Jalview 2.8.2

Manual and Introductory Tutorial

David Martin, James Procter, Andrew Waterhouse, Saif Shehata, Nancy Giang, Suzanne Duce and Geoff Barton

College of Life Sciences, University of Dundee

Dundee, Scotland DD1 5EH, UK

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 platforms that support Java). Jalview is 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 standalone 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] and includes a javascript API to allow customisable display of alignments for web sites such as **Pfam**.[2]

Jalview 2.8.2 was released in December 2014. The Jalview Desktop in this version provides access to protein and nucleic acid sequence, alignment and structure databases, and includes the Jmol[3] viewer for molecular structures, and the VARNA[4] program for the visualization of RNA secondary structure. A Distributed Annotation System (DAS) client[5] which facilitates the retrieval and display of third party sequence annotation in association with sequences and any associated structure. It also provides a graphical user interface for the multiple sequence alignment, conservation analysis and protein disorder prediction methods provided as **Ja**va **B**ioinformatics **A**nalysis **W**eb **S**ervices

---

[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.xfam.org

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

[4]Provided under GPL licence at http://varna.lri.fr

[5]jDAS - released under Apache license (v2.0) at http://code.google.com/p/jdas

(JABAWS). JABAWS[6] is a system for running bioinformatics programs that you can download and run on your own machine or cluster, or install on compute clouds.

1.1.2   Jalview's Capabilities

Figure 1.1 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 into the program. Web services enable Jalview to access online 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[7] style figure generation capabilities for the preparation of alignments for publication.
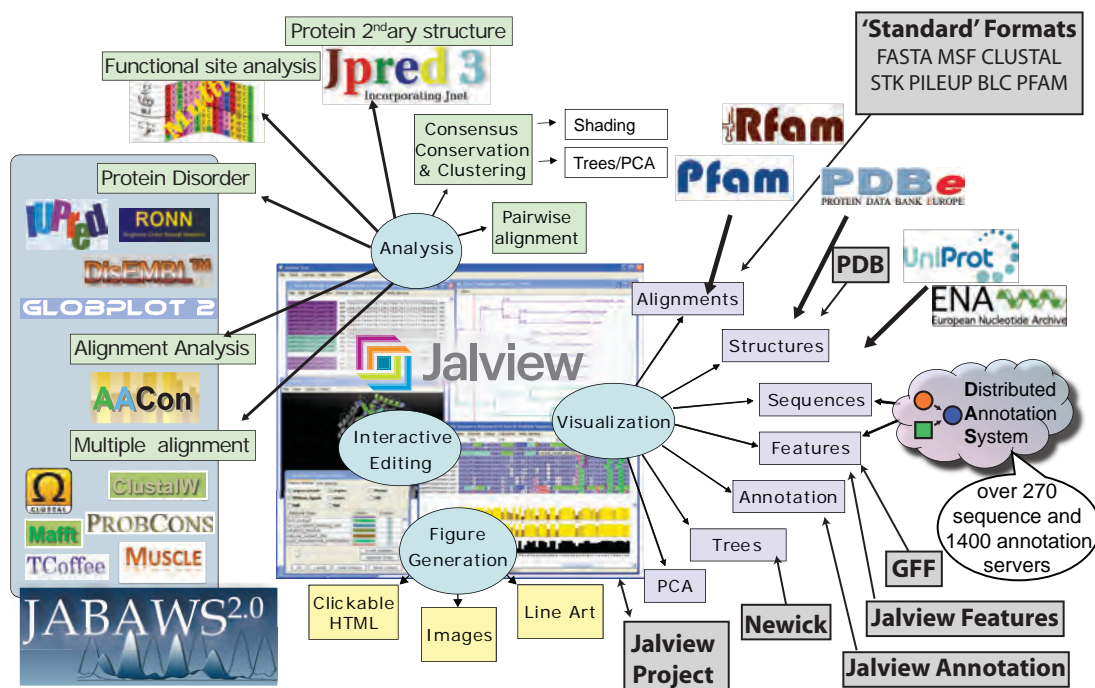

Figure 1.1: **Capabilities of the Jalview Desktop.** The Jalview Desktop Application provides a stable environment for the creation, editing and analysis of alignments and the generation of figures.

---

[6]released under GPL at `http://www.compbio.dundee.ac.uk/jabaws`
[7]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[8] 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 has been supported from 2009 by awards from the BBSRC's Tools and Resources fund, and, since 2014, a Wellcome Trust Biomedical Resource grant[9]. In 2010, 2011, and 2012, Jalview benefitted from the Google Summer of Code, when Lauren Lui and Jan Engelhardt introduced new features for handling RNA alignments and secondary structure annotation, in collaboration with Yann Ponty.[10]

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 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 Jmol molecular structure viewer, building and viewing trees and PCA plots, and using trees for sequence conservation analysis. An overview of the Jalview Desktop's webservices is given in Section 2.3, and the alignment and secondary structure prediction services are described in detail in Sections 2.4 and 2.6. Following this, Section 2.8 details the creation and visualization of sequence and alignment annotation, and the retrieval of sequences and annotation from databases and DAS Servers. Finally, Section 2.10 discusses specific features of use when working with nucleic acid sequences, such as translation and linking to protein coding regions, and the display and analysis of RNA secondary structure.

---

[8]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.

[9]Wellcome grant number 101651/Z/13/Z

[10]http://www.lix.polytechnique.fr/~ponty/

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 at www.jalview.org.**

This tutorial is based on the Jalview Desktop application. 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 webstart, or as an application loaded onto your hard drive. The webstart version is launched from the pink 'Launch Jalview Desktop button' at the top right hand side of pages of the website (www.jalview.org). To download the locally installable version, follow the links on the download page (www.jalview.org/download) (Figure 1.2). These links will launch the latest stable release of Jalview.

When the application is launched with webstart, two dialogs may appear before the application starts. If your browser is not set up to handle webstart, then clicking the launch link may download a file that needs to be opened manually, or prompt you to select the correct program to handle the webstart file. If that is the case, then you will need to locate the **javaws** program on your system[11]. Once java webstart has been launched, you may also be prompted to accept a security certificate signed by the Barton Group.[12] 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 at `http://www.jalview.org`.

Figure 1.3: **Jalview splash screen.**

When Jalview starts it will automatically load an example alignment from the Jalview site. This behaviour can be switched off in the Jalview Desktop preferences dialog by unchecking the open file option. This alignment will look like the one in Figure 1.4 (taken from Jalview version 2.7).

Jalview News RSS Feed

Announcements are made available to users of the Jalview Desktop *via* the Jalview Newsreader. This window will open automatically when new news is available, and can also be accessed *via* the Desktop's *Tools ⇒ Show Jalview News* menu entry.

---

[11]The file that is downloaded will have a type of **application/x-java-jnlp-file** or **.jnlp**. The **javaws** program that can run this file is usually found in the **bin** directory of your Java installation

[12]On some systems, the certificate may be signed by 'UNKNOWN'. In this case, clicking through the dialogs to look at the detailed information about the certificate should reveal it to be a Barton group certificate.
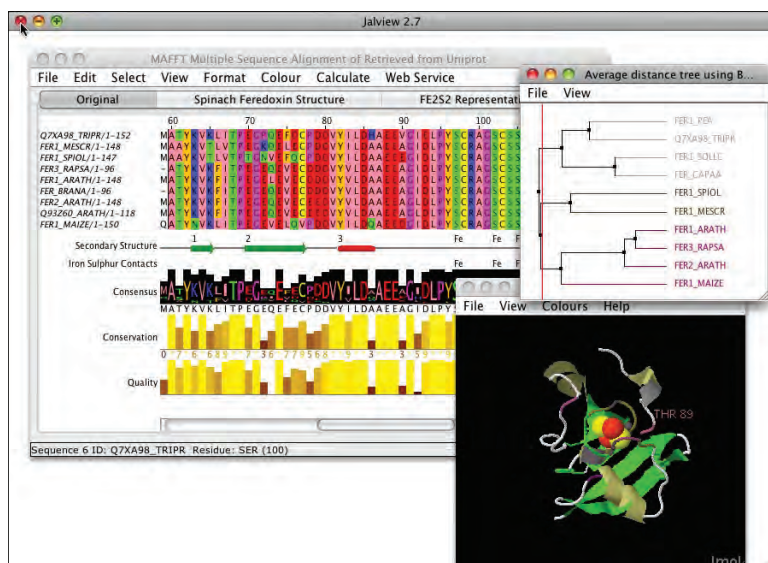
Figure 1.4: **Default startup for Jalview.**

Figure 1.5: **The Jalview News Reader.** The newsreader opens automatically when new articles are available from the Jalview Desktop's news channel.

Exercise 1: Launching Jalview from the Jalview website

1.a. Open the Jalview web site (www.jalview.org) in your web browser. Launch Jalview by clicking on the pink 'Launch Jalview' Desktop button in the top right hand corner. This will download and open a jalview.jnlp webstart file.

1.b. Dialogue boxes will open and ask if you want to open the jalview.jnlp file as the file is an application downloaded from the Internet, click Open. (Note you maybe asked to update Java, if you agree then it will automatically update the Java software). As Jalview opens, four demo Jalview windows automatically load.

1.c. If you are having trouble, it may help changing the browser you are using, as the browsers and it's version may affect this process.

1.d. To deactivate the opening of the 4 demo sequences during the launch, go to the *Tools ⇒ Preferences...* menu on the desktop. A 'Preference' dialogue box opens, untick the box adjacent to the 'Open file' entry in the 'Visual' preferences tab. Click OK to save the preferences.

1.e. Launch another Jalview workbench from the web site by clicking on the pink Launch button. The example alignment should not be loaded as Jalview starts up.

1.f. To reload the original demo file select the *File⇒ From URL* entry in the Desktop menu. Click on the URL history button on the right hand side of the dialog box to view the files, select exampleFile_2_7.jar, then click OK.

Note: Should you want to reload the example alignment or load your own sequence during the launch process, then go to the *Tools ⇒ Preferences...* menu on the desktop. The tick the 'Open file' entry of 'Visual' preferences tab, type in the URL of the sequence you want to load. This file will load during the start up process.

As the jalview.jnlp file launches Jalview on your desktop, you may want to move this from the downloads folder to another folder. Opening from this file will allow Jalview to be launched offline.

Help launching Jalview is available in a video on the Getting Started page of the website.

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 1.6). 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.

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.
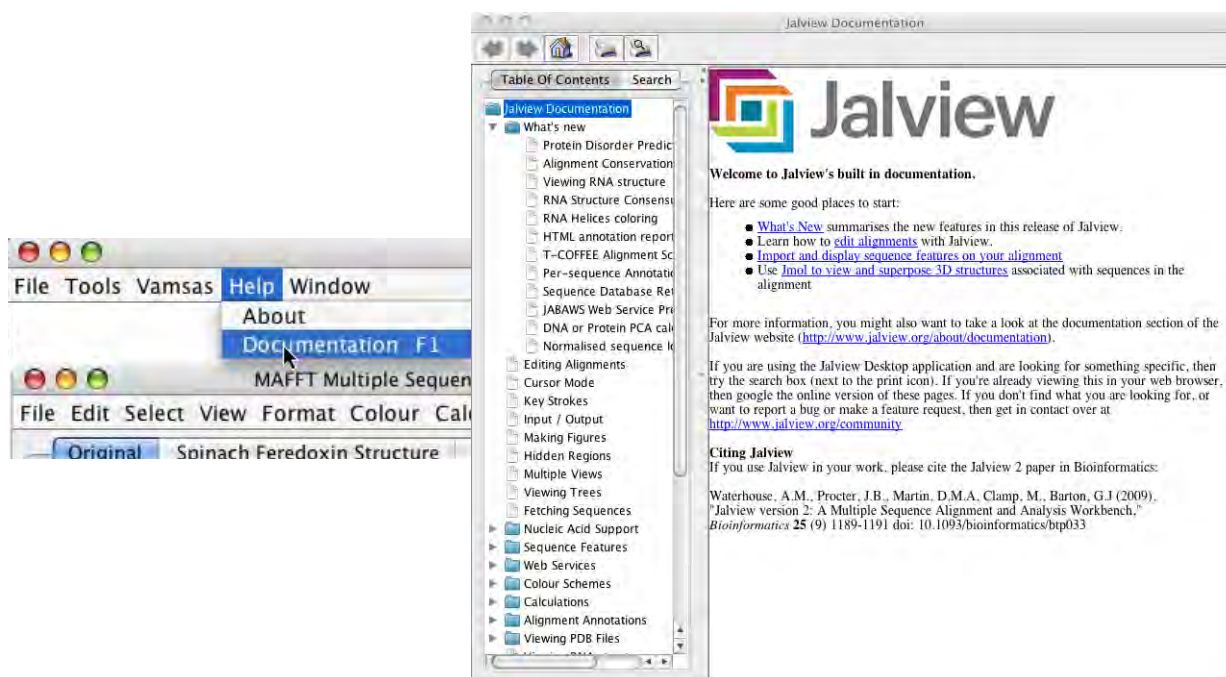
Figure 1.6: **Accessing the built in Jalview documentation.**

Archives and mailing list subscription details can be found in the Jalview web site's community section.

1.3   Navigation

The major features of the Jalview Desktop are illustrated in Figure 1.7. 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. With a Mac as the F2 is often assigned to screen brightness, one may often need to type **function [Fn] key with F2**.

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.
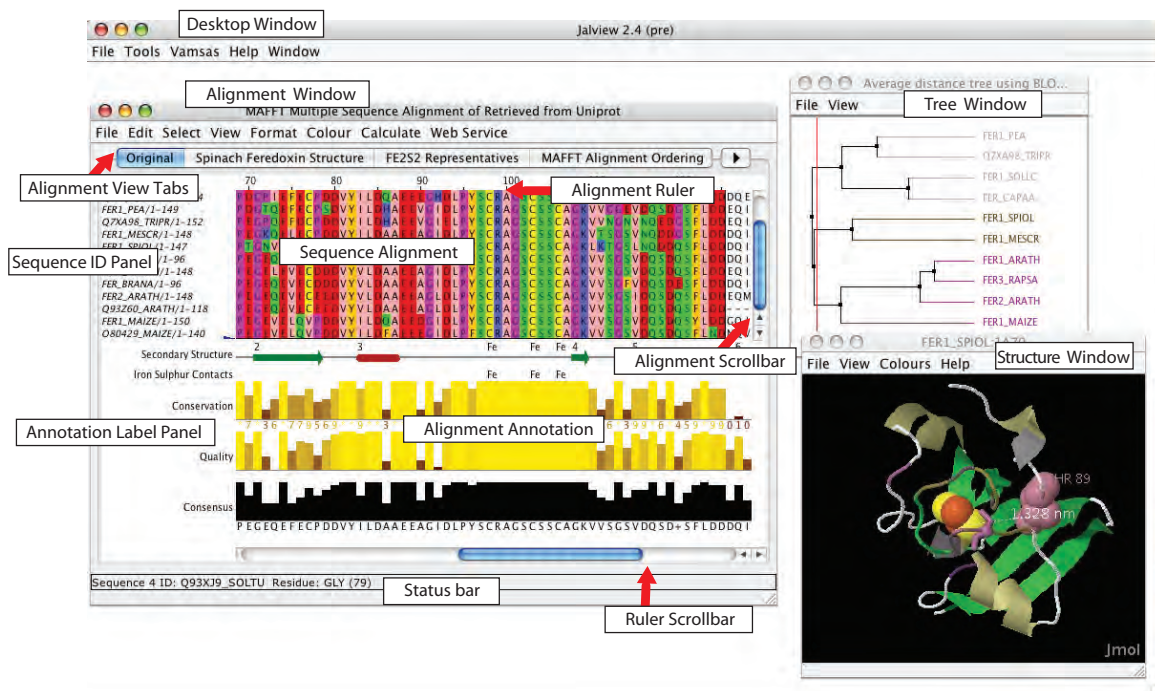
Figure 1.7: **The anatomy of Jalview.** The major features of the Jalview Desktop Application are labeled.

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 Alignment window menu bar (Figure 1.8[13]).

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 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!

---

[13]the menu shown in this figure is from Jalview 2.2, later versions have more options.
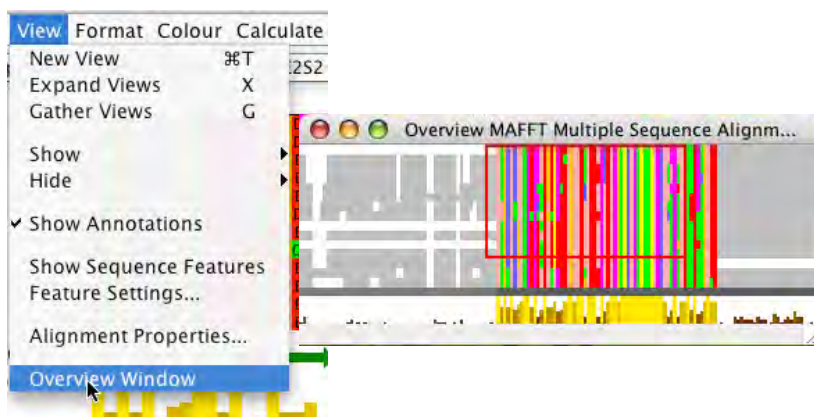
Figure 1.8: **Alignment Overview Window.** The overview window for a view is opened from
the *View* menu.

1.3.2   Navigation in Cursor Mode

Cursor mode navigation enables the 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 ($\uparrow$, $\downarrow$, $\leftarrow$, $\rightarrow$).
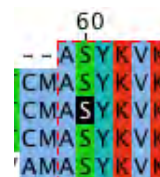
Rapid movement to specific positions is accomplished as listed below:

- **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

Jalview has two navigation and editing modes: **normal** mode (where editing and navigation are via the mouse) and the **cursor** mode (where editing and navigation are via the keyboard). The **F2 key** is used to switch between these two modes. With a Mac often need to type function **Fn key and F2**, as button is often assigned to screen brightness. Jalview always starts up in **normal mode**.

2.a. Load an example alignment from its URL (`http://www.jalview.org/examples/exampleFile_2_7.jar`) via the Desktop using *File ⇒ Input Alignment ⇒ From URL* dialog box. (The URL should be stored in its history and clicking on the *down arrow* is an easy way to access it.)

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

2.c. Find the Overview Window, *Views ⇒ Overview Window* and open it. Move around the alignment by clicking and dragging the red box in the overview window.

2.d. Look at the status bar (lower left hand corner of the alignment window) as you move the mouse over the alignment. It indicates information about the sequence and residue under the cursor.

2.e. Press [F2] to enter **Cursor mode**. Use the direction keys to move the cursor around the alignment.

2.f. 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]**.

Note: To view Jalview's comprehensive on-line help documentations select Help on desktop menu, clicking on Documentation will open a Documentation window. Select topic from the navigation panel on the left hand side or use the Search tab to select specific key words **Help navigating Jalview is available in a video on the Getting Started page of the website.**

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) occurrence 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 most operating systems 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. Drag and
drop also works when loading data from a URL - simply drag the link or url from the address panel
of your browser on to an alignment or the Jalview desktop background and Jalview will load data
from the URL directly.

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 1.9).
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.
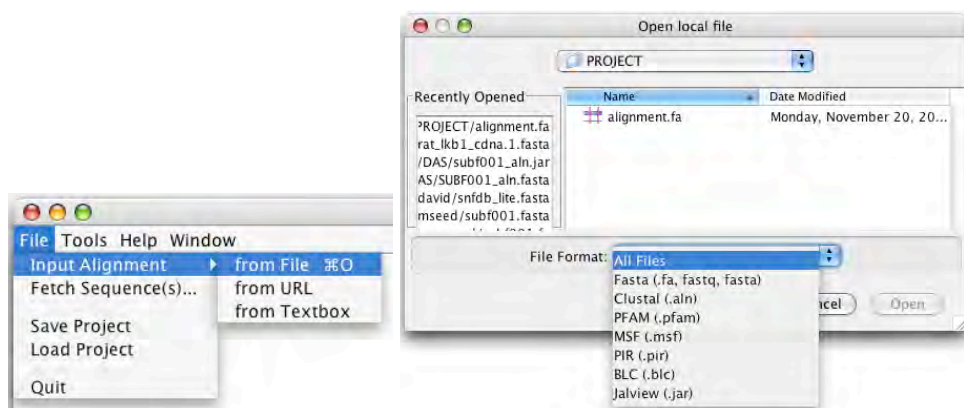
Figure 1.9: **Opening an alignment from a file saved on disk.**

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 - an 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 (Figure 1.10). Jalview will attempt to automatically discover the file format.

Figure 1.10: **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 1.11). In both cases, presuming that they are in the right format, Jalview will happily read them into a new alignment window.
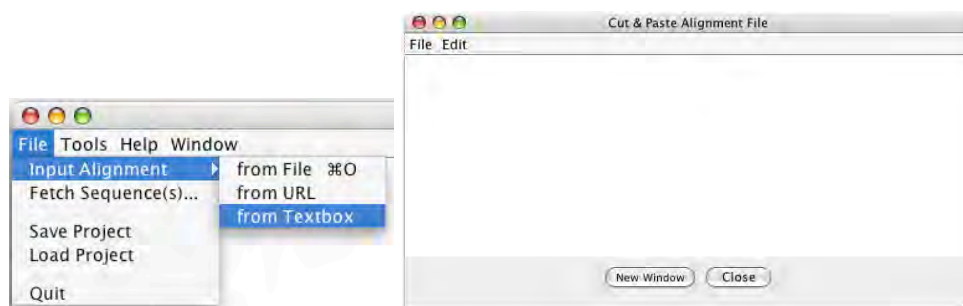
Figure 1.11: **Opening an alignment from pasted text.**

1.4.5 From a Public Database

Jalview can retrieve sequences and sequence alignments from the public databases housed at the European Bioinformatics Institute, including Uniprot, Pfam, Rfam and the PDB, as well as any DAS sequence server registered at the configured DAS registry. Jalview's sequence fetching capabilities allow you to avoid having to manually locate and save sequences from a web page before loading them into Jalview. It also 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 1.12). Pressing the database selection button in the dialog box opens a new window showing all the database sources Jalview can access (grouped by the type of database). Once you've selected the appropriate database, hit OK close the database selection window, and then enter one or several database IDs or accession numbers separated by a semicolon and press OK. Jalview will then attempt to retrieve them from the chosen database. Example queries are provided for some databases 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.[14]

---

[14]Most DAS sources support *range queries* that can be used to download just a particular range from a sequence database record.
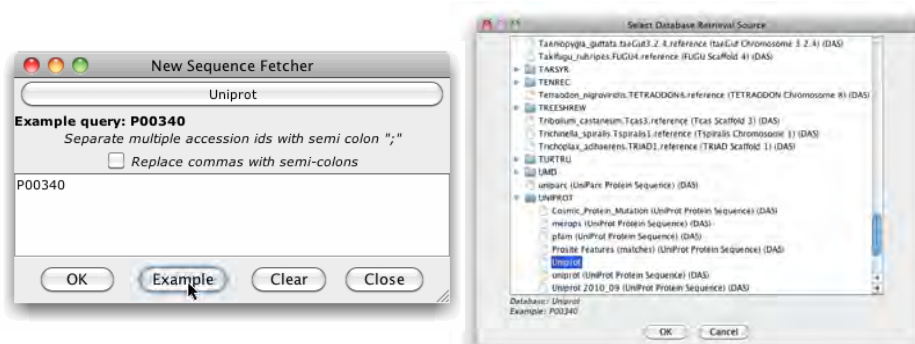
Figure 1.12: **Retrieving sequences from a public database.**

Exercise 3: Loading sequences

3.a. Use *Window* ⇒ *Close All* from the Desktop window menu to close all windows.

3.b. **Loading sequences from URL:** Selecting File ⇒ *Input Alignment* ⇒ *From URL* from the Desktop and enter http://www.jalview.org/tutorial/alignment.fa in the box. Click OK to load the alignment.

3.c. **Loading sequences from a file:** Close all windows using the *Window* ⇒ *Close All* menu option from the Desktop. Then type the same URL (http://www.jalview. org/tutorial/alignment.fa) into your web browser and **save** the file to your desktop.

Open the file you have just saved in Jalview by selecting *File* ⇒ *Input Alignment* ⇒ *From File* from the desktop menu and selecting this file. Click OK to load the alignment.

3.d. **Loading sequences by 'Drag and Drop' / 'Cut and Paste':**
(i) Select *Desktop* ⇒ *Window* ⇒ *Close All*. Then drag the alignment.fa file from the desktop and drop it onto the Jalview window, the alignment should open.
Test the differences between (a) dragging the sequence onto the empty Jalview desktop and (b) dragging the sequence onto an existing alignment window.
(ii) Open http://www.jalview.org/tutorial/alignment.fa in a web browser. Drag the URL directly from browser window onto Jalview.
(If the URL is downloaded instead of opened in the browser, then locate the file in your download directory and open it in a text editor.)
(iii) Open the alignment.fa file using text editor. Copy the sequence text from the file into the clipboard and paste it into the desktop background by right-clicking and selecting Paste to new alignment option.
(iv) In the text editor, copy the sequence text from alignment.fa into the clipboard (usually *via* the browser's *Edit* ⇒ *Copy* menu option). In the Desktop menu, select *File* ⇒ *Input Alignment* ⇒ *From Textbox*. Paste the clipboard into the large window using the *Edit* ⇒ *Paste* text box menu option. Click *New Window* and the alignment will be loaded.

3.e. **Loading sequences from Public Database:** Select *File* ⇒ *Fetch Sequence(s)...* from the Desktop to open up new window called New Sequence Fetcher. Press database selection button at the top of the dialog box, this opens another window called Select Database Retrieval Source showing all the database sources.
Select the **PFAM seed** database and click ok, then enter the accession number **PF03460** and click OK. An alignment of about 107 sequences should load.
Several database IDs or accession numbers can be loaded by using semicolons to separate them.

1.4.6 Memory Limits

Jalview is a Java program. One unfortunate implication of this is that Jalview cannot 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 the built in documentation (opened by selecting *Help* ⇒ *Documentation*) and on the JVM memory parameters page (`http://www.jalview.org/jvmmemoryparams.html`).

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 1.13). 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 save the file during or after editing by using the *File* ⇒ *Save* entry.

Jalview offers several different formats in which an alignment can be saved. The jalview format (.jar) is the only format which will preserve the colours, groupings and other additional 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 program 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.

1.5.2 Jalview Projects

If you wish to save a complete Jalview session rather than just a single alignment (e.g. because you have calculated trees or multiple different alignments) then save your work as a Jalview Project file (.jvp).[15] 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.

---

[15]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.
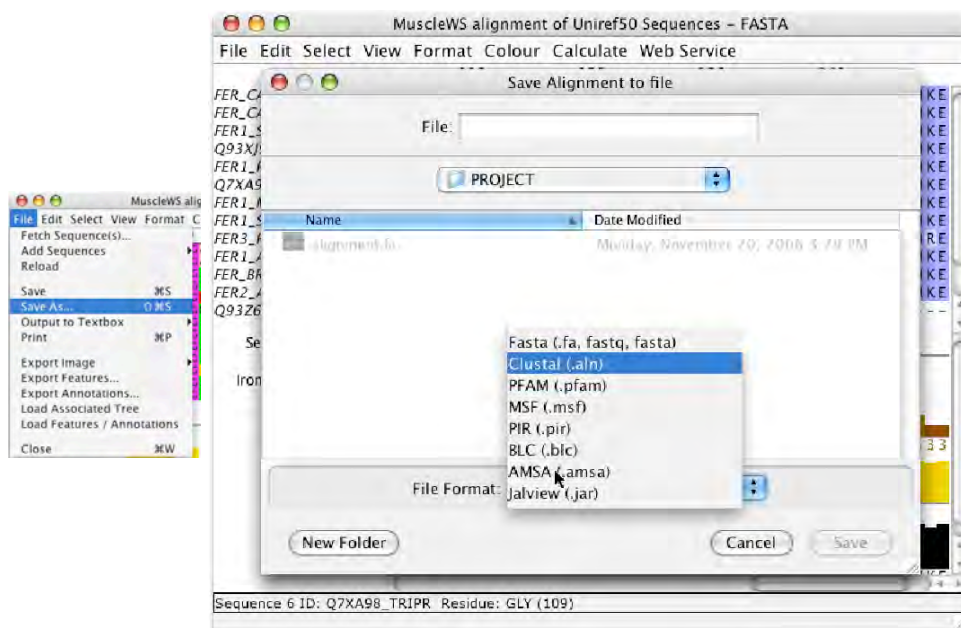
Figure 1.13: **Saving alignments in Jalview to disk.**

Exercise 4: Saving Alignments

4.a. Launch Jalview, or use close all windows. Load the ferredoxin alignment from PFAM (seed) data base using the PFAM seed 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 your preferred format. All formats except *Jalview* jvp 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 closing all windows and opening it with Jalview, or by browsing to it with your web browser.

4.c. Repeat the previous step saving the files in different file formats.

4.d. Select *File ⇒ Output to Textbox ⇒ FASTA*. 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 active in Jalview. Open the overview window *View ⇒ Overview Window* and scroll red box to any part of the alignment. Select *File ⇒ Save Project* from the main menu and save the project in a suitable folder.

Close all windows and then load the project *via* the *File ⇒ Load Project* menu option. Observe 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 ⇒ To New Alignment* in the 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 and hold the mouse button and drag the mouse to the bottom right corner of the chosen region then release the mouse button. A dashed red box appears around the selected region (Figure 1.14).
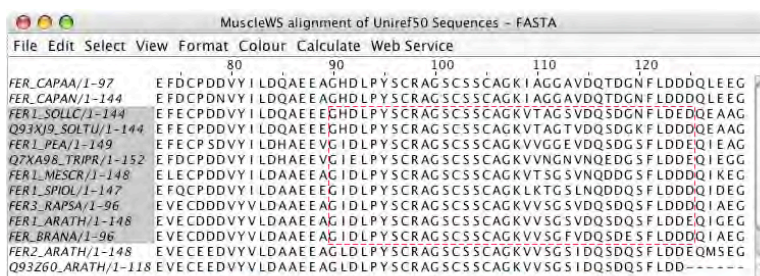
Figure 1.14: **Selecting a region in an alignment.**

Selecting Columns

To select the same residues in all sequences, click and drag along the alignment ruler. This selects the entire column of the alignment. Ranges of positions from the alignment ruler can also be selected by clicking on the first position and 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 1.15).

Selecting Sequences

To select multiple complete sequences, click and drag the mouse down the sequence ID panel. The same techniques as used for columns (above) can be used with [SHIFT]-Click for continuous and [CTRL]-Click to select discontinuous ranges of sequences (Figure 1.16).
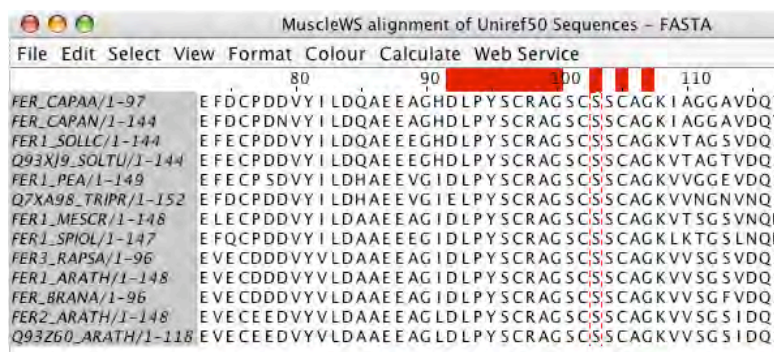
Figure 1.15: **Selecting multiple columns in an alignment.** The red highlighting on the align-
ment ruler marks the selected columns. Note that only the most recently selected
column has a dashed-box around it to indicate a region selection.
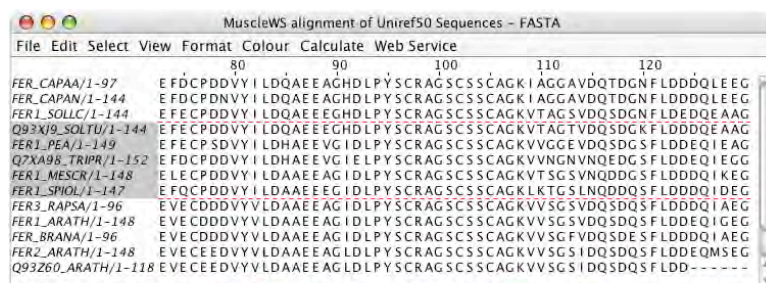
Figure 1.16: **Selecting multiple sequences in an alignment.** Use [CTRL] or [SHIFT] to se-
lect many sequences at once.

Making Selections in Cursor Mode

To define a selection in cursor mode (which is enabled by pressing [F2] when the alignment window
is selected), navigate to the top left corner of the proposed selection (using the mouse, the arrow
keys, or the keystroke commands described in Section 1.3.2). Pressing the [Q] key marks this as the
corner. A red outline appears around the cursor (Figure 1.17)

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.
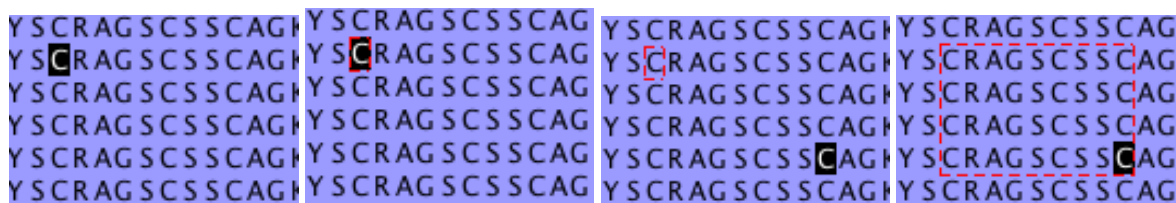
Figure 1.17: **Making a selection in cursor mode.** Navigate to the top left corner (left), press
[Q] (left center), navigate to the bottom right corner (right center) and press [M]
(right).
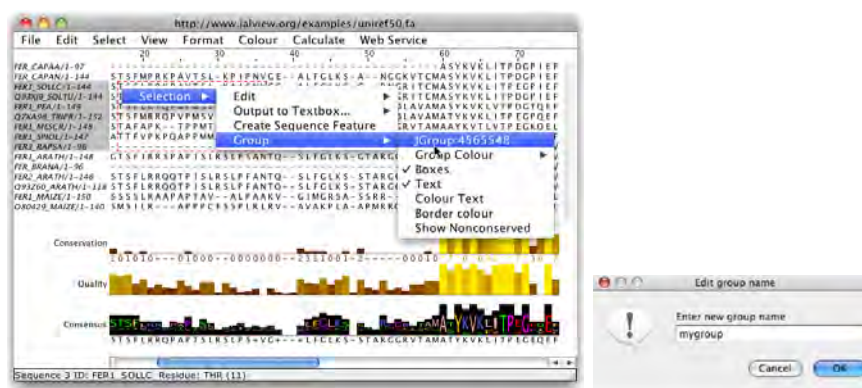
Figure 1.18: **Creating a new group from a selection.**

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 useful when selecting large regions in an alignment, simply select the region that is to be kept unselected, and then invert the selection. This may also be useful when hiding large regions in an alignment (see Section 1.6.5 below). Instead of selecting the columns and rows that are to be hidden, simply select the region that is to be kept visible, invert the selection, then select *View ⇒ Hide ⇒ Selected Region* .

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 ⇒ Edit name and description of current group*[16] 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 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 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.

---

[16]In earlier versions of Jalview, this entry was variously 'Group', 'Edit Group Name', or 'JGroupXXXXX' (Where XXXXX was some serial number).

Exercise 5: Making Selections and Groups

5.a. Close all windows in Jalview. Load the ferredoxin alignment (PFAM ID PF03460 from PFAM seed database). Choose a residue and place the mouse cursor on it. Note residue information will show in alignment window status bar. 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. Press [ESC] to clear the selection.

5.b. Select one sequence by clicking on the sequence ID panel. Note that the sequence ID takes on a highlighted background and a red box appears around the selected sequence. 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.
Hold down [CTRL] and then 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. Select column by clicking on the Alignment Ruler. Note whole column is highlighted with a red box. Hold down [SHIFT] and click column beyond. Note the selection expands to include all the sequences between the two positions on which you clicked.

5.d. To selecting arbitrary regions in alignment, place the mouse at the top left corner of the region you wish to select. Press and hold the mouse button. Drag the mouse to the bottom right corner of the chosen region and release the mouse button and a dashed red box appears around the selected region

5.e. Enter Cursor mode using [F2] (or [Fn]-F2 for Macs). Navigate to column 59, row 1 by pressing **5 9 , 1 [RETURN]**. Press **Q** to mark this position. Navigate to column 65, row 8 by pressing **6 5 , 8 [RETURN]**. Press **M** to complete the selection. Note to clear the selection press the **[ESC]** key.

5.f. To create a **group** from the selected the region, click the right mouse button when mouse is on the selection, this opens a pop-up menu in the alignment window.
Open the *Selection ⇒ Edit New Group ⇒ Group Colour* menu and select 'Percentage Identity'. This will turn the selected region into a group and colour it accordingly.

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

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

5.i. The current selection can be **exported** and saved by right clicking on the text area to open the Sequence ID pop-up menu. Follow the menus and pick an output format (eg BLC) from the *Selection ⇒ Output to Textbox . . .* submenu.

5.j. In the alignment output window, try manually editing the alignment, importing group into a new alignment window by clicking 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 be moved then press the up or down arrow keys as appropriate (Figure 1.19). 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.

Figure 1.19: **Reordering the alignment.** The selected sequence moves up one position on pressing the ↑ key.

Exercise 6: Reordering the Alignment

6.a. Close all windows in Jalview from desktop. Load the ferredoxin alignment (e.g.the PFAM domain PF03460 from PFAM seed). Select one of the sequence in the sequence ID panel, use the up and down arrow keys to alter the sequence's position in the alignment. (Note that this will not work in cursor mode)

6.b. To select and move multiple sequences, use hold [SHIFT] and [CTRL], and select two sequences separated by one or more un-selected sequences. Note how multiple sequences are grouped together when they are re-ordered using the up and down arrow keys.

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 the several different views in the example alignment 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 pop-up menu. Select *Hide Sequences* and the sequences will be concealed, with a small blue triangle indicating their position (Figure 1.20). To unhide (reveal) the sequences, right click on the triangle and select *Reveal Sequences* from the context menu.
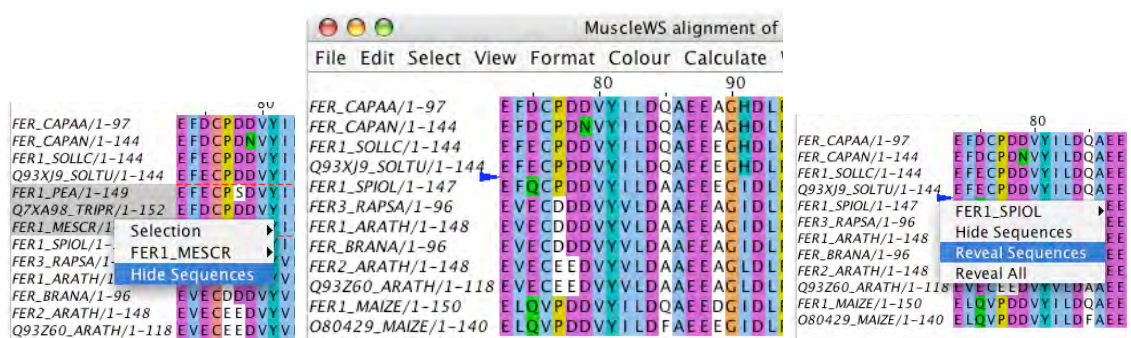
Figure 1.20: **Hiding Sequences** Hidden sequences are represented by a small blue triangle in the sequence ID panel.

A similar mechanism applies to columns (Figure 1.21). Selected columns (indicated by a red marker) can be hidden and revealed in the same way *via* the context pop-up menu by right clicking on the ruler bar. The hidden column selection is indicated by a small blue triangle in the ruler bar.
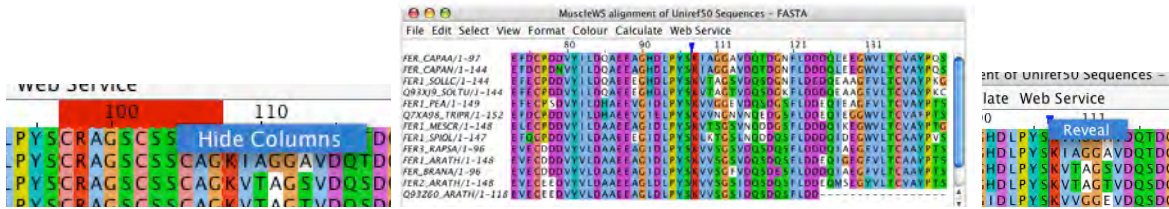
Figure 1.21: **Hiding Columns** Hidden columns are represented by a small blue triangle in the ruler bar.

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, select the required region and use the *View ⇒ Hide ⇒ All but Selected Region* menu entry, or press [Shift]+[Ctrl]+H to hide the unselected region.

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, 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 to bring up the sequence ID pop-up menu, select *Hide Sequences*.

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

7.c. Repeat the process but use 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 using columns to hide and reveal columns instead of sequences.

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

7.f. Select some sequences and pick one to represent the rest by hovering the mouse over this sequence. Bring up the Sequence ID pop-up menu by right clicking and then select *(Sequence ID name) ⇒ Represent group with (Sequence ID name)*. To reveal these hidden sequences, right click on the Sequence ID and in the pop-up menu select Reveal All.
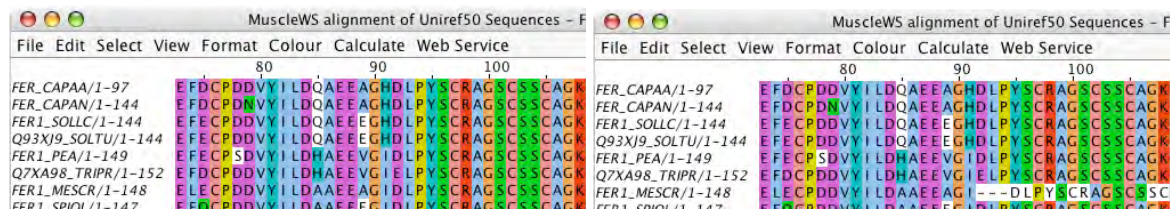
Figure 1.22: **Introducing gaps in a single sequence.** Gaps are introduced as the selected sequence is dragged to the right while pressing and holding [SHIFT].
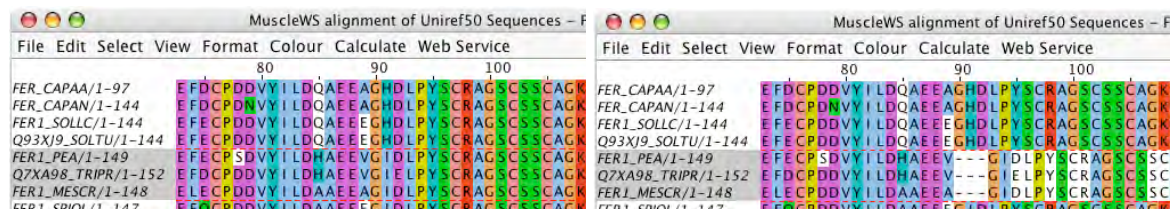
Figure 1.23: **Introducing gaps in a group.** Gaps are introduced as the selected group is dragged to the right with [CTRL] pressed.

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 shifting 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 until 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). Slides occur regardless of the region selection - which, for example, allows you to easily reposition misaligned subfamilies within a larger alignment.

Undoing Edits

Jalview supports the undoing of edits *via* the *Edit ⇒ Undo Edit* alignment window menu option, or CTRL-Z. An edit, if undone, may be re-applied with *Edit ⇒ Redo Edit*, or CTRL-Y. Note, however, that the *Undo* function only works for edits to the alignment or sequence ordering. Colouring of the alignment, showing and hiding of sequences or modification of annotation cannot be undone.

Exercise 8: Editing Alignments

You are going to manually reconstruct part of the example Jalview alignment available at http://www.jalview.org/examples/exampleFile.jar.

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.

If you are using OSX, and a key combination - such as [CTRL]+A - does not work, then try pressing the [CMD] key instead of [CTRL].

8.a. Load the URL http://www.jalview.org/tutorial/unaligned.fa which contains part of the ferredoxin alignment from PF03460.

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 pop-up 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: you can do this by pressing [CTRL]-I to invert the sequence selection and then deselect 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 (those not hidden) in the block by pressing [CTRL]-A. Insert a single gap in all selected sequences at column 38 of the alignment by holding [CTRL] and clicking on the R at column 38 in the FER1_SPIOL, then drag one column to right. Insert another gap at column 47 in all sequences in the same way.

8.f. Correct the ferredoxin domain alignment for FER1_SPIOL by inserting two additional gaps after the gap at column 47: First press [ESC] to clear the selection, then 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 in column 47 of FER1_SPIOL 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 (25C in cursor mode). 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. Press [ESC] to clear the selection, and then 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 the gap in O80429_MAIZE at column 56 using [SHIFT]-drag to the left on 56C.

8.j. Use the *Edit* ⇒ *Undo Edit* and *Edit* ⇒ *Redo Edit* menu option, or their keyboard shortcuts ([CTRL]+Z and [CTRL]+Y) to step backwards and replay the edits you have made.

Editing in Cursor mode

Gaps can be easily inserted when in cursor mode (toggled with [F2]) by pressing [SPACE]. Gaps will be inserted at the cursor, shifting 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]. First make sure you have everything unselected by pressing ESC. 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, press [CTRL]-[BACKSPACE] or [SHIFT]-[BACKSPACE] (both keys held down together). Note that the deletion will only occur if the gaps are in the same columns in all sequences in the selected group, and those columns are to the right of the selected residue.

Exercise 9: Keyboard Edits

This continues on from exercise 8, and recreates the final part of the example ferredoxin alignment from the unaligned sequences using Jalview's keyboard editing mode.

Note for Windows Users: The [SHIFT]-[SPACE] command has the same effect as the [CTRL]-[SPACE] command mentioned in this exercise, and you should use [SHIFT]-[SPACE] in order to avoid opening the window menu.

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 press *[RETURN]*. Now insert 6 gaps in all the sequences. Type *6* then hold down *[CTRL]* and press *[SPACE]*.

9.e. Now insert one gap at column 34 and another at 38. Insert 3 gaps at 47. Press *34C* then *[CTRL]-[SPACE]*. Press *38C* then [CTRL]-[SPACE]. Press *47C* 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. Press *[ESC]* to clear the selection. 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 *31C [BACKSPACE]* .

9.g. Similarly delete the gap now at column 34, then insert two gaps at column 38. Press *34C [BACKSPACE] 38C 2 [SPACE]*. Delete three gaps at column 44 and insert one at column 47 by pressing *44C 3 [BACKSPACE] 47C [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 2.8). 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. **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.
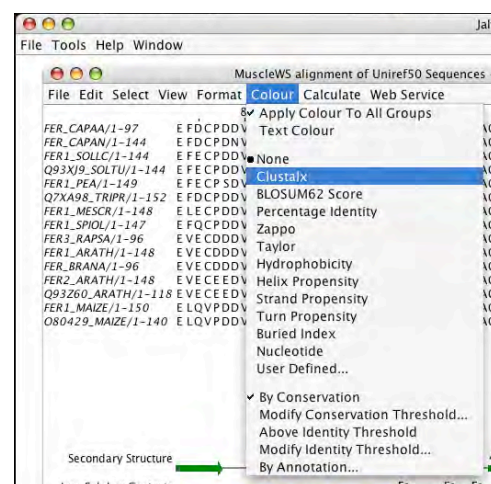
1.7.2 Colouring a Group or Selection

Selections or groups can be coloured in two ways. The first is *via* the Alignment Window's *Colour* menu as stated above, after first ensuring that the *Apply Colour To All Groups* flag is not selected. This must be turned *off* specifically as it is *on* by default. When unticked, selections from the Colours menu will only change the colour for residues in the current selection, or the alignment view's "background colourscheme" when no selection exists.

The second method is to select sequences and right click mouse to open pop-up menu and select *Selection ⇒ Edit New Group ⇒ Group Colour* from context menu options (Figure 1.24). This only changes the colour of the current selection or group.
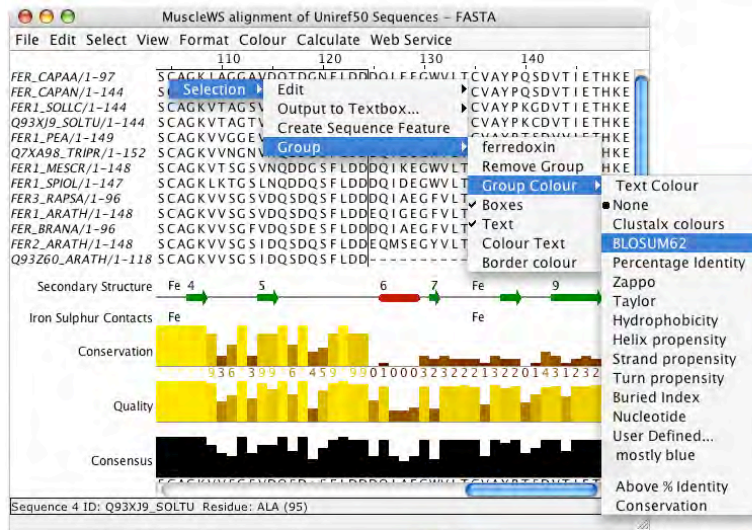
Figure 1.24: **Colouring a group via the context menu.**

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* enables this mode, and *Modify Conservation Threshold...* brings up a selection box (the *Conservation Colour Increment* dialog box) allowing the alignment colouring to be modified. Selecting a higher value limits colouring to more highly conserved columns (Figure 1.25).

Figure 1.25: **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.

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 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.[17]

The *Colour ⇒ By Annotation* option opens a di-
alog which allows you to select which annota-
tion to use, the minimum and maximum shading
colours or whether the original colouring should
be thresholded (the 'Use original colours' option).
Default settings for minimum and maximum
colours can be set in the Jalview Desktop's pref-
erences.
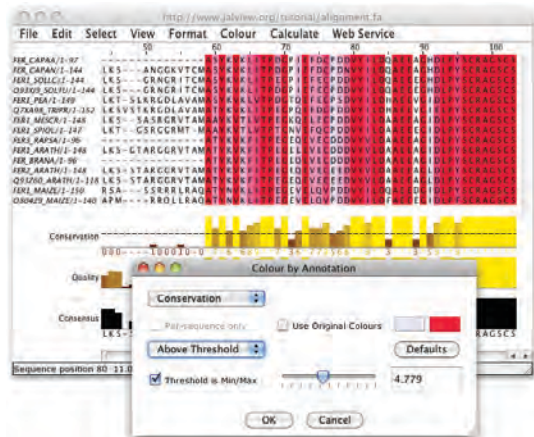
The **per Sequence** option in the **Colour By Annotation** dialog allows each sequence to be shaded
according to sequence associated annotation rows, such as protein disorder scores. This functionality
is described further in Section 2.7.

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 ClustalX, 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.
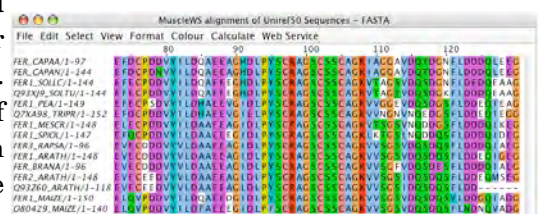
Blosum62

Gaps are coloured white. If a residue matches the con-
sensus sequence residue at that position it is coloured
dark blue. If it does not match the consensus residue
but the Blosum62 matrix gives a positive score, it is
coloured light blue.

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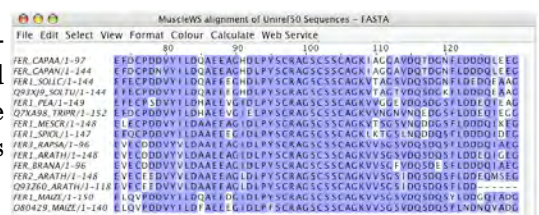
[17]Please remember to turn off Sequence Feature display to see the shading

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.
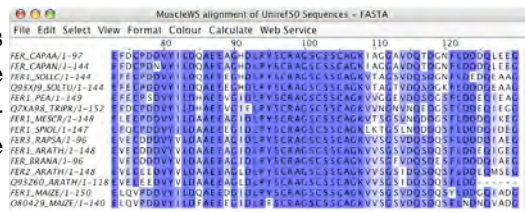
Zappo

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.
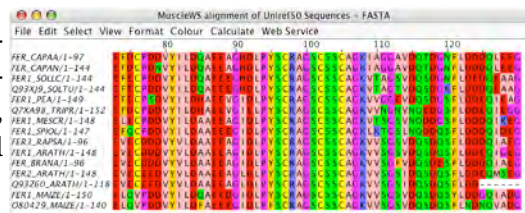
Taylor

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)
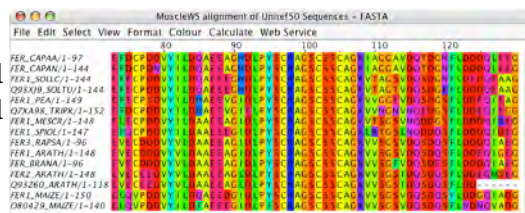
Hydrophobicity

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.
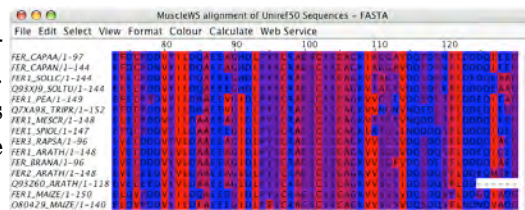
Helix Propensity

The residues are coloured according to their Chou-Fasman[18] helix propensity. The highest propensity is magenta, the lowest is green.

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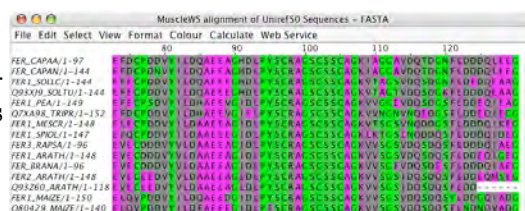
[18]Chou, PY and Fasman, GD. Annu Rev Biochem. 1978;47:251-76.

Strand Propensity

The residues are coloured according to their Chou-Fasman[18] Strand propensity. The highest propensity is Yellow, the lowest is blue.
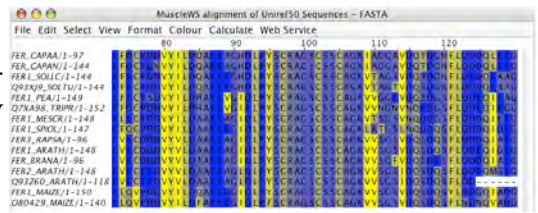
Turn Propensity

The residues are coloured according to their Chou-Fasman[18] turn propensity. The highest propensity is red, the lowest is cyan.

Buried Index

The residues are coloured according to their Chou-Fasman[18] burial propensity. The highest propensity is blue, the lowest is green.
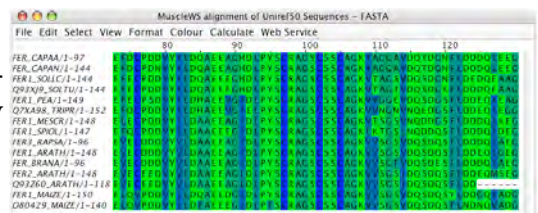
Nucleotide

Residues are coloured with four colours corresponding to the four nucleotide bases. All non ACTG residues are uncoloured. See Section 2.10 for further information about working with nucleic acid sequences and alignments.
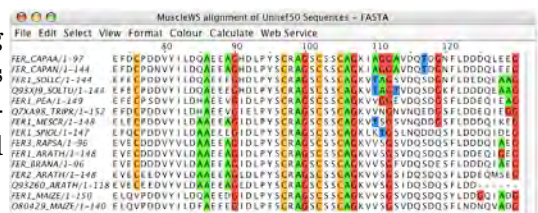
Purine Pyrimidine

Residues are coloured according to whether the corresponding nucleotide bases are purine (magenta) or pyrimidine (cyan) based. All non ACTG residues are uncoloured. For further information about working with nucleic acid sequences and alignments, see Section 2.10 .

RNA Helix colouring

Columns are coloured according to their assigned RNA helix as defined by a secondary structure annotation line on the alignment. Colours for each helix are randomly assigned, and option only available when an RNA secondary structure row is present on the alignment.
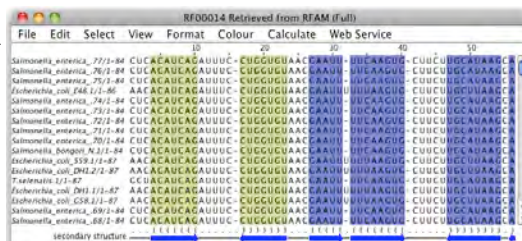
Exercise 10: Colouring Alignments

10.a. Ensuring that the *Apply Colour To All Groups* flag is not selected in View sequence alignment menu. This must be turned *off* specifically as it is *on* by default.

10.b. Open a sequence alignment, for example the PFAM domain PF03460 in PFAM seed. 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.c. 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 exercise 5 during the group selection step).

10.d. 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 1.26).

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 for the colourscheme 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.
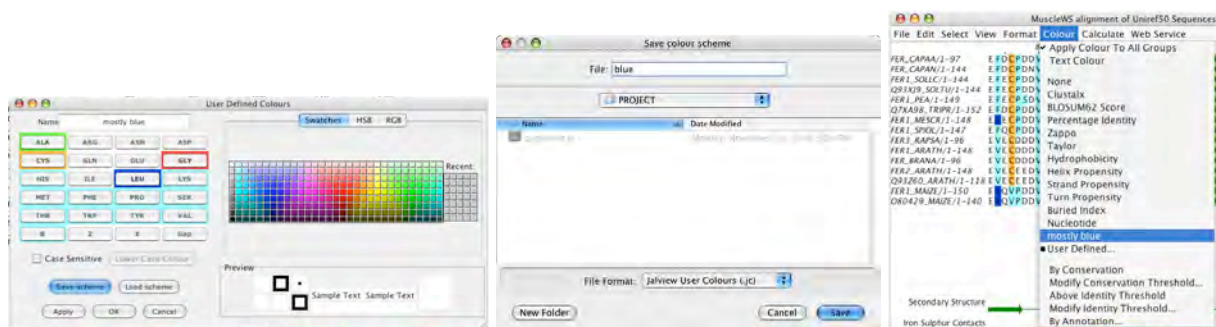
Figure 1.26: **Creation of a user defined colour scheme.** Residue types are assigned colours (left). The profile is saved (center) and can then be accessed *via* the *Colour* menu (right).

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 outputted 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 or by Pressing [CTRL]-T. This will create a new view with the same groups, alignment layout and display options as the current one. Pressing G will gather together Views as named tabs on the alignment window, and pressing X will expand gathered Views so they can be viewed simultaneously in their own separate windows. To delete a group, press [CTRL]-W.
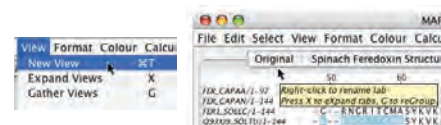
1.8.2 Alignment Layout

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 1.27). 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.

If annotations are not all visible in wrapped mode, expand the alignment window to view them. Note that alignment annotation (see Section 2.8) cannot be interactively created or edited in wrapped mode, and selection of large regions is difficult.
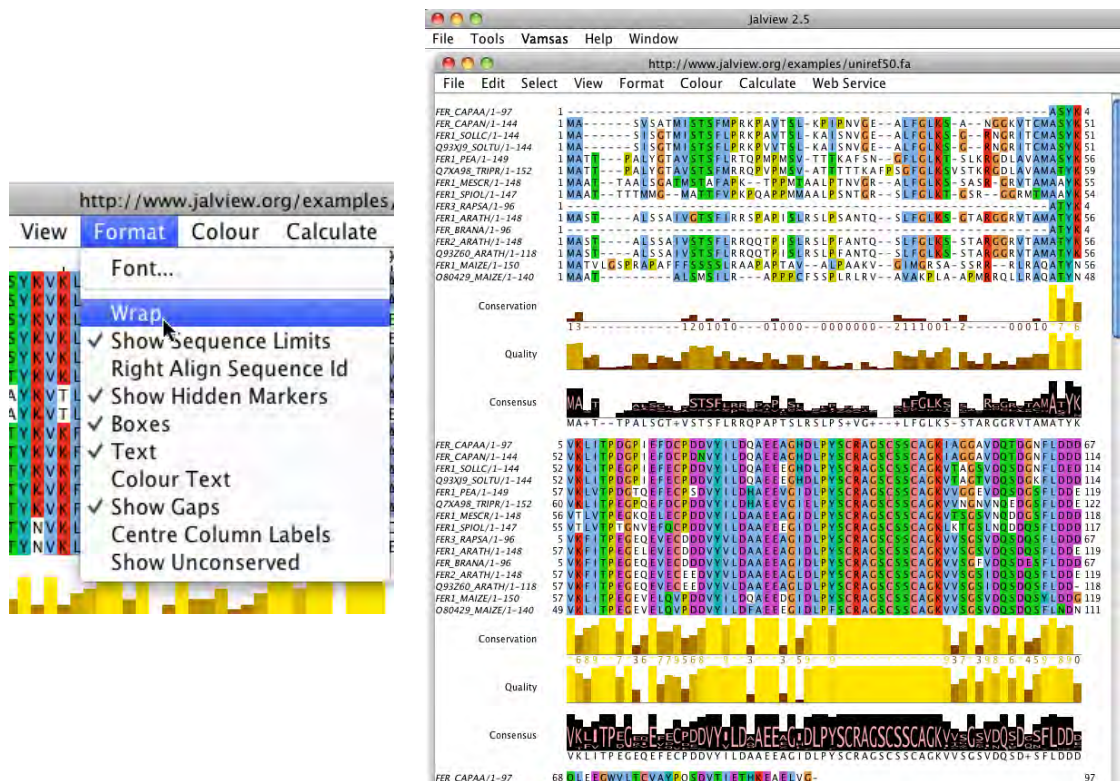
Figure 1.27: **Wrapping the alignment.**

Fonts

The text appearance in a view can be modified *via* the *Format ⇒ Font...* alignment window menu. This setting applies for all alignment 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.
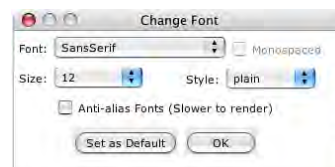
Numbering and Label Justification

Options in the *Format* menu are provided to control the alignment view, and provide 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 *Selection ⇒ Group ⇒ Show Nonconserved* by right clicking on a group). 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 '.'.

1.8.3 Annotation Ordering and Display

The annotation lines which appear below the sequence alignment are described in detail in Section 2.8. 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 pop-up context menu on the annotation name panel (Figure 1.28). 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 on the left. When this is displayed, Click-dragging up and down provides more space in the alignment window for viewing the annotations, and less space for the sequence alignment.
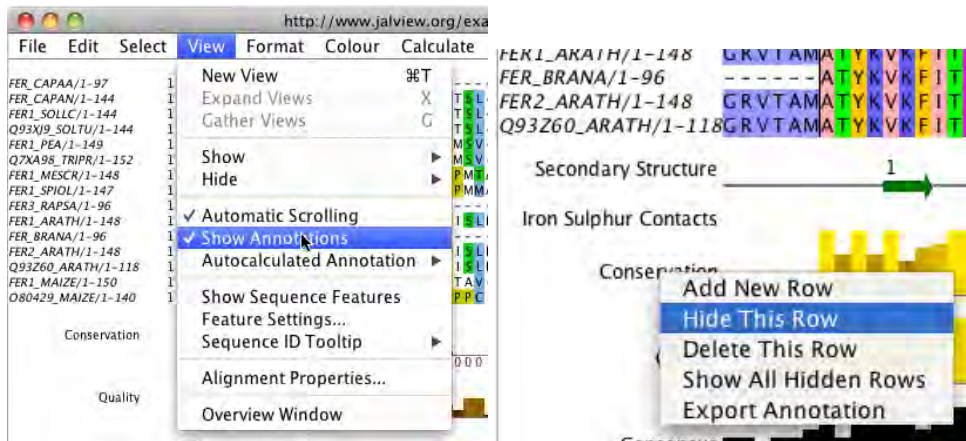
Figure 1.28: **Hiding Annotations** Annotations can either be hidden from the *View* menu (left)
or individually from the context menu (right).

Exercise 12: Alignment Layout

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

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

12.c. Deselect *Format ⇒ Wrap*. Right click on the annotation row labels to bring up the
pop-up context menu, then select *Hide This Row*. Bring up the pop-up 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 2.8 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 dif-
ferent formats, each of which is suited to a particular purpose.
Image export is *via* the *File ⇒ Export Image ⇒ . . .* alignment
window menu option.
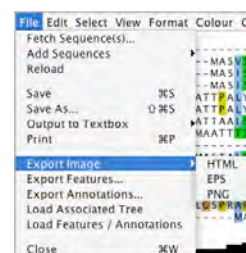
HTML

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 publications and posters** as it gives the highest quality output of any of the image types. It can be scaled to any size, 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 or Inkscape.

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[19].

Zoom Detail of PNG image.

---

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

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 favoured web browser.

13.b. Wrap the alignment 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 such as Paint or Photoshop (Windows), or Preview (Mac OS X) and zoom in. Notice that the image is a bitmap and it becomes pixelated very quickly. (Note 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 Photoshop, Illustrator, Inkscape, Ghostview, Powerpoint (Windows), or Preview (Mac OS X). Zoom in and note that the image has near-infinite resolution.

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 descriptions and exercises on building and displaying trees, PCA analysis, alignment redundancy removal, pairwise alignments and alignment conservation analysis. Section 2.3 introduces the various web based services available to Jalview users, and Section 2.3.3 explains how to configure the Jalview Desktop for access to new JABAWS servers. Section 2.4 describes how to use the range of multiple alignment programs provided by JABAWS, and Section 2.5 introduces JABAWS' AACon service for protein multiple alignment conservation analysis. Section 2.6 explains how to perform protein secondary structure predictions with JPred, and JABAWS' protein disorder prediction services are introduced in Section 2.7.

Section 2.8 describes the mechanisms provided by Jalview for interactive creation of sequence and alignment annotation, and how they can be displayed, imported and exported and used to reorder the alignment. Section 2.9 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. Section 2.10 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 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 European Protein Databank (PDBe) using

---

[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.

the Sequence Fetcher (see 1.4.5).

2.1.1   Automatic association of PDB structures with sequences

Jalview can automatically determine which structures are associated with a sequence in a number
of ways.

Discovery of PDB IDs from sequence database cross-references

If a 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 *Structure* ⇒ *Associate Structure with Sequence*
⇒ *Discover PDB IDs* from the context menu (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.
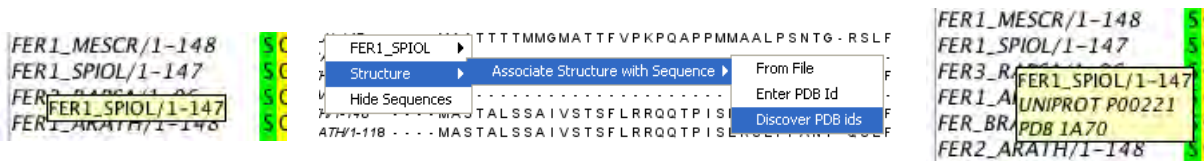
Figure 2.1: **Automatic PDB ID discovery.** The tooltip (left) indicates that no PDB structure
has been associated with the sequence. After PDB ID discovery (center) the tool tip
now indicates the Uniprot ID and any associated PDB structures (right)

Drag-and-drop association of PDB files with sequences by filename match

If one or more PDB files stored on your computer are dragged from their location on the file browser
onto an alignment window, Jalview will search the alignment for sequences with IDs that match any
of the files dropped onto the alignment. If it discovers matches, a dialog like the one in Figure 2.2 is
shown, giving the option of creating associations for the matches.

If no associations are made, then sequences extracted from the structure will be simply added to the
alignment. However, if only some of the PDB files are associated, jalview will raise another dialog
box giving you the option to add any remaining sequences from the PDB structure files not present
in the alignment. This allows you to easily decorate sequences in a newly imported alignment with
any corresponding structures you've already collected in a directory accessible from your computer.[4]

---

[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.

[4]We plan to extend this facility in future so Jalview will automatically search for PDB files matching your sequence
within a local directory. Check out Jalview issue 801
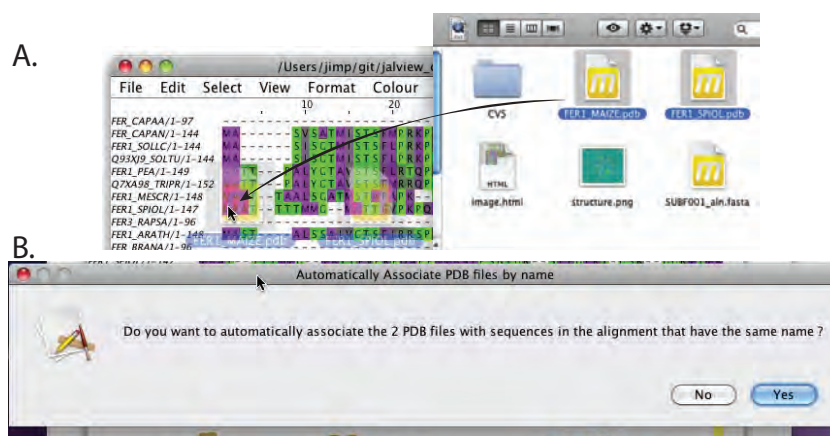
Figure 2.2: **Associating PDB files with sequences by drag-and-drop.** Dragging PDB files onto an alignment of sequences with names matching the dragged files names (A), results in a dialog box (B) that gives the option to associate each file with any sequences with matching IDs.

2.1.2   Viewing Structures

The structure viewer can be launched in two ways from the sequence ID context menu. To view a particular structure associated with a sequence in the alignment, simply select it from popup menu's associated structures submenu in *Structure ⇒ View Structure ⇒ <PDB ID>*. The second way is most useful if you want to view all structural data available for a set of sequences in an alignment. If any of the **currently selected** sequences have structures associated, the *Structure* submenu of the sequence ID popup menu will include an option to *View* **N** *structures*. Selecting this option will open a new structure view containing the associated structures superposed according to the alignment.

In both cases, each structure to be displayed 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 2.3 (right)). The structure can be rotated by clicking and dragging in the structure window. The structure can be zoomed using the mouse scroll wheel or by [SHIFT]-dragging the structure. Moving the mouse cursor over a sequence to which the structure is linked in the alignment view 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. Clicking the alpha carbon or phosphorous backbone atom will toggle the highlight and residue label on and off. Often, the position highlighted in the sequence may not be in the visible portion of the current alignment view and the sliders will scroll automatically to show the position. If the alignment window's *View ⇒ Automatic Scrolling* option is not selected, however, then the automatic adjustment will be disabled for the current view.

2.1.3   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.
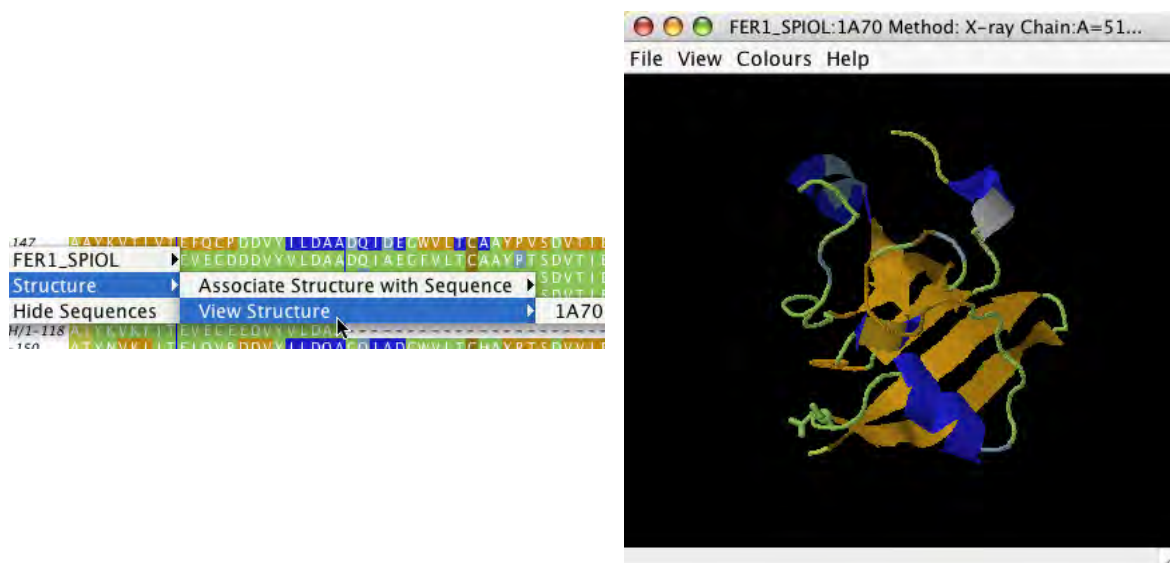
Figure 2.3: **Structure visualization** The structure viewer is launched from the sequence ID
            context menu (left) and allows the structure to be visualized using the embedded
            Jmol molecular viewer (right).

By default, the structure will be coloured in the same way as the associated sequence(s) in the align-
ment view from which it was launched. The structure can be coloured independently of the sequence
by selecting an 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
saved as an EPS or PNG with the *File ⇒ Save As ⇒ ...* submenu, which also allows the raw data to
be saved as PDB format. 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[5]-like" scripting language, which is described
elsewhere[6]. 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.

If you would prefer to use Jmol to manage structure colours, then select the *Colours ⇒ Colour with
Jmol* option. This will disable any automatic application of colour schemes when new structure data

---

[5]See http://www.rasmol.org

[6]Jmol Wiki: http://wiki.jmol.org/index.php/Scripting

Jmol Scripting reference: http://www.stolaf.edu/academics/chemapps/jmol/docs/

is added, or when associated alignment views are modified.

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* ⇒ *Structure* ⇒ *Associate Structure with Sequence* ⇒ *Discover PDB IDs*. Jalview will now attempt to find PDB structures for the sequences in the alignment.
Note: If you are using Jalview v2.8 - use the *Uniprot* source from the *Web services* ⇒ *Fetch DB References* ⇒ *..* submenu of the Alignment Window to retrieve the PDB IDs.

14.b. Right-click on the sequence ID for *FER1_SPIOL*. Select *FER1_SPIOL* ⇒ *Structure* ⇒ *View Structure* ⇒ *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 a suitable program.

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 (n)* ⇒ *All* command (where *n* is the number of residues selected), 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.

2.1.4 Superimposing structures

Many comparative biomolecular analysis investigations aim to determine if the biochemical properties of a given molecule are significantly different to its homologues. When structure data is available, comparing the shapes of molecules by superimposing them enables substructure that may

impart different behaviour to be quickly identified. The identification of optimal 3D superpositions involves aligning 3D data rather than sequence symbols, but the result can still be represented as a sequence alignment, where columns indicate positions in each molecule that should be superposed to recreate the optimal 3D alignment.

Jalview can employ Jmol's 3D fitting routines[7] to recreate 3D structure superpositions based on the correspondences defined by one or more sequence alignments involving structures shown in the Jmol display. Superposition based on the currently displayed alignment view happens automatically if a structure is added to an existing Jmol display using the *Structure ⇒ View PDB Structure ⇒ ...* A new Jmol view containing superposed structures can also be created using the *Structure ⇒ View all* **N** *PDB Structures* option (when *N* > 1) if the current selection contains two or more sequences with associated structures.

Obtaining the RMSD for a superposition

The RMSD (Root Mean Square Deviation) is a measure of how similar the structures are when they are superimposed. Figure 2.4 shows a superposition created during the course of Exercise 15. The parts of each molecule used to construct the superposition are rendered using the cartoon style, with other parts of the molecule drawn in wireframe. The Jmol console, which has been opened after the superposition was performed, shows the RMSD report for the superposition. Full information about the superposition is also output to the Jalview console.[8] This output also includes the precise atom pairs used to superpose structures.

Choosing which part of the alignment is used for structural superposition

Jalview uses the visible part of each alignment view to define which parts of each molecule are to be superimposed. Hiding a column in a view used for superposition will remove that correspondence from the set, and will exclude it from the superposition and RMSD calculation. This allows the selection of specific parts of the alignment to be used for superposition. Only columns that define a complete set of correspondences for all structures will be used for structural superposition, and as a consequence, the RMSD values generated for each pair of structures superimposed can be directly compared.

In order to recompute a superposition after changing a view or editing the alignment, select the *Jmol ⇒ Align sequences* menu option. The *Jmol ⇒ Superpose with ..* submenu allows you to choose which of the associated alignments and views are to be used to create the set of correspondences. This menu is useful when composing complex superpositions involving multi-domain and multi-chain complexes, when correspondences may be defined by more than one alignment.

Note that these menu options appear when you have two or more structures in one Jmol viewer.

---

[7]See http://chemapps.stolaf.edu/jmol/docs/?ver=12.2#compare for more information.
[8]The Jalview Java Console is opened from *Tools ⇒ Java Console* option in the Desktop's menu bar
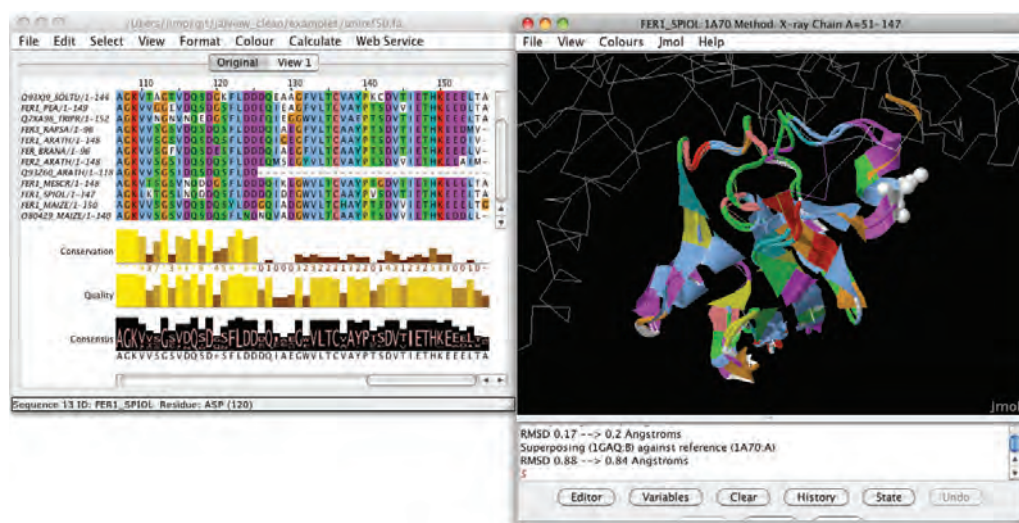
Figure 2.4: **Superposition of two ferredoxin structures.** The alignment on the left was used by jalview to superpose structures associated with the FER1_SPIOL and FER1_MAIZE sequences in the alignment. Parts of each structure used for superposition are rendered as a cartoon, the remainder rendered in wireframe. The RMSD between corresponding positions in the structures before and after the superposition is shown in the Jmol console.

Exercise 15: Aligning structures using the ferredoxin sequence alignment.

15.a. Continue with the Jalview project created in exercise 14. Use the *Discover PDB IDs* function to retrieve PDB IDs associated with the FER1_MAIZE sequence.

15.b. Once discovery has completed, use the *View PDB Structure* submenu to view the PDB file associated with FER1_MAIZE. Jalview will give you the option of aligning the structure to the one already open. To superimpose the structure associated with FER1_MAIZE with the one associated with FER1_SPIOL, press the **Yes** button.
The Jmol view will update to show both structures, and one will be moved on to the other. If this doesn't happen, use the Align function in the Jmol submenu

15.c. Create a new view on the alignment, and hide all but columns 121 through to 132.

15.d. Use the *Jmol* submenu to recompute the superposition using just columns 121-132 of the alignment.
Note how the molecules shift position when superposed using a short part of the two structures.

15.e. Compare the initial and final RMSDs for superimposing molecules with the small section and with the whole alignment. Which view do you think give the best 3D superposition, and why ?

2.1.5 Colouring structure data associated with multiple alignments and views

Normally, the original view from which a particular structure view was opened will be the one used to colour structure data. If alignments involving sequences associated with structure data shown in a Jmol have multiple views, Jalview gives you full control over which alignment, or alignment view,

is used to colour the structure display. Sequence-structure colouring associations are changed *via* the *View ⇒ Colour by ..* menu, which lists all views associated with data shown in the embedded Jmol view. A tick is shown beside views currently used as colouring source, and moving the mouse over each view will bring it to the front of the alignment display, allowing you to browse available colour sources prior to selecting one. If the *Select many views* option is selected, then multiple views can be selected as sources for colouring the structure data. *Invert selection* and *Select all views* options are also provided to quickly change between multi-view selections.

Note that the *Select many views* option is useful if you have different views that colour different areas or domains of the alignment. This option is further explored in Section 2.1.5.
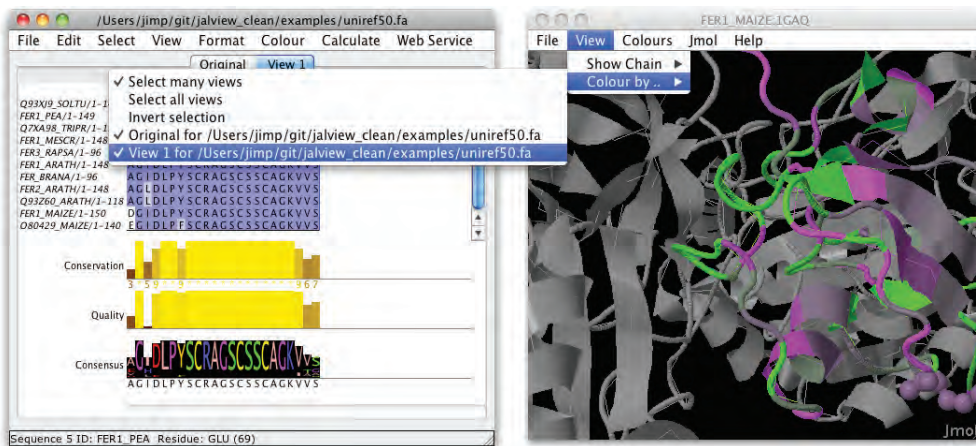
Figure 2.5: **Choosing a different view for colouring a structure display** Browsing the *View ⇒ Colour by ..* menu provides full control of which alignment view is used to colour structures when the *Colours ⇒ By Sequence* option is selected.

Colouring complexes

The ability to control which multiple alignment view is used to colour structural data is essential when working with data relating to multidomain biomolecules and complexes.

In these situations, each chain identified in the structure may have a different evolutionary history, and a complete picture of functional variation can only be gained by integrating data from different alignments on the same structure view. An example of this is shown in Figure 2.6, based on data from Song et. al[9]

---

[9]Structure of DNMT1-DNA Complex Reveals a Role for Autoinhibition in Maintenance DNA Methylation. Jikui Song, Olga Rechkoblit, Timothy H. Bestor, and Dinshaw J. Patel. *Science* 2011 **331** 1036-1040 DOI:10.1126/science.1195380
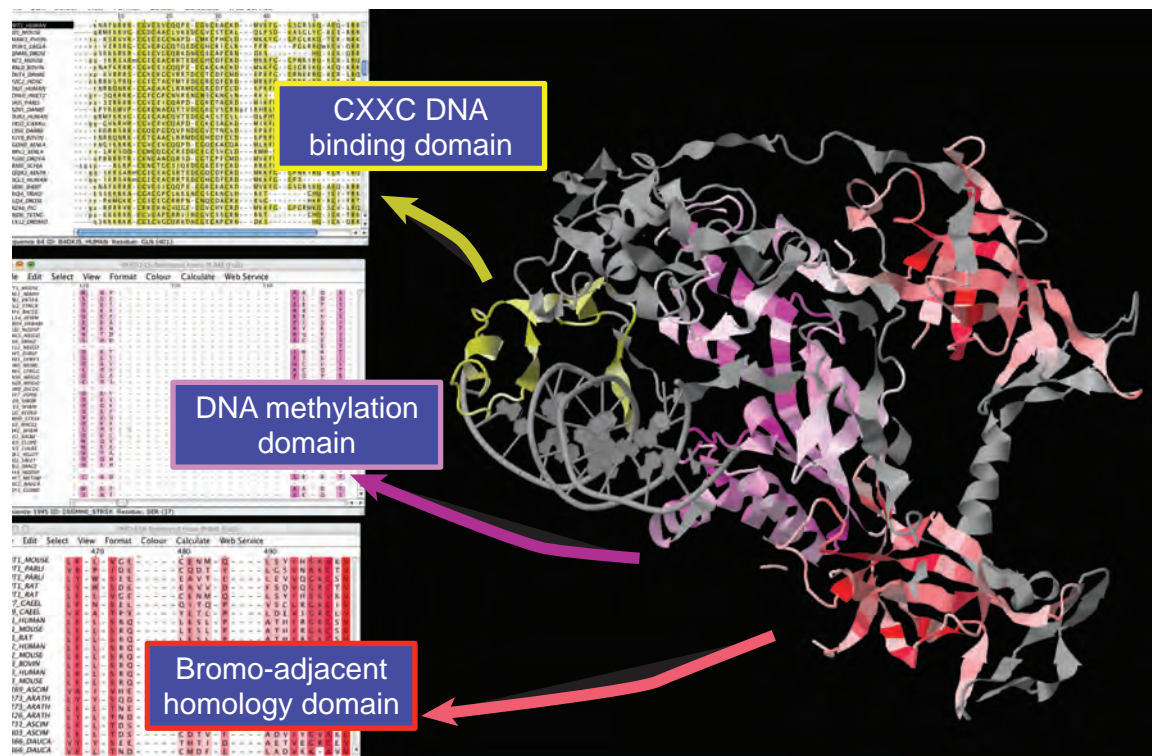
Figure 2.6: **The biological assembly of Mouse DNA Methyltransferase-1 coloured by Pfam alignments for its major domains** Alignments for each domain within the Uniprot sequence DNMT1_MOUSE have been used to visualise sequence conservation in each component of this protein-DNA complex. Instructions for recreating this figure are given in exercise 16.

Exercise 16: Colouring a protein complex to explore domain-domain interfaces

16.a. Download the PDB file at `http://www.jalview.org/tutorial/DNMT1_MOUSE.pdb` to your desktop.

This is the biological unit for PDB ID 3pt6, as identified by the PDBe's PISA server.

16.b. Launch the Jalview desktop and ensure you have at least 256MB of free memory available.

Use the following webstart link:

http://www.jalview.org/services/launchApp?jvm-max-heap=1G.

16.c. Retrieve the following **full** PFAM alignments: PF02008, PF00145, PF01426 (make sure you select the *PFAM* **(Full)** source). These will each be retrieved into their own alignment window.

16.d. Drag the structure you downloaded in step 1 onto one of the alignments to associate it with the mouse sequence in that Pfam domain family.

16.e. For every DNMT1_MOUSE sequence in the alignment, use the sequence ID popup menu's *Structure* submenu to view the DNMT1_MOUSE structure for the associated mouse sequence. When given the option, **view all of the structures in the same Jmol viewer**. Check the contents of the *View ⇒ Colour by ..* submenu to see what alignments can be used to colour the sequence.

16.f. Repeat the previous two steps for each of the other alignments. In each case, when performing the 'View DNMT1_MOUSE.pdb' step, Jalview will ask if you wish to create a new Jmol view. You should respond 'No', **ensuring that each sequence fragment is associated with the same Jmol view**.

16.g. Pick a different colourscheme for each alignment, and use the *Colour by ..* submenu to ensure they are all used to colour the complex shown in the Jmol window.

16.h. The final step needed to reproduce the shading in Figure 2.6 is to use the *Colour ⇒ By Annotation* option in each alignment window to shade the alignment by the **Conservation** annotation row. This function was described in section 1.7.5.

Ensure that you first disable the *View ⇒ Show Features* menu option, or you may not see any colour changes in the associated structure.

Note: Choose a different shading scheme for each alignment so that the regions of strong physicochemical conservation are highlighted. This kind of shading will reveal conserved regions of interaction between domains in the structure.

16.i. Save your work as a Jalview project and verify that it can be opened again by starting another Jalview Desktop instance, and dragging the saved project into the desktop window.

2.2   Analysis of alignments

Jalview provides support for sequence analysis in two ways. A number of analytical methods are 'built-in', these are accessed from the *Calculate* alignment window menu. Computationally intensive analyses are run outside Jalview *via* web services - these are typically accessed *via* the *Web Service* menu, and described in 2.3 and subsequent sections. 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, and how to overcome them, 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 pairwise substitution 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* [10] and implemented at the SeqSpace server at the EBI.

Jalview provides two different options for the PCA calculation. Protein PCAs are by default computed using BLOSUM 62 pairwise substitution scores, and nucleic acid alignment PCAs are computed using a score model based on the identity matrix that also treats Us and Ts as identical, to support analysis of both RNA and DNA alignments. The *Change Parameters* menu also allows the calculation method to be toggled between SeqSpace and a variant calculation that is detailed in Jalview's built in documentation.[11]

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 2.7). Each sequence is represented by a small 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.

---

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

[11] See http://www.jalview.org/help/html/calculations/pca.html.
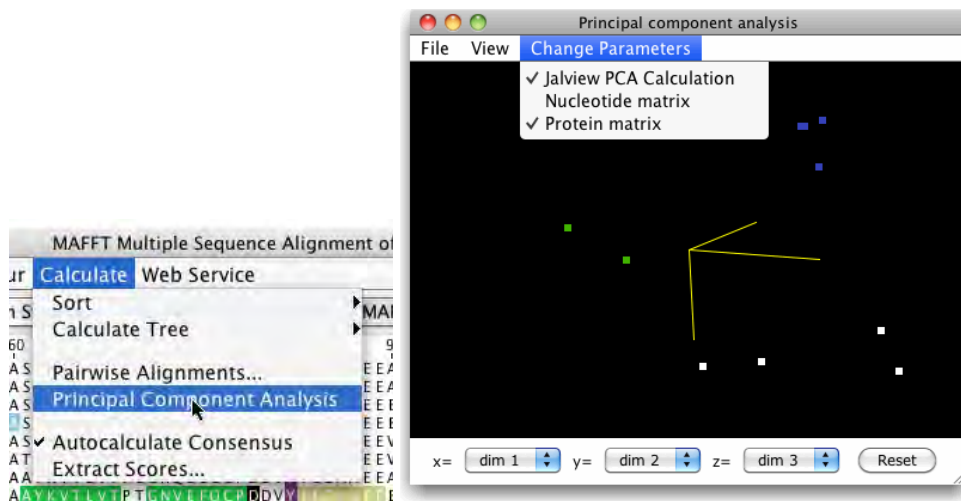
Figure 2.7: **PCA Analysis**

Exercise 17: Principle Component Analysis

17.a. Load the alignment at http://www.jalview.org/examples/exampleFile.jar and press
[ESC] to clear any selections. Alternatively, select *Select ⇒ Undefine Groups* to re-
move all groups and colourschemes.

17.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.

17.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.

PCA data export

Although the PCA viewer supports export of the current view, the plots produced are rarely suit-
able for direct publication. The PCA viewer's *File* menu includes a number of options for exporting
the PCA matrix and transformed points as comma separated value (CSV) files. These files can be
imported by tools such as **R** or **gnuplot** in order to graph the data.

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 is either the selected region or the whole alignment, excluding any hidden regions.
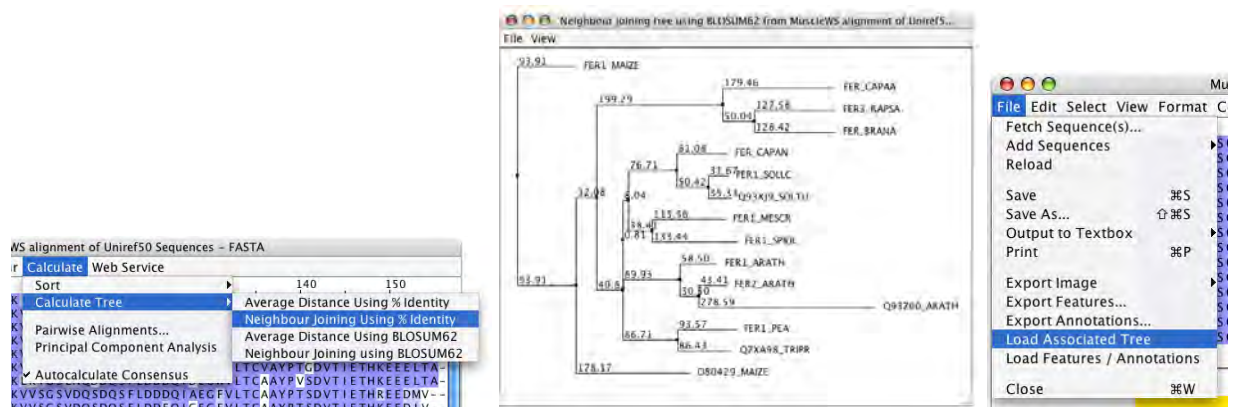
Figure 2.8: **Calculating Trees** Jalview provides four built in models for calculating trees. Jalview can also load precalculated trees in Newick format (right).
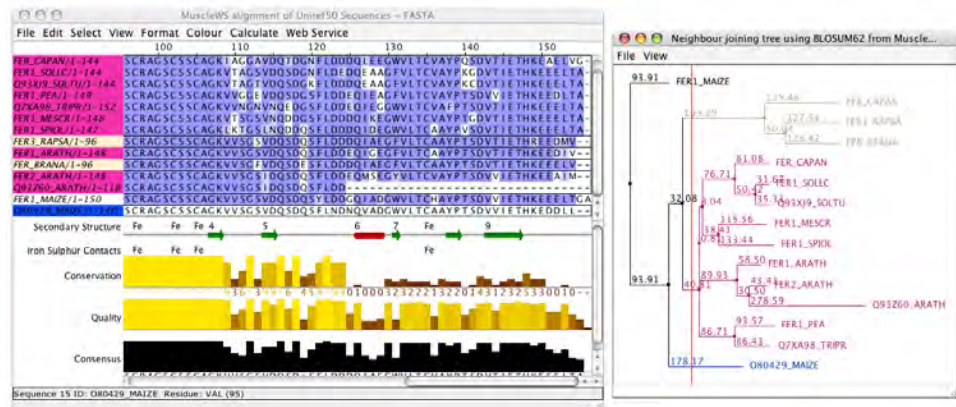
Figure 2.9: **Interactive Trees** The tree level cutoff can be used to designate groups in Jalview

On calculating a tree, a new window opens (Figure 2.8) which contains the tree. Various display settings can be found in the tree window *View* menu, including font, scaling and label display options, and the *File ⇒ Save As* submenu contains options for image and Newick file export. 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 alignment - unmatched leaves will still be displayed, and can be highlighted using the *View ⇒ Mark 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 2.9). To group them together, select the *Calculate ⇒ Sort ⇒ By Tree Order ⇒ . . .* alignment window menu option and choose 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. Note that tree partitioning will also remove any groups and colourschemes on a view, so create a new view ([CTRL-T]) if you wish to preserve these.

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 sub-
sequently changed.
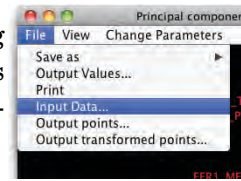
Changing the associated view for a tree or PCA viewer

The *View ⇒ Associated Nodes With ⇒ ..* submenu is shown
when the viewer is associated with an alignment that is in-
volved 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 according to the
colouring and selection state of the newly associated view.
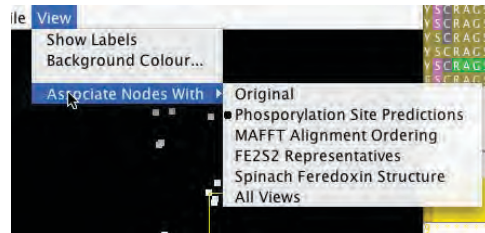
Exercise 18: Trees

18.a. Ensure that you have at least 1G memory available in Jalview (start with this link:
http://www.jalview.org/services/launchApp?jvm-max-heap=1G).

18.b. 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.

18.c. Click on the tree window. A cursor will appear. Note that placing this cursor di-
vides 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.

18.d. 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.

18.e. 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 three trees illustrates the importance of selecting appropriate regions of
the alignment for the calculation of trees.

18.f. Recover the *Input Data* for the tree you just calculated from the *File* menu. Check the
Edit ⇒ Pad Gaps option is *not* ticked, and insert one gap anywhere in the alignment.
Now select *Calculate ⇒ Calculate Tree ⇒ Neighbour Joining Using BLOSUM62*.
A warning dialog box **"Sequences not aligned"** appears because the sequences
input to the tree calculation are of different lengths.

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

Exercise 19: Tree Based Conservation Analysis

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

19.b. Build a Neighbourjoining tree using BLOSUM62 and use the *Sort Alignment By Tree* option in the tree viewer submenu to order alignment using the calculated tree.

19.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 if you first uncheck the *View ⇒ Show Annotations* option, and open the Overview Window to aid navigation.

19.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[12].

---

[12]Which can usually be undone. A future version of Jalview may allow redundant sequences to be hidden, or represented by a chosen sequence, rather than deleted.
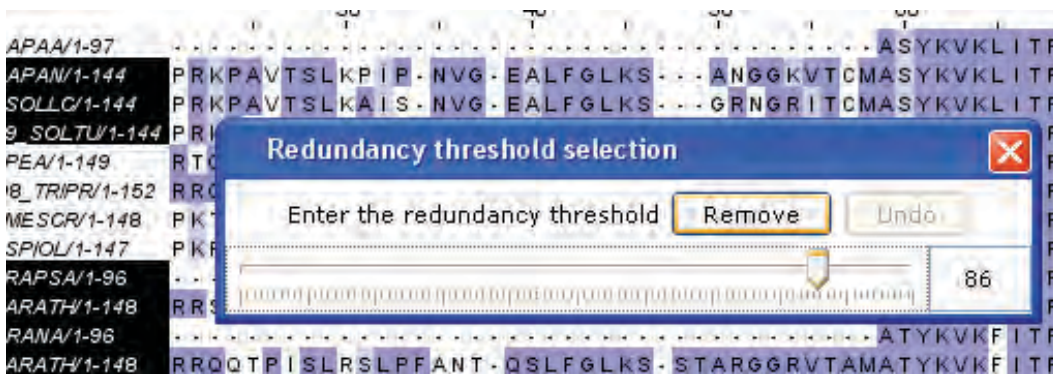
Figure 2.10: The Redundancy Removal dialog box opened from the edit menu. Sequences that exceed the current percentage identity threshold and are to be removed are highlighted in black.

Exercise 20: Remove redundant sequences

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

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

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

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

20.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[13] 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 BLOSUM 62) and physicochemical conservation will also be shown. Conservation is calculated according to Livingstone and Barton[14]. 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 toggle to autocalculate a consensus sequence upon editing. This is normally selected by default, but can be turned off for large alignments *via* the *Calculate ⇒ Autocalculate Consensus* menu option if the interface is too slow.

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 ⇒ Autocalculated 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.

---

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

[14]*"Protein Sequence Alignments: A Strategy for the Hierarchical Analysis of Residue Conservation."* Livingstone C.D. and Barton G.J. (1993) *CABIOS* **9**, 745-756
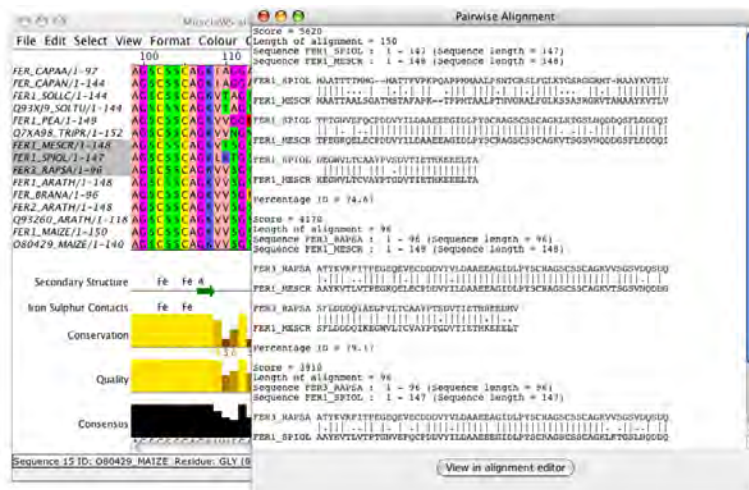
Figure 2.11: **Pairwise alignment of sequences.** Pairwise alignments of three selected sequences are shown in a textbox.

Exercise 21: Group conservation analysis

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

21.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.

21.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*.

21.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.

21.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.

21.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.

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

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.[15]

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. Most of these two-way services are provided by **Ja**va **B**ioinformatics **A**nalysis **W**eb **S**ervice (JABAWS) servers[16], which provides an easily installable system for performing a range of bioinformatics analysis tasks.
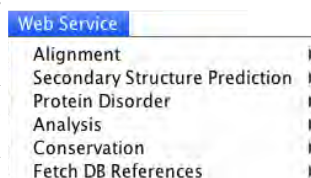
2.3.1 One-way web services

There are two 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 Section 2.9. A second type of one way service is provided by Jalview's DAS sequence feature retrieval system, which is described in Section 2.9.2.

2.3.2 Remote Analysis Web Services

Remote analysis services enable Jalview to use external computational facilities. There are curently three types of service - multiple sequence alignment, protein secondary structure prediction, and alignment analysis. Many of these are provided by JABA servers, which are described at the end of this section. In all 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[17], 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.

---

[15]HTTP: Hyper-Text Transfer Protocol.

[16]See http://www.compbio.dundee.ac.uk/jabaws for more information and to download your own server.

[17]This may be rectified in future versions.

2.3.3 JABA Web Services for sequence alignment and analysis

JABA stands for "JAva Bioinformatics Analysis", which is a system developed by Peter Troshin and Geoff Barton at the University of Dundee for running computationally intensive bioinformatics analysis programs. A JABA installation typically provides a range of JABA web services (JABAWS) for use by other programs, such as Jalview.

Exercises in the remainder of this section will demonstrate the simplest way of installing JABA on your computer, and configuring Jalview so it can access the JABA services. If you need any further help or more information about the services, please go to the JABAWS home page.

2.3.4 Changing the Web Services menu layout

If you are working with a lot of different JABA services, you may wish to change the way Jalview lays out the web services menu. You can do this from the Web Services tab of the *Preferences* dialog box.

Exercise 22: Changing the Layout of the Web Services Menu

22.a. Make sure you have loaded an alignment into Jalview, and examine the current layout of the alignment window's *Web Service* menu.

22.b. Open the preferences dialog box and select the web services tab.

22.c. Ensure the *Enable JABAWS services* checkbox is selected, and unselect the *Enable Enfin Services* checkboxes.

22.d. Hit *Refresh Services* to update the web services menu – once the progress bar has completed, open the *Web Service* menu to view the changes.

22.e. Select the *Index by host* checkbox and refresh the services once again.
      Observe the way the layout of the JABAWS Alignment submenu changes.

22.f. Do the same with the *Index by type* checkbox.

Jalview provides these options for configuring the layout of the *Web Service* menu because different Jalview users may have access to a different number of JABA services, and each will have their own preference regarding the layout of the menu.

Testing JABA services

The JABAWS configuration dialog shown in Figure 2.12 has colour codes to indicate whether the Desktop can access the server, and whether all services advertised by the server are functional. The colour codes are:

- Red - Server cannot be contacted or reports a connection error.

- Amber - Jalview can connect, but one or more services are non-functional.

- Green - Server is functioning normally.
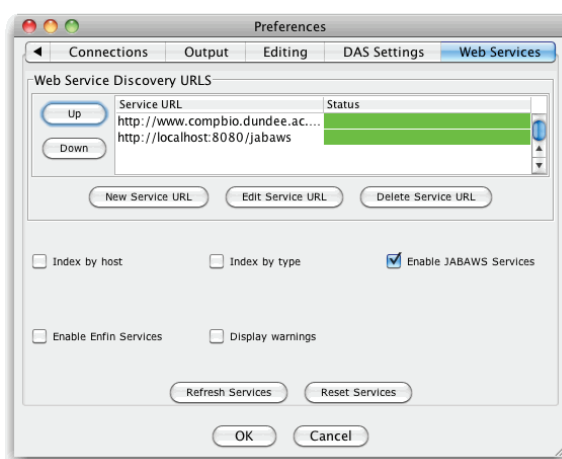
Figure 2.12: **The Jalview Web Services preferences panel.** Options are provided for configuring the list of JABA servers that Jalview will use, enabling and disabling Enfin services, and configuring the layout of the web services menu.

Test results from JABAWS are reported on Jalview's console output (opened from the Tools menu). Tests are re-run every time Jalview starts, and when the [Refresh Services] button is pressed on the Jalview JABAWS configuration panel.

Resetting the JABA services setting to their defaults

Once you have configured a JABAWS server and selected the OK button of the preferences menu, the settings will be stored in your Jalview preferences file, along with any preferences regarding the layout of the web services menu. If you should ever need to reset the JABAWS server list to its defaults, use the 'Reset Services' button on the Web Services preferences panel.

2.3.5 Running your own JABA server

You can download and run JABA on your own machine using the 'VMWare' or VirtualBox virtual machine environments. If you would like to learn how to do this, there are full instructions at the JABA web site.

Exercise 23: Installing a JABA Virtual Machine on your computer
This tutorial will demonstrate the simplest way of installing JABA on your computer, and configuring Jalview so it can access the JABA services.
Prerequisites
You will need a copy of VMWare Player/Workstation/Fusion on your machine.
23.a. If you do not have VMWare player installed, download it from www.vmware.com (this takes a few minutes – you will need to register and wait for an email with a download link).
23.b. Download the JABA virtual appliance archive called 'jaba-vm.zip' from `http://www.compbio.dundee.ac.uk/jabaws/archive/jabaws-vm.zip`
WARNING: This is large (about 300MB) and will take some time to download.
23.c. Unpack the archive's contents to a place on your machine with at least 2GB of free space.
(On Windows, right click on the archive, and use the 'Extract archive..' option).
23.d. Open the newly extracted directory and double click the VMWare virtual machine configuration file (jabaws.vcf). This will launch the VMWare player.
23.e. Once VMWare player has started up, it may ask the question "Did you move or copy this virtual appliance?" – select 'Copy'.
23.f. You may be prompted to download the VMWare linux tools. These are not necessary, so close the window or click on 'Later'.
23.g. You may also be prompted to install support for one or more devices (USB or otherwise). Say 'No' to these options.
23.h. Once the machine has loaded, it will display a series of IP addresses for the different services provided by the VM. Make a note of the JABAWS URL – this will begin with 'http:' and end with '/jabaws".

Exercise 24: Configuring Jalview to access your new JABAWS virtual appliance
24.a. Start Jalview (If you have not done so already).
24.b. Enable the Jalview Java Console by selecting its option from the Tools menu.
Alternately, use the System Java console if you have configured it to open when Jalview is launched, via your system's Java preferences (under the 'Advanced' tab on Windows).
24.c. Open the *Preferences* dialog and locate the Web Services tab.
24.d. Add the URL for the new JABAWS server you started in Exercise 23 to the list of JABAWS urls using the 'New Service URL' button.
24.e. You will be asked if you want to test the service. Hit 'Yes' to do this – you should then see some output in the console window.
Take a close look at the output in the console. What do you think is happening?
24.f. Hit OK to save your preferences – you have now added a new JABA service to Jalview!
24.g. Try out your new JABA services by loading the ferredoxin sequences from http://www.jalview.org/tutorial/alignment.fa
24.h. Launch an alignment using one of the JABA methods provided by your server. It will be listed under the JABAWS Alignment submenu of the *Web Service* menu on the alignment window.
Note: You can watch the JABA VM appliance's process working by opening the process monitor on your system. (On Windows XP, this involves right-clicking the system clock and opening the task manager – then selecting the 'Processes' tab and sort by CPU).

2.4 Multiple Sequence Alignment

Sequences can be aligned using a range of algorithms provided by JABA web services. These include ClustalW[18], Muscle[19], MAFFT[20], ProbCons,[21] T-COFFEE[22] and Clustal Omega.[23] Of these, T-COFFEE is the slowest, but also the most accurate. ClustalW is historically the most widely used. Muscle is faster than ClustalW and probably the most accurate for smaller alignments and MAFFT is probably the best for large alignments, however Clustal Omega, which was released in 2011, is arguably the fastest and most accurate tool for protein multiple alignment.

To run an alignment web service, select the appropriate method from the *Web Service ⇒ Alignment ⇒ . . .* submenu (Figure 2.13). For each service you may either perform an alignment with default settings, use one of the available presets, or customise the parameters with the '*Edit and Run ..*' dialog box. Once the job is submitted, 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 *Calculate ⇒ Sort* sub menu.
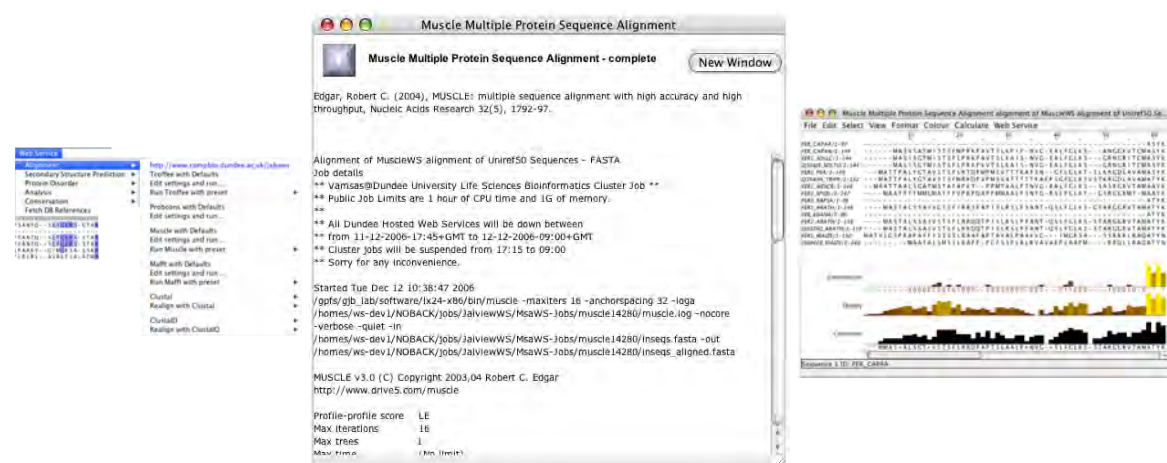
Figure 2.13: **Multiple alignment via web services** The appropriate method is selected from the menu (left), a status box appears (centre), and the results appear in a new window (right)

[18]*"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

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

[20]*"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.

[21]PROBCONS: Probabilistic Consistency-based Multiple Sequence Alignment. Do, C.B., Mahabhashyam, M.S.P., Brudno, M., and Batzoglou, S. (2005) *Genome Research* **15** 330-340.

[22]T-Coffee: A novel method for multiple sequence alignments. (2000) Notredame, Higgins and Heringa *JMB* **302** 205-217

[23]Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Sievers F, Wilm A, Dineen DG, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG (2011) *Molecular Systems Biology* **7** 539 doi:10.1038/msb.2011.75

Realignment

The re-alignment option is currently only supported by ClustalW and Clustal Omega. 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.

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 in 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), and 2) hidden columns can be used to preserve existing parts of an alignment whilst the visible parts are locally refined.

Exercise 25: Multiple Sequence Alignment

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

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

25.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.

25.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.

25.e. Select columns 60 to 125 in the original MAFFT alignment and hide them. Select *Web Services ⇒ Alignment ⇒ Mafft with Defaults* 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.

25.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.

2.4.1 Customising the parameters used for alignment

JABA web services allow you to vary the parameters used when performing a bioinformatics analysis. For JABA alignment services, this means you are usually able to modify the following types of parameters:

- Amino acid or nucleotide substitution score matrix

- Gap opening and widening penalties

- Types of distance metric used to construct guide trees

- Number of rounds of re-alignment or alignment optimisation

Getting help on the parameters for a service

Each parameter available for a method usually has a short description, which jalview will display as a tooltip, or as a text pane that can be opened under the parameter's controls. In the parameter shown in Figure 2.14, the description was opened by selecting the button on the left hand side. Online help for the service can also be accessed, by right clicking the button and selecting a URL from the pop-up menu that will open.
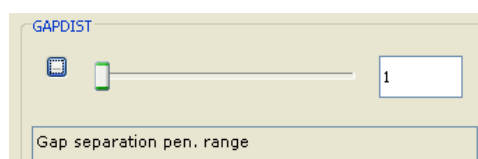
Figure 2.14: **ClustalW parameter slider detail**. From the ClustalW *Clustal ⇒ Edit settings and run ...* dialog box.

2.4.2 Alignment Presets

The different multiple alignment algorithms available from JABA vary greatly in the number of adjustable parameters, and it is often difficult to identify what are the best values for the sequences that you are trying to align. For these reasons, each JABA service may provide one or more presets – which are pre-defined sets of parameters suited for particular types of alignment problem. For instance, the Muscle service provides the following presets:

- Huge

- Protein alignments (fastest speed)

- Nucleotide alignments (fastest speed)

The presets are displayed in the JABA web services submenu, and can also be accessed from the parameter editing dialog box, which is opened by selecting the '*Edit settings and run ...*' option from

the web services menu. If you have used a preset, then it will be mentioned at the beginning of the job status file shown in the web service job progress window.

Alignment Service Limits

Multiple alignment is a computationally intensive calculation. Some JABA server services and service presets only allow a certain number of sequences to be aligned. The precise number will depend on the server that you are using to perform the alignment. Should you try to submit more sequences than a service can handle, then an error message will be shown informing you of the maximum number allowed by the server.

2.4.3   User defined Presets

Jalview allows you to create your own presets for a particular service. To do this, select the '*Edit settings and run ...*' option for your service, which will open a parameter editing dialog box like the one shown in Figure 2.15.

The top row of this dialog allows you to browse the existing presets, and when editing a parameter set, allows you to change its nickname. As you adjust settings, buttons will appear at the top of the parameters dialog that allow you to Revert or Update the currently selected user preset with your changes, Delete the current preset, or Create a new preset, if none exists with the given name. In addition to the parameter set name, you can also provide a short description for the parameter set, which will be shown in the tooltip for the parameter set's entry in the web services menu.
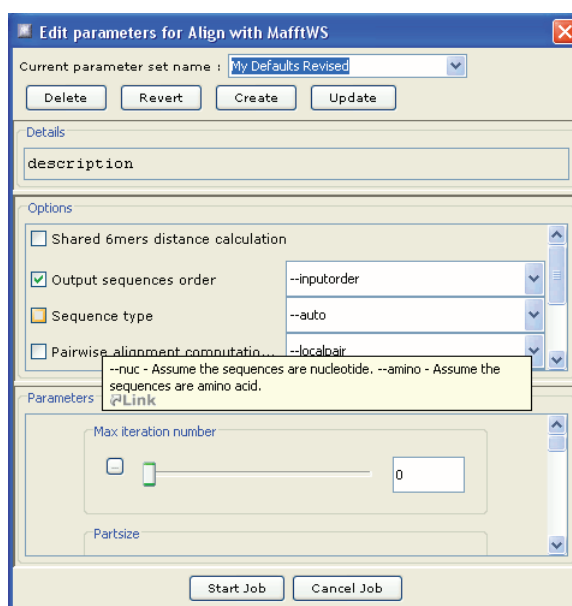
Figure 2.15: **Jalview's JABA alignment service parameter editing dialog box**.

Saving parameter sets

When creating a custom parameter set, you will be asked for a file name to save it. The location of the file is recorded in the Jalview user preferences in the same way as a custom alignment colourscheme, so when Jalview is launched again, it will show your custom preset amongst the options available for running the JABA service.

2.5 Protein alignment conservation analysis

The *Web Service ⇒ Conservation* menu controls the computation of up to 17 different amino acid conservation measures for the current alignment view. The JABAWS AACon Alignment Conservation Calculation Service, which is used to calculate these scores, provides a variety of standard measures described by Valdar in 2002[24] as well as an efficient implementation of the SMERFs score developed by Manning et al. in 2008.[25]

Enabling and disabling AACon calculations

When the AACon Calculation entry in the *Web Services ⇒ Conservation* menu is ticked, AACon calculations will be performed every time the alignment is modified. Selecting the menu item will enable or disable automatic recalculation.

Configuring which AACon calculations are performed

The *Web Services ⇒ Conservation ⇒ Change AACon Settings ...* menu entry will open a web services parameter dialog for the currently configured AACon server. Standard presets are provided for quick and more expensive conservation calculations, and parameters are also provided to change the way that SMERFS calculations are performed. AACon settings for an alignment are saved in Jalview projects along with the latest calculation results.

Changing the server used for AACon calculations

If you are working with alignments too large to analyse with the public JABAWS server, then you will most likely have already configured additional JABAWS servers. By default, Jalview will chose the first AACon service available from the list of JABAWS servers available. If available, you can switch to use another AACon service by selecting it from the *Web Services ⇒ Conservation ⇒ Switch Server* submenu.

---

[24]Scoring residue conservation. Valdar (2002) *Proteins: Structure, Function, and Genetics* **43** 227-241.
[25]SMERFS Score Manning et al. *BMC Bioinformatics* 2008, **9** 51 doi:10.1186/1471-2105-9-51

2.6   Protein Secondary Structure Prediction

Protein secondary structure prediction is performed using the Jpred[26] server at the University of
Dundee[27]. The behaviour of this calculation depends on the current selection:

- 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 JPred 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 JPred prediction.

- If just one sequence (or a region in one sequence) has been selected, it will be submitted to the
  automatic JPred 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 JPred 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*[28] from the alignment window menu
(Figure 2.16). 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.
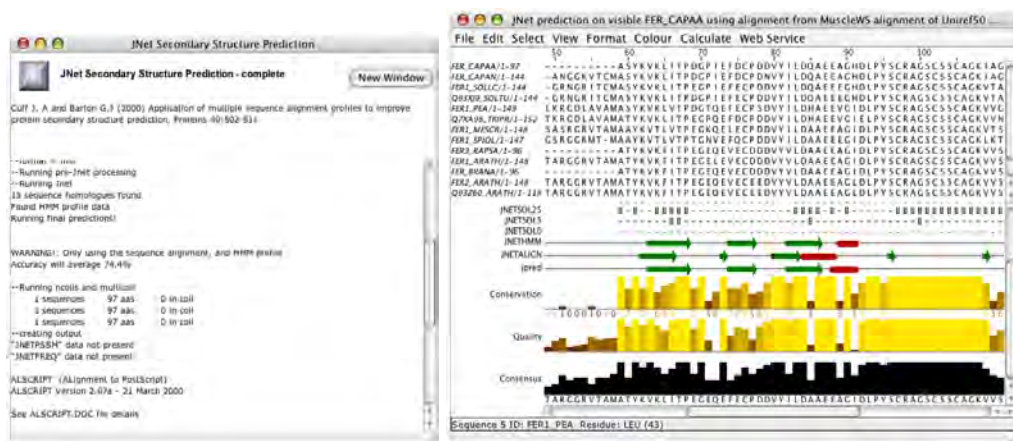
Figure 2.16: **Secondary Structure Prediction** Status (left) and results (right) windows for
                JNet predictions.

---

[26]*"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
[27]http://www.compbio.dundee.ac.uk/www-jpred/
[28]JNet is the Neural Network based secondary structure prediction method that the JPred server uses.

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 can produce different results. In some cases, these secondary structure predictions can be more reliable for sequence on either side of the insertion[29]. 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.

Exercise 26: Secondary Structure Prediction

26.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 (about 2-4 min) 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.

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

26.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 annotations are examples of **sequence-associated alignment annotation**.

26.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.

26.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 29.

26.f. In the original alignment that you loaded in step 1, **select all** sequences, then open the **Sequence ID ⇒ Selection** submenu and select the **Add Reference Annotation** option.

The JNet predictions for the sequences should now be visible in the original alignment.

2.7 Protein Disorder Prediction

Disordered regions in proteins were classically thought to correspond to 'linkers' between distinct protein domains, but disorder can also play a role in function. The *Web Services ⇒ Disorder* menu in the alignment window allows access to protein disorder prediction services provided by the configured JABAWS servers.

---

[29]This, of course, cannot be guaranteed.

2.7.1   Disorder prediction results

Each service operates on sequences in the alignment to identify regions likely to be unstructured or flexible, or alternately, fold to form globular domains. As a consequence, disorder predictor results include both sequence features and sequence associated alignment annotation rows. Section 2.8 describes the manipulation and display of these data in detail, and **Figure 2.17** demonstrates how sequence feature shading and thresholding (described in Section 2.9.3) can be used to highlight differences in disorder prediction across aligned sequences.
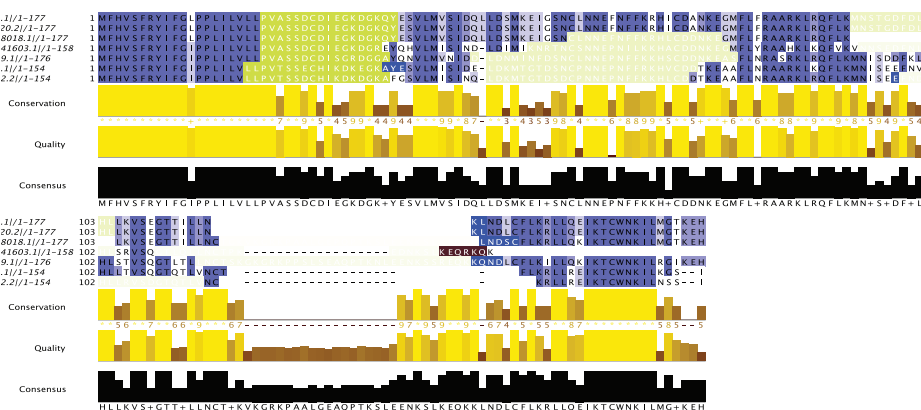
Figure 2.17: **Shading alignment by sequence disorder**. Alignment of Interleukin IV homologs coloured with Blosum62 with protein disorder prediction sequence features overlaid, shaded according to their score. Borderline disordered regions appear white, reliable predictions are either Green or Brown depending on the type of disorder prediction.

Navigating large sets of disorder predictions

Figure 2.18 shows a single sequence annotated with a range of disorder predictions. Disorder prediction annotation rows are associated with a sequence in the same way as secondary structure prediction results. When browsing an alignment containing large numbers of disorder prediction annotation rows, clicking on the annotation row label will highlight the associated sequence in the alignment display, and double clicking will select that sequence.

2.7.2   Disorder predictors provided by JABAWS 2.0

For full details of each predictor and the results that Jalview can display, please consult Jalview's protein disorder service documentation. Short descriptions of the methods provided in JABAWS 2.0 are given below:
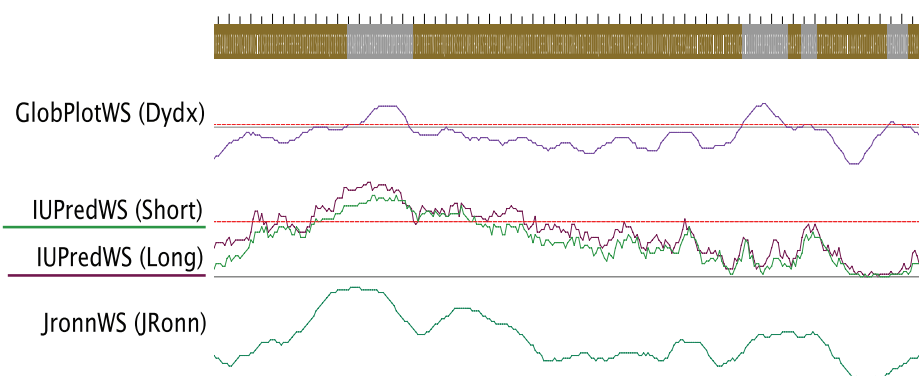
Figure 2.18: **Annotation rows for several disorder predictions on a sequence**. A zoomed out view of a prediction for a single sequence. The sequence is shaded to highlight disordered regions (brown and grey), and the line plots below the Sequence show the raw scores for various disorder predictors. Horizontal lines on each graph mark the level at which disorder predictions become significant.

DisEMBL

DisEMBL (Linding et al., 2003) is a set of machine-learning based predictors trained to recognise disorder-related annotation found on PDB structures.

COILS Predicts loops/coils according to DSSP definitions[30]. Features mark range(s) of residues predicted as loops/coils, and annotation row gives raw value for each residue. Value over 0.516 indicates loop/coil.

HOTLOOPS constitute a refined subset of **COILS**, namely those loops with a high degree of mobility as determined from $C\alpha$ temperature factors (B factors). It follows that highly dynamic loops should be considered protein disorder. Features mark range(s) of residues predicted to be hot loops and annotation row gives raw value for each residue. Values over 0.6 indicates hot loop.

REMARK465 "Missing coordinates in X-ray structure as defined by remark465 entries in PDB. Nonassigned electron densities most often reflect intrinsic disorder, and have been used early on in disorder prediction." Features give range(s) of residues predicted as disordered, and annotation rows gives raw value for each residue. Values over 0.1204 indicates disorder.

RONN *a.k.a.* **Regional Order Neural Network**

RONN employs an approach known as the 'bio-basis' method to predict regions of disorder in sequences based on their local similarity with a gold-standard set of disordered protein sequences. It yields a set of disorder prediction scores, which are shown as sequence annotation below the alignment.

---

[30]DSSP Classifications of secondary structure are: α-helix (H), 310-helix (G), β-strand (E) are ordered, and all other states (β-bridge (B), β-turn (T), bend (S), π-helix (I), and coil (C)) considered loops or coils.

JRonn[31] Annotation Row gives RONN score for each residue in the sequence. Scores above 0.5 identify regions of the protein likely to be disordered.

IUPred

IUPred employs an empirical model to estimate likely regions of disorder. There are three different prediction types offered, each using different parameters optimized for slightly different applications. It provides raw scores based on two models for predicting regions of 'long disorder' and 'short disorder'. A third predictor identifies regions likely to form structured domains.

Long disorder Annotation rows predict context-independent global disorder that encompasses at least 30 consecutive residues of predicted disorder. A 100 residue window is used for calculation. Values above 0.5 indicates the residue is intrinsically disordered.

Short disorder Annotation rows predict for short, (and probably) context-dependent, disordered regions, such as missing residues in the X-ray structure of an otherwise globular protein. Employs a 25 residue window for calculation, and includes adjustment parameter for chain termini which favors disorder prediction at the ends. Values above 0.5 indicate short-range disorder.

Structured domains are marked with sequence Features. These highlight likely globular domains useful for structure genomics investigation. Post-analysis of disordered region profile to find continuous regions confidently predicted to be ordered. Neighbouring regions close to each other are merged, while regions shorter than the minimal domain size of at least 30 residues are ignored.

GLOBPLOT

GLOBPLOT defines regions of globularity or natively unstructured regions based on a running sum of the propensity of residues to be structured or unstructured. The propensity is calculated based on the probability of each amino acid being observed within well defined regions of secondary structure or within regions of random coil. The initial signal is smoothed with a Savitzky-Golay filter, and its first order derivative computed. Residues for which the first order derivative is positive are designated as natively unstructured, whereas those with negative values are structured.

Disordered region sequence features are created marking mark range(s) of residues with positive first order derivatives, and **Globular Domain** features mark long stretches of order. **Dydx** annotation rows gives the first order derivative of smoothed score. Values above 0 indicates residue is disordered.

Smoothed Score and Raw Score annotation rows give the smoothed and raw scores used to create the differential signal that indicates the presence of unstructured regions. These are hidden by default, but can be shown by right-clicking on the alignment annotation panel and selecting **Show hidden annotation**.

---

[31]JRonn denotes the score for this server because JABAWS runs a Java port of RONN developed by Peter Troshin and distributed as part of Biojava 3

Exercise 27: Protein Disorder Prediction

27.a. Open the alignment at `http://www.jalview.org/tutorial/interleukin7.fa`.

27.b. Run the DisEMBL disorder predictor the Services ⇒ Disorder Prediction submenu.

27.c. Use *Sequence ID ⇒ Structure ⇒ Discover PDB IDs* to retrieve all the PDB structures for the sequences.

27.d. Open and align the structures for all sequences.
  Hint: see 2.1.2 to see how to do this.

27.e. Compare the disorder predictions to the structure data by mapping any available temperature factors to the alignment *via* the *Sequence ID Popup ⇒ Selection ⇒ Add reference annotation* option.

27.f. Apply the IUPred disorder prediction method

27.g. Use the *Per sequence option* in the *Colour ⇒ By annotation . . .* dialog to shade the sequences by the long and short disorder predictors. Do the two methods agree with the structure ?

2.8 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 shown 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, so as the alignment changes, they change along with it. Conversely, sequence features are properties of the individual sequences, so 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 JNet (see 2.16 above) can be used to analyse a given sequence or alignment and generate annotation for it.

2.8.1 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 2.19). 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.

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.
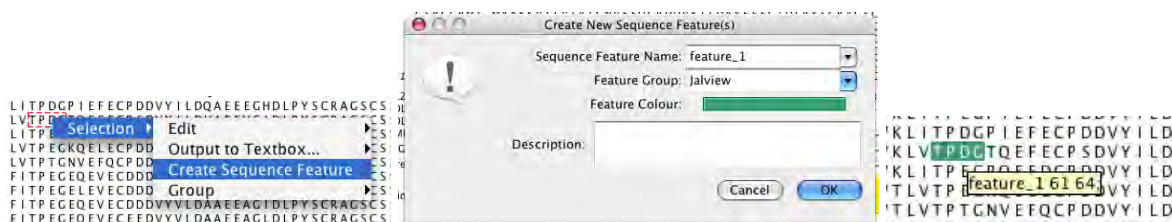
Figure 2.19: **Creating sequence features.** Features can readily be created from selections via
the context menu and are then displayed on the sequence.

2.8.2   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 2.21) 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 2.9.3 and 2.9.4.
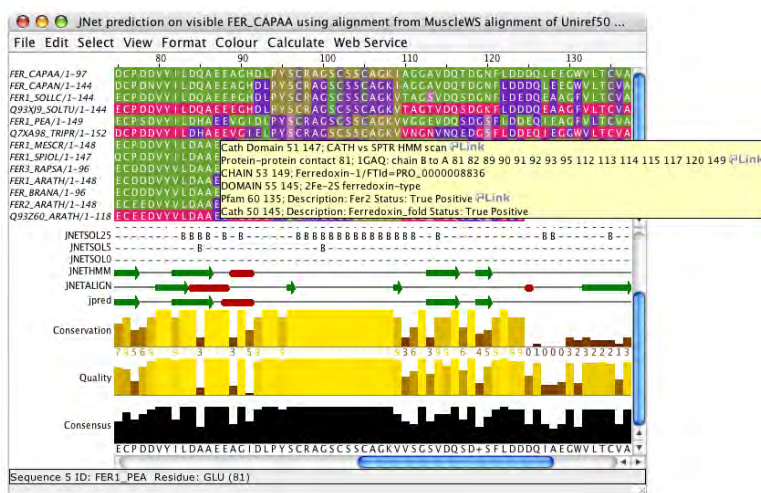
Figure 2.20: **Multiple sequence features.** An alignment with JPred secondary structure pre-
diction annotation below it, and many sequence features overlaid onto the aligned
sequences. The tooltip lists the features annotating the residue below the mouse-
pointer.

Figure 2.21: **Customising sequence features.** Features can be recoloured, switched on or off and have the rendering order changed.

2.8.3   Sequence Feature File Formats

Jalview supports the widely used GFF tab delimited format[32] 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.

---

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

---

Exercise 28: Creating features

28.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.

28.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.

28.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 column 95. Roll the mouse over the features and you will see that the
feature has moved with the sequence. Delete the gap you created.

28.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.

28.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.

2.8.4   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 2.22). A dialogue box appears. Enter the label to use
for this row and a new row will appear.
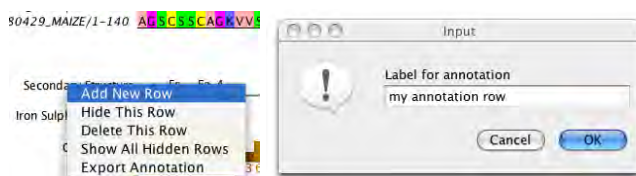
Figure 2.22: **Creating a new annotation row.** Annotation rows can be reordered by dragging
them to the desired place.

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 2.23). 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[33]. Annotations can be coloured or deleted as desired.

Figure 2.23: **Creating a new annotation.** Annotations are created from a selection on the annotation row and can be coloured as desired.

Exercise 29: Annotating alignments

29.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.

29.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.

29.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.

29.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.

29.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.

29.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.*

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

29.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.*

29.i. Recover or recreate the secondary structure prediction that you made in exercise 26. 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.

---

[33]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.

2.9   Importing features from databases

Jalview supports feature retrieval from public databases either directly or *via* the Distributed Anno-
tation System (DAS[34]). It includes built in parsers for Uniprot and ENA (or EMBL) records retrieved
from the EBI. Sequences retrieved from these sources using the sequence fetcher (see Section 1.4.5)
will already possess features.

2.9.1   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 po-
sition. If the start/end positions do not match the coordinate system from which the features were
defined, then the features will be displayed incorrectly.

Viewing and exporting a sequence's database annotation

You can export all the database cross references and annotation terms shown in the sequence ID
tooltip for a sequence by right-clicking and selecting the *[Sequence ID] ⇒ Sequence details . . .* option
from the popup menu. A similar option is provided in the *Selection* sub-menu allowing you to obtain
annotation for the sequences currently selected.

The *Sequence Details . . .* option will open a window
containing the same text as would be shown in the
tooltip window, including any web links associated
with the sequence. The text is HTML, and options
on the window allow the raw code to be copied and
pasted into a web page.

---

[34]http://www.biodas.org/
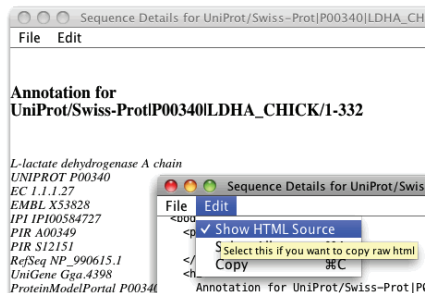
Automatically discovering a sequence's database references

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 *Webservice ⇒ 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).

Exercise 30: Retrieving Database References

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

30.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.

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

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

30.e. Once the process has finished, save the alignment as a Jalview Project.

30.f. 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

30.g. View the *Sequence details . . .* report for the FER1_SPIOL sequence and for the whole alignment. Which sequences have web links associated with them ?

2.9.2 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 Sequence Feature Settings Window (Figure 2.24). A list of DAS sources compiled from the currently configured DAS registry[35] is shown in the left hand pane. Highlighting an entry on the left brings up information about that source in the right hand panel.

---

[35]By default, this will be the major public DAS server registry maintained by the Sanger Institute: http://www.dasregistry.org
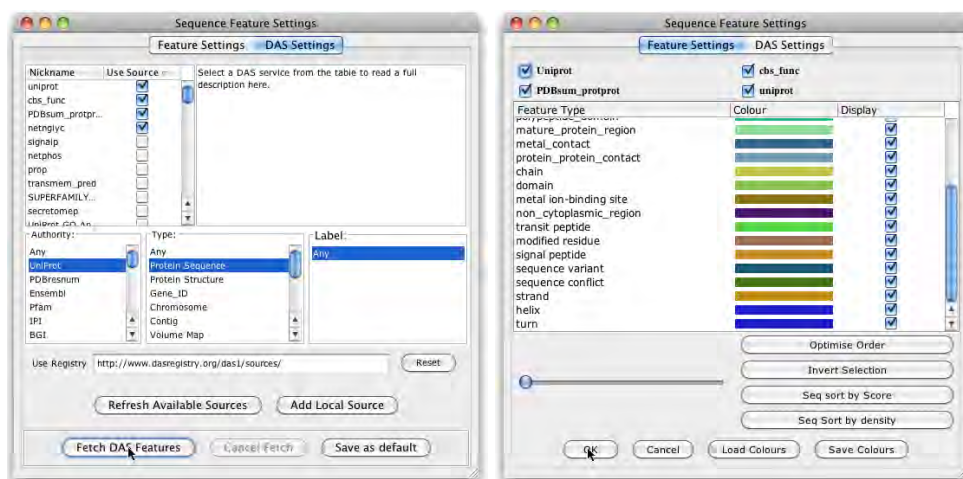
Figure 2.24: **Retrieving DAS annotations.** DAS features are retrieved using the *DAS Settings* tab (left) and their display customised using the *Feature Settings* tab (right).

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 (Figure 2.24 (right)). Each data source is listed and groups of features from one data source can be selected/deselected by checking the labelled 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 Service ⇒ Fetch Database References* function described in Section 2.9.1. 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 31: Retrieving features with DAS**
>
> *31.a.* Load the alignment at `http://www.jalview.org/tutorial/alignment.fa`. Select *View ⇒ Feature Settings . . .* 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 netnglyc. 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.
>
> *31.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 Sequence Feature Settings window.
>
> *31.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.
>
> *31.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.
>
> *31.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 are drawn on top of those below, obscuring them in the alignment view where they overlap. 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.
>
> *31.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.
>
> *31.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.
>
> *31.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.

2.9.3 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* or *Color* in the *Sequence Feature Settings* dialog 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 colourschemes 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 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.

2.9.4   Using features to re-order the alignment

The presence of sequence features on certain sequences 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 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 Sequence 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.

2.10   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. Jalview recognises nucleotide sequences and alignments 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 nucleotide 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 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.

Program	NA support	Notes
ClustalW	Yes	Default is to autodetect nucleotide sequences. Editable parameters include nucleotide substitution matrices and distance metrics.
Muscle	Yes (treat U as T)	Default is to autodetect nucleotide sequences. Editable parameters include nucleotide substitution matrices and distance metrics.
MAFFT	Yes	Will autodetect nucleotide sequences and use a hardwired substitution model (all amino-acid sequence related parameters are ignored). Unknown whether substitution model treats Uracil specially.
ProbCons	No	ProbCons has no special support for aligning nucleotide sequences. Whilst an alignment will be returned, it is unlikely to be reliable.
T-COFFEE	Yes	Sequence type is automatically detected and an appropriate parameter set used as required. A range of nucleotide specific score models are available.

Table 2.1: **JABAWS Alignment programs suitable for aligning nucleic acid sequences.** All JABAWS alignment services will return an alignment if provided with RNA or DNA sequences, with varying reliability.

2.10.1 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 *Calculate ⇒ Pairwise Alignment . . .* option will utilise an identity score matrix to calculate alignment score when aligning two nucleotide sequences.

Aligning Nucleic Acid Sequences

Jalview has limited knowledge of the capabilities of the programs that are made available to it *via* web services, so it is up to you, the user, to decide which service to use when working with nucleic acid sequences. The table below shows which alignment programs are most appropriate for nucleotide alignment. Generally, all will work, but some may be more suited to your purposes than others. We also note that none of these include support for taking RNA secondary structure prediction into account when aligning sequences (but will be providing services for this in the future!).

2.10.2 Translate cDNA

The *Calculate ⇒ 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 documentation 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.

2.10.3   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 that have been fetched 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.
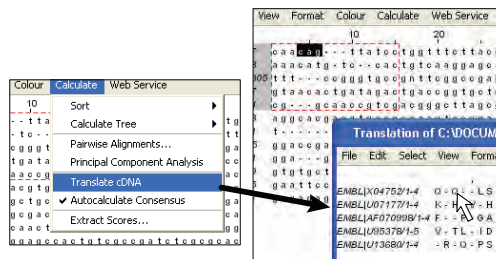
2.10.4   Coding regions from EMBL records

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 *Calculate ⇒ 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 protein 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 accession numbers associated with the sequence to retrieve features, and then map them onto the nucleotide sequence's coordinate system using the coding region location.
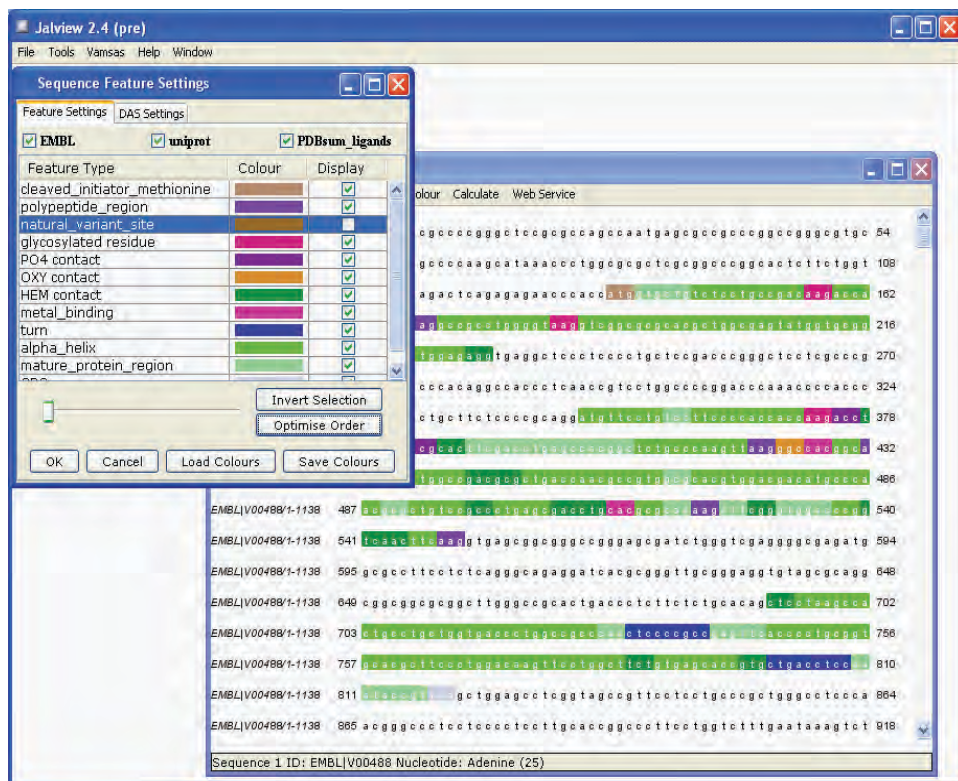
Figure 2.25: Uniprot and PDB sum features retrieved *via* DAS and mapped onto coding regions of EMBL record V00488 (an earlier version of Jalview is shown here).

Exercise 32: Visualizing protein features on coding regions

32.a. Use the sequence fetcher to retrieve EMBL record D49489.

32.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.

32.c. Open the *DAS Settings* tab in the *Sequence Feature Settings. . .* window and fetch features for D49489 from the Uniprot reference server, and any additional servers that work with the Uniprot coordinate system.

32.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.

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

2.10.5   Working with RNA

Jalview allows the creation of RNA secondary structure annotation, and includes the VARNA secondary structure viewer for the display of RNA base pair diagrams. It also allows the extraction of RNA secondary structure from 3D data when available.

2.10.6   Performing RNA secondary structure predictions

Secondary structure consensus calculations can be performed by enabling the VIENNA service *via* the *Web Services ⇒ Secondary Structure* menu. These consensus structures are created by analysing the covariation patterns in all visible sequences on the alignment. For more information see the VIENNA documentation.
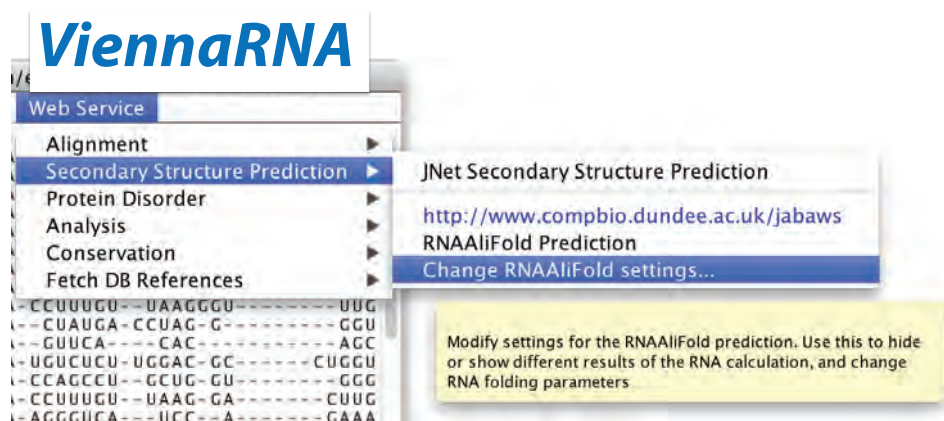

Figure 2.26: Secondary structure consensus calculations can be performed by enabling the VIENNA service *via* the *Web Services ⇒ Secondary Structure* menu.
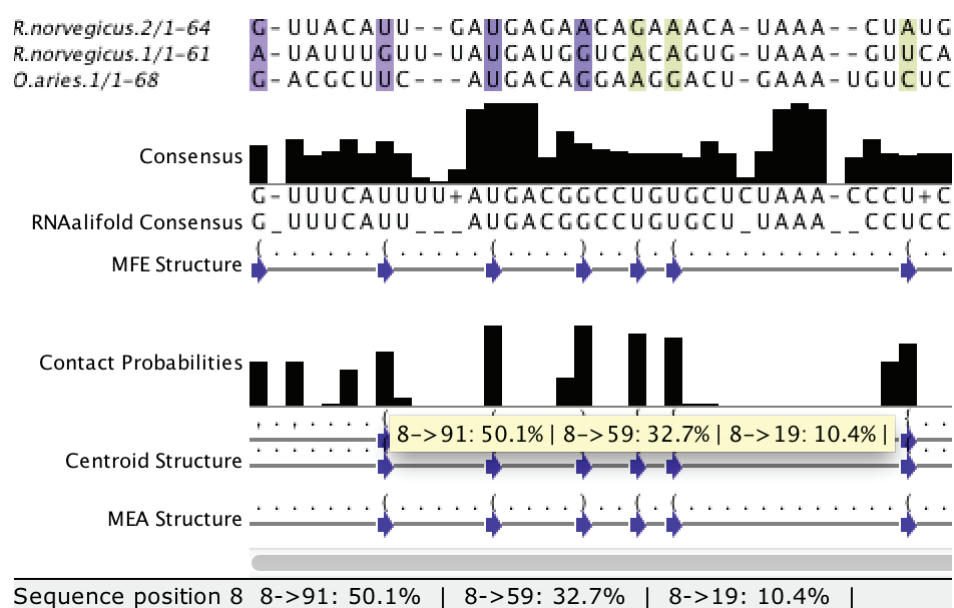
Figure 2.27: VIENNA can calculate alternate RNA base pairing probabilities. These are shown
in Jalview as tool-tips on the RNA secondary structure probability score.

Exercise 33: Viewing RNA structures

33.a. Import RF00162 from Rfam (Full).

33.b. Select *Colour by RNA Helices* to shade the alignment by the secondary structure
annotation provided by Rfam.

33.c. Open VARNA with *Structure ⇒ View Structure ⇒ RNA Secondary Structure.*
Explore the difference between trimmed and untrimmed views for the structure.

33.d. Add and link a Jmol structure view for Bacillus_amyloliquef.9 for 3NPB (from the
PDB). Display the secondary structure along-side the consensus structure for the
alignment by adding reference annotation from the 3D structure.
Hint: You need to make sure the RNAview service is enabled in your Structure pref-
erences to obtain RNA secondary structure annotation from PDB files.

33.e. Perform a secondary structure prediction. Enable the VIENNA consensus calculation
via the *Web Services* menu. Compare this with the annotation line provided by Rfam.

33.f. Edit the VIENNA calculation settings to show Base Pair probabilities.
Explore how editing the alignment affects the consensus calculation.