustalW/X distributions. Manuals for ClustalW and ClustalX are available on the Web at *http://www-igbmc.u-strasbg.fr/BioInfo/ClustalW/Top.html* and *http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html* respectively.

GUIDELINES FOR UNDERSTANDING RESULTS

Once an alignment has been carried out, the main output is the alignment itself. This is usually contained in a file called `x.aln` if `x.pep`, for example, was the name of the input sequence file. This is a simple text file that can be viewed using any text editor (e.g., Windows Notepad) or word-processing software. An example output file for seven globin sequences is shown in Figure 2.3.19. This is a simple text file and the user must view it using Courier or some other fixed-space font. It may be necessary to adjust the font size or margins to prevent line wrapping in the middle of the alignment. The stars indicate columns of identical residues (as explained in the Basic Protocol) and the colons and dots indicate columns where there is some conservation of the biochemical character of the side chains. A more immediate and visually striking representation of column conservation is of course provided by the ClustalX window display. Interpretation of this alignment is usually up to the user and depends on what one is looking for. These text alignments are useful for importing into other packages such as PHYLIP (see Internet Resources; *UNIT 6.3*) for further analysis.

One general problem of interpretation is in deciding if a set of sequences are well aligned or if, indeed, they are related to each other at all. This is sometimes phrased informally as: "is this alignment significant?" Significance or otherwise of these alignments is a very difficult thing to decide in a statistical sense, but it is possible to take some simple steps to check if the alignment is reasonable and if all of the sequences belong in the alignment. First, check the overall look of the alignment. Real alignments of homologous sequences will have relatively neat-looking blocks of alignment separated by sections that are full of gaps. This is perfectly normal, and the gaps usually just indicate loop regions with no conserved core secondary structure. An examination of the pattern of conservation in the conserved blocks will usually indicate some runs of partially or weakly conserved columns. These can be seen by runs of stars or dots in the text output or from neat columns of color in the ClustalX display. For example, the sequences shown in Figure 2.3.17, profile 2, all belong to the C-2 type subfamily of the immunoglobulins and share more than 25% residue identity. The alignment contains a number of completely conserved columns, indicated by stars in the ClustalX display. An example of more distantly related sequences is shown in Figure 2.3.6. These five sequences all belong to the immunoglobulin superfamily and share the same 3-dimensional fold, although their sequence similarity is low (less than 22% residue identity between any two sequences).

Figure 2.3.19 A sample text output file (`x.aln`) showing the alignment (obtained with default parameters) of seven globin sequences. The stars, dots and colons below the alignment indicate degree of conservation in the columns.

By contrast, if the sequences are not all homologous, there will be very few stars or dots in the text output and these will not be found in short runs (normally corresponding, e.g., to active sites or binding sites). There will be gaps everywhere, indicating that there is no pattern of conserved core regions separated by variables loops. Finally, the use of the Quality menu items in the main menu of ClustalX will provide simple and striking visual guides to columns, residues, or sequences that are very dubious. Of course, in reality, it is possible to have a mixture of well aligned regions and regions where the alignment is effectively random, as will happen with multidomain proteins which share just one or two homologous domains. This can, however, also happen if one or more proteins have frameshift mutations or mistakes from the sequencing of their coding regions. This will cause a sudden shift from well conserved blocks to nonsense alignment.

The second output file from most analyses contains the dendrogram. An example is shown in Figure 2.3.20. This is a description of the approximate relationships between the input sequences, in the format of a phylogenetic tree. The word dendrogram is used to help remind users that these are not intended to be used as phylogenetic trees. Rather, these are used by ClustalX and ClustalW to carry out progressive alignments. Nonetheless, these can be viewed using Manolo Gouy's NJplot program, which is supplied with Clustal (also available by anonymous FTP from *ftp://pbil.univ-lyon1.fr/pub/mol_phylogeny/njplot*) or with Rod Page's Treeview program (*UNIT 6.2*). Normally, it is not worthwhile bothering

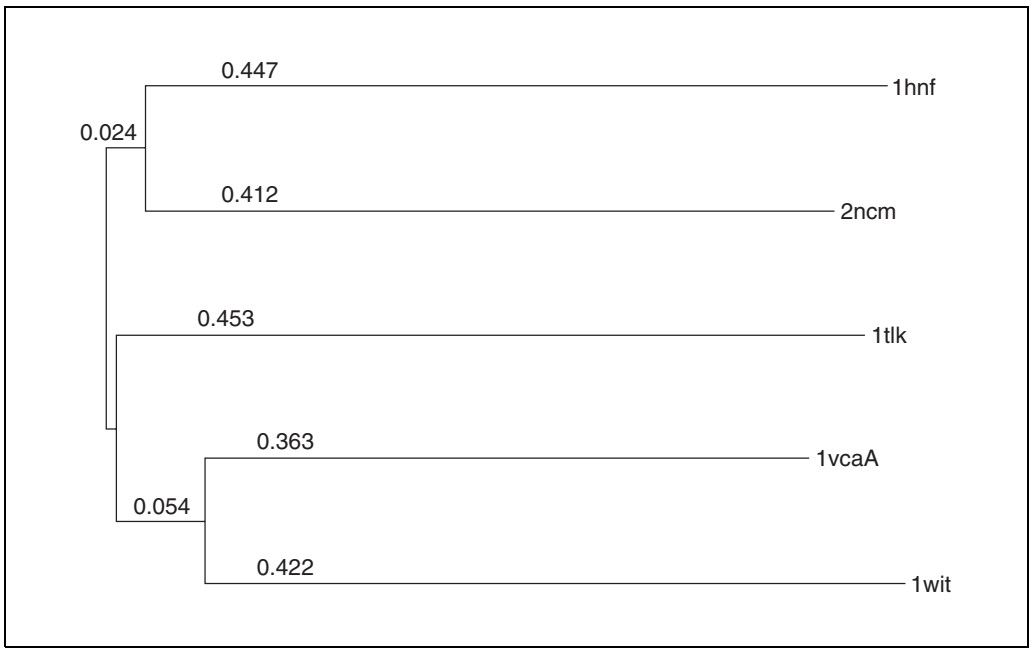

Figure 2.3.20 Dendrogram of the alignment shown in Figure 2.3.6.

with these files, but they can be very instructive if there is a suspicious sequence. An outlier can be seen in the dendrogram when viewed as a tree. It will be on a very long branch from the roots of the tree. Sequences with frameshifts (in the underlying DNA sequence) will be seen as unusually long branches, but not necessarily from the root.

COMMENTARY

Background Information

Progressive alignment

All of the Clustal programs carry out what is called progressive alignment. This term was introduced by Feng and Doolittle (1987), but the first clear description of the method in the Clustal programs comes from Taylor (1988). An earlier method by Hogeweg and Hesper (1984) also described the essential elements of progressive alignment. Overall, the approach is based on gradually building up the multiple alignment by merging larger and larger subalignments. Each merge is carried out using standard dynamic programming (Needleman and Wunsch, 1970; Gotoh, 1982; *UNIT 3.1*) which finds an alignment that is guaranteed to have the best score given gap penalties and an amino acid weight matrix (*UNIT 3.5*). A number of programs are available that perform progressive alignments. A comparison study (Thompson et al., 1999) of some of the more widely used ones showed that Clustal generally performed better for a wide variety of different alignment sets. This, together with Clustal's portability and ease of use, has made Clustal one of the most popular alignment programs in use today.

The order in which the sequences are merged is determined, most simply, by following the branching order of a dendrogram. Starting at the tips, the program first aligns the closest two sequences. These two sequences are then kept fixed to each other and any gaps that were introduced in either sequence cannot be moved later. Next, the program moves down the tree and either aligns two new sequences or aligns the first alignment with a new sequence to give a subalignment of three sequences. This process follows the branching order in the tree from the tips to the root and at each step merges two sequences, two sub-alignments, or a sequence with a subalignment. All alignments are carried out by taking full account of all of the amino acids at all positions in the sequences to be aligned next. Once the tree is given at the start, this progressive alignment is very fast, even with huge numbers of sequences or very long sequences.

Once there is a dendrogram of the sequences, it is possible to carry out progressive

alignments as described above. These trees do not have to be very accurate for the procedure to work, although we do expect the quality of the alignment to be poor if the tree is markedly wrong. Even if the dendogram has the correct branching order, there is no guarantee that the alignment will be correct. There are always at least some positions that are not ideally aligned or where the alignment is ambiguous. This is especially true when the sequences are highly divergent. The goal is to build the alignment, starting with the easiest alignments. This is why the algorithm starts at the tips of the dendrogram, where the sequences are closely related. By the time it gets to the harder alignments between the more distantly related sequences, the alignment already contains some information about conservation or lack of it at each position in the subalignments. In general, progressive alignment methods are widely used because they are so fast and because the quality of the alignments is very high.

One problem that arises is how to derive the dendrogram in the first place. Trees are normally calculated from multiple alignments, but the multiple alignment does not exist before the progressive alignment. The dendrogram is calculated by the simple trick of first comparing all the unaligned sequences to each other. This provides a similarity score (percent identity) between each pair of sequences and these can be used to make a simple distance based tree using the Neighbor-Joining method (Saitou and Nei, 1987; *UNIT 6.3*). This tree is written to the dendrogram (`.dnd`) file and used to drive the progressive alignment. For N sequences, this requires the calculation of $(N \times N - 1)/2$ pairwise alignments. For large N, this can require thousands of alignments, so ClustalW and ClustalX do offer the chance to calculate these using a fast approximate method (Wilbur and Lipman, 1983) instead of the more accurate but slower dynamic programming (see Basic Protocol, step 8; Myers and Miller, 1988).

Versions of Clustal

The first Clustal programs were run on PCs only and were written as a series of stand-alone Fortran programs (Clustal1-4) that were run one after another in order to produce the multiple alignment (Higgins and Sharp, 1988, 1989). These were later replaced by a single Fortran program that could be run on Unix or VAX/VMS machines and was simply called Clustal. The current menu style of ClustalW dates from this program. ClustalV (Higgins et al., 1992) was the first version to be written in C and this featured the ability to produce phylogenetic trees (*UNIT 6.1*), with bootstrap confidence measures (Felsenstein, 1985) from alignments, using the Neighbor-Joining method (Saitou and Nei, 1987; *UNIT 6.3*). This version was a single program that could be run on all platforms (e.g., Mac, PC, and Unix) and also featured a simple command line as well as a text menu interface.

ClustalW (Thompson et al., 1994) was derived from ClustalV by the addition of numerous new features for improving the sensitivity of protein alignments and for extending the functionality of the interface. This was the first version to be actively maintained and to feature version numbers. The most recent version number (January 2002) is 1.81. The program can read and write in many different file formats and there are dozens of parameters for controlling the details of the alignments. There are extensive facilities for adding sequences to old alignments, thus allowing users to maintain alignments of their sequences. This program is, essentially, the one still in use today. ClustalX (Thompson et al., 1997) was based directly on ClustalW but featured a user-friendly graphical user interface and extensive graphical features for annotating alignments. ClustalX and ClustalW programs with the same version numbers are expected to produce identical alignments and use the same underlying code.

Critical Parameters and Troubleshooting

The quality of the multiple alignment will depend heavily on the sequences included in the alignment set. When the sequences are closely related, almost any set of alignment parameters will find the correct solution. With very divergent sequences, however, the parameters used will become critically important (Doolittle, 1986; Rost, 1999). For example, the Negative Matrix option should be used when the sequences to be aligned do not align well globally because they only have domains in common. Nevertheless, it has been shown (Thompson et al., 1999) that overall alignment quality improves when more sequences are included in the alignment. Thus, it is important to include as many sequences as possible in order to provide more information about the patterns of residue conservation for the family. For DNA sequences coding for protein, it is almost always better to compare the protein translations than to compare DNA directly

(Pearson, 2000; States et al., 1991) because after only a small amount of evolutionary change, the DNA sequences contain less information with which to detect homology.

There are three main groups of parameters that can be set to control the alignments: pairwise parameters, multiple alignment parameters and protein gap parameters. These are found under the Multiple Alignment, Alignment Parameters option of ClustalX (see Fig. 2.3.7). The first group control the way the initial alignments that are used to generate the dendrogram are made. It is not usually worth changing these except to choose between slow accurate alignments (the default) or fast/approximate alignments, which use the method of Wilbur and Lipman (1983). This will have a huge affect on the speed of alignment, but this will not be noticed unless you have many long sequences. In terms of changing the alignment, the most that these parameters can do is to change the branching order in the dendrogram. This can have an effect on the final multiple alignment quality, but the changes will be hard to predict from the choices made in the menu. Further discussion of pairwise alignments, and DNA and protein scoring matrices, can be found in Chapter 3.

The second group of parameters will control the alignments that are used to build up the multiple alignment. These allow you to set the main gap penalties and weight matrix (*UNIT 3.5*), for example. These can be used to change the alignment by making gaps happen more easily or by encouraging long gaps, but the effects can be complicated. These parameters are modified in various complicated ways by the final set of parameters (the protein gap parameters). One parameter here that is very important is the one that allows you to use a negative matrix or not. This controls whether the amino acid weight matrix will contain positive values only or positive and negative. The former is the default, but it is sometimes necessary to choose the latter, especially if you have large terminal deletions or fragments of sequences. The protein gap parameters are used by ClustalX to control the way gaps are placed in protein alignments.

Literature Cited

Doolittle, R.F. 1986. Of URFs and ORFs: A primer on how to analyze derived amino acid sequences. University Science Books, Mill Valley, Ca.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783-791.

Feng, D.-F. and Doolittle, R.F. 1987. Progressive sequence alignment as a pre-requisite to correct phylogenetic trees. *J. Mol. Evol.* 25:351-360.

Gotoh, O. 1982. An improved algorithm for matching biological sequences. *J. Mol. Biol.* 162:705-708.

Gribskov, M., McLachlan, A.D., and Eisenberg, D. 1987. Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad. Sci. U.S.A.* 84:4355-4358.

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

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

Higgins, D.G., Bleasby, A.J., and Fuchs R. 1992. CLUSTAL V: Improved software for multiple sequence alignment. *Comp. Appl. Biosci.* 8:189-191.

Hogeweg, P. and Hesper, B. 1984. The alignment of sets of sequences and the construction of phyletic trees: an integrated method. *J. Mol. Evol.* 20:175-186.

Myers, E.W. and Miller, W. 1988. Optimal alignments in linear space. *CABIOS* 4:11-17.

Needleman, S.B. and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48:443-453.

Pearson, W.R. 2000. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* 132:185-219.

Rost, B. 1999. Twilight zone of protein sequence alignments. *Protein Eng.* 12:85-94.

Saitou, N. and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4: 406-425.

States, D.J., Gish, W., and Altschul, S.F. 1991. Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. *Methods* 3:66-70.

Taylor, WR. 1988. A flexible method to align large numbers of biological sequences. *J. Mol. Evol.* 28:161-169.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673-4680.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. 1997. The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25:4876-4882.

Thompson, J.D., Plewniak, F. and Poch, O. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.* 27:2682-2690.

Wilbur, W.J. and Lipman, D.J. 1983. Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. U.S.A.* 80:726-730.

Key References

Jeanmougin, F., Thompson, J.D., Gouy, M., Higgins, D.G., and Gibson, T.J. 1998. Multiple sequence alignment with ClustalX. *Trends Biochem Sci.* 23:403-405.

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

Both of these articles give extensive background and descriptive details as to what exactly happens when you try to use Clustal and what all of the parameters mean. They are intended for a lay, nontechnical audience.

Internet Resources

http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html

Get information on or download ClustalX.

http://www.ebi.ac.uk/clustalw/

Run ClustalW at the EBI using the Web.

http://cmgm.stanford.edu/phylip/

PHYLIP (Phylogeny Inference Package) version 3.5c., by J. Felsenstein. Department of Genetics, University of Washington, Seattle.

Contributed by Julie D. Thompson
Institut de Génétique et de Biologie Moléculaire et Cellulaire
Illkirch Cedex, France

Toby. J. Gibson
European Molecular Biology Laboratory
Heidelberg, Germany

Des G. Higgins
University College
Cork, Ireland