Jalview 2.5

A manual and introductory tutorial

David Martin, James Procter, Andrew Waterhouse and Geoff Barton

College of Life Sciences, University of Dundee

Dundee, Scotland DD1 5EH, UK

Manual version 1.2

8th June 2010

Contents

Chapter 1

Basics

1.1 Introduction

1.1.1 Jalview

Jalview is a multiple sequence alignment viewer, editor and analysis tool. Jalview is designed to be platform independent (running on Mac, MS Windows, Linux and any other platform that supports Java), capable of editing and analysing large alignments (thousands of sequences) with minimal degradation in performance, and able to show multiple integrated views of the alignment and other data. Jalview can read and write many common sequence formats including FASTA, Clustal, MSF(GCG) and PIR.

There are two types of Jalview program. The **Jalview Desktop** is a stand alone application that provides powerful editing, visualization, annotation and analysis capabilities. The **JalviewLite** applet has the same core visualization, editing and analysis capabilities as the desktop, without the desktop's webservice and figure generation capabilities. It is designed to be embedded in a web page[1], to allow customisable display of alignments for web sites such as **pfam**[2].

Jalview 2.5 was released in May 2010. The Jalview Desktop in this version provides access to sequence, alignment and protein structure databases, and alignment and analysis web services, and includes the Jmol[3] protein structure viewer. It is also a Distributed Annotation System (DAS) client[4] which facilitates the retrieval and display of third party sequence annotation in association with sequences and any associated structure.

[1] A demonstration version of Jalview (Jalview Micro Edition) also runs on a mobile phone but the functionality is limited to sequence colouring.

[2] `http://pfam.sanger.ac.uk`

[3] Provided under the LGPL licence at `http://www.jmol.org`

[4] with thanks to Andreas Prlic

1.1.2 Jalview's Capabilities

Figure 1.1.2 gives an overview of the main features of the Jalview desktop application. Its primary function is the editing and visualization of sequence alignments, and their interactive analysis. Tree building, principal components analysis, physico-chemical property conservation and sequence consensus analyses are built in to the program. Web services enable Jalview to access remote alignment and secondary structure prediction programs, as well as to retrieve protein and nucleic acid sequences, alignments, protein structures and sequence annotation. Sequences, alignments, trees, structures, features and alignment annotation may also be exchanged with the local filesystem. Multiple visualizations of an alignment may be worked on simultaneously, and the user interface provides a comprehensive set of controls for colouring and layout. Alignment views are dynamically linked with Jmol structure displays, a tree viewer and spatial cluster display, facilitating interactive exploration of the alignment's structure. The application provides its own Jalview project file format in order to store the current state of an alignment and analysis windows. Jalview also provides WYSIWIG[5] style figure generation capabilities for the preparation of alignments for publication.

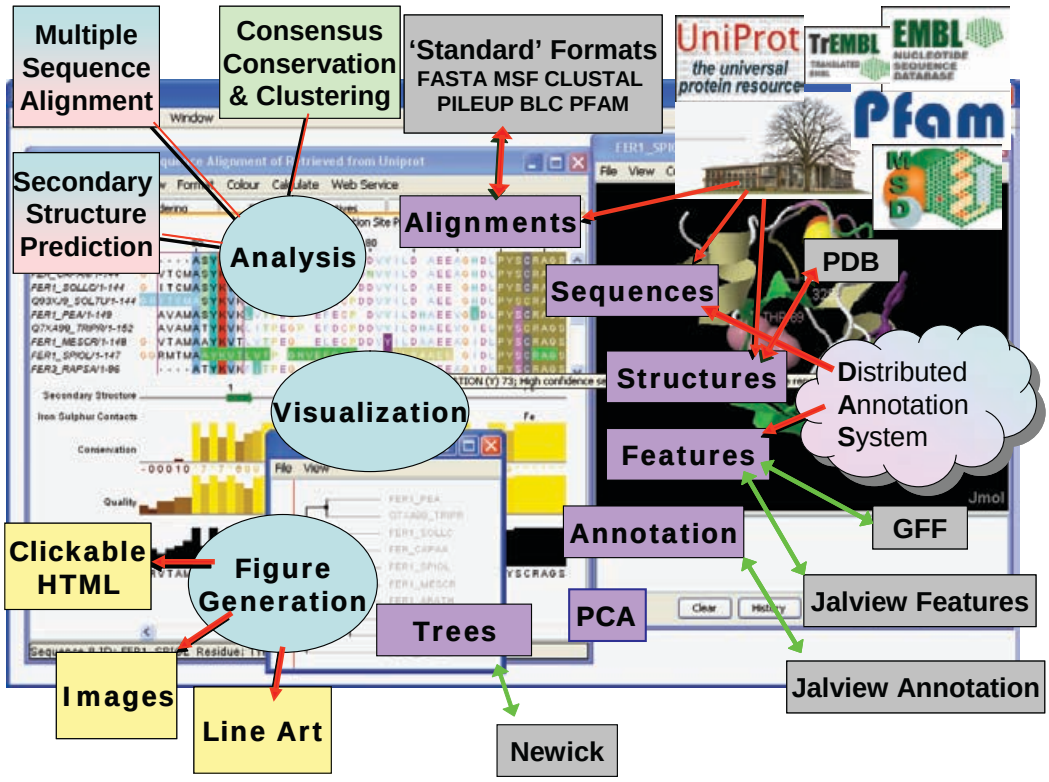


Figure 1.1: **Capabilities of the Jalview Desktop.** The Jalview Desktop Application provides a stable environment for the creation, editing and analysis of alignments and the generation of figures.

[5]WYSIWIG: What You See Is What You Get.

Jalview History

Jalview was initially developed in 1996 by Michele Clamp, James Cuff, Steve Searle and Geoff Barton at the University of Oxford and then the European Bioinformatics Institute. Development of Jalview 2 was made possible with eScience funding from the BBSRC[6] in 2004, enabling Andrew Waterhouse and Jim Procter to re-engineer the original program to introduce contemporary developments in bioinformatics and take advantage of the latest web and Java technology. Jalview's development is now supported for a further 5 years from October 2009 by an award from the BBSRC's Tools and Resources fund.

Citing Jalview

If you use Jalview in your work you should cite:
"Jalview Version 2 - a multiple sequence alignment editor and analysis workbench"
Waterhouse, A.M., Procter, J.B., Martin, D.M.A, Clamp, M. and Barton, G. J. (2009)
Bioinformatics doi: 10.1093/bioinformatics/btp033

This paper supersedes the original Jalview publication:
"The Jalview Java alignment editor"
Michele Clamp, James Cuff, Stephen M. Searle and Geoffrey J. Barton (2004)
Bioinformatics **20** 426-427.

1.1.3 About this tutorial

This tutorial is written in a manual format with short exercises where appropriate, typically at the end of each section. This chapter concerns the basic operation of Jalview and should be sufficient for those who just want to load Jalview (Section 1.2), open an alignment (Section 1.4), perform basic editing and colouring (Section 1.6 and Section 1.7), and produce publication and presentation quality graphical output (Section 1.8).

Chapter 2 covers the additional visualization and analysis techniques that Jalview provides. This includes working with the embedded PDB structure viewer, building and viewing trees and PCA plots, and using trees for sequence conservation analysis. The use of the Jalview webservices for alignment and secondary structure prediction is described in Section 2.3. Following this, Section 2.4 details the creation and visualization of sequence and alignment annotation, and the retrieval of sequences and annotation from databases and DAS Servers. Finally, Section 2.6 discusses specific features of use when working with nucleic acid sequences and protein coding regions.

[6]Biotechnology and Biological Sciences Research Council grant *"VAMSAS: Visualization and Analysis of Molecules, Sequence Alignments and Structures"*, a joint project to enable interoperability between Jalview, TOPALi and AstexViewer.

Typographic Conventions

Keystrokes using the special non-symbol keys are represented in the tutorial by enclosing the pressed keys with square brackets (*e.g.* [RETURN] or [CTRL]). Keystroke combinations are combined with a '-' symbol (*e.g.* [CTRL]-C means press [CTRL] and the 'C' key). Menu options are given as a path from the menu that contains them - for example *File* ⇒ *Input Alignment* ⇒ *From URL* means to select the 'From URL' option from the 'Input Alignment' submenu of a window's 'File' dropdown menu.

1.2 Obtaining and starting The Jalview Desktop Application

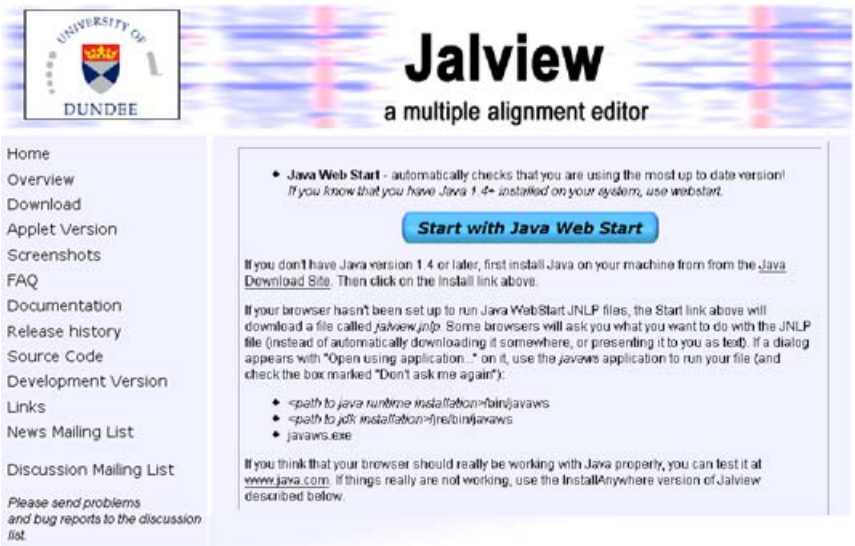


Figure 1.2: **Download page on the Jalview web site**

This tutorial is based on the application version of Jalview, the Jalview Desktop. Much of the information will also be useful for users of the JalviewLite applet, which has the same core editing, analysis and visualization capabilities (see the JalviewLite Applet Examples page for examples). The Jalview Desktop, however, is much more powerful, and includes additional support for interaction with external web services, and production of publication quality graphics.

The Jalview Desktop can be run in two ways; as an application launched from the web via Java Web Start, or as an application loaded onto your hard drive. Both versions are obtained from the Download page at the the Jalview web site (http://www.jalview.org/).

Jalview can be started directly with webstart by navigating to the Download page (via the menu on the left hand side), and clicking the 'Start with Java Webstart' button. (Figure 1.2). This will always launch the latest stable release of Jalview.

The application will start automatically though you may be prompted to accept a security certificate signed by the Barton Group. You can always trust us, so click trust or accept as appropriate. The splash screen (Figure 1.3) gives information about the version and build date that you are running,

information about later versions (if available), and the paper to cite in your publications. This information is also available on the Jalview web site and from the *Help* ⇒ *About* menu option.

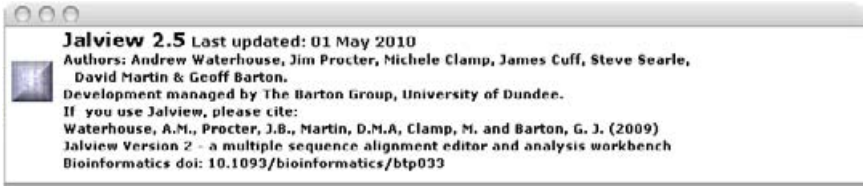


Figure 1.3: **Jalview splash screen**

When Jalview starts it will automatically load an example alignment from the Jalview site. This behaviour can be changed in the Jalview Desktop preferences dialog opened from the Desktop's *Tools* ⇒ *Preferences..* menu. This alignment will look like the one in Figure 1.4 (this is taken from the Jalview 2.4 manual).

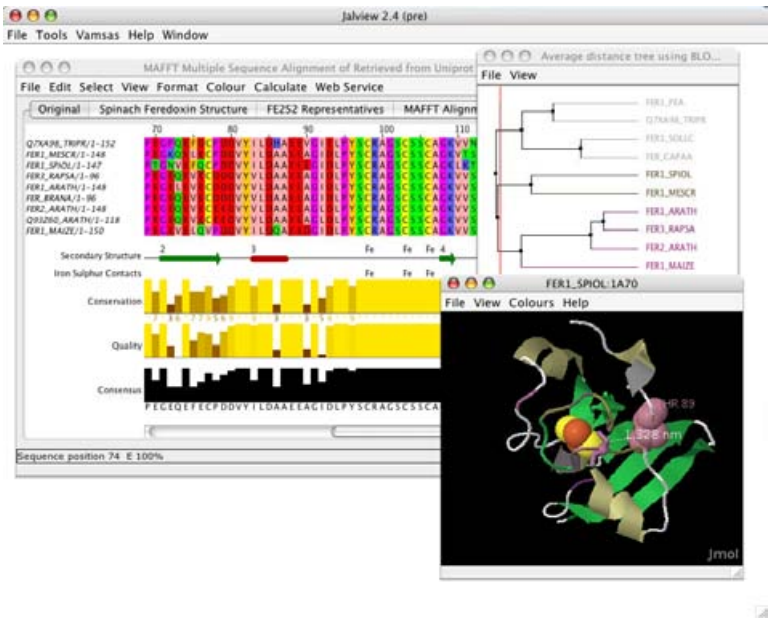


Figure 1.4: **Default startup for Jalview**

Exercise 1: Starting Jalview

1.a. Point your web browser at the Jalview web site and start Jalview by clicking on the 'Start with Java WebStart' button.

1.b. Open the Jalview Desktop's user preferences dialog (from the Tools menu), and untick the checkbox adjacent to the 'Open file' entry in the 'Visual' preferences tab.

1.c. Click OK to save the preferences, then close Jalview and launch it again. The example alignment should not be loaded when Jalview starts up.

1.2.1 Getting Help

Built in documentation

Jalview has comprehensive on-line help documentation. Select *Help* ⇒ *Documentation* from the main window menu and a new window will open (Figure 1.5). The appropriate topic can then be selected from the navigation panel on the left hand side. To search for a specific topic, click the 'search' tab and enter keywords in the box which appears.

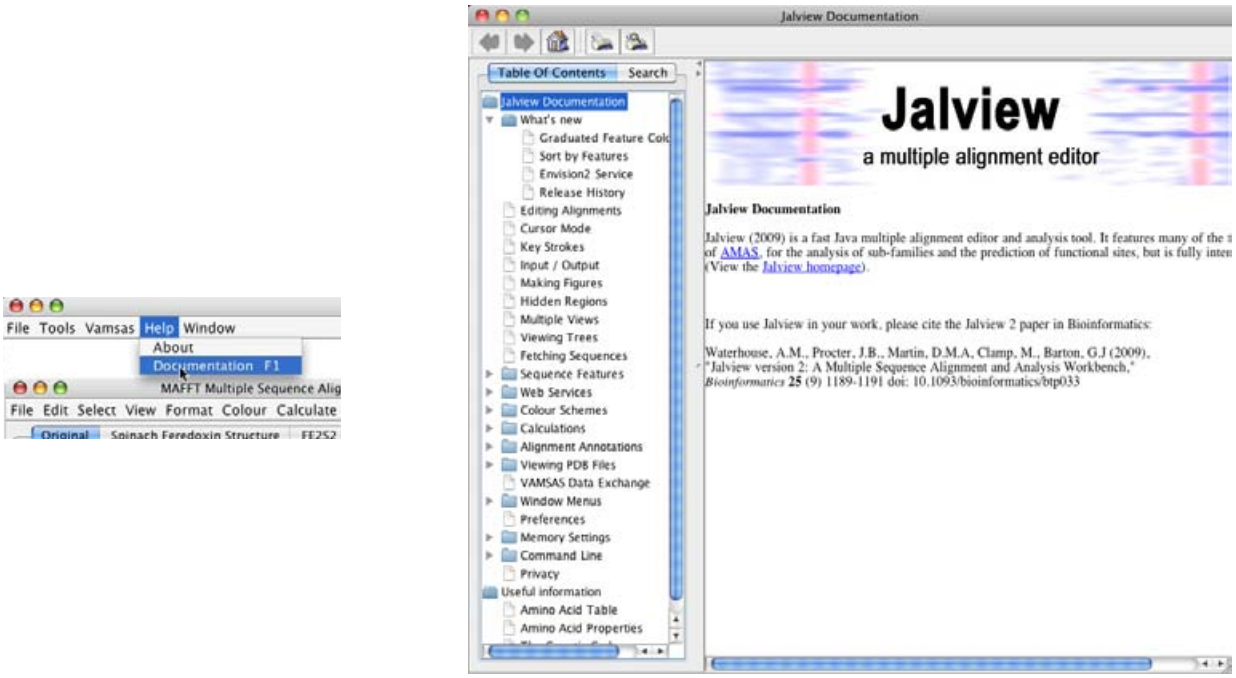


Figure 1.5: **Accessing the built in Jalview documentation**

Email lists

The Jalview Discussion list `jalview-discuss@jalview.org` provides a forum for Jalview users and developers to raise problems and exchange ideas - any problems, bugs, and requests for help should be raised here. The `jalview-announce@jalview.org` list can also be subscribed to if you wish to be kept informed of new releases and developments. Archives and mailing list subscription details can be found on the Jalview web site.

1.3 Navigation

The major features of the Jalview Desktop are illustrated in Figure 1.6. The alignment window is the primary window for editing and visualization, and can contain several independent views of the alignment being worked with. The other windows (Trees, Structures, PCA plots, etc) are linked to

a specific alignment view. Each area of the alignment window has a separate context menu accessed by clicking the right mouse button.

Jalview has two navigation and editing modes: normal mode, where editing and navigation is performed using the mouse, and cursor mode where editing and navigation are performed using the keyboard. The F2 key is used to switch between these two modes.

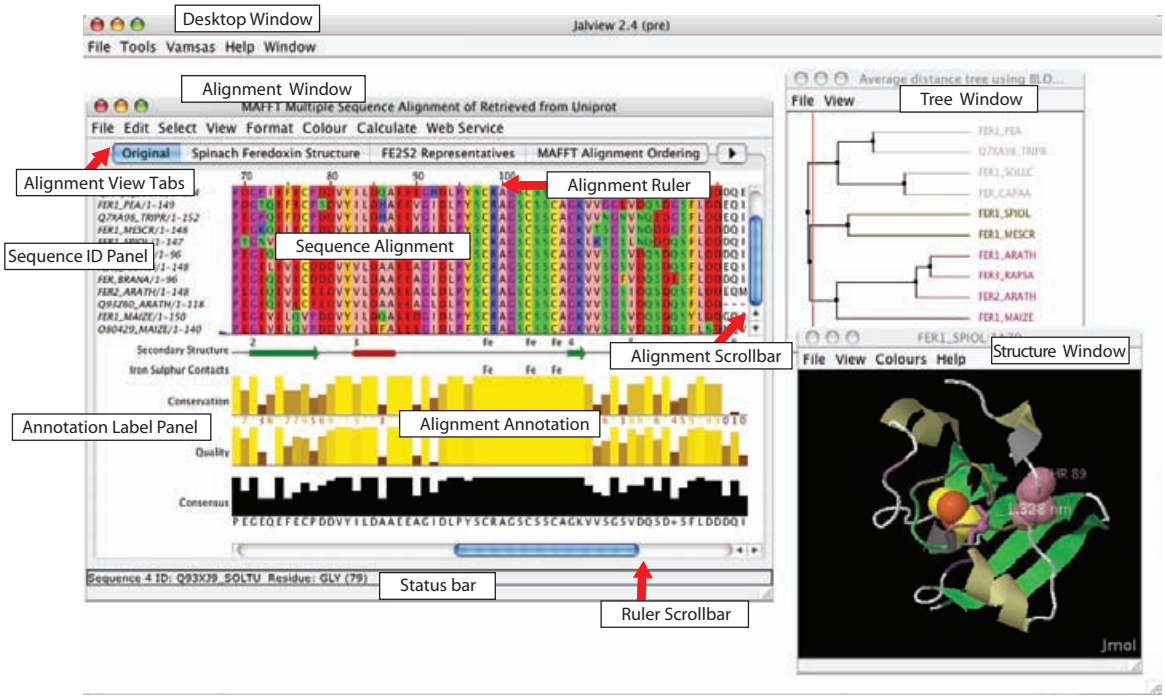


Figure 1.6: **The anatomy of Jalview.** The major features of the Jalview Desktop Application are labelled.

1.3.1 Navigation in Normal mode

Jalview always starts up in Normal mode, where the mouse is used to interact with the displayed alignment view. You can move about the alignment by clicking and dragging the ruler scroll bar to move horizontally, or by clicking and dragging the alignment scroll bar to the right of the alignment to move vertically. If all the rows or columns in the alignment are displayed, the scroll bars will not be visible.

Each alignment view shown in the alignment window presents a window onto the visible regions of the alignment. This means that with anything more than a few residues or sequences, alignments can become difficult to visualize on the screen because only a small area can be shown at a time. It can help, especially when examining a large alignment, to have an overview of the whole alignment. Select *View* ⇒ *Overview Window* from the window menu (Figure 1.7).

The red box in the overview window shows the current view in the alignment window. A percent identity histogram is plotted below the alignment overview. Shaded parts indicate rows and columns

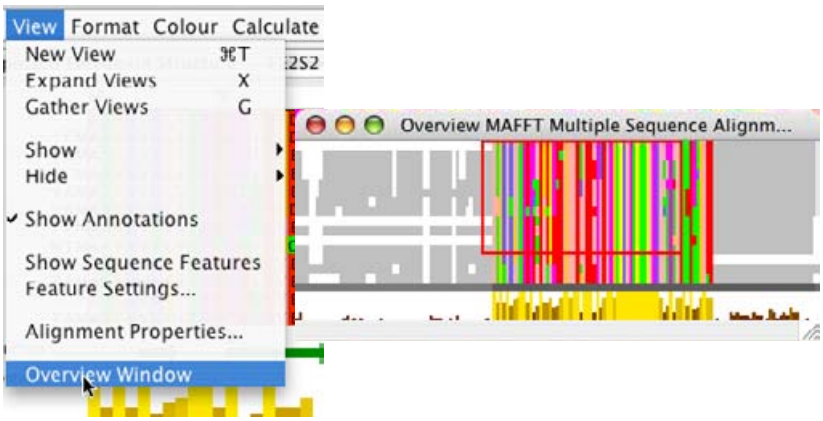


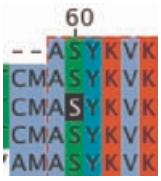
Figure 1.7: **Alignment Overview Window**

of the alignment that are hidden (in this case, a single row at the bottom of the alignment - see Section 1.6.5). You can navigate around the alignment by dragging the red box.

Alignment and analysis windows are closed by clicking on the usual 'close' icon (indicated by arrows on Mac OS X). If you want to close all the alignments and analysis windows at once, then use the *Window* ⇒ *Close All* option from the Jalview desktop (*warning: make sure you have saved your work because this cannot be undone !).*

1.3.2 Navigation in Cursor mode

Cursor mode navigation enables the experienced user to quickly and precisely navigate, select and edit parts of an alignment. On pressing F2 to enter cursor mode the position of the cursor is indicated by a black background and white text. The cursor can be placed using the mouse or moved by pressing the arrow keys (↑, ↓, ←, →).



Rapid movement to specific positions is accomplished as listed below:

- **Jump to Sequence** *n***:** Type a number *n* then press [S] to move to sequence (row) *n*
- **Jump to Column** *n***:** Type a number *n* then press [C] to move to column *n* in the alignment.
- **Jump to Residue** *n***:** Type a number *n* then press [P] to move to residue number *n* in the current sequence.
- **Jump to column** *m* **row** *n***:** Type the column number *m*, a comma, the row number *n* and press [RETURN].

Exercise 2: Navigation

2.a. Scroll around the alignment using the alignment (vertical) and ruler (horizontal) scroll bars.

2.b. Find and open the Overview Window. Move around the alignment by clicking and dragging the red box in the overview window.

2.c. Look at the status bar as you move the mouse over the alignment. It should indicate information about the sequence and residue under the cursor.

2.d. Press [F2] to enter Cursor mode. Use the arrow keys to move the cursor around the alignment. Move to sequence 7 by pressing *7 S*. Move to column 18 by pressing *1 8 C*. Move to residue 18 by pressing *1 8 P*. Note that these can be two different positions if gaps are inserted into the sequence. Move to sequence 5, column 13 by typing *1 3 , 5 [RETURN]*.

1.3.3 The Find Dialog Box

A further option for navigation is to use the *Select* ⇒ *Find...* function. This opens a dialog box into which can be entered regular expressions for searching sequences and sequence IDs, or sequence numbers. Hitting the [Find next] button will highlight the first (or next) occurence of that pattern in the sequence ID panel or the alignment, and will adjust the view in order to display the highlighted region. The Jalview help provides comprehensive documentation for this function, and a quick guide to the regular expressions that can be used with it.

1.4 Loading your own sequences

Jalview provides many ways to load your own sequences.

1.4.1 Drag and Drop

In some operating systems (Mac OS X, Windows XP) you can just drag a file icon from a file browser window and drop it on an open Jalview application window. The file will then be opened as a new alignment window. If you drop an alignment file onto an open alignment window it will be appended to that alignment.

1.4.2 From a File

Jalview can read sequence alignments from a sequence alignment file. This is a text file, not a word processor document. For entering sequences from a wordprocessor document see Cut and Paste (Section 1.4.4) below. Select *File* ⇒ *Input Alignment* ⇒ *From File* from the main menu (Figure 1.8). You will then get a file selection window where you can choose the file to open. Remember to select the appropriate file type. Jalview can automatically identify some sequence file formats.

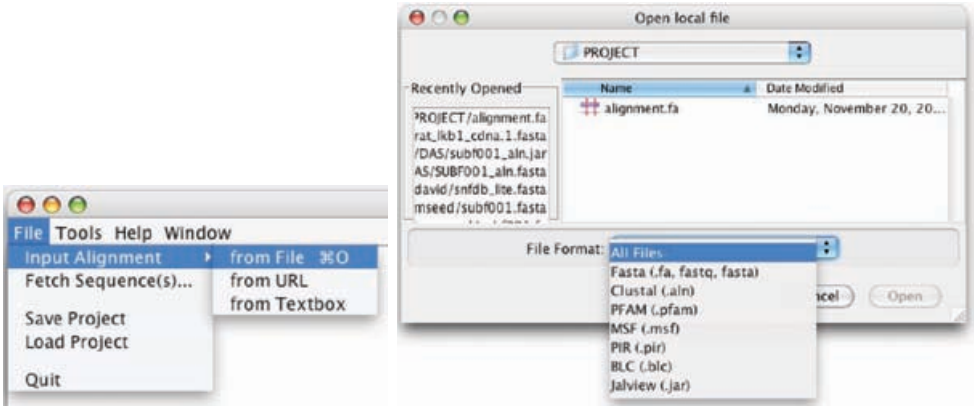


Figure 1.8: **Opening an alignment from a file saved on disk.**

1.4.3 From a URL

Jalview can read sequence alignments directly from a URL. Please note that the files must be in a sequence alignment format - a pretty HTML alignment or graphics file cannot be read by Jalview. Select *File* ⇒ *Input Alignment* ⇒ *From URL* from the main menu and a window will appear asking you to enter the URL (Figure1.9). Jalview will attempt to automatically discover the file format.

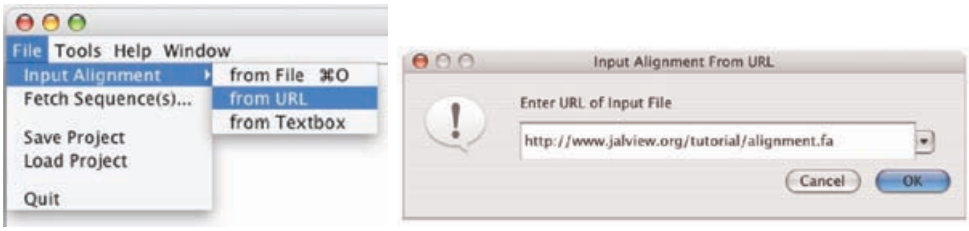


Figure 1.9: **Opening an alignment from a URL**

1.4.4 Cut and Paste

Documents such as those produced by Microsoft Word cannot be readily understood by Jalview. The way to read sequences from these documents is to select the data from the document and copy it to the clipboard. There are two ways to do this. One is to right-click on the desktop background, and select the 'Paste to new alignment' option in the menu that appears. The other is to select *File* ⇒ *Input Alignment* ⇒ *From Textbox* from the main menu, and paste the sequences into the textbox window that will appear (Figure 1.10). In both cases, presuming that they are in the right format, Jalview will happily read them into a new alignment window.

1.4.5 From a public database

Jalview can retrieve sequences and sequence alignments from the public databases housed at the European Bioinformatics Institute, such as Uniprot, Pfam and the PDB, as well as any DAS

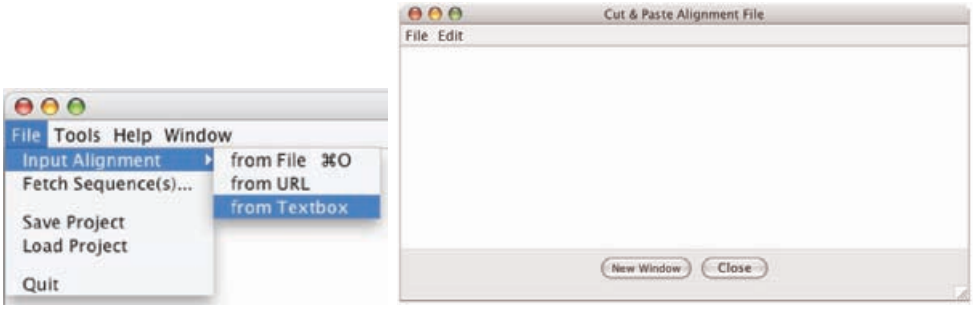


Figure 1.10: **Opening an alignment from pasted text**

sequence server registered at the configured DAS registry. This facility avoids having to manually locate, save and load the sequences, and allows Jalview to gather additional metadata provided by the source, such as annotation and database cross references. Select *File* ⇒ *Fetch Sequence(s)* ... from the main menu and a window will appear (Figure 1.11). Use the menu box to select the appropriate database, enter a sequence ID/accession number, or several separated by a semicolon and Jalview will attempt to retrieve it/them from the chosen database source. Example queries are provided to test that a source is operational, and can also be used as a guide for the type of accession numbers understood by the source.

Figure 1.11: **Retrieving sequences from a public database**

Exercise 3: Loading sequences

3.a. Start Jalview then close all windows by selecting *Window* ⇒ *Close All* from the main menu

3.b. Select *File* ⇒ *Input Alignment* ⇒ *From URL* from the main menu and enter http://www.jalview.org/tutorial/alignment.fa in the box. Click *OK* and the alignment should load.

3.c. Close all windows using the *Window* ⇒ *Close All* main menu option. Point your web browser to http://www.jalview.org/tutorial/alignment.fa and save the file to your desktop. Open this file in Jalview by selecting *File* ⇒ *Input Alignment* ⇒ *From File* from the main menu and browsing to the appropriate location. Click OK and load the alignment

3.d. Drag the alignment.fa file from the desktop onto the Jalview window. The alignment should open. Try dragging onto an empty Jalview and onto an existing alignment and observe the results.

3.e. Select *File* ⇒ *Fetch Sequence(s)..* from the main menu. Select the *PFAM (seed)* database and enter the accession number PF03460. Click OK. An alignment of about 107 sequences should load.

3.f. Open the URL http://www.jalview.org/tutorial/alignment.fa in a web browser. Select and copy the entire text to the clipboard (usually via the browser's *Edit* ⇒ *Copy* menu option). Ensure Jalview is running and select *File* ⇒ *Input Alignment* ⇒ *From Textbox* . Paste the clipboard into the large window using the *Edit* ⇒ *Paste* text box menu option. Click *Close* and the alignment will be loaded.

1.4.6 Memory Limits

Jalview is a Java program. One unfortunate implication of this is that it does not allow Jalview to dynamically request additional memory from the operating system. It is important, therefore, that you ensure that you have allocated enough memory to work with your data. On most occasions, Jalview will warn you when you have tried to load an alignment that is too big to fit in to memory (for instance, some of the PFAM alignments are **very** large). You can find out how much memory is available to Jalview with the desktop window's ⇒ *Tools* ⇒ *Show Memory Usage* function, which enables the display of the currently available memory at the bottom left hand side of the Desktop window's background. Should you need to increase the amount of memory available to Jalview, full instructions are given in both the built in documentation and on the JVM memory parameters page (http://www.jalview.org/jvmmemoryparams.html) on the website.

1.5 Writing sequence alignments

1.5.1 Saving the alignment

Jalview allows the current sequence alignments to be saved to file so they can be restored at a later date, passed to colleagues or analysed in other programs. From the alignment window menu select *File* ⇒ *Save As* and a dialog box will appear (Figure 1.12). You can navigate to an appropriate directory in which to save the alignment. Jalview will remember the last filename and format used

to save (or load) the alignment, enabling you to quickly update the file after editing by using the *File* ⇒ *Save* entry.

Jalview offers several different formats in which an alignment can be saved. The jalview format is the only one which will preserve the colours, groupings and similar information in the alignment. The other formats produce text files containing just the sequences with no visualization information, although some allow limited annotation and sequence features to be stored (e.g. AMSA). Unfortunately only Jalview can read Jalview files. The *File* ⇒ *Output To Textbox* menu option allows the alignment to be copied and pasted into other documents or web servers.

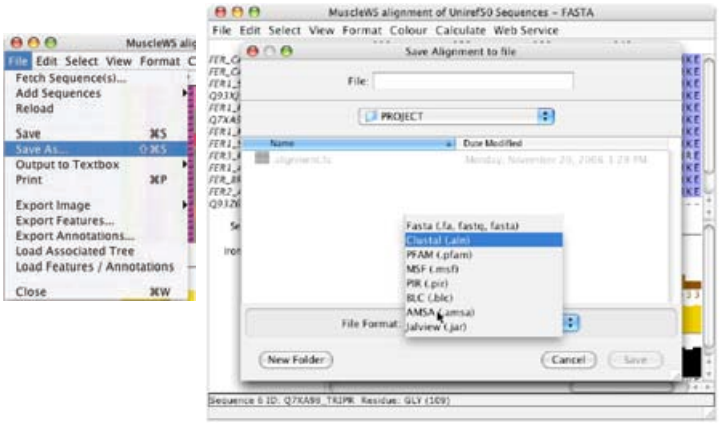
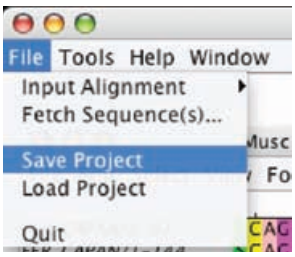


Figure 1.12: **Saving alignments in Jalview to disk**

1.5.2 Jalview Projects

If you wish to save the complete Jalview session rather than just one alignment (e.g. because you have calculated trees or multiple different alignments) then your work should be saved as a Jalview Project file[7]. From the main menu select *File* ⇒ *Save Project* and a file save dialog box will appear. Loading a project will restore Jalview to exactly the view at which the file was saved, complete with all alignments, trees, annotation and displayed structures rendered appropriately.



[7]Tip: Ensure that you have allocated plenty of memory to Jalview when working with large alignments in Jalview projects. See Section 1.4.6 above for how to do this.

Exercise 4: Saving Alignments

4.a. Start Jalview, close all windows and load the ferrodoxin alignment from pFam (accession number PF03460 (see Exercise 3).

4.b. Select *File* ⇒ *Save As* from the alignment window menu. Choose a location into which to save the alignment and select a format. All formats except *Jalview* can be viewed in a normal text editor (e.g. Notepad) or in a web browser. Enter a file name and click *Save*. Check this file by browsing to it with your web browser or by closing all windows and opening it with Jalview.

4.c. Repeat the previous step trying different file formats.

4.d. Select *File* ⇒ *Output to Textbox* ⇒ *FASTA*. You can select and copy this alignment to the clipboard using the textbox menu options *Edit* ⇒ *Select All* followed by *Edit* ⇒ *Copy*. The alignment can then be pasted into any application of choice, e.g. a word processor or web form.

4.e. Ensure at least one alignment window is shown in Jalview. Open the overview window and scroll to any part of the alignment. Select *File* ⇒ *Save Project* from the main menu and save in a suitable place. Close all windows and then load the project via the *File* ⇒ *Save Project* menu option. Note how all the windows and positions are exactly as they were when they were saved.

1.6 Selecting and editing sequences

Jalview makes extensive use of selections - most of the commands available from its menus operate on the currently selected region of the alignment, either to change their appearance or perform some kind of analysis. This section illustrates how to make and use selections and groups.

1.6.1 Selecting parts of an alignment

Selections can be of arbitrary regions in an alignment, one or more complete columns, or one or more complete sequences.

A selected region can be copied and pasted as a new alignment using the *Edit* ⇒ *Copy* and *Edit* ⇒ *Paste* ⇒ *As New Alignment* alignment window menu options.

To clear (unselect) the selection press the [ESC] (escape) key.

Selecting arbitrary regions

To select part of an alignment, place the mouse at the top left corner of the region you wish to select. Press the mouse button and drag the mouse to the bottom right corner of the chosen region before releasing the mouse button. A dashed red box appears around the selected region (Figure 1.13).

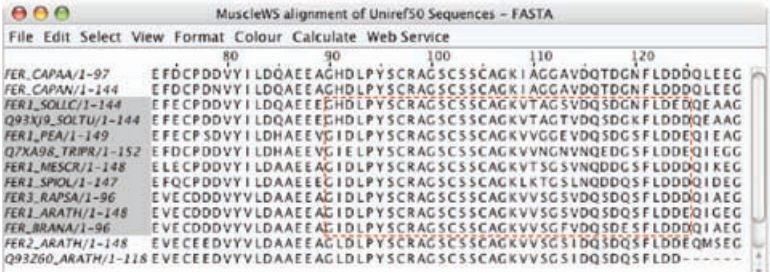


Figure 1.13: **Selecting a region in an alignment**

Selecting columns

To select the same residues in all sequences, click and drag along the alignment ruler. This selects the entire height of the alignment. Ranges of positions can also be selected by clicking on the first position then holding down the [SHIFT] key whilst clicking the other end of the selection. Discontinuous regions can be selected by holding down [CTRL] and clicking on positions to add to the column selection. Note that each [CTRL]-Click changes the current selected sequence region to that column, but adds to the column selection. Selected columns are indicated by red highlighting in the ruler bar (Figure 1.14).

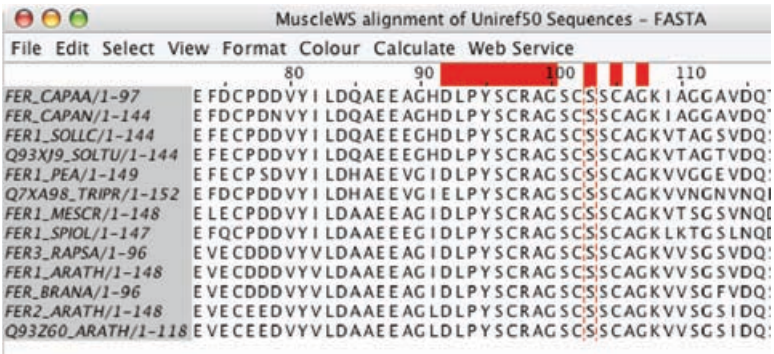


Figure 1.14: **Selecting multiple columns in an alignment.** The red highlighting on the alignment ruler marks the selected columns. Note that only the most recently selected column has a dashed-box around it to indicate a region selection.

Selecting sequences

To select multiple complete sequences, click and drag the mouse down the sequence ID panel. The same technique as used for columns above can be used with [SHIFT]-Click for continuous and [CTRL]-Click to select discontinuous ranges of sequences (Figure 1.15).

Making selections in Cursor mode

To define a selection in cursor mode, navigate to the top left corner of the proposed selection. Pressing the [Q] key marks this as the corner. A red outline appears around the cursor (Figure

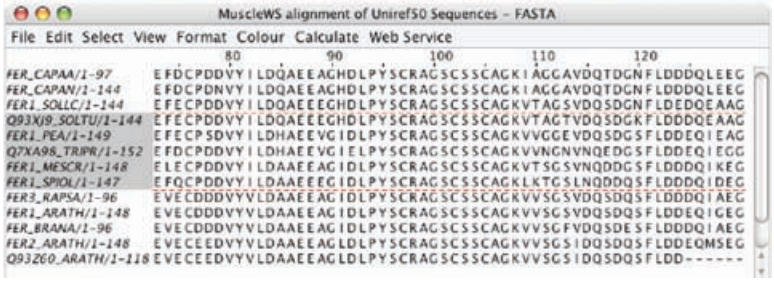


Figure 1.15: **Selecting multiple sequences in an alignment.** Use [CTRL] or [SHIFT] to select many sequences at once.

1.16)

Navigate to the bottom right corner of the proposed selection and press the [M] key. This marks the bottom right corner of the selection. The selection can then be treated in the same way as if it had been created in normal mode.

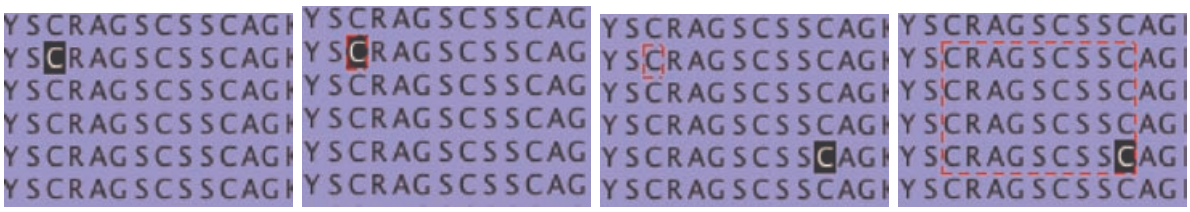


Figure 1.16: **Making a selection in cursor mode.** Navigate to the top left corner (left), press [Q] (left center), navigate to the bottom right corner (right center) and press [M] (right)

Inverting the current selection

The current sequence or column selection can be inverted, using Select ⇒ Invert **Sequence/Column** Selection in the Alignment window. Inverting the selection is particularly useful when hiding regions in a large alignment (see Section 1.20 below). Instead of selecting the columns and rows that are to be hidden, simply select the region that is to be kept visible, and then invert the selection.[8]

1.6.2 Creating groups

Selections are lost as soon as a different region is selected. Groups can be created which are labeled regions of the alignment. To create a group, first select the region which is to comprise the group. Then click the right mouse button on the selection to bring up a context menu. Select *Selection* ⇒ *Group* ⇒ *Group* then enter a name for the group in the dialogue box which appears.

By default the new group will have a box drawn around it. The appearance of the group can be

[8] It is also possible to hide everything but the selected region using the *View* ⇒ *Hide* ⇒ *All but selected region* menu entry.

Figure 1.17: **Creating a new group from a selection**

changed (see Section 1.7 below). This group will stay defined even when the selection is removed.

1.6.3 Exporting the current selection

The current selection can be copied to the system clipboard (in PFAM format). It can also be output to a textbox using the output functions in the pop-up menu obtained by right clicking the current selection. The textbox enables quick manual editing of the alignment prior to importing it into a new window (using the [New Window] button) or saving to a file with the *File* ⇒ *Save As* pulldown menu option from the text box.

Exercise 5: Making selections and groups

5.a. Close all windows in Jalview and load the ferrodoxin alignment (PFAM ID PF03460). Choose a residue and place the mouse cursor on it. Click and drag the mouse cursor to create a selection. As you drag, a red box will 'rubber band' out to show the extent of the selection. Release the mouse button and a red box should border the selected region. Now press [ESC] to clear the selection.

5.b. Select one sequence by clicking on the id panel. Note that the sequence ID takes on a highlighted background and a red box appears around the selected sequence. Now hold down [SHIFT] and click another sequence ID a few positions above or below. Note how the selection expands to include all the sequences between the two positions on which you clicked. Now hold down [CTRL] and click on several sequences ID's both selected and unselected. Note how unselected IDs are individually added to the selection and previously selected IDs are individually deselected.

5.c. Repeat the step above but selecting columns by clicking on the ruler bar instead of selecting rows by clicking on the sequence ID.

5.d. Press [F2] to enter Cursor mode. Navigate to column 59, row 1 by pressing *5 9 , 1 [RETURN]*. Press *Q* to mark this position. Now navigate to column 65, row 8 by pressing *6 5 , 8 [RETURN]*. Press *M* to complete the selection.

5.e. Open the popup menu by right-clicking the selected region with the mouse. Open the *Selection* ⇒ *Group* ⇒ *Group Colour* menu and select 'Percentage Identity'. This will turn the selected region into a group.

5.f. Hold down [CTRL] and use the mouse to select and deselect sequences by clicking on their Sequence ID label. Note how the group expands to include newly selected sequences, and the 'Percentage Identity' colouring changes.

5.g. Use the mouse to click and drag the right-hand edge of the selected group. Note again how the group resizes.

5.h. Right click on the text area to open the selection popup-menu. Follow the menus an pick an output format from the *Selection* ⇒ *Output to Textbox ...* submenu.

5.i. Try manually editing the alignment and then press the [New Window] button to import the file into a new alignment window.

1.6.4 Reordering the alignment

Sequence reordering is simple. Highlight the sequences to move then press the up or down arrow keys as appropriate (Figure 1.18). If you wish to move a sequence up past several other sequences it is often quicker to select the group past which you want to move it and then move the group rather than the individual sequence.

Exercise 6: Reordering the alignment

6.a. Open an alignment (e.g.the PFAM domain PF03460). Select one sequence. Using the up and down arrow keys, alter its position in the alignment.

6.b. Hold [CTRL] and select two sequences separated by one or more un-selected sequences. Note how multiple sequences are grouped together when they are reordered using the up and down arrow keys.

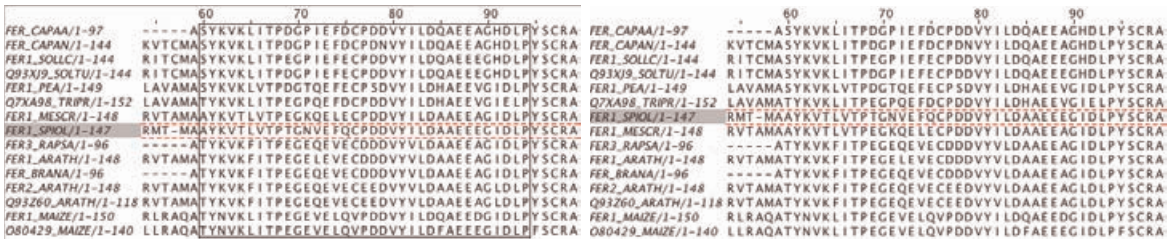


Figure 1.18: **Reordering the alignment.** The selected sequence moves up one position on pressing the ↑ key

1.6.5 Hiding regions

It is sometimes convenient to exclude some sequences or residues in the alignment without actually deleting them. Jalview allows sequences or alignment columns within a view to be hidden, and this facility has been used to create several different views of the example alignment in the file that is loaded when Jalview is first started (See Figure 1.4).

To hide a set of sequences, select them and right-click the mouse on the selected sequence IDs to bring up the context menu. Select *Hide Sequences* and the sequences will be concealed, with a small triangle indicating their position (Figure 1.19). To unhide (reveal) the sequences, right click on the triangle and select *Reveal Sequences* from the context menu.

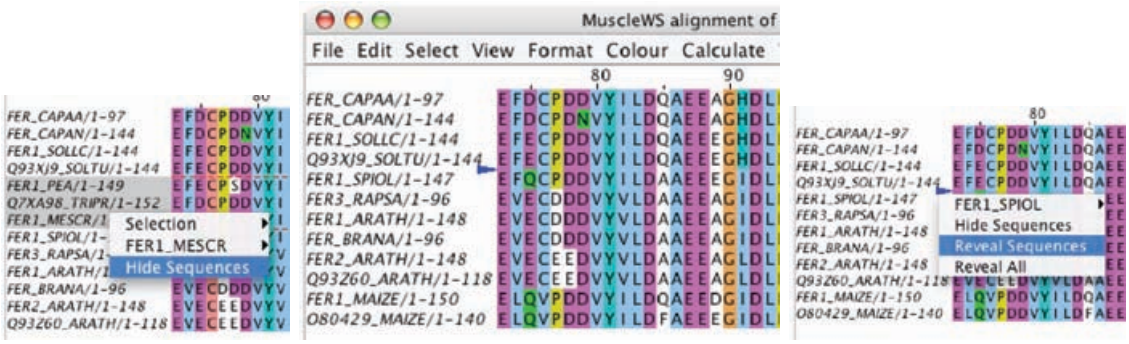


Figure 1.19: **Hiding Sequences** Hidden sequences are represented by a blue triangle in the sequence ID panel

A similar mechanism applies to columns (Figure 1.20). Selected columns (indicated by a red marker) can be hidden and revealed in the same way via the context menu (right click) on the ruler bar. The hidden column selection is indicated by a blue triangle in the ruler bar.

It is often easier to select the region that you intend to work with, rather than the regions that you want to hide. In this case, use the *View* ⇒ *Hide* ⇒ *All but selected region* menu entry, or press [Shift]+[Ctrl]+H to hide the unselected region.

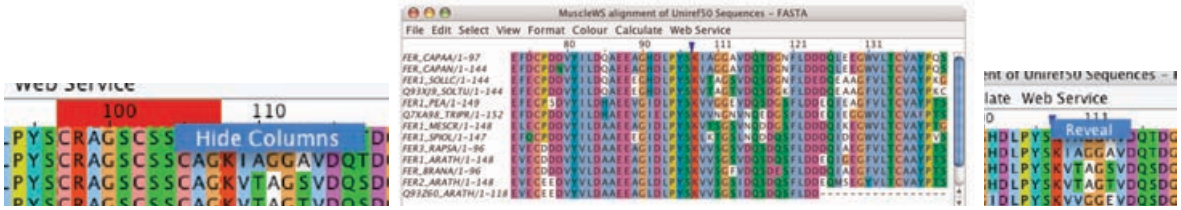


Figure 1.20: **Hiding Columns** Hidden columns are represented by a blue triangle in the ruler bar

Representing a group with a single sequence

Instead of hiding a group completely, it is sometimes useful to work with just one representative sequence. The *<Sequence ID> ⇒ Represent group with <Sequence ID>* option from the sequence ID pop-up menu enables this variant of the hidden groups function. The remaining representative sequence can be visualized and manipulated like any other. However, any alignment edits that affect the sequence will also affect the whole sequence group.

Exercise 7: Hiding and revealing regions

7.a. Close all windows then open the PFAM accession PF03460. Select a contiguous set of sequences by clicking and dragging on the sequence ID panel. Right click on the selected sequence IDs and select *Hide Sequences*.

7.b. Right click on the blue triangle indicating hidden sequences and select *Reveal Sequences*. (If you have hidden all sequences then you will need to use the alignment window menu option *View ⇒ Show ⇒ All Sequences*.)

7.c. Repeat but using a non-contiguous set of sequences. Note that when multiple regions are hidden there are two options, *Reveal Sequences* and *Reveal All*.

7.d. Repeat the above but hiding and revealing columns instead of sequences.

7.e. Select a region of the alignment, add in some additional columns to the selection, and experiment with the 'Hide all but selected region' function.

7.f. Select some sequences and pick one to represent the rest. Bring up the sequence ID pop-up menu for that sequence and select the *Represent group with <Sequence ID>* option. Use the pop-up menu again to reveal the hidden sequences that you just picked a representative for.

1.6.6 Introducing and removing gaps

The alignment view provides an interactive editing interface, allowing gaps to be inserted or deleted to the left of any position in a sequence or sequence group. Alignment editing can only be performed whilst in keyboard editing mode (entered by pressing [F2]) or by clicking and dragging residues with the mouse when [SHIFT] or [CTRL] is held down (which differs from earlier versions of Jalview).

Locked Editing

The Jalview alignment editing model is different to that used in other alignment editors. Because edits are restricted to the insertion and deletion of gaps to the left of a particular sequence position, editing has the effect of shiting the rest of the sequence(s) being edited down or up-stream with respect to the rest of alignment. The *Edit* ⇒ *Pad Gaps* option can be enabled to eliminate 'ragged edges' at the end of the alignment, but does not avoid the 'knock-on' effect which is sometimes undesirable. However, its effect can be limited by performing the edit within a selected region. In this case, gaps will only be removed or inserted within the selected region. Edits are similarly constrained when they occur adjacent to a hidden column.

Introducing gaps in a single sequence

To introduce a gap, place the cursor on the residue to the immediate right of where the gap should appear. Hold down the SHIFT key and the left mouse button, then drag the sequence to the right till the required number of gaps has been inserted.

One common error is to forget to hold down [SHIFT]. This results in a selection which is one sequence high and one residue long. Gaps cannot be inserted in such a selection. The selection can be cleared and editing enabled by pressing the [ESC] key.

Introducing gaps in all sequences of a group

To insert gaps in all sequences in a selection or group, place the mouse cursor on any residue in the selection or group to the immediate right of the position in which a gap should appear. Hold down the CTRL key and the left mouse button, then drag the sequences to the right until the required number of gaps has appeared.

Gaps can be removed by dragging the residue to the immediate right of the gap leftwards whilst holding down [SHIFT] (for single sequences) or [CTRL] (for a group of sequences).

Sliding Sequences

Pressing the [←] or [→] arrow keys when one or more sequences are selected will "slide" the selected sequences to the left or right (respectively).

Undoing edits

Jalview supports the undoing of edits via the *Edit* ⇒ *Undo Edit* alignment window menu option. Each editing action is stored and can be reversed in sequence. Colouring of the alignment is not reversible via the *Undo* option.

Figure 1.21: **Introducing gaps in a single sequence.** Gaps are introduced as the selected sequence is dragged to the right with [SHIFT] pressed.

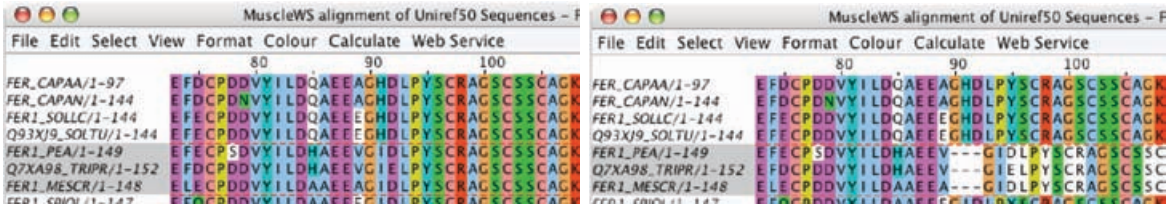


Figure 1.22: **Introducing gaps in a group.** Gaps are introduced as the selected group is dragged to the right with [CTRL] pressed.

Exercise 8: Editing alignments

8.a. Load the URL http://www.jalview.org/tutorial/unaligned.fa which contains part of the ferrodoxin alignment from PF03460.
You are going to manually reconstruct the alignment that jalview loads by default. Remember to use [CTRL]+Z to undo an edit, or the File Reload function to revert the alignment back to the original version if you want to start again.

8.b. Select the first 7 sequences, and press H to hide them (or right click on the sequence IDs to open the sequence ID popup menu, and select Hide Sequences).

8.c. Select FER3_RAPSA and FER_BRANA. Slide the sequences to the left so the initial **A** lies at column 57 using the ⇒ key.

8.d. Select FER1_SPIOL, FER1_ARATH, FER2_ARATH, Q93Z60_ARATH and O80429_MAIZE (Hint: press [CTRL]-I to invert the sequence selection and then deleset FER1_MAIZE), and use the ⇒ key to slide them to so they begin at column 5 of the alignment view.

8.e. Select all the visible sequences in the block by pressing [CTRL]-A. Insert a single gap in all selected sequences at column 38 by holding [CTRL] and clicking on the R in FER1_SPIOL and dragging one column to right. Insert another gap at column 47 in all sequences in the same way.

8.f. Correct the feredoxin domain alignment for FER1_SPIOL by insert two additional gaps after the gap at column 47: hold [SHIFT] and click and drag on the G and move it two columns to the right.

8.g. Now complete the alignment of FER1_SPIOL with a **locked edit** by pressing [ESC] and select columns 47 to 57 of the FER1_SPIOL row. Move the mouse onto the G at column 50, hold [SHIFT] and drag the G to the left by one column to insert a gap at column 57.

8.h. In the next two steps you will complete the alignment of the last two sequences.
Select the last two sequences (FER1_MAIZE and O80429_MAIZE), then press [SHIFT] and click and drag the initial methionine of O80429_MAIZE 5 columns to the right so it lies at column 10. Keep holding [SHIFT] and click and drag to insert another gap at the proline at column 25 (16P). Remove the gap at column 44, and insert 4 gaps at column 47 (after AAPM).

8.i. Hold [SHIFT] and drag the I at column 39 of FER1_MAIZE 2 columns to the right. Remove the gap at FER1_MAIZE column 49 by [SHIFT]+click and drag left by one column. Insert three gaps in FER1_MAIZE at column 47 by holding [SHIFT] and click and drag the S in FER1_MAIZE to the right by three columns. Finally, remove

Editing in Cursor mode

Gaps can be be easily inserted when in cursor mode (toggled with [F2]) by pressing [SPACE]. Gaps will be inserted at the cursor, pushing the residue under the cursor to the right. To insert *n* gaps type *n* and then press [SPACE]. To insert gaps into all sequences of a group, use [CTRL]-[SPACE] or [SHIFT]-[SPACE] (both keys held down together).

Gaps can be removed in cursor mode by pressing [BACKSPACE]. The gap under the cursor will be removed. To remove *n* gaps, type *n* and then press [BACKSPACE]. Gaps will be deleted up to the number specified. To delete gaps from all sequences of a group, use [CTRL]-[BACKSPACE] or [SHIFT]-[BACKSPACE] (both keys held down together).

Exercise 9: Keyboard edits

9.a. Load the sequence alignment at http://www.jalview.org/tutorial/unaligned.fa, or continue using the edited alignment from exercise 8. If you continue from the previous exercise, then first right click on the sequence ID panel and select Reveal All. Now, enter cursor mode by pressing [F2]

9.b. Insert 58 gaps at the start of the first sequence (FER_CAPAA). Press *58* then *[SPACE]*.

9.c. Go down one sequence and select rows 2-5 as a block. Click on the second sequence ID (FER_CAPAN). Hold down shift and click on the fifth (FER1_PEA).

9.d. Insert 6 gaps at the start of this group. Go to column 1 row 2 by typing *1,2* then pressing *[RETURN]*. Now insert 6 gaps. Type *6* then hold down [CTRL] and press the space bar.

9.e. Now insert one gap at column 34 and another at 38. Insert 3 gaps at 47. Press *3 4 C* then [CTRL]-[SPACE]. Press *3 8 C* then [CTRL]-[SPACE]. Press *4 7 C* then *3 [CTRL-SPACE]* the first through fourth sequences are now aligned.

9.f. The fifth sequence (FER1_PEA) is poorly aligned. We will delete some gaps and add some new ones. Navigate to the start of sequence 5 and delete 3 gaps. Press *1 , 5 [RETURN]* then *3 [BACKSPACE]* to delete three gaps. Go to column 31 and delete the gap. Press *3 1 C [BACKSPACE]* .

9.g. Similarly delete the gap now at column 34, then insert two gaps at column 38 . Press *3 4 C [BACKSPACE] 3 8 C 2 [SPACE]*. Delete three gaps at 44 and insert one at 47 by pressing *4 4 C 3 [BACKSPACE] 4 7 C [SPACE]*. The top five sequences are now aligned.

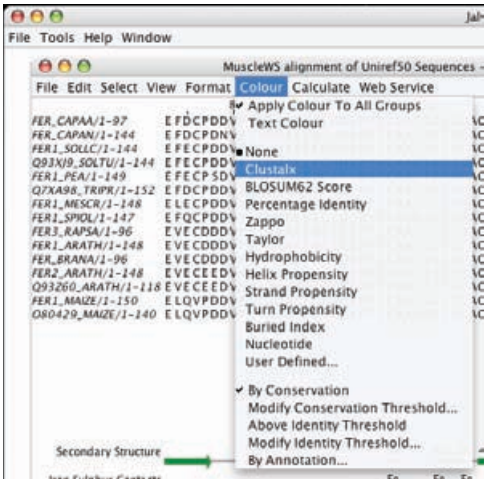
1.7 Colouring sequences

Colouring sequences is a key aspect of alignment presentation. Jalview allows you to colour the whole alignment, or just specific groups. Alignment and group colours are rendered *below* any other colours, such as those arising from sequence features (these are described in Section 2.4). This means that if you try to apply one of the colourschemes described in this section, and nothing appears to happen, it may be that you have sequence feature annotation displayed, and you may have to disable it using the *View* ⇒ *Show Features* option before you can see your colourscheme.

There are two main types of colouring styles: simple static residue colourschemes and dynamic schemes which use conservation and consensus analysis to control colouring. A hybrid colouring is also possible, where static residue schemes are modified using a dynamic scheme. The individual schemes are described in Section 1.7.6 below.

1.7.1 Colouring the whole alignment

The alignment can be coloured via the *Colour* menu option in the alignment window. Selecting the colour scheme causes all residues to be coloured. The menu is divided into three sections. The first section gives options for the behaviour of the menu options, the second lists static and dynamic colourschemes available for selection. The last gives options for making hybrid colourschemes using conservation shading or colourscheme thresholding.



1.7.2 Colouring a group or selection

Selections or groups can be coloured in two ways. The first is via the Alignment Window's *Colour* menu, after first ensuring that the Apply to all groups flag is not selected. This must be turned off specifically as it is on by default.

The second method is to use the *Selection* ⇒ *Group* ⇒ *Group Colour* context menu option obtained by right clicking on the group (Figure 1.23).

1.7.3 Shading by conservation

For many colour schemes, the intensity of the colour in a column can be scaled by the degree of amino acid property conservation. Selecting *Colour* ⇒ *By Conservation* brings up a selection box (the *Conservation Threshold dialog box*) allowing the alignment colouring to be modified. Selecting a higher value limits colouring to more highly conserved columns (Figure 1.24).

1.7.4 Thresholding by percentage identity

'Thresholding' is another hybrid colour model where a residue is only coloured if it is not excluded by an applied threshold. Selecting *Colour* ⇒ *Above Identity Threshold* brings up a selection box with a slider controlling the minimum percentage identity threshold to be applied. Selecting a

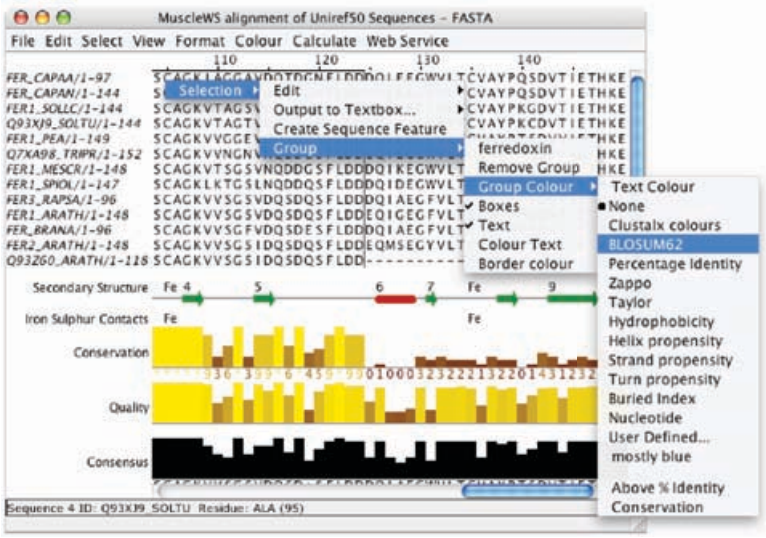


Figure 1.23: **Colouring a group via the context menu.**

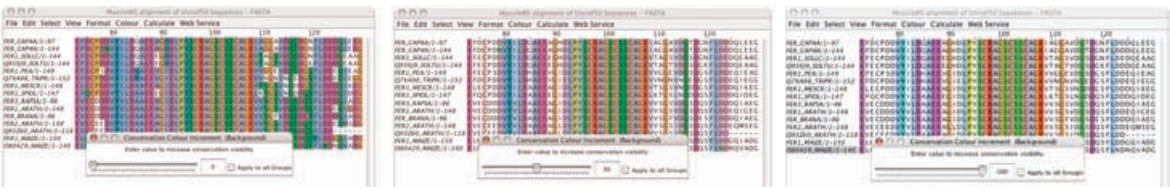
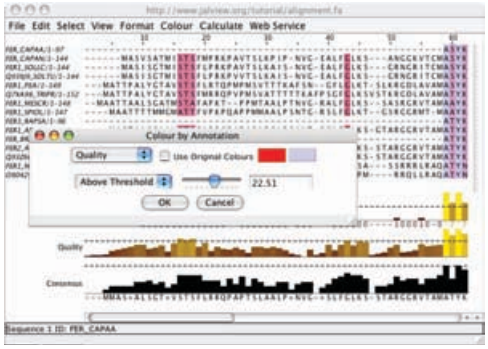


Figure 1.24: **Conservation Shading** The density of the ClustalX style residue colouring is controlled by the conservation threshold. The effect of 0% (left), 50% (center) and 100% (right) thresholds are shown.

higher threshold (by sliding to the right) limits the colouring to columns with a higher percentage identity (as shown by the Consensus histogram in the annotation panel).

1.7.5 Colouring by Annotation

Any of the quantitative annotations shown on an alignment can be used to threshold or shade the whole alignment[9]. The *Colour ⇒ By Annotation* options opens a dialog which allows you to select which annotation to use, the minimum and maximum shading colours or whether the original colouring should be thresholded (the 'Use original colours' option).

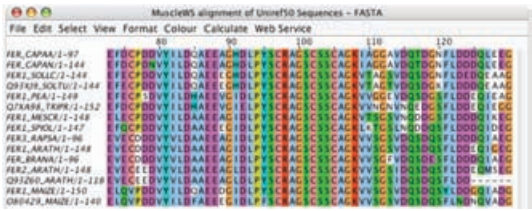


1.7.6 Colour schemes

Full details on each colour scheme can be found in the Jalview on-line help. A brief description of each one is provided below:

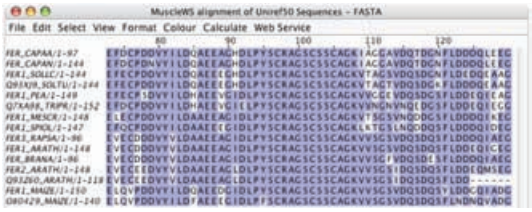
ClustalX

This is an emulation of the default colourscheme used for alignments in Clustal X, a graphical interface for the ClustalW multiple sequence alignment program. Each residue in the alignment is assigned a colour if the amino acid profile of the alignment at that position meets some minimum criteria specific for the residue type.



Blosum62

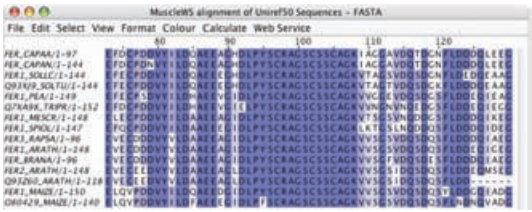
Gaps are coloured white. If a residue matches the consensus sequence residue at that position it is coloured dark blue. If it does not match the consensus residue but the Blosum 62 matrix gives a positive score, it is coloured light blue.



[9]Please remember to turn off Sequence Feature display to see the shading

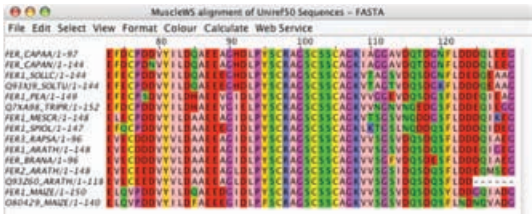
Percentage Identity

The Percent Identity option colours the residues (boxes and/or text) according to the percentage of the residues in each column that agree with the consensus sequence. Only the residues that agree with the consensus residue for each column are coloured.



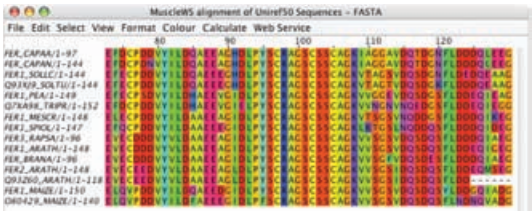
Zappo

The residues are coloured according to their physicochemical properties. The physicochemical groupings are Aliphatic/hydrophobic, Aromatic, Positive, Negative, Hydrophillic, conformationally special, and Cyst(e)ine.



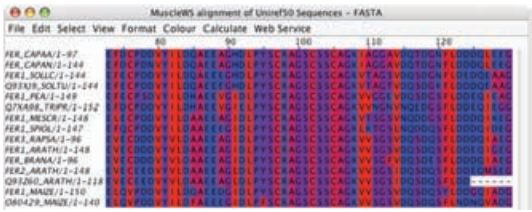
Taylor

This colour scheme was devised by Willie Taylor and an entertaining description of it's origin can be found in Protein Engineering, Vol 10 , 743-746 (1997)



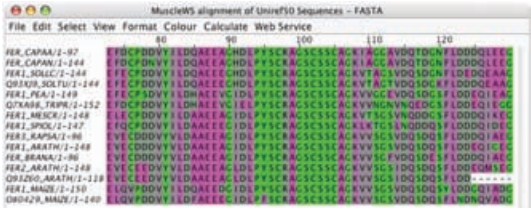
Hydrophobicity

Residues are coloured according to the hydrophobicity table of Kyte, J., and Doolittle, R.F., J. Mol. Biol. 1157, 105-132, 1982. The most hydrophobic residues are coloured red and the most hydrophilic ones are coloured blue.



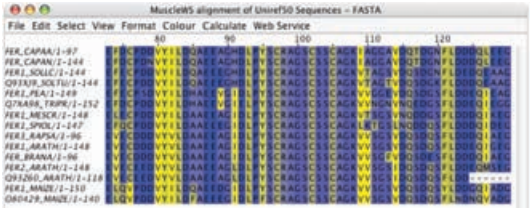
Helix Propensity

The residues are coloured according to their Chou-Fasman[10] helix propensity. The highest propensity is magenta, the lowest is green.



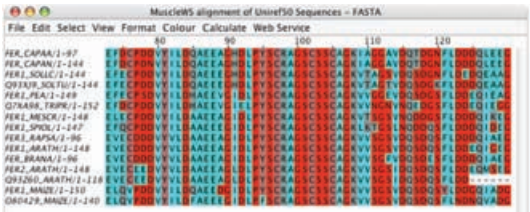
Strand Propensity

The residues are coloured according to their Chou-Fasman[10] Strand propensity. The highest propensity is Yellow, the lowest is blue.



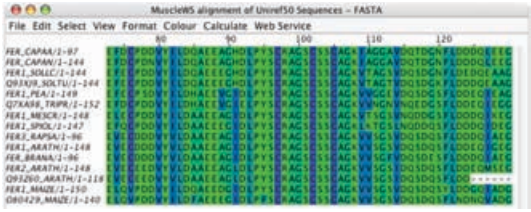
Turn Propensity

The residues are coloured according to their Chou-Fasman[10] turn propensity. The highest propensity is red, the lowest is cyan.



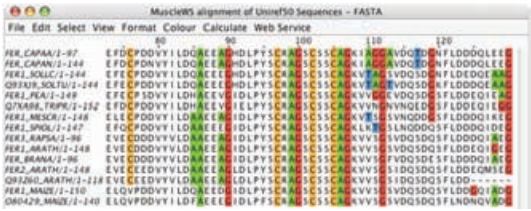
Buried Index

The residues are coloured according to their Chou-Fasman[10] burial propensity. The highest propensity is blue, the lowest is green.



Nucleotide

Residues are coloured with four colours corresponding to the four nucleotide bases. All non ACTG residues are uncoloured. See Section 2.6 for further information about working with nucleic acid sequences and alignments.



[10] Chou, PY and Fasman, GD. Annu Rev Biochem. 1978;47:251-76.

Exercise 10: Colouring Alignments

10.a. Open a sequence alignment, for example the PFAM domain PF03460. Select the alignment menu option *Colour* ⇒ *ClustalX*. Note the colour change. Now try all the other colour schemes in the *Colour* menu. Note that some colour schemes do not colour all residues.

10.b. Colour the alignment using *Colour* ⇒ *Blosum62*. Select a group of around 4 similar sequences. Use the context menu (right click on the group) option *Selection* ⇒ *Group* ⇒ *Group Colour* ⇒ *Blosum62* to colour the selection. Notice how some residues which were not coloured are now coloured. The calculations performed for dynamic colouring schemes like Blosum62 are based on the group being coloured, not the whole alignment (this also explains the colouring changes observed in the group selection exercise step 5).

10.c. Keeping the same selection as before, colour the complete alignment using *Colour* ⇒ *Taylor*. Select the menu option *Colour* ⇒ *By Conservation*. Slide the selector from side to side and observe the changes in the alignment colouring in the selection and in the complete alignment.

User Defined

This dialogue allows the user to create any number of named colour schemes at will. Any residue may be assigned any colour. The colour scheme can then be named. If you save the colour scheme, this name will appear on the Colour menu

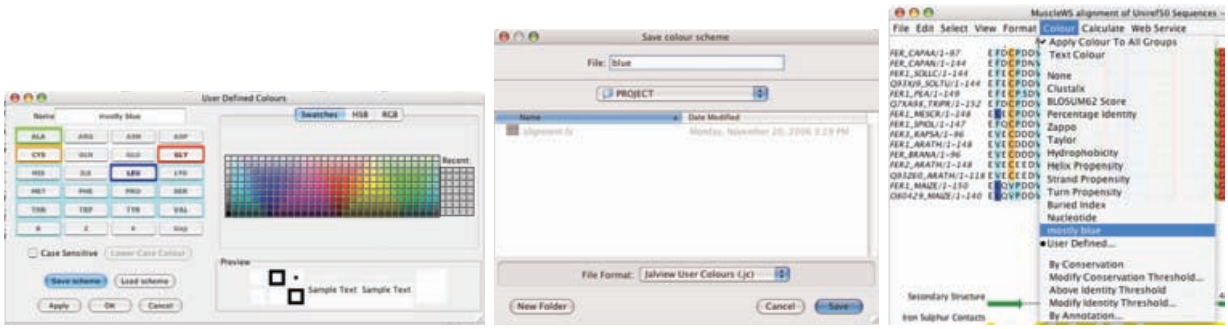


Figure 1.25: **Creation of a user defined colour scheme.** Residue types are assigned colours (left). The profile is saved (center) and can then be accessed via the *Colour* menu (right).

Exercise 11: User defined colour schemes

11.a. Load a sequence alignment. Select the alignment menu option *Colour* ⇒ *User Defined*. A dialogue window will open.

11.b. Click on an amino acid button, then select a colour for that amino acid. Repeat till all amino acids are coloured to your liking.

11.c. Insert a name in the appropriate field and click *Save Scheme*. You will be prompted for a file name in which to save the colour scheme. The dialogue window can now be closed.

11.d. The new colour scheme appears in the list of colour schemes in the *Colour* menu and can be selected in future Jalview sessions.

1.8 Alignment formatting and graphics output

Jalview is a WYSIWIG alignment editor. This means that for most kinds of graphics output, the layout that is seen on screen will be the same as what is output in an exported graphics file. It is therefore important to pick the right kind of display layout prior to generating figures.

1.8.1 Multiple Alignment Views

Jalview is able to create multiple independent visualizations of the same underlying alignment - these are called Views. Because each view displays the same underlying data, any edits performed in one view will update the alignment or annotation visible in all views.

Alignment views are created using the View ⇒ New View option of the alignment window. This will create a new view with the same groups, alignment layout and display options as the current one. Views may be gathered (by pressing G) together as named tabs on the alignment window, or displayed simultaneously in their own window (by pressing X).

1.8.2 Alignment layout

Jalview provides two screen layout modes, unwrapped (the default) where the alignment is in one long line across the window, and wrapped, where the alignment is on multiple lines, each the width of the window. Most layout options are controlled by the Format menu option in the alignment window, and control the overall look of the alignment in the view (rather than just a selected region).

Wrapped alignments

Wrapped alignments can be toggled on and off using the *Format* ⇒ *Wrap* menu option (Figure 1.26). Note that the annotation lines are also wrapped. Wrapped alignments are great for publications and presentations but are of limited use when working with large numbers of sequences. Furthermore, alignment annotation (see Section 2.4) cannot be interactively created or edited in wrapped mode, and selection of large regions is difficult.

Fonts

The text appearance in a view can be modified via the *Format* ⇒ *Font...* alignment window menu. This setting applies for all alignnment and annotation text except for that displayed in tool-tips. Additionally, font size and spacing can be adjusted rapidly by clicking the middle mouse button and dragging across the alignment window.

Figure 1.26: **Wrapping the alignment**

Numbering and label justification

Options in the *Format* menu are provided to control the also provides a range of options to control the display of sequence and alignment numbering, the justification of sequence IDs and annotation row column labels on the annotation rows shown below the alignment.

Alignment and Group colouring and appearance

The display of hidden row/column markers and gap characters can be turned off with *Format* ⇒ *Hidden Markers* and *Format* ⇒ *Show Gaps*, respectively. The *Text* and *Colour Text* option controls the display of sequence text and the application of alignment and group colouring to it. *Boxes* controls the display of the background area behind each residue that is coloured by the applied coloursheme.

Highlighting nonconserved symbols

The alignment layout and group sub-menu both contain an option to hide conserved symbols from the alignment display (*Format* ⇒ *Show nonconserved* in the alignment window or *Popup Menu* ⇒ *Group* ⇒ *Show nonconserved*). This mode is useful when working with alignments that exhibit a high degree of homology, because Jalview will only display gaps or sequence symbols that differ from the consensus for each column, and render all others with a '.'.

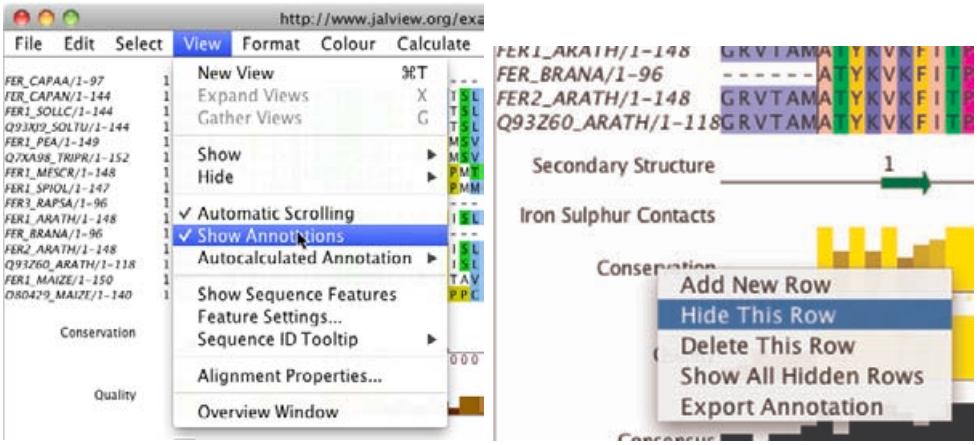


Figure 1.27: **Hiding Annotations** Annotations can either be hidden from the *View* menu (left) or individually from the context menu (right)

1.8.3 Annotation ordering and display

The annotation lines which appear below the sequence alignment are described in detail in Section 1.27. They can be hidden by toggling the *View* ⇒ *Show Annotations* menu option. Additionally, each annotation line can be hidden and revealed in the same way as sequences via the context menu on the annotation name panel (Figure 1.27). Annotations can be reordered by dragging the annotation line label on the annotation label panel. Placing the mouse over the top annotation label brings up a resize icon. When this is displayed, Click-dragging up and down alters the relative size of the sequence alignment and annotation alignment panels.

Exercise 12: Alignment Layout

12.a. Start Jalview and open the URL http://www.jalview.org/examples/exampleFile.jar. Select *Format* ⇒ *Wrap* from the alignment window menu. Experiment with the various options from the *Format* menu. to adjust the ruler placement, sequence ID format and so on.

12.b. Hide all the annotation rows by selecting *View* ⇒ *Show Annotations* from the alignment window menu. Reveal the annotations by selecting the same menu option.

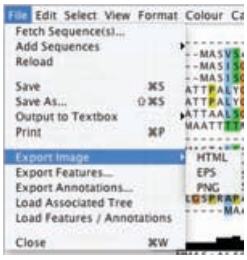
12.c. Right click on the annotation row labels to bring up the context menu. Select *Hide This Row*. Bring up the context menu again and select *Show All Hidden Rows* to reveal them

12.d. Annotations can be reordered by clicking and dragging the row to the desired position. Click on the *Consensus* row and drag it upwards to just above *Quality*. The rows should now be reordered. Features and annotations are covered in more detail in Section 2.4 below.

12.e. Move the mouse to the top left hand corner of the Secondary Structure annotation row label - a grey up/down arrow symbol should appear - when this is shown, the height of the *Annotation Area* can be changed by Clicking and dragging the mouse up or down.

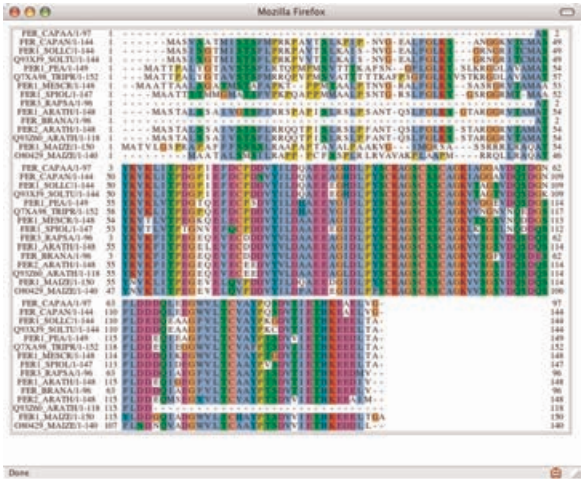
1.8.4 Graphical output

Jalview allows alignments figures to be exported in three different formats, each of which is suited to a particular purpose. Image export is via the *File* ⇒ *Export Image* ⇒ ... alignment window menu option.



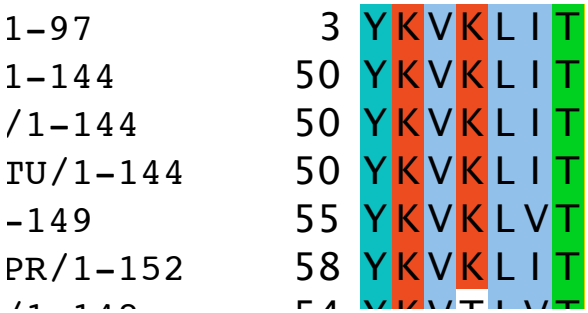
HTML

HTML is the format used by web pages. Jalview outputs the alignment as an HTML table with all the colours and fonts as seen. Any additional annotation will also be embedded as sensitive areas on the page, such as URL links for each sequence's ID label. This file can then be viewed directly with any web browser. Each residue is placed in an individual table cell. Unwrapped alignments will produce a very wide page.



EPS

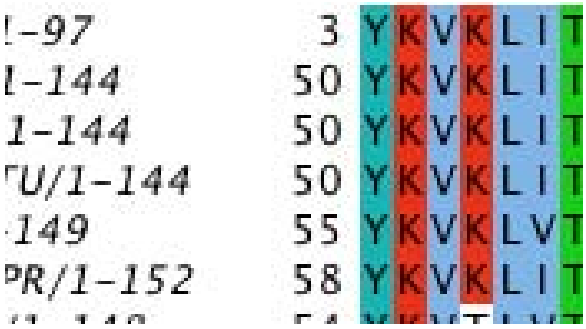
EPS is Encapsulated Postscript. It is the format of choice for publication and posters as it gives the highest quality output of any of the image types. It can be scaled indefinitely so will still look good on an A0 poster. This format can be read by most good presentation and graphics packages such as Adobe Illustrator.



Zoom Detail of EPS image.

PNG

PNG is Portable Network Graphics. This output option produces an image that can be easily included in web pages and incorporated in presentations using e.g. Powerpoint or Open Office. It is a bitmap image so does not scale and is unsuitable for use on posters, or in publications.
For submission of alignment figures to journals, please use EPS[11].



Zoom Detail of PNG image.

Exercise 13: Graphical Output

13.a. Load the example Jalview Jar file in Exercise 12. Customise it how you wish but leave it unwrapped. Select *File* ⇒ *Export Image* ⇒ *HTML* from the alignment menu. Save the file and open it in your favourite web browser.

13.b. Now wrap the alignment (Exercise 12) and export the image to HTML again. Compare the two images. Note that the exported image matches the format displayed in the alignment window but annotations are not exported.

13.c. Export the alignment using the *File* ⇒ *Export Image* ⇒ *PNG* menu option. Open the file in an image viewer that allows zooming (eg. Paint or Photoshop on Windows, Preview on Mac OS X) and zoom in. Notice that the image is a bitmap and it becomes pixelated very quickly. Note also that the annotation lines are included in the image.

13.d. Export the alignment using the *File* ⇒ *Export Image* ⇒ *EPS* menu option. Open the file in a suitable program such as Ghostview or Preview (Mac OS X). Zoom in and note that the image is indefinitely scalable.

[11] If the journal complains, *insist*.

Chapter 2

Analysis and Annotation

This chapter describes the annotation, analysis, and visualization tasks that the Jalview Desktop can perform. Section 2.1 introduces the structure visualization capabilities of Jalview. In Section 2.2, you will find details of the Tree building, viewing and PCA capabilities, alignment redundancy removal, pairwise alignments and alignment conservation analysis. Subsequently, in Section 2.3, programs available remotely for multiple sequence alignment and secondary structure prediction are described.

Section 2.4 describes the mechanisms provided by Jalview for interactive creation of sequence and alignment annotation and how they are displayed, imported and exported. Section 2.5 discusses the retrieval of database references and establishment of sequence coordinate systems for the retrieval and display of features from databases and DAS annotation services. Finally, Section 2.6 describes functions and visualization techniques relevant to working with nucleotide sequences, coding region annotation and nucleotide sequence alignments.

2.1 Working with structures

Jalview facilitates the use of protein structures for the analysis of alignments by providing a linked view of structures associated with protein sequences in the alignment. The Java based molecular viewing program Jmol[1] has been incorporated[2] which enables sophisticated molecular visualizations to be prepared and investigated alongside an analysis of associated sequences. PDB format files can be imported directly or structures can be retrieved from the the Macromolecular Structure Database (MSD) using the Sequence Fetcher (see 1.4.5).

[1]See the Jmol homepage `http://www.jmol.org` for more information.

[2]Earlier versions of Jalview included MCView - a simple main chain structure viewer. Structures are visualized as an alpha carbon trace and can be viewed, rotated and coloured in a structure viewer and the results interpreted on a sequence alignment.

2.1.1 Automatic association of PDB structures with sequences

Jalview can automatically determine which structures are associated with a sequence if that sequence has an ID from a public database that contains cross-references to the PDB, such as Uniprot. Right-click on any sequence ID and select <*Sequence ID*> ⇒ *Associate Structure with Sequence* ⇒ *Discover PDB IDs* from the context menu (where <*Sequence ID*> is the ID of the sequence on which you clicked) (Figure 2.1). Jalview will attempt to associate the sequence with a Uniprot sequence and from there discover any associated PDB structures. This takes a few seconds and applies to all sequences in the alignment which have valid Uniprot IDs. On moving the cursor over the sequence ID the tool tip[3] now shows the Uniprot ID and any associated PDB structures.

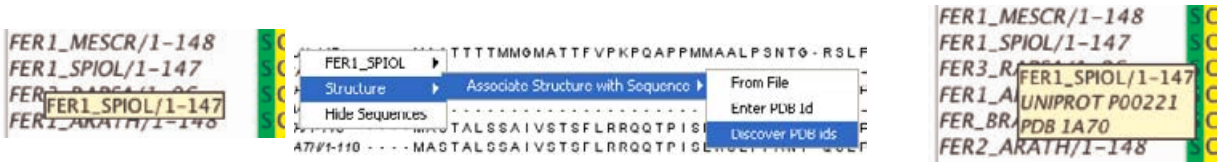


Figure 2.1: **Automatic PDB ID discovery.** The tooltip (left) indicates that no PDB structure has been associated with the sequence. After PDB ID discovery (center) the tool tip now indicates the Uniprot ID and any associated PDB structures (right)

2.1.2 Viewing Protein Structures

The structure viewer can be launched through the sequence ID context menu. Select *Structure* ⇒ *View PDB entry* ⇒ <*PDB ID*>. The structure will be downloaded or loaded from the local file system, and shown as a ribbon diagram coloured according to the associated sequence in the current alignment view (Figure 2.2 (right)). The structure can be rotated by clicking and dragging in the structure window. The structure can be zoomed using the mouse scroll wheel (if available). Moving the mouse cursor over a sequence to which the structure is linked in the alignment panel highlights the respective residue's sidechain atoms. The sidechain highlight may be obscured by other parts of the molecule. Similarly, moving the cursor over the structure shows a tooltip and highlights the corresponding residue in the alignment. Often, the position highlighted may not be in the visible portion of the current alignment view. If the alignment window's *View* ⇒ *Automatic Scrolling* option is not selected, then you may have to manually move the alignment scroll bars to see the highlighted region.

Customising structure display

Structure display can be modified using the *Colour* and *View* menus in the structure viewer. The background colour can be modified by selecting the *Colours* ⇒ *Background Colour...* option.

By default, the structure will be coloured in the same way as the sequence in the associated alignment view. The structure can be coloured independently of the sequence by selecting an

[3]Tip: The sequence ID tooltip can often become large for heavily cross referenced sequence IDs. Use the *View Sequence ID Tooltip* ⇒ submenu to disable the display of database cross references or non-positional features.

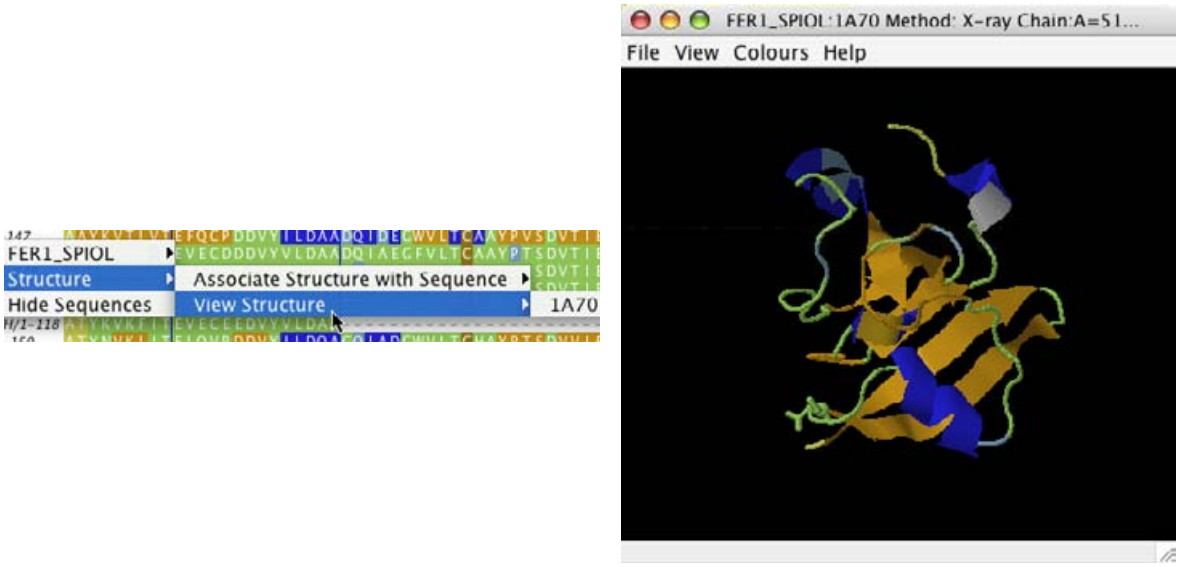


Figure 2.2: **Structure visualization** The structure viewer is launched from the sequence ID context menu (left) and allows the structure to be visualized using the embedded Jmol molecular viewer (right).

appropriate colour scheme from the *Colours* menu. It can be coloured according to the alignment using the *Colours* ⇒ *By Sequence* option. The image in the structure viewer can be output to EPS or PNG format via the *File* ⇒ *Save As* ⇒ ... submenu. The mapping between the structure and the sequence (How well and which parts of the structure relate to the sequence) can be viewed with the *File* ⇒ *View Mapping* menu option.

Using the Jmol visualization interface

Jmol has a comprehensive set of selection and visualization functions that are accessed from the Jmol popup menu (by right-clicking in the Jmol window or by clicking the Jmol logo). Molecule colour and rendering style can be manipulated, and distance measurements and molecular surfaces can be added to the view. It also has its own "Rasmol[4]-like" scripting language, which is described elsewhere[5]. Jalview utilises the scripting language to interact with Jmol and to store the state of a Jmol visualization within Jalview archives, in addition to the PDB data file originally loaded or retrieved by Jalview. To access the Jmol scripting environment directly, use the *Jmol* ⇒ *Console* menu option.

[4]see http://www.rasmol.org
[5]Jmol Wiki: `http://wiki.jmol.org/index.php/Scripting`
Jmol Scripting reference: `http://www.stolaf.edu/academics/chemapps/jmol/docs/`

Exercise 14: Viewing Structures

14.a. Load the alignment at http://www.jalview.org/examples/exampleFile.jar. Right-click on the sequence ID label for any of the sequences (e.g. *FER1_SPIOL*) to bring up the context menu. Select *FER1_SPIOL* ⇒ *Associate Structure with Sequence* ⇒ *Discover PDB ids*. Jalview will now attempt to find PDB structures for the sequences in the alignment.

14.b. Right-click on the sequence ID for *FER1_SPIOL*. Select *FER1_SPIOL* ⇒ *View PDB Entry: 1A70* A structure viewing window appears. Rotate the molecule by clicking and dragging in the structure viewing box. Zoom with the mouse scroll wheel.

14.c. Roll the mouse cursor along the *FER1_SPIOL* sequence in the alignment. Note that if a residue in the sequence maps to one in the structure, a label will appear next to that residue in the structure viewer. Move the mouse over the structure. Placing the mouse over a part of the structure will bring up a tool tip indicating the name and number of that residue. The corresponding residue in the sequence is highlighted in black. Clicking the alpha carbon toggles the highlight and residue label on and off. Try this by clicking on a set of three or four adjacent residues so that the labels are persistent, then finding where they are in the sequence.

14.d. Select *Colours* ⇒ *Background Colour...* from the structure viewer menu and choose a suitable colour. Press *OK* to apply this. Select *File* ⇒ *Save As* ⇒ *PNG* and save the image. View this with your web browser.

14.e. Select *File* ⇒ *View Mapping* from the structure viewer menu. A new window opens showing the residue by residue alignment between the sequence and the structure.

14.f. Select *File* ⇒ *Save* ⇒ *PDB file* and choose a new filename to save the PDB file. Once the file is saved, open the location in your file browser (or explorer window) and drag the PDB file that you just saved on to the Jalview desktop (or load it from the *Jalview Desktop* ⇒ *Input Alignment* ⇒ *From File* menu). Verify that you can open and view the associated structure from the sequence ID pop-up menu's *Structure* submenu in the new alignment window.

14.g. Right click on the structure to bring up the Jmol window. Explore the menu options. Try to change the style of molecular display - by first using the *Jmol* ⇒ *Select* ⇒ *all* command, and then the *Jmol* ⇒ *Style* ⇒ *Scheme* ⇒ *Ball and stick* command.

14.h. Use the *File* ⇒ *Save As ..* function to save the alignment as a Jalview Project. Now close the alignment and the structure view, and load the project file you just saved.
Verify that the Jmol display is as it was when you just saved the file.

2.2 Analysis of alignments

Jalview provides support for sequence analysis in two ways. A number of analytical methods are 'built-in' and run inside Jalview itself and are mostly accessed from the *Calculate* alignment window menu. Computationally intensive analyses are run outside Jalview via web services - these are typically accessed via the *Web Services* menu, and described in Section 2.3. In this section, we describe the built-in analysis capabilities common to both the Jalview Desktop and the JalviewLite applet.

2.2.1 PCA

This calculation creates a spatial representation of the similarities within the current selection or the whole alignment if no selection has been made. After the calculation finishes, a 3D viewer displays the each sequence as a point in 3D 'similarity space'. Sets of similar sequences tend to lie near each other in this space. Note: The calculation is computationally expensive, and may fail for very large sets of sequences - because the JVM has run out of memory. Memory issues were discussed in Section 1.4.6.

What is PCA?

Principal components analysis is a technique for examining the structure of complex data sets. The components are a set of dimensions formed from the measured values in the data set, and the principle component is the one with the greatest magnitude, or length. The sets of measurements that differ the most should lie at either end of this principle axis, and the other axes correspond to less extreme patterns of variation in the data set. In this case, the components are generated by an eigenvector decomposition of the matrix formed from the sum of BLOSUM scores at each aligned position between each pair of sequences. The basic method is described in the 1995 paper by *G. Casari, C. Sander* and *A. Valencia* [6] and implemented at the SeqSpace server at the EBI.

The PCA Viewer

PCA analysis can be launched from the *Calculate* ⇒ *Principle Component Analysis* menu option. PCA requires a selection containing at least 4 sequences. A window opens containing the PCA tool (Figure 2.3). Each sequence is represented by a square, coloured by the background colour of the sequence ID label. The axes can be rotated by clicking and dragging the left mouse button and zoomed using the ↑ and ↓ keys or the scroll wheel of the mouse (if available). A tool tip appears if the cursor is placed over a sequence. Sequences can be selected by clicking on them. [CTRL]-Click can be used to select multiple sequences. Labels will be shown for each sequence by toggling the *View* ⇒ *Show Labels* menu option, and the plot background colour changed via the *View* ⇒ *Background Colour..* dialog box. A graphical representation of the PCA plot can be exported as an EPS or PNG image via the *File* ⇒ *Save As* ⇒ ... submenu.

[6] *Nature Structural Biology* (1995) **2**, 171-8. PMID: 7749921

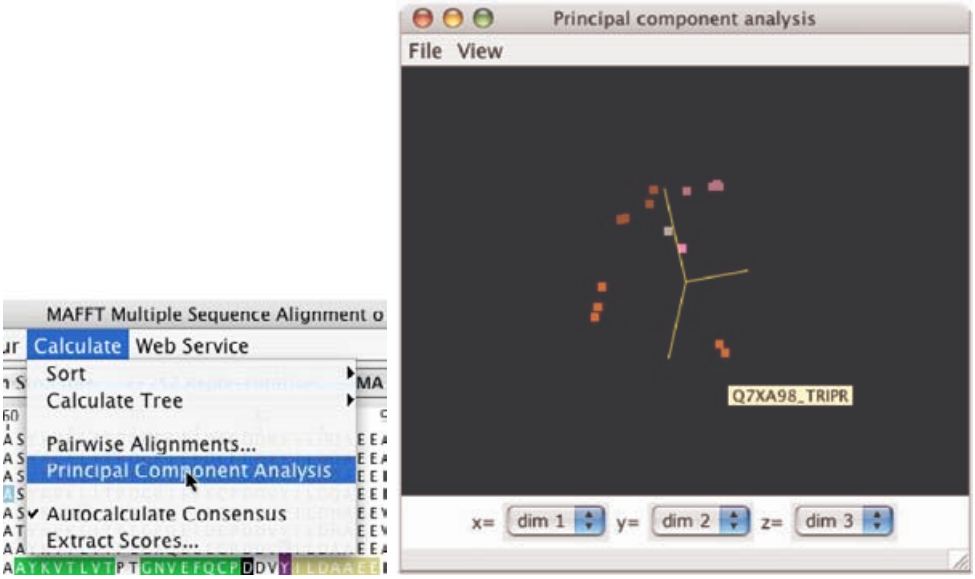


Figure 2.3: **PCA Analysis**

Exercise 15: Principle Component Analysis

15.a. Load the alignment at http://www.jalview.org/examples/exampleFile.jar and select *Edit ⇒ Undefine Groups*.

15.b. Select the menu option *Calculate ⇒ Principle Component Analysis*. A new window will open. Move this window so that the tree, alignment and PCA viewer window are all visible. Try rotating the plot by clicking and dragging the mouse on the plot in the PCA window. Note that clicking on points in the plot will highlight them on the alignment and tree.

15.c. Click on the tree window. Careful selection of the tree partition location will divide the alignment into a number of groups, each of a different colour. Note how the colour of the sequence ID label matches both the colour of the partitioned tree and the points in the PCA plot.

2.2.2 Trees

Jalview can calculate and display trees, providing interactive tree-based grouping of sequences though a tree viewer. All trees are calculated via the *Calculate ⇒ Calculate Tree ⇒ ...* submenu. Trees can be calculated from distance matrices determined from % identity or aggregate BLOSUM 62 score using either average distance (UPGMA) or Neighbour joining algorithms. The input data for a tree calculation is either the visible portions of the current selection, or the whole alignment if no selection is present.

On calculating a tree, a new window opens (Figure 2.4) which contains the tree. Various display options can be found in the tree window *View* menu, and export options in the *File ⇒ Save As* submenu. Newick format is a standard file format for trees which allows them to be exported to other programs. Jalview can also read in external trees in Newick format via the *File ⇒ Load Associated Tree* menu option. Leaf names on imported trees will be matched to the associated

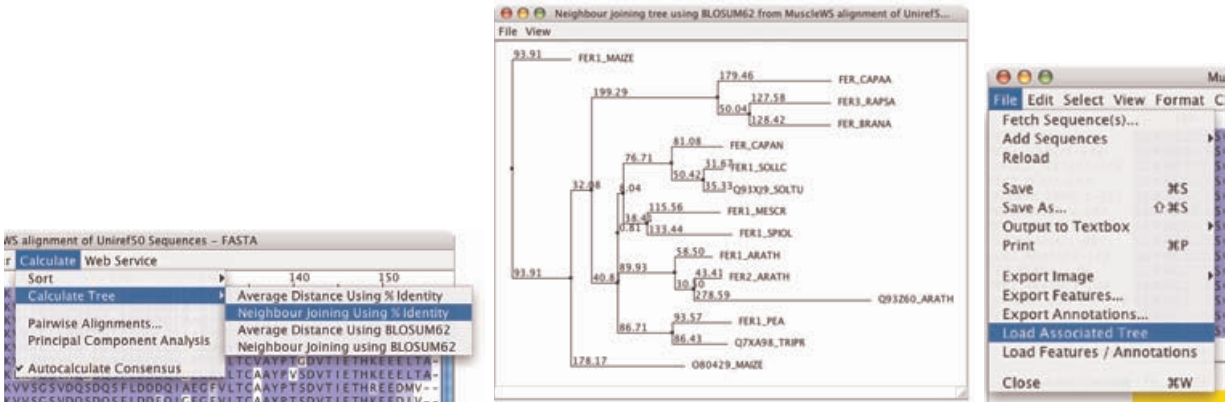


Figure 2.4: **Calculating Trees** Jalview provides four built in models for calculating trees. Jalview can also load precalculated trees in Newick format (right).

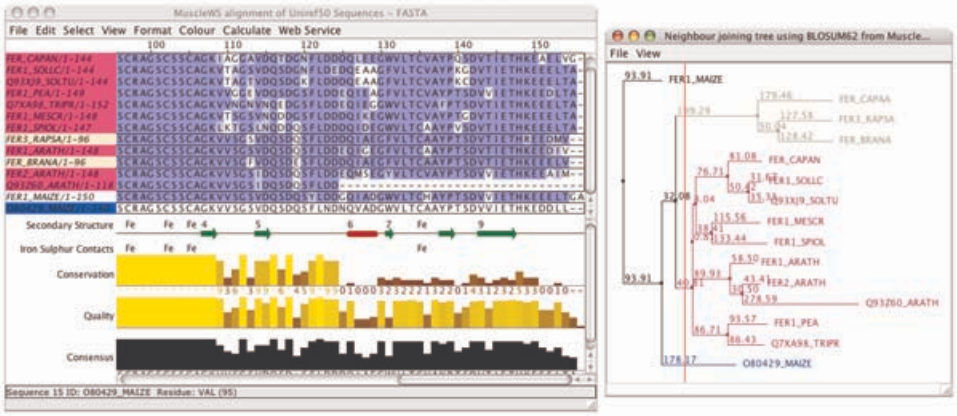


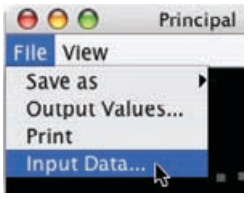
Figure 2.5: **Interactive Trees** The tree level cutoff can be used to designate groups in Jalview

alignment - unmatched leaves will still be displayed, and can be highlighted using the *View* ⇒ *Show Unlinked Leaves* menu option.

Clicking on the tree brings up a cursor across the height of the tree. The sequences are automatically partitioned and coloured (Figure 2.5). To group them together, select the *Calculate* ⇒ *Sort* ⇒ *By Tree Order* alignment window menu option and the correct tree. The sequences will then be sorted according to the leaf order currently shown in the tree view. The coloured background to the sequence IDs can be removed with *Select* ⇒ *Undefine Groups* from the alignment window menu.

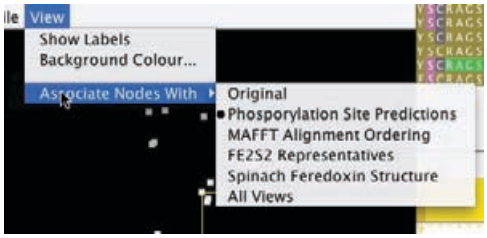
Recovering input data for a tree or PCA plot calculation

The *File* ⇒ *Input Data* option will open a new alignment window containing the original data used to calculate the tree or PCA plot (if available). This function is useful when a tree has been created and then the alignment subsequently changed.



Changing the associated view for a tree or PCA viewer

The *View ⇒ Associated View ⇒ ..* submenu is shown when the viewer is associated with an alignment that is involved in multiple views. Selecting a different view does not affect the tree or PCA data, but will change the colouring and display of selected sequences in the display in accord with the colouring and selection state of the newly associated view.



Exercise 16: Trees

16.a. Open the alignment at http://www.jalview.org/tutorial/alignment.fa. Select *Calculate ⇒ Calculate Tree ⇒ Neighbour Joining Using BLOSUM62.* A new tree window will appear.

16.b. Click on the tree window. A cursor will appear. Note that placing this cursor divides the tree into a number of groups by colour. Place the cursor to give about 4 groups, then select *Calculate ⇒ Sort ⇒ By Tree Order ⇒ Neighbour Joining Tree using BLOSUM62 from ...* . The sequences are reordered to match the order in the tree and groups are formed implicitly.

16.c. Select *Calculate ⇒ Calculate Tree ⇒ Neighbour Joining Using % Identity.* A new tree window will appear. The group colouring makes it easy to see the diferences between the two trees, calculated using different methods.

16.d. Select from sequence 2 column 60 to sequence 12 column 123. Select *Calculate ⇒ Calculate Tree ⇒ Neighbour Joining Using BLOSUM62.* A new tree window will appear. It can be seen that the tree contains 11 sequences. It has been coloured according to the already selected groups from the first tree and is calculated purely from the residues in the selection. Comparing the location of individual sequences between the two trees illustrates the importance of selecting appropriate regions of the alignment for the calculation of trees.

16.e. Recover the input data for the tree you just calculated. Check the *Edit ⇒ Pad Gaps* option is not ticked, and insert one gap in the alignment. Now select *Calculate ⇒ Calculate Tree ⇒ Neighbour Joining Using BLOSUM62.*
A warning dialog box **"Sequences must be aligned"** appears because the sequences input to the tree calculation are of different lengths.

16.f. Now select *Edit ⇒ Pad Gaps* and try to perform the tree calculation again - this time a new tree should appear.
This demonstrates the use of the *Pad Gaps* editing preference, which ensures that all sequences are the same length after editing.

2.2.3 Tree Based Conservation Analysis

Trees reflect the pattern of global sequence similarity exhibited by the alignment or region within the alignment that was used for their calculation. The Jalview tree viewer enables sequences to be partitioned into groups based on the tree, by clicking within the tree viewer window. Once subdivided, the conservation between and within groups can be visually compared in order to better understand the pattern of similarity revealed by the tree, and the variation within the clades partitioned by the grouping. The conservation based colourschemes and the group associated

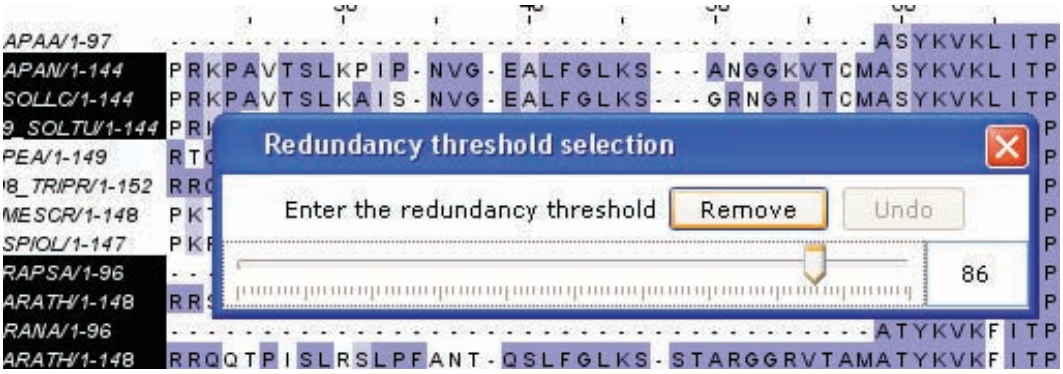


Figure 2.6: The Redundancy Removal dialog box opened from the edit menu. Sequences that exceed the current percentage identity threshold and are to be removed are highlighted in black.

conservation and consensus annotation (enabled using the alignment window's *View* ⇒ *Automatic Annotation* ⇒ *Group Consensus* and *Group Conservation* options) can help when working with larger alignments.

Exercise 17: Tree Based Conservation Analysis

17.a. Load the PF03460 Seed alignment using the sequence fetcher. Colour it with the *Taylor colourscheme*, and apply *Conservation* shading.

17.b. Build a Neighbourjoining tree using BLOSUM62 and use the sort submenu to order alignment using the calculated tree.

17.c. Select a point on the tree to partition the alignment, and examine the variation in colouring between different groups.
You may find it easier to browse the alignment by if you uncheck the *Show Annotations* view option, and open the overview window to aid navigation.

17.d. Try changing the colourscheme to BLOSUM62 (whilst ensuring that *Apply Colour to All Groups* is selected)

Note: You may want to save the alignment and tree as a project file, since it is used in the next few exercises.

2.2.4 Redundancy Removal

The redundancy removal dialog box is opened using the *Edit* ⇒ *Remove Redundancy...* option in the alignment menu. As its menu option placement suggests, this is actually an alignment editing function, but it is convenient to describe it here. The redundancy removal dialog box presents a percentage identity slider which sets the redundancy threshold. Aligned sequences which exhibit a percentage identity greater than the current threshold are highlighted in black. The [Remove] button can then be used to delete these sequences from the alignment as an edit operation[7].

[7]Which can usually be undone. A future version of Jalview may allow redundant sequences to be hidden, or represented, rather than deleted.

Exercise 18: Remove redundant sequences

18.a. Re-use or recreate the alignment and tree which you worked with in the tree based conservation analysis exercise (exercise 2.2.3)

18.b. Open the Remove Redundancy dialog and adjust the threshold to 90%. Remove the sequences that are more than 90% similar under this alignment.

18.c. Select the Tree viewer's *View* ⇒ *Show Linked Leaves* option, and note that the removed sequences are now prefixed with a * in the tree view.

18.d. Use the [Undo] button on the dialog to recover the sequences. Note that the * symbols disappear from the tree display.

18.e. Experiment with the redundancy removal and observe the relationship between the percentage identity threshold and the pattern of unlinked nodes in the tree display.

2.2.5 Subdividing the alignment according to specific mutations

It is often necessary to explore variations in an alignment that may correlate with mutations observed in a particular region; for example, sites exhibiting single nucleotide polymorphism, or residues involved in substrate recognition in an enzyme. One way to do this would be to calculate a tree using the specific region, and subdivide it in order to partition the alignment. However, calculating a tree can be slow for large alignments, and the tree may be difficult to partition when complex mutation patterns are being analysed. The *Select* ⇒ *Make groups for selection* function was introduced to make this kind of analysis easier. When selected, it will use the characters in the currently selected region to subdivide the alignment. For example, if a single column is selected, then the alignment (or each group defined on the alignment) will be divided into groups based on the residue or nucleotide found at that position. These new groups are annotated with the characters in the selected region, and Jalview's group based conservation analysis annotation and colourschemes can then be used to reveal any associated pattern of sequence variation across the whole alignment.

2.2.6 Automated annotation of Alignments and Groups

On loading a sequence alignment, Jalview will normally[8] calculate a set of automatic annotation rows which are shown below the alignment. For nucleotide sequence alignments, only an alignment consensus row will be shown, but for amino acid sequences, alignment quality (based on BLOSUM62) and physicochemical conservation will also be shown. Conservation is calculated according to Livingstone and Barton[9]. Consensus is the modal residue (or + where there is an equal top residue). The inclusion of gaps in the consensus calculation can be toggled by right-clicking on the the Consensus label and selecting *Ignore Gaps in Consensus* from the context menu. Quality is a measure of the inverse likelihood of unfavourable mutations in the alignment. Further details on these calculations can be found in the on-line documentation.

These annotations can be hidden and deleted but are only created on loading an alignment. If they are deleted then the alignment should be saved and reloaded to restore them. Jalview provides a

[8] Automatic annotation can be turned off in the *Visual* tab in the *Tools* ⇒ *Preferences* dialog box.

[9] *"Protein Sequence Alignments: A Strategy for the Hierarchical Analysis of Residue Conservation."* Livingstone C.D. and Barton G.J. (1993) *CABIOS* **9**, 745-756

toggle to autocalculate a consensus sequence upon editing. This is normally left selected but for large alignments can be turned off via the *Calculate* ⇒ *Autocalculate Consensus* menu option if the interface is too sluggish.

Group Associated Annotation

Group associated consensus and conservation annotation rows reflect the sequence variation within a particular group. Their calculation is enabled by selecting the *Group Conservation* or *Group Consensus* options in the *View* ⇒ *Automatic Annotation* submenu of the alignment window.

Alignment and Group Sequence Logos

The consensus annotation row that is shown below the alignment can be overlaid with a sequence logo that reflects the symbol distribution at each column of the alignment. Right click on the Consensus annotation row and select the 'Show Logo' option to display the Consensus profile for the group or alignment. Sequence logos can be enabled by default for all new alignments *via* the Visual tab in the Jalview desktop's preferences dialog box.

Exercise 19: Group conservation analysis

19.a. Re-use or recreate the alignment and tree which you worked with in the tree based conservation analysis exercise (exercise 2.2.3)

19.b. Create a new view, and ensure the annotation panel is displayed, and enable the display of *Group Consensus*, and the display of sequence logos to make it easier to see the different residue populations within each group.

19.c. Select a column exhibiting about 50% conservation that lies within the central conserved region of the alignment. Subdivide the alignment according to this selection using *Select* ⇒ *Make groups for selection.*

19.d. Re-order the alignment according to the new groups that have been defined. Click on the group annotation row IDs to select groups exhibiting a specific mutation.

19.e. Select another column exhibiting about 50% conservation overall, and subdivide the alignment further. Note that the new groups inherit the names of the original groups, allowing you to identify the combination of mutations that resulted in the subdivision.

19.f. Clear the groups, and try to subdivide the alignment using two non-adjacent columns. *Hint: You may need to hide the intervening columns before you can select both of the columns that you wish to use to subdivide the alignment.*

19.g. Switch back to the original view, and experiment with subdividing the tree groups made in the previous excercise.

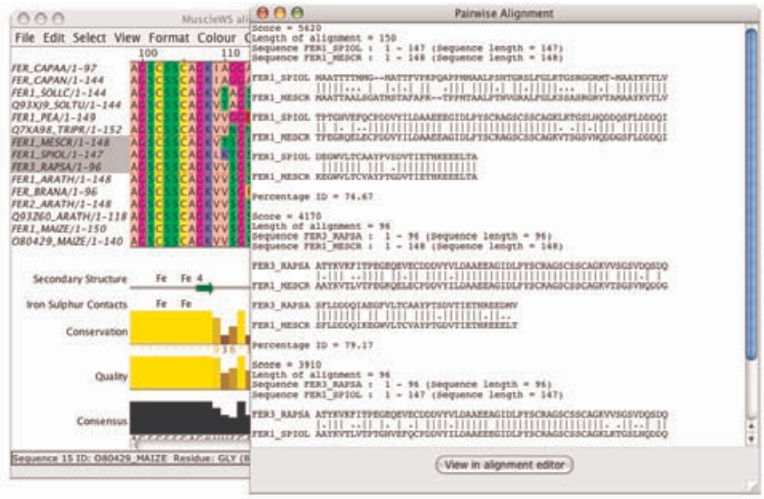


Figure 2.7: **Pairwise alignment of sequences.** Pairwise alignments of three selected sequences are shown in a textbox.

2.2.7 Other Calculations

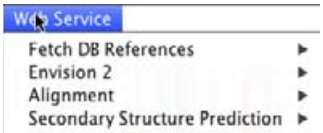
Pairwise Alignments

Jalview can calculate optimal pairwise alignments between arbitrary sequences via the *Calculate ⇒ Pairwise Alignments...* menu option. Global alignments of all pairwise combinations of the selected sequences are performed and the results returned in a text box.

2.3 Webservices

The term "Webservices" refers to a variety of data exchange mechanisms based on HTTP.[10]

Jalview can exploit public webservices to access databases remotely, and also submit data to public services by opening pages with your web browser. These types of services are 'one-way', *i.e.* data is either sent to the webservice or retrieved from it by Jalview. The desktop application can also interact with 'two-way' remote analysis services in order to offload computationally intensive tasks to High Performance Computing facilities.



2.3.1 One way web services

There are three types of one way service in jalview. Database services, which were introduced in in Section 1.4.5, provide sequence and alignment data. They can also be used to add sequence IDs to an alignment imported from a local file, prior to further annotation retrieval, as described in

[10]HTTP: Hyper-Text Transfer Protocol.

Section 2.5. A second type of one way service is provided by Jalview's DAS sequence feature retrieval system, which is described in Section 2.5.2. The final type of one way service are sequence and ID submission services, exemplified by the 'Envision2 Services' provided by the ENFIN Consortium[11].

One-way submission services

Jalview can use the system's web browser to submit sets of sequences and sequence IDs to web based applications. Single sequence IDs can be passed to a web site using the user definable URL links listed under the *Links* submenu of the sequence ID popup menu. These are configured in the *Connections* tab of the *Preferences* dialog box.

The Envision2 services presented in the webservice menu provides are the first example of one way services where multiple sequences or sequence IDs can be sent. The *Web services* ⇒ *Envision2 Services* menu entry provides two sub-menus that enable you to submit the sequences or IDs associated with the alignment or just the currently selected sequences to one of the Envision2 workflows. Selecting any one will open a new browser window on the Envision2 web application. The menu entries and their tooltips provide details of the Envision2 workflow and the dataset set that will be submitted (*i.e.* the database reference type, or associated sequence subset). Please note, due to technical limitations, Jalview can currently only submit small numbers of sequences to the workflows - if no sequence or ID submissions are presented in the submenus, then try to select a smaller number of sequences to submit.

2.3.2 Remote Analysis Services

Remote analysis services enable Jalview to use external computational facilities. There are curently two types of service - multiple sequence alignment and protein secondary structure prediction. In both cases, Jalview will construct a job based on the alignment or currently selected sequences, ask the remote server to run the job, monitor status of the job and, finally, retrieve the results of the job and display them. The Jalview user is kept informed of the progress of the job through a status window.

Currently, web service jobs and their status windows are not stored in Jalview Project Files[12], so it is important that you do not close Jalview whilst a job is running. It is also essential that you have a continuous network connection in order to successfully use Web Services from Jalview, since it periodically checks the progress of running jobs.

[11] ENFIN is the European Network for Functional INtegration. Please see http://www.enfin.org for more information.

[12] This may be rectified in future versions.

2.3.3 Multiple Sequence Alignment

Sequences can be aligned using any of three algorithms: ClustalW[13], Muscle[14] or MAFFT[15] Of these, ClustalW is the slowest but is historically the most widely used. Muscle is fast and probably the most accurate for smaller alignments and MAFFT is probably the best for large alignments.

To run an alignment web service, select the appropriate method from the *Web Service* ⇒ *Alignment* ⇒ ... submenu (Figure 2.8). A progress window will appear giving information about the job and any errors that occur. After successful completion of the job, a new window is opened with the results, in this case an alignment. By default, the new alignment will be ordered in the same way as the input sequences; however, many alignment programs re-order the input to place homologous sequences close together. This ordering can be recovered using the 'Original ordering' entry within the *Calculation* ⇒ *Sort* sub menu.

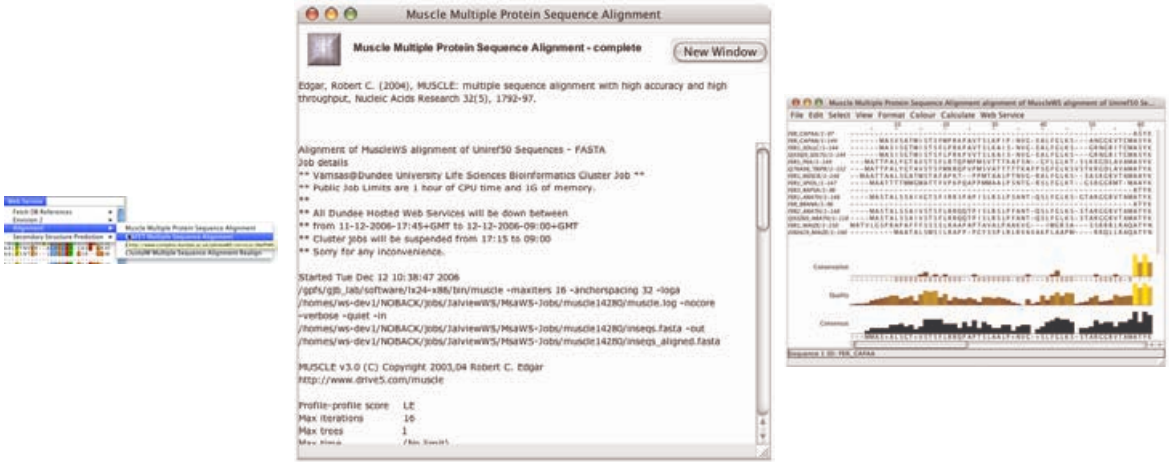


Figure 2.8: **Multiple alignment via web services** The appropriate method is selected from the menu (left), a status box appears (centre), and the results appear in a new window (right)

Realignment

The re-alignment option is currently only supported by ClustalW. When performing a re-alignment, Jalview submits the current selection to the alignment service complete with any existing gaps. This approach is useful when one wishes to align additional sequences to an existing alignment without any further optimisation to the existing alignment. The Re-alignment service provided by ClustalW in this case is effectively a simple form of profile alignment.

[13] *"CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice."* Thompson JD, Higgins DG, Gibson TJ (1994) *Nucleic Acids Research* **22**, 4673-80

[14] *"MUSCLE: a multiple sequence alignment method with reduced time and space complexity"* Edgar, R.C. (2004) *BMC Bioinformatics* **5**, 113

[15] *"MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform"* Katoh, K., Misawa, K., Kuma, K. and Miyata, T. (2002) *Nucleic Acids Research* **30**, 3059-3066. and *"MAFFT version 5: improvement in accuracy of multiple sequence alignment"* Katoh, K., Kuma, K., Toh, H. and Miyata, T. (2005) *Nucleic Acids Research* **33**, 511-518.

Alignments of sequences that include hidden regions

If the view or selected region that is submitted for alignment contains hidden regions, then only the visible sequences will be submitted to the service. Furthermore, each contiguous segment of sequences will be aligned independently (resulting a number of alignment 'subjobs' appearing in the status window). Finally, the results of each subjob will be concatenated with the hidden regions in the input data prior to their display in a new window. This approach ensures that 1) hidden column boundaries in the input data are preserved in the resulting alignment - in a similar fashion to the constraint that hidden columns place on alignment editing (see Section 1.6.6). 2) hidden columns can be used to preserve existing parts of an alignment whilst the visible parts are locally refined.

Exercise 20: Multiple Sequence Alignment

20.a. Close all windows and open the alignment at http://www.jalview.org/tutorial/unaligned.fa. Select *Web Service* ⇒ *Alignment* ⇒ *Muscle Multiple Protein Sequence Alignment*. A window will open giving the job status. After a short time, a second window will open with the results of the alignment.

20.b. Select the first sequence set by clicking on the window and try running ClustalW and MAFFT (from the *Web Services* ⇒ *Alignment* menu) on the same initial alignment. Compare them and you should notice small differences.

20.c. Select the last three sequences in the MAFFT alignment, and de-align them with *Edit* ⇒ *Remove All Gaps*. Press [ESC] to deselect them and then submit the view for re-alignment with ClustalW.

20.d. Use [CTRL]-Z to recover the alignment of the last three sequences in the MAFFT alignment. Once the ClustalW re-alignment has completed, compare the results of re-alignment of the three sequences with their alignment in the original MAFFT result.

20.e. Select columns 60 to 125 in the original MAFFT alignment and hide them. Select *Web Services* ⇒ *Alignment* ⇒ *MAFFT* to submit the visible portion of the alignment to MAFFT. When the web service job pane appears, note that there are now two alignment job status panes shown in the window.

20.f. When the MAFFT job has finished, compare the alignment of the N-terminal visible region in the result with the corresponding region of the original alignment. If you wish, select and hide a few more columns in the N-terminal region, and submit the alignment to the service again and explore the effect of local alignment on the non-homologous parts of the N-terminal region.

2.3.4 Protein Secondary Structure Prediction

Protein secondary structure prediction is performed using the Jpred[16] server at the University of Dundee[17]. The behaviour of this calculation depends on the current selection:

[16] *"The Jpred 3 Secondary Structure Prediction Server"* Cole, C., Barber, J. D. and Barton, G. J. (2008) *Nucleic Acids Research* **36**, (Web Server Issue) W197-W201

"Jpred: A Consensus Secondary Structure Prediction Server" Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M. and Barton, G. J. (1998) *Bioinformatics* **14**, 892-893

[17] http://www.compbio.dundee.ac.uk/www-jpred/

- If nothing is selected, Jalview will check the length of each alignment row to determine if the visible sequences in the view are aligned.
  - If all rows are the same length (often due to the application of the *Edit ⇒ Pad Gaps* option), then a JNet prediction will be run for the first sequence in the alignment, using the current alignment as the profile to use for prediction.
  - Otherwise, just the first sequence will be submitted for a full JNet prediction.
- If just one sequence (or a region on one sequence) has been selected, it will be submitted to the automatic JNet prediction server for homolog detection and prediction.
- If a set of sequences are selected, and they appear to be aligned using the same criteria as above, then the alignment will be used for a Jnet prediction on the first sequence in the set (that is, the one that appears first in the alignment window).

Jpred is launched in the same way as the other web services. Select *Web Services ⇒ Secondary Structure Prediction ⇒ JNet Secondary Structure Prediction* from the alignment window menu (Figure 2.9). A status window opens to inform you of the progress of the job. Upon completion, a new alignment window opens and the Jpred predictions are included as annotations. Consult the Jpred documentation for information on interpreting these results.

Figure 2.9: **Secondary Structure Prediction** Status (left) and results (right) windows for JNet predictions.

Hidden Columns and JNet Predictions

Hidden columns can be used to exclude parts of a sequence or profile from the input sent to the JNet service. For instance, if a sequence is known to include a large loop insertion, hiding that section prior to submitting the JNet prediction may result in a more reliable[18] secondary structure prediction either side of the insertion. Prediction results returned from the service will be mapped back onto the visible parts of the sequence, to ensure a single frame of reference is maintained in your analysis.

[18]This, of course, cannot be guaranteed, but the profile calculated by JNet will at least be different.

Exercise 21: Secondary Structure Prediction

21.a. Open the alignment at http://www.jalview.org/tutorial/alignment.fa. Select the sequence *FER_MESCR* by clicking on the sequence ID. Then select *Web Services* ⇒ *Secondary Structure Prediction* ⇒ *JNet Secondary Structure Prediction* from the alignment window menu. A status window will appear and after some time a new window with the JPred prediction will appear. Note that the number of sequences in the results window is many more than in the original alignment as JNet performs a PSI-BLAST search to expand the prediction dataset.

21.b. Select a different sequence and perform a JNet prediction in the same way. There will probably be minor differences in the predictions.

21.c. Select the second sequence prediction, and copy and paste it into the first prediction window. You can now compare the two predictions. Jnet secondary structure prediction annotation are examples of **sequence associated alignment annotation**.

21.d. Select and hide some columns in one of the profiles that were returned from the JNet service, and then submit the profile for prediction again.

21.e. When you get the result, verify that the prediction has not been made for the hidden parts of the profile, and that the JPred reliability scores differ from the prediction made on the full profile.
Note: you may want to keep this data for use in exercise 23.

2.4 Features and Annotation

Features and annotations are additional information that is overlaid on the sequences and the alignment. Generally speaking, annotations are associated with columns in the alignment. Features are associated with specific residues in the sequence.

Annotations are rendered below the alignment, in the annotation panel, and often reflect properties of the alignment as a whole. The conservation, consensus and quality scores are examples of dynamic annotation. As the alignment changes, these annotations will change along with it. Conversely, sequence features are properties of the individual sequences. They do not change with the alignment, but are shown mapped on to specific residues within the alignment.

Features and annotation can be interactively created, or retrieved from external data sources. DAS (the Distributed Annotation System) is the primary source of sequence features, whilst webservices like JPred (see 2.9 above) can be used to analyse a given sequence or alignment and generate annotation for it.

2.4.1 Creating sequence features

Sequence features can be created simply by selecting the area in a sequence (or sequences) to form the feature and selecting *Selection* ⇒ *Create Sequence Feature* from the right-click context menu (Figure 2.10). A dialogue box allows the user to customise the feature with respect to name, group, and colour. The feature is then associated with the sequence. Moving the mouse over a residue associated with a feature brings up a tool tip listing all features associated with the residue.

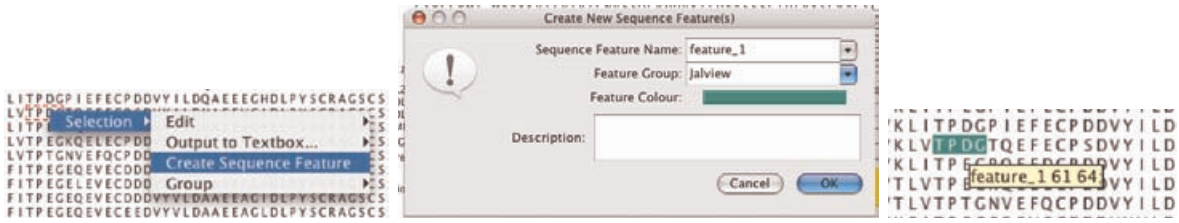


Figure 2.10: **Creating sequence features.** Features can readily be created from selections via the context menu and are then displayed on the sequence.

Creation of features from a selection spanning multiple sequences results in the creation of one feature per sequence. Each feature remains associated with it's own sequence.

2.4.2 Customising feature display

Feature display can be toggled on or off by selecting the *View* ⇒ *Show Sequence Features* menu option. When multiple features are present it is usually necessary to customise the display. Jalview allows the display, colour, rendering order and transparency of features to be modified via the *View* ⇒ *Feature Settings...* menu option. This brings up a dialogue window (Figure 2.12) which allows the visibility of individual feature types to be selected, colours changed (by clicking on the colour of each sequence feature type) and the rendering order modified by dragging feature types to a new position in the list. Dragging the slider alters the transparency of the feature rendering. The Feature Settings dialog also includes functions for more advanced feature shading schemes and buttons for sorting the alignment according to the distribution of features. These capabilities are described further in sections 2.5.3 and 2.5.4.

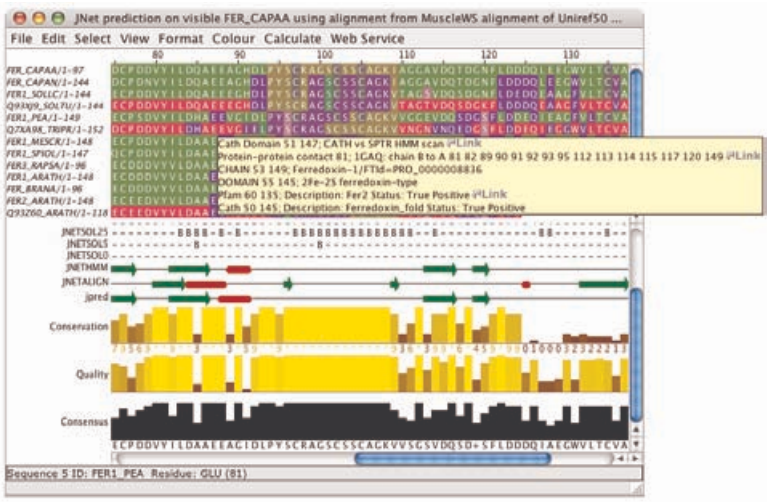


Figure 2.11: **Multiple sequence features.** An alignment with JPred secondary structure prediction annotation below it, and many sequence features overlaid onto the aligned sequences. The tooltip lists the features annotating the residue below the mouse-pointer.

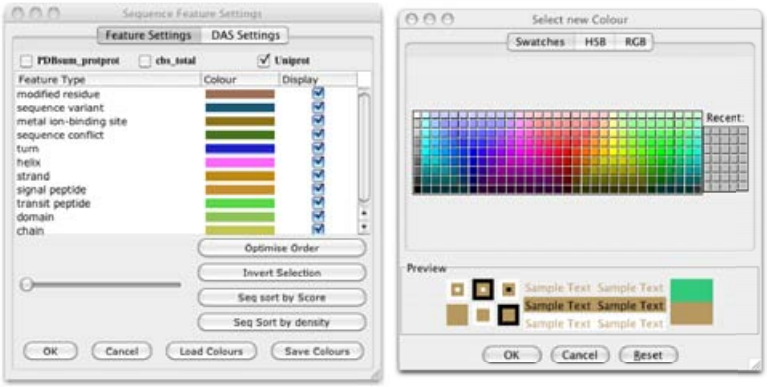


Figure 2.12: **Customising sequence features.** Features can be recoloured, switched on or off and have the rendering order changed.

2.4.3 Sequence Feature File Formats

Jalview supports the widely used GFF tab delimited format[19] and its own Jalview Features file format for the import of sequence annotation. Features and alignment annotation are also extracted from other formats such as Stockholm, and AMSA. URL links may also be attached to features. See the online documentation for more details of the additional capabilities of the jalview features file.

[19]see http://www.sanger.ac.uk/resources/software/gff/spec.html

Exercise 22: Creating features

22.a. Open the alignment at http://www.jalview.org/tutorial/alignment.fa. We know that the Cysteine residues at columns 97, 102, 105 and 135 are involved in iron binding so we will create them as features. Navigate to column 97, sequence 1. Select the entire column by clicking in the ruler bar. Then right-click on the selection to bring up the context menu and select *Selection* ⇒ *Create Sequence Feature*. A dialogue box will appear.

22.b. Enter a suitable Sequence Feature Name (e.g. "Iron binding site") in the appropriate box. Click on the Feature Colour bar to change the colour if desired, add a short description ("One of four Iron binding Cysteines") and press OK. The features will then appear on the sequences.

22.c. Roll the mouse cursor over the new features. Note that the position given in the tool tip is the residue number, not the column number. To demonstrate that there is one feature per sequence, clear all selections by pressing [ESC] then insert a gap in sequence 3 at position 95. Roll the mouse over the features and you will see that the feature has moved with the sequence. Delete the gap you created.

22.d. Add a similar feature to column 102. When the feature dialogue box appears, clicking the Sequence Feature Name box brings up a list of previously described features. Using the same Sequence Feature Name allows the features to be grouped.

22.e. Select *View* ⇒ *Feature Settings...* from the alignment window menu. The Sequence Feature Settings window will appear. Move this so that you can see the features you have just created. Click the check box for "Iron binding site" under *Display* and note that display of this feature type is now turned off. Click it again and note that the features are now displayed. Close the sequence feature settings box by clicking *OK* or *Cancel*.

2.4.4 Creating user defined annotation

Annotations are properties that apply to the alignment as a whole and are visualized on rows in the annotation panel. To create a new annotation row, right click on the annotation label panel and select the *Add New Row* menu option (Figure 2.13). A dialogue box appears. Enter the label to use for this row and a new row will appear.

Figure 2.13: **Creating a new annotation row.** Annotation rows can be reordered by dragging them to the desired place.

To create a new annotation, first select all the positions to be annotated on the appropriate row. Right-clicking on this selection brings up the context menu which allows the insertion of graphics for secondary structure (*Helix* or *Sheet*), text *Label* and the colour in which to present the annotation (Figure 2.14). On selecting *Label* a dialogue box will appear, requesting the text to place at that position. After the text is entered, the selection can be removed and the annotation becomes clearly

visible[20]. Annotations can be coloured or deleted as desired.

Figure 2.14: **Creating a new annotation.** Annotations are created from a selection on the annotation row and can be coloured as desired.

Exercise 23: Annotating alignments

23.a. Load the alignment at http://www.jalview.org/tutorial/alignment.fa. Right-click on the annotation label for *Conservation* to bring up the context menu and select *Add New Row*. A dialogue box will appear asking for *Label for annotation*. Enter "Iron binding site" and click *OK*. A new, empty, row appears.

23.b. Navigate to column 97. Select column 97 on the new annotation row. Right click on the selection and select *Label* from the context menu. Enter "Fe" in the box and click *OK*. Right-click on the selection again and select *Colour*. Choose a colour from the colour chooser dialogue and click *OK*. Press [ESC] to remove the selection.

23.c. Select columns 70-77 on the annotation row. Right-click and choose *Sheet* from the context menu. You will be prompted for a label. Enter "B" and press OK. A new line showing the sheet as an arrow appears. The colour of the label can be changed but not the colour of the sheet arrow.

23.d. Right click on the annotation row that you just created. Select *Export Annotation* and, in the **Export Annotation** dialog box that will open, select the Jalview format and click the [To Textbox] button.
The format for this file is given in the Jalview help. Press [F1] to open it, and find the "Annotations File Format" entry in the "Alignment Annotations" section of the contents pane.

23.e. Export the file to a text editor and edit the file to change the name of the annotation row. Save the file and drag it onto the alignment view.

23.f. Try to add an additional helix somewhere along the row by editing the file and re-importing it. *Hint: Use the* **Export Annotation** *function to view what helix annotation looks like in a jalview annotation file.*

23.g. Use the *Alignment Window* ⇒ *File* ⇒ *Export Annotation..* function to export all the alignment's annotation to a file.

23.h. Open the exported annotation in a text editor, and use the **Annotation File Format** documentation to modify the style of the Conservation, Consensus and Quality annotation rows so they appear as several lines on a single line graph. *Hint: You need to change the style of annotation row in the first field of the annotation row entry in the file, and create an annotation row grouping to overlay the three quantitative annotation rows.*

23.i. Recover or recreate the secondary structure prediction that you made in exercise 21. Use the *File* ⇒ *Export Annotation* function to view the jnet secondary structure prediction annotation row. Note the **SEQUENCE_REF** statements surrounding the row specifying the sequence association for the annotation.

[20]When annotating a block of positions, the text can be partly obscured by the selection highlight. Pressing the [ESC] key clears the selection and the label is then visible.

2.5 Importing features from databases

Jalview supports feature retrieval from public databases either directly or *via* the Distributed Annotation System (DAS[21]). It includes built in parsers for Uniprot and EMBL records retrieved from the EBI. Sequences retrieved from these sources using the sequence fetcher (see Section 1.4.5) will already posess features.

2.5.1 Sequence Database Reference Retrieval

Jalview maintains a list of external database references for each sequence in an alignment. These are listed in a tooltip when the mouse is moved over the sequence ID when the *View ⇒ Sequence ID Tooltip ⇒ Show Database Refs* option is enabled. Sequences retrieved using the sequence fetcher will always have at least one database reference, but alignments imported from an alignment file generally have no database references.

Database References and Sequence Coordinate Systems

Jalview displays features in the local sequence's coordinate system which is given by its 'start' and 'end'. Any sequence features on the sequence will be rendered relative to the sequence's start position. If the start/end positions do not match the coordinate system from which the features were defined, then the features will be displayed incorrectly.

Automatically discovering a sequence's database references

Jalview includes a function to automatically verify and update each sequence's start and end numbering against any of the sequence databases that the *Sequence Fetcher* has access to. This function is accessed from the *Webservices ⇒ Fetch DB References* sub-menu in the Alignment window. This menu allows you to query either the set of *Standard Databases*, which includes EMBL, Uniprot, the PDB, and the currently selected DAS sequence sources, or just a specific datasource from one of the submenus. When one of the entries from this menu is selected, Jalview will use the ID string from each sequence in the alignment or in the currently selected set to retrieve records from the external source. Any sequences that are retrieved are matched against the local sequence, and if the local sequence is found to be a sub-sequence of the retrieved sequence then the local sequence's start/end numbering is updated. A new database reference mapping is created, mapping the local sequence to the external database, and the local sequence inherits any additional annotation retrieved from the database sequence.

The database retrieval process terminates when a valid mapping is found for a sequence, or if all database queries failed to retrieve a matching sequence. Termination is indicated by the disappearance of the moving progress indicator on the alignment window. A dialog box may be shown once it completes which lists sequences for which records were found, but the sequence retrieved from the database did not exactly contain the sequence given in the alignment (the *"Sequence not 100% match" dialog box*).

[21] http://www.biodas.org/

Exercise 24: Retrieving Database References

24.a. Load the example alignment at http://www.jalview.org/tutorial/alignment.fa

24.b. Verify that there are no database references for the sequences by first checking that the *View* ⇒ *Sequence ID Tooltip* ⇒ *Show Database IDs* option is selected, and then mousing over each sequence's ID.

24.c. Use the *Webservices* ⇒ *Fetch DB References* menu option to retrieve database IDs for the sequences.

24.d. Examine the tooltips for each sequence in the alignment as the retrieval progresses - note the appearance of new database references.

24.e. Once the process has finished, save the alignment as a Jalview Project.
Now close all the windows and open the project again, and verify that the database references and sequence features are still present on the alignment

2.5.2 Retrieving Features *via* DAS

Jalview includes a client to retrieve features from DAS annotation servers. To retrieve features, select *View* ⇒ *Feature Settings...* from the alignment window menu. Select the *DAS Settings* tab in the Feature Settings Window (Figure 2.15). A list of DAS sources compiled from the currently configured DAS registry[22] is shown in the left hand pane. Highlighting an entry on the left brings up information about that source in the right hand panel.

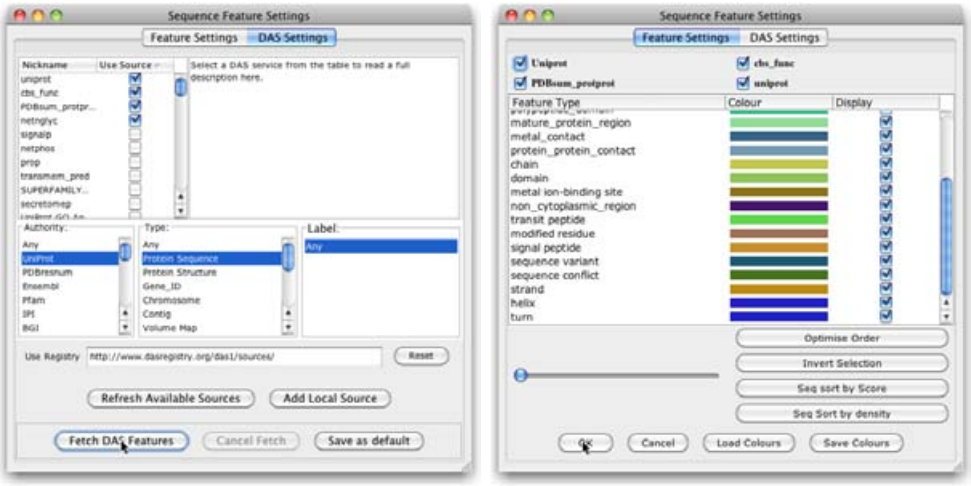


Figure 2.15: **Retrieving DAS annotations.** DAS features are retrieved using the *DAS Settings* tab (left) and their display customised using the *Feature Settings* tab (right).

Select appropriate DAS sources as required then click on *Fetch DAS Features*. If you know of additional sources not listed in the configured registry, then you may add them with the *Add Local Source* button. Use the *Authority*,*Type*, and *Label* filters to restrict the list of sources to just those that will return features for the sequences in the alignment.

Following DAS feature retrieval, the *Feature Settings* panel takes on a slightly different appearance

[22]By default, this will be the major public DAS server registry maintained by the Sanger Institute: http://www.dasregistry.org

(Figure 2.15 (right)). Each data source is listed and groups of features from one data source can be selected/deselected by checking the labeled box at the top of the panel.

The Fetch Uniprot IDs dialog box

If any sources are selected which refer to Uniprot coordinates as their reference system, then you may be asked if you wish to retrieve Uniprot IDs for your sequence. Pressing [OK] instructs Jalview to verify the sequences against Uniprot records retrieved using the sequence's ID string. This operates in much the same way as the *Web Services* ⇒ *Fetch Database References* function described in Section 2.5.1. If a sequence is verified, then the start/end numbering will be adjusted to match the Uniprot record to ensure that features retrieved from the DAS source are rendered at the correct position.

Rate of feature retrieval

Feature retrieval can take some time if a large number of sources is selected and if the alignment contains a large number of sequences. This is because Jalview only queries a particular DAS source with one sequence at a time, to avoid overloading it. As features are retrieved, they are immediately added to the current alignment view. The retrieved features are shown on the sequence and can be customised as described previously.

Exercise 25: Retrieving features with DAS

25.a. Load the alignment at http://www.jalview.org/tutorial/alignment.fa. Select *View* ⇒ *Sequence Features...* from the alignment window menu. Select the *DAS Settings* tab. A long list of available DAS sources is listed. Select a small number, eg Uniprot, DSSP, signalP and netoglyc. Click *OK*. A window may prompt whether you wish Jalview to map the sequence IDs onto Uniprot IDs. Click *Yes*. Jalview will start retrieving features. As features become available they will be mapped onto the alignment.

25.b. If Jalview is taking too long to retrieve features, the process can be cancelled with the *Cancel Fetch* button. Rolling the mouse cursor over the sequences reveals a large number of features annotated in the tool tip. Close the Feature Settings window.

25.c. Move the mouse over the sequence ID panel. Non-positional features such as literature references and protein localisation predictions are given in the tooltip, below any database cross references associated with the sequence.

25.d. Search through the alignment to find a feature with a link symbol next to it. Right click to bring up the alignment view popup menu, and find a corresponding entry in the *Link* sub menu.

25.e. Select *View* ⇒ *Feature Settings...* to reopen the Feature Settings window. All the loaded feature types should now be displayed. Those at the top of the list sit on top of and obscure those below. Move the feature settings window so that the alignment is visible and uncheck some of the feature types by clicking the tick box in the display column. Observe how the alignment display changes. Note that unselected feature types do not appear in the tool tip.

25.f. Reorder the features by dragging feature types up and down the order in the Feature Settings panel. e.g. Click on *CHAIN* then move the mouse downwards to drag it below *DOMAIN*. Note that *DOMAIN* is now shown on top of *CHAIN* in the alignment window. Drag *METAL* to the top of the list. Observe how the cysteine residues are now highlighted as they have a *METAL* feature associated with them.

25.g. Press the *Optimise Order* button. The features will be ordered according to increasing length, placing features that annotate shorter regions of sequence higher on the display stack.

25.h. Select *File* ⇒ *Export Features* from the Alignment window. You can choose to export the retrieved features as a GFF file, or Jalview's own Features format.

2.5.3 Colouring features by score or description text

Sometimes, you may need to visualize the differences in information carried by sequence features of the same type. This is most often the case when features of a particular type are the result of a specific type of database query or calculation. Here, they may also carry information within their textual description, or most commonly for calculations, a score related to the property being investigated. Jalview can shade sequence features using a graduated colourscheme in order to highlight these variations. In order to apply a graduated scheme to a feature type, select the 'Graduated colour' entry in the Sequence feature type's popup menu, which is opened by right-clicking the feature type's color in the settings dalog box. Two types of colouring styles are currently supported: the default is quantitative colouring, which shades each feature based on its score, with the highest scores receiving the 'Max' colour, and the lowest scoring features coloured with the 'Min' colour. Alternately, you can select the 'Colour by label' option to create feature colours according to the description text associated with each feature. This is useful for general feature

types - such as Uniprot's 'DOMAIN' feature - where the actual type of domain is given in the feature's description.

Graduated feature colour schemes can also be used to exclude low or high-scoring features from the alignment display. This is done by choosing your desired threshold type (either above or below), using the drop-down menu in the dialog box. Then, adjust the slider or enter a value in the text box to set the threshold for displaying this type of feature.

The feature settings dialog box allows you to toggle between a graduated and simple feature colourscheme using the pop-up menu for the feature type. When a graduated scheme is applied, it will be indicated by in the colour column for that feature type - with coloured blocks or text to indicate the colouring style and a greater than (>) or less than (<) symbol to indicate when a threshold has been defined.

2.5.4 Using features to re-order the alignment

The presence of sequence features on certain seqences or in a particular region of an alignment can quantitatively identify important trends in the aligned sequences. In this case, it is more useful to re-order the alignment based on the number of features or their associated scores, rather than simply re-colour the aligned sequences. The sequence feature settings dialog box provides buttons two buttons, 'Seq sort by Density' and 'Seq sort by Score' that allow you to reorder the alignment according to the number of sequence features present on each sequence, and also according to any scores associated with a feature. Each of these buttons uses the currently displayed features to determine the ordering, but if you wish to re-order the alignment using a single type of feature, then you can do this from the feature type's popup menu. Simply right-click the type's style in the Feature Settings dialog box, and select one of the *Sort by score* and *Sort by density* options to re-order the alignment. Finally, if a specific region is selected, then only features found in that region of the alignment will be used to create the new alignment ordering.

Exercise 26: Shading and sorting alignments using sequence features

26.a. Re-load the alignment from 25.

26.b. Open the feature settings panel, and, after first clearing the current selection, press the *Seq Sort by Density* button a few times.

26.c. Use the DAS fetcher to retrieve the Kyte and Doolittle Hydrophobicity scores for the protein sequences in the alignment. *Hint: the nickname for the das source is 'kd_hydrophobicity'.*

26.d. Change the feature settings so only the hydrophobicity features are displayed. Mouse over the annotation and also export and examine the GFF and Jalview features file to better understand how the hydrophobicity measurements are recorded.

26.e. Apply a graduated colourscheme to the hydrophobicity annotation to reveal the variation in average hydrophobicity across the alignment.

26.f. Select a range of alignment columns, and use one of the sort by feature buttons to order the alignment according to that region's average hydrophobicity.

26.g. Save the alignment as a project, for use in exercise 27.

Exercise 27: Shading alignments with combinations of graduated feature colourschemes

27.a. Reusing the annotated alignment from exercise 26, experiment with the colourscheme threshold to highlight the most, or least hydrophobic regions. Note how the colour scheme icon for the feature type changes when you change the threshold type.

27.b. Change the colourscheme so that features at the threshold are always coloured grey, and the most hydrophobic residues are coloured red, regardless of the threshold value (*hint - there is a switch on the dialog to do this for you*).

27.c. Enable the Uniprot *chain* annotation in the feature settings display and re-order the features so it is visible under the hydrophobicity annotation.

27.d. Apply a graduated colourscheme to the *chain* annotation so that it distinguishes the different canonical names associated with the mature polypeptide chains.

27.e. Export the alignment's sequence features using the Jalview sequence feature file format, to see how the different types of graduated feature colour styles are encoded.

2.6 Working with DNA

Jalview was originally developed for the analysis of protein sequences, but now includes some specific features for working with nucleic acid sequences and alignments. Nucleotide sequences and alignments are recognised based on the presence of nucleotide symbols [ACGT] in greater than 85% of the sequences. Built in codon-translation tables can be used to translate ORFs into peptides for further analysis. EMBL records retrieved *via* the sequence fetcher (see Section 1.4.5) are also parsed in order to identify codon regions and extract peptide products. Furthermore, Jalview records mappings between protein sequences that are derived from regions of a nucleotide sequence. Mappings are used to to transfer annotation between nucleic acid and protein sequences, and to dynamically highlight regions in one sequence that correspond to the position of the mouse pointer in another.

2.6.1 Alignment and Colouring

Jalview provides a simple colourscheme for DNA bases, but does not apply any specific conservation or substitution score model for the shading of nucleotide alignments. However, pairwise alignments performed using the *Alignment Window* ⇒ *Calculations* ⇒ *Pairwise Alignment* ... option will utilise an identity score matrix to calculate alignment score when aligning two nucleotide sequences.

Aligning Nucleic Acid Sequences

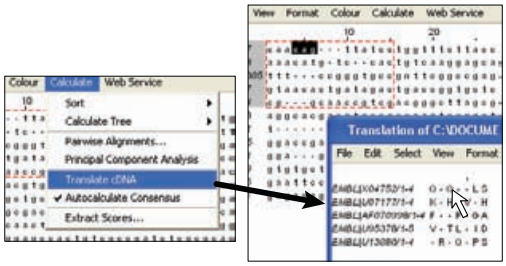
Jalview only has limited knowledge of the capabilities of the programs that are made available to it *via* web services. In particular, only the ClustalW and MAFFT programs will successfuly recognise and align nucleic acid sequences. MAFFT will also choose an appropriate parameter model. Whilst Muscle may appear to align DNA, it simply treats the base symbols as amino-acids, often leading to a poor quality alignment. Furthermore, it will almost certainly fail to align RNA containing Uracil bases, since 'U' is not a valid one-letter amino acid code.

2.6.2 Translate cDNA

The *Calculations* ⇒ *Translate cDNA* function in the alignment window is only available when working with a nucleic acid alignment. It uses the standard codon translation table given in the online help to translate a nucleotide alignment, or the currently selected region, into a set of aligned peptide sequences. Any features or annotation present on the nucleotide alignment will also be translated, allowing DNA alignment analysis results to be transferred on to peptide products for further investigation.

2.6.3 Linked DNA and Protein Views

Views of alignments involving DNA sequences are linked to views of alignments containing their peptide products in a similar way to views of protein sequences and views of their associated structures. Peptides translated from cDNA and extracted from EMBL records for DNA contigs are linked to their 'parent' coding regions. Mousing over a region of the peptide highlights codons in views showing the original coding region.



2.6.4 Coding regions from EMBL records

Many EMBL records that can be retrieved with the sequence fetcher contain exons. Coding regions will be marked as features on the EMBL nucleotide sequence, and Uniprot database cross references will be listed in the tooltip displayed when the mouse hovers over the sequence ID. Uniprot database cross references extracted from EMBL records are sequence cross references, and associate a Uniprot sequence's coordinate system with the coding regions annotated on the EMBL sequence. Jalview utilises cross-reference information in two ways.

Retrieval of Protein or DNA Cross References

The *Calculations* ⇒ *Get Cross References* function is only available when Jalview recognises that there are protein/DNA cross-references present on sequences in the alignment. When selected, it retrieves the cross references from the alignment's dataset (a set of sequence and annotation metadata shared between alignments) or using the sequence database fetcher. This function can be used for EMBL sequences containing coding regions to open the Uniprot protein products in a new alignment window. The new alignment window that is opened to show the protein products will also allow dynamic highlighting of codon positions in the EMBL record for each residue in the product(s).

Retrieval of protein DAS features on coding regions

The Uniprot cross-references derived from EMBL records can be used by Jalview to visualize protein sequence features directly on nucleotide alignments. This is because the database cross references include the sequence coordinate mapping information to correspond regions on the protein sequence with that of the nucleotide contig. Jalview will use the Uniprot accessions associated with the sequence to retrieve features, and then map them onto the nucleotide sequence's coordinate system using the coding region location.

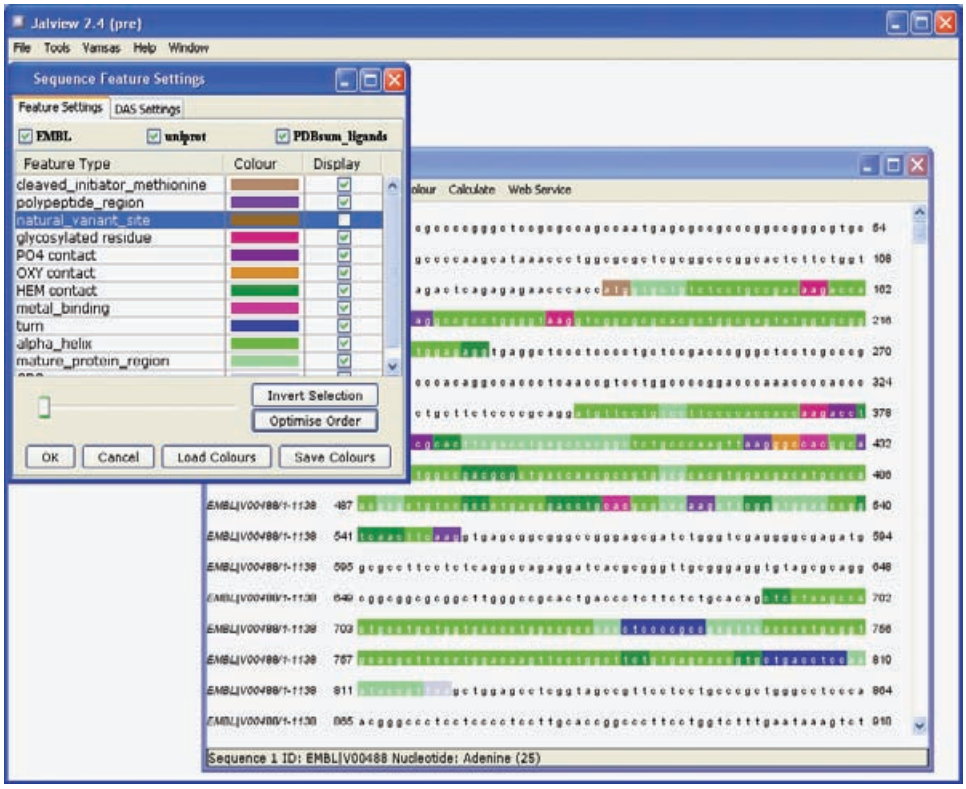


Figure 2.16: Uniprot and PDB sum features retrieved *via* DAS and mapped onto coding regions of EMBL record V00488 (an earlier version of Jalview is shown here).

Exercise 28: Visualizing protein features on coding regions

28.a. Use the sequence fetcher to retrieve EMBL record V00488.

28.b. Ensure that *View* ⇒ *Show Sequence Features* is checked and change the alignment view format to Wrapped mode so the distinct exons can be seen.

28.c. Open the DAS sequence feature fetcher window and fetch features for V00488 the Uniprot reference server, and any additional servers that work with the Uniprot coordinate system.

28.d. Mouse over the features retrieved, note that they have been mapped onto the coding regions, and in some cases broken into several parts to cover the distinct exons.

28.e. Open a new alignment view containing the Uniprot protein product with *Calculations* ⇒ *Get Cross References* ⇒ *Uniprot* and examine the database references and sequence features. Experiment with the interactive highlighting of codon position for each residue.