The manual

Version 3.5

Complex Systems and Translational Bioinformatics,
The ICube laboratory
Centre de Recherches Biomédicales de Strasbourg (CRBS)
1, rue Eugène Boeckel
67 000 Strasbourg
FRANCE

Contents

1 Installing Ordalie

1.1 Requirements

Ordalie runs on Windows, MacOS and Linux 32/64 bits platforms. In order to display 3D structures Ordalie uses the OpenGL library which is generally provided by the graphic card in most computers.

On MacOS, since the *Mojave OS version, the OpenGL library is not included in MacOS distribution anymore. It could still be installed freely through the* App Store.

Ordalie can be run without network connection. Nevertheless, accessing Internet is required in order to benefit of all functionalities such as accessing and querying sequences databases (PDB, UniProt, NCBI), aligning sequences on-the-fly, or using web services.

1.2 Installation

Ordalie is freely available at *http: //www.lbgi.fr/ordalie. Clicking on a given platform (Windows 64 bits, MacOS and Linux 64 bits) downloads an installer for that platform.*

1.2.1 Windows

Run the installer, accept the license and follow the instructions. After install is completed, the installer will automatically launch Ordalie. A desktop icon will be also created.

1.2.2 Linux and Mac versions

The Ordalie installer comes as a zipped tar file. To install Ordalie, please follow these instructions:

```
% tar -zxf setup-linux-x86_64.tar.gz
% cd setup-linux-x86_64
% ./install.sh
```

By default, installation is made in `/usr/local/`*. By sure to have privileges before installing. Alternatively Ordalie can be installed in the user directory or in any specified directory.*

2 Introduction

2.1 Context

Proteins, one of the fundamental building blocks of life, can be classified into various hierarchical categories based on their structural and functional similarities. This classification helps scientists understand protein evolution, function, and relationships. The concept of protein family has been established in the 70's where few protein sequences and structures were known and most of them were

small and constituted of a single domain. Since then, the massive increase of protein 3D structures and sequences led to more subtle definitions, like superfamily or sub-family organizations.

- *Super family : A protein superfamily is the broadest level of classification in the hierarchy of protein classification. Members of a superfamily share distant evolutionary ancestry and often have a common structural fold or domain, but they can perform a wide range of functions. Superfamilies may include proteins that perform vastly different biological roles but still exhibit similarities in their protein structure or specific domains.*
- *Family : Within a protein superfamily, proteins are further organized into families. Protein families are more closely related in terms of sequence, structure, and function compared to superfamily members. Members of a protein family typically share a common evolutionary origin and structural features. While they may perform similar functions, there can still be some functional diversity within a family due to subtle variations in sequence or structure.*
- *Sub-family : The most specific level of classification is the protein subfamily. Subfamilies are groups of proteins within a family that are even more closely related in terms of sequence, structure, and function. Members of a subfamily often perform very similar or identical functions and may have evolved more recently from a common ancestor.*

This introduces a granularity in the protein family concept, providing several scales for analysis that allow for the identification of the zones or residues responsible for this granularity.

2.2 Ordalie

Ordalie (ORDered ALignment Information Explorer) is an interactive tool designed for the exploration of the informational content of a Multiple Sequence Alignment (MSA) into a hierarchical manner, and within different contexts, such as phylogeny or 3D structure.

The Ordalie philosophy (see fig. 1) resides in its capacity to perform a concomitant multi-scale analysis across three axes: the amino acids sequence axis, the taxa axis, and the contexts axis.

The information running along the amino acid sequence (horizontal axis in figure 1) can be considered according to several scales:

- *Large-scale features: domain organization, conserved regions,*
- *Middle-scale features: low-complexity regions, secondary structures, recognition patches, motifs,*
- *Small-scale or local features: post-translational signals, mutation positions, residue conservation.*

Another analysis axis resides in the way the different taxa present in the alignment are handled. The study can be done at a global level (all taxa) to characterize the whole family through different features, such as conserved motifs or key signature, it can also be done on a particular taxon to identify and

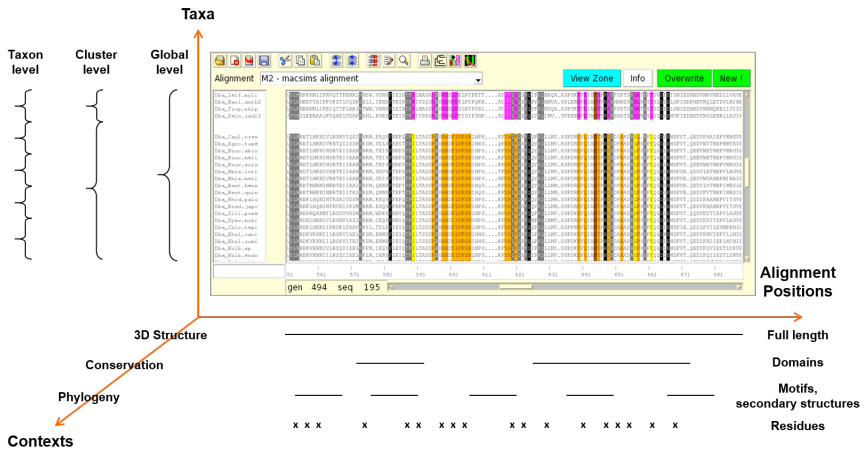

Figure 1: Diagram of the Ordalie philosophy

specify point mutation positions, or at an intermediary level to study the features allowing sub-family identification, such as differentially conserved residues between the sub-family and the other taxa.

As a third analysis axis, Ordalie embeds tools allowing different analysis contexts: residue conservation computation, phylogenetic tree computation and rendering, external features mapping, a 3D structure viewer, etc. ... All analyses can be done in a structural context, as all available features can be mapped and compared on the available 3D structures present in the alignment. As a conclusion of this short introduction, the strength of Ordalie for a protein family analysis resides in the cross-comparison of all information seen in different contexts and at different scales. By adjusting the coarseness of the scale (all taxa, a subgroup of taxa, or a taxon alone for example), the resulting information will help in deciphering different aspects of the sequence - structure - function - evolution relationships for the protein family under study.

3 Ordalie Basics

This section will shortly present some of the fundamental aspects of Ordalie. Most of the following sub-sections will be treated in more detail in subsequent sections of this manual.

3.1 Alignments

Ordalie can read and write FASTA, MSF, RSF, ClustalW ALN, Macsim/XML and ORD (Ordalie file format) file formats. When loading the alignment file Ordalie check for consistency between file extension and the inner file format, and raise an error if the two differ. Upon alignment loading basics information are logged in the "Log Console" for the global alignment and for each sequence: length, isoelectric point, ... The pairwise identity scores are also logged for the global and for each group of sequence if present.

3.2 Snapshots

For a given alignment loaded in Ordalie, it is easy to understand that many different instances of the same alignment may exist. One instance could have a given set of sequence clustered with a given sequence conservation computation, and another instance could have another set of clusters, in order to estimate different hypotheses. These instances are called "snapshots" in Ordalie and can be annotated, saved and retrieved at any time (see 3.5.2). This is made possible thanks to the database embedded in Ordalie.

3.3 Sequences names

Once the alignment is loaded, Ordalie tries to recognize if the sequences names are UniProt, RefSeq, or Protein Data Bank (PDB) accessions names. If a sequence name is prefixed by a database identifier (for example, `sw` *for SwissProt,* `gi` *for Gene Identifier,* `PDB` *for PDB) the prefix will be removed by default. Thus, the sequence name* `>sw|P12345` *will appear as* `P12345` *in Ordalie. The list of recognized bank prefixes and their separator can be changed through the 'Preferences' menu item.*

If sequence names are proper databases accession, Ordalie can then fetch information on these databases upon request (see 5.4.4). Ordalie is dedicated to the analysis of protein multiple sequence alignments. Although it can read DNA/RNA alignments, most of its functionalities will be disabled for such sequence types. Ordalie can still be used to view or edit such alignments.

3.4 Conventions

3.4.1 Mouse Buttons

In this manual, the mouse left, middle and right buttons will be designed as <B1> or <Button-1>, <B2> or <Button-2>, <B3> or <Button-3> respectively. Any words enclosed by '<' and '>' refer to the corresponding keyboard key.

3.4.2 Selections

Sequence names selection and amino acid sequence range selection are always achieved using the same mechanisms:

3.4.2.1 Sequence names selections

Sequences names can be selected by left-clicking on their names. The selection mechanism obeys standard rules:

Keys	Action
<Button-1>	Selects the sequence under the mouse pointer
<Control> + <Button-1>	If the sequence name under the mouse pointer is UNSELECTED, a If the sequence name under the mouse pointer is SELECTED, rem
<Shift> + <Button-1>	Adds all sequences from the previously selected one up to the curre
<Control + a>	Selects all sequences
<Control> + <x>	Cut selected sequence(s)
<Control> + <c>	Copy selected sequence(s)
<Control> + <v>Paste sequence(s)	

Table 1: Keys combination to manipulate sequences names

Sequences Cut/Copy/Paste is available at any time, and allows the user to duplicate, remove or change sequence order.

If a sequence is duplicated using Cut, Copy then Paste, its name will be suffixed by _ _ <n> where n is the copy number.

3.4.2.2 Selecting a residue range

By default, no residue selection or edition is allowed. This can only be achieved within particular tools, like 'Editor', 'Cluster', 'Phylogenetic Tree', or 'Superposition' tools. In such mode, zones of residues are selected by:

Keys	Action
<Button-1>	Sets the starting point of the zo
<Button-3>	Sets the end of the zone, the sel
<Control> + <Button-3>Unselects the zone under the mouse pointer.	
<Control> + <Button-1>	Selects the feature under the mo
<Control> + <Button-3>	Unselect the feature under the n

Table 2: Keys for amino acid sequence zone selection

Several zones can be defined one after the other, either by left/right clicks and/or feature selection.

It is possible to select the zone corresponding to a feature item (for example a PFAM domain) by clicking on this feature item with <Control + B1>.

3.4.3 The database and the Ordalie file format

In order to manage snapshots, features, 3D structures, etc... Ordalie internally embeds a SQLite database [3]. This database is lightweight, and can easily be copied or moved around. The Ordalie file format (. ord extension) is in fact the SQLite database itself.

The scheme of the database can be found in Appendix 6.1.

In short, the database contains:

- *All Ordalie state variables, thus allowing to restart an Ordalie session with the same settings.*
- *All sequence information not linked to the amino acid sequence (length, isoelectric point, description, ...).*
- *The snapshots, with their sequence composition and associated clustering.*
- *All features attached to sequences.*
- *All information related to the PDB entries present in the original alignment or added later on (headers, atomic coordinates, superposition matrices, ...) and all the associated 3D objects.*

As an Ordalie files (the SQLite database) contains all the information, it should be preferred as being the default working format.

3.5 The Main Window

The Ordalie main window can be separated in several parts, from top to bottom (see fig. 2).

3.5.1 The Menus and the icons bar

All the different menus are described in detail in section 5 of this manual. In short, the "File" menu manages input/output files, as well as printing. The "View" menu controls the appearance of the user interface. It contains options to toggle on or off parts of the main window, to change the font size, or to toggle the full-screen mode. The "Sequence" menu allows changing the sequence names, browse, edit or retrieve sequence information, search for sequence motif, compute sequences identity. The "Alignment" menu gives access to all tools linked to the alignment: alignment editor, clustering, phylogenetic tree, features editor, ... The "Structure" menu is dedicated to the structural analysis of the alignment if any sequence corresponding to a 3D structure is present. The menu gives access to a structure superposition module, the 3D viewer, a secondary structure colouring scheme according to sequence conservation, and allows saving PDB

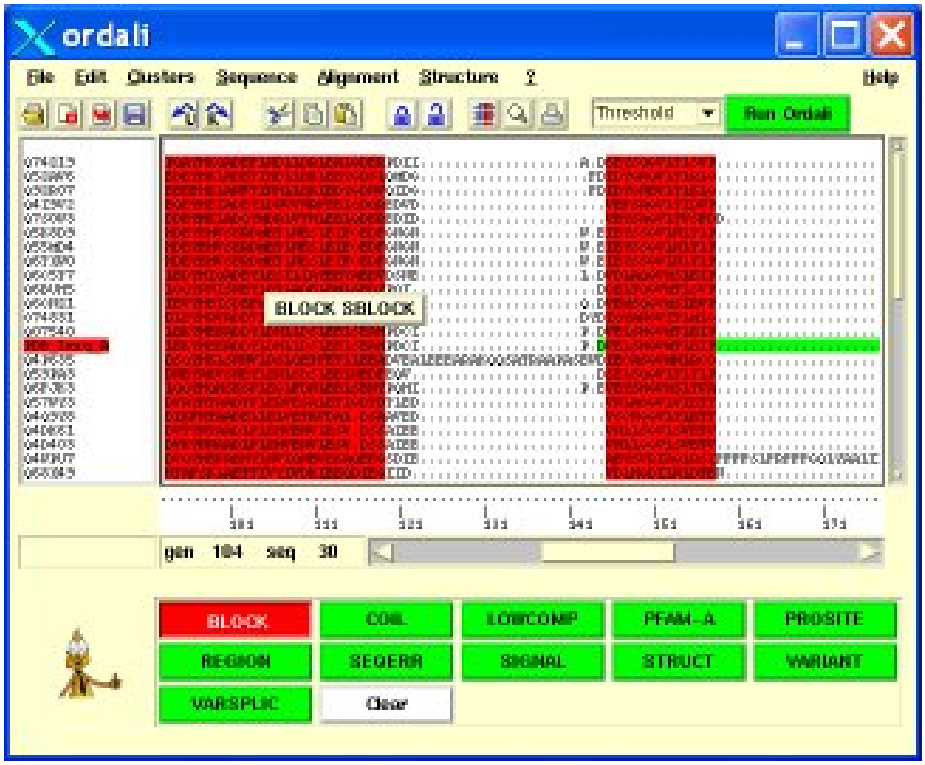

Figure 2: Ordalie main window

files. Finally, the "?" menu allows the access to the on-line documentation and Ordalie version information. Below the menus, the icon bar gives direct access to some of the most useful menu items. When the mouse pointer is above a button, a small message box describing the button's action appears.

3.5.2 The Snapshot bar

As previously mentioned, working with an alignment may lead to several trials in terms of sequence clustering or even amino acid alignments. A trial can be saved as a snapshot of the loaded alignment. A given snapshot can also contain a different set of sequences than the original loaded alignment in case of deletion or addition of sequences.

In the snapshot bar from left to right, the combobox allows selecting a given snapshot. The "Annotation" button shows or hides the annotation of the current snapshot if they exist. Annotations are created through the "Annotate snapshot" item in the Alignment menu (see 4.12). The "View Zone" button toggles the zone used to make the clustering of the given snapshot if it has been clustered. The "Info" button pops up a window the information relative to the snapshot. This information is sought when creating the snapshot. The "Reset" button will reload the current snapshot which will erase all changes made so far. The "Overwrite" button saves the current changes to the current snapshot while the "New" button creates a new snapshot.

3.5.3 The Snapshot Frame

The sequence names are displayed on the left part of the frame, the amino acid sequences on the right part.

<Mouse-wheel> scrolls names and sequences up and down.
<Control> + <Mouse wheel> scrolls the amino acid sequences horizontally.

3.5.3.1 Sequence names

The sequence names highlighted in red correspond to PDB sequences. If there is information associated to a given sequence (present in Macsim/XML, ORD files or retrieved on-line, see 5.4.4) a yellow message window containing a description of the current sequence appears above the sequence pointed by the mouse pointer. A right-click (mouse button-3) on a given sequence name displays a more detailed message window containing the accession, the bank ID, the organism, the length and the description of the sequence. Below the sequence names, an entry box allows the user to search a sequence by its name, or part of its name. After hitting <Return> the first sequence found will be displayed as the top sequence in the window.

3.5.3.2 Amino acid sequences

The right part of the frame contains the alignment itself (amino acids sequences), the ruler, indicating the position of the column, the horizontal and vertical scrollbars and the position counter. Any mouse motion above the amino acid sequences will update the position counter that shows two positions for the residue below the mouse pointer: the 'seq' position is the position of the residue inside its sequence, the 'gen' position corresponds to the position of that residue inside the snapshot.

The position within the sequence is referred to as the *local position*, the position within the snapshot is referred to as the *global position.*

When a given feature is displayed, moving the mouse over the feature will display the note associated with it, for example, in the case of a PFAM domain, the description of the domain will be shown. If there are several features superposed, the first description corresponds to the top feature.

3.5.3.3 Moving along the amino acid sequence

As mentioned above, while the mouse wheel and the horizontal slider allows moving along the amino acid sequence, it is also possible to jump from position to position using the numeric keypad and the left and right arrows. For example, by typing '200' + <Right Arrow> key, the window will go 200 positions to the right. Similarly, typing '500' + <Left Arrow> key will scroll the alignment 500 positions to the left.

3.5.4 The Scores frame

This frame is not shown by default, but it can be toggled on or off using "View -> Show/Hide scores" menu. When residue conservation has been computed, a score is assigned to each column of the snapshot at the global level and at the groups level if available. The Scores frame shows these normalized scores (between 0 and 100) for each column, the colour of the score line corresponding to the group colour, the black line corresponding to the whole snapshot.

3.5.5 The Control Panel

The Control Panel is at the bottom of the main window. By default, the frame only contains a welcome message. When available, this frame contains buttons corresponding to the available features of the current snapshot, one button per feature. Pressing a button will colour the button in red, and display the feature on the snapshot. Pressing the button again will turn it to green and remove the feature.

When changing tool, the content of the Control panel will change according to the tool. The content of the Control panel will be described in each tool section.

The features are displayed in the order the buttons are pressed.
To put a feature over another one, play with the buttons !

3.6 Features

Features *are a central concept in Ordalie. A* Feature *can be defined as a characteristic attached to a zone of a sequence, a group of sequences or to the global snapshot. A sequence / group / snapshot feature can contain several items (for example, a sequence feature can contain several PFAM domains). One of the strength of Ordalie is its ability to investigate these features in different contexts, for example in the structural context of the protein.*

Features are imported into Ordalie through the Macsims program XML output file [15], or using a dedicated feature file format (see section 6.4) or defined by the user through the Features Editor tool (see 4.10).

4 Ordalie Tools

Ordalie contains a collection of tools that can be called at any time, and that can stay alive whatever happens in the main window. It is necessary to leave a tool before using a new one.

4.1 Snapshot Overview

The 'Snapshot Overview' is one of the available schematic representation of a snapshot. When launched, a window appears with the schematic snapshot at the top and a control panel at the bottom. Any number of Overviews can be launched for a given snapshot.

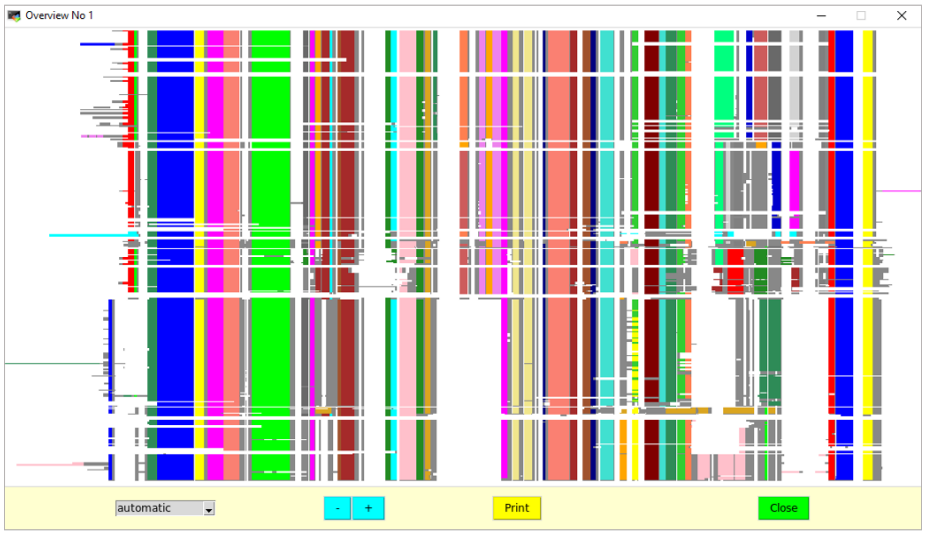

Figure 3: The Overview tool window

4.1.1 The snapshot frame

In this frame, the alignment is schematized by replacing every residue by a grey pixel over a white background. It is then possible to map any feature on top of this scheme. Clicking anywhere on the scheme will automatically centre the main snapshot window on the corresponding position. On the scheme, a stippled rectangle encompasses the region shown by the main window.

4.1.2 Control panel

Below the snapshot frame, a control panel allows interaction with the scheme.

The combobox on the left is a feature selector. By default, it is set to 'automatic', meaning that all features drawn or removed in the main window will automatically be drawn or removed in the Overview window. Selecting any feature in the combobox will simply display it on the Overview.

The '+' and '-' buttons will zoom in and out the scheme.

The 'Print' button will output a PNG file of the schematic snapshot in its current state. The user is prompted to give an output file name.

The 'Close' button will close the 'Overview' window.

4.2 Editor

A fruitful exploitation of multiple sequence alignments (MSAs), which ensures high-quality data usage by analysis tools and feature mapping, is directly dependent on the alignment's accuracy. Although research on dedicated alignment algorithms is still intensive and the resulting software is becoming increasingly accurate, the need for manual MSA inspection, curation, and editing remains essential. This is why Ordalie integrates a high-performance sequence editor, inspired by SeqLab from the GCG Wisconsin Package.

Entering the Editor tool will first clear the sequence from any displayed feature and colour the sequences according to physicochemical properties. The default colouring scheme is:

Figure 4: Amino acids colouring scheme in the Editor

The default scheme can be changed inside the 'Edit -> Preferences' menu item.

As the sequences are changing upon edition, some functionalities will not be available in the Editor.

4.2.1 Control panel

At the bottom of the Ordalie window, the Control panel will display the following buttons from left to right:

Figure 5: The control panel of the Editor tool.

4.2.1.1 Clear

Clears any current sequence selection.

4.2.1.2 Group

This will group the selected sequences. The names of the grouped sequences will be coloured in a different colour, a unique colour for a given group. If one or more sequences already belong to a group, the user should decide through a

dialogue box, if the sequences should be merged with an existing group, or if a new group should be created.

Grouped sequences will behave as a single sequence.

4.2.1.3 Ungroup

The selected sequences will be removed from any group. If a group consists of only one sequence, the group is automatically destroyed.

4.2.1.4 Lock/Unlock

By default, only gaps ('. ') can be deleted or inserted, it is not allowed to insert/delete amino acids. Unlocking the sequences allows the insertion/deletion of residues.

4.2.1.5 Rem. Col. Gap.

"Rem. Col. Gap." stands for "Remove columns of gaps". Ordalie runs through all snapshot columns and removes those containing only gaps. A 'Rem. Col. Gap. ' is automatically done when leaving the 'Editor' tool.

4.2.1.6 Temp. Save

This creates a TFA file copy of the current snapshot under edition. The user is asked for a file name the first time, and successive 'Save' will use this file name to output the snapshot.

4.2.1.7 Save & Return

Leaves the 'Editor' with all the changes made. A dialogue box will ask the user if the current snapshot should be overwritten, or if a new snapshot should be created with the current changes.

4.2.1.8 Cancel

Leaves the 'Editor' and restore the original snapshot.

4.2.2 Edition actions

Clicking inside the snapshot frame will show a yellow blinking cursor. The following actions are then available:

- *The <Left>, <Right>, <Up> and <Down> arrows move the cursor accordingly.*
- *<Backspace> deletes the gap/character before the blinking cursor. In lock mode, only gap characters can be deleted.*
- *<Shift-Right> pushes contiguous characters placed immediately at the right of the blinking cursor up to the next gap character. When dealing with grouped sequences, the next gap is the position which contains a gap for ALL sequences of the group.*

- *<Shift-Left> pushes contiguous characters placed at the left of the blinking cursor up to the next gap character. When dealing with grouped sequences, the next gap is the position which contains a gap for ALL sequences of the group.*

In order to speed up editing, pressing keypad or keyboard digits [0-9] stores the number in a buffer, i.e. pressing '1' then '2' will store number 12 in the buffer. Pressing then the <Left> arrow will move the cursor 12 characters to the left, and will empty the buffer. All the actions previously described can take advantage of this mechanism.

4.3 The Identity tool

This tool is used to query information on identity percentages between sequences or between zones of sequences.

4.3.1 Control panel

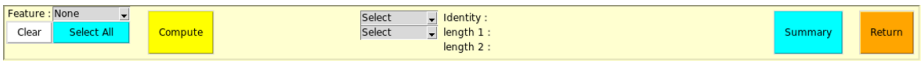

Figure 6: The control panel of the Identity tool

.

4.3.1.1 Selection

The identity percentage can be computed for all or some selected sequences and over all or a user defined residue range. The left part of the Control panel deals with sequences and residue range selection.

- *Feature : displays the available feature. The user can then select one or more feature items as a residue range.*
- *Clear : clears all residue ranges and all sequences previously selected.*
- *Select All : selects residues from the whole snapshot.*

4.3.1.2 Computation and Results

The 'Compute' button calculates the identity percentage between selected sequences for the selected residue range. A summary of the computation is logged and is available through the Log console. The selection of two sequences for which the identity percentage is desired is done with the following two comboboxes. The identity percentage and the length of the two ungapped sequences is then given.

The 'Summary' button will make a window appear that will give for the whole sequence and for each group :

- *The average identity percentage,*
- *The standard deviation,*

- *The pairs having the maximum and minimum identity percentage.*

The 'Return' button will leave the Identity tool.

4.4 The Search motif tool

This tool allows the user to search for a particular sequence motif inside the current snapshot.

4.4.1 The search pattern syntax

The syntax of the search pattern follows the rules of the `FindPatterns` *program of the GCG Wisconsin Package [20]. A detailed description of the syntax is available in Appendix 6.2.*

4.4.2 Control panel

Figure 7: The Control panel of the Search motif tool

The Control panel is limited to the motif entry box in which the pattern should be entered, the 'Search' button to launch the search, the 'Find Next' button to go to the next occurrence of the motif, and the 'Return' button to leave the search tool.

When a motif is found, the background of the snapshot window will become black, and all the instances of the motif will be highlighted in red. The "search motif" is the only tool that can be called within the "editor" tool.

4.5 The Clustering tool

The 'Clustering' tool allows the creation of clusters (or groups) of sequences based on numerical criterions characterizing the sequences to be clustered. The computation can be done using all or part of the sequence as well as all or part of the snapshot columns. The user chooses one or more numerical criterions as the basis for the computation and a clustering algorithm. The computation can then be launched and the newly created sequence clusters are automatically displayed in the main window. The different clustering trials are temporally kept while using the clustering tool, but a given clustering can be saved as a new snapshot.

4.5.1 Criterions

At present, the available criterions are:

- *Identity percentage : for each sequence, the identity percentage computed over the selected residue range against all other selected sequences.*
- *Length : the sequence length.*
- *Hydrophobicity : only available for Macsim alignments.*

- *Isoelectric point (pI) : the isoelectric point is computed using EMBOSS pKa values for amino acids,*
- *Amino acid composition : the relative percentage of the 20 amino acids for a given sequence.*

4.5.2 Algorithms

Ordalie clusters and automatically defines the number of groups. The clustering algorithms along with the algorithms that define the number of clusters are taken from the Cluspack package. The available methods are:

- *kmeans / DPC (Density of Points Clustering) [18]: The clusters identification is done using a point density criteria. The actual cluster selection is done according to the k-means algorithm.*
- *Hierarchic / Secator [19] : The groups are identified through an ascendant hierarchical classification. The cluster selection is done using an inertia loss criterion.*
- *Mixture Model [11] / AIC [1] or BIC [12] : After a Gaussian modelling of the data distribution, the clustering is done according to AIC or BIC criteria.*

4.5.3 Control panel

Figure 8: The control panel of the Clustering tool.

4.5.3.1 Selections

The left part of the Control panel deals with sequence and residue range selection.

- *Feature : displays the selected feature. The user can then select one or more items of this feature as a residue range.*
- *Clear : clears all residue ranges and all sequences previously selected.*
- *Select All : selects all columns of the current snapshot.*

If no sequence names are selected, the clustering will use ALL sequences. If some sequences are selected (more than 3), then the clustering will only apply to these selected sequences. The remaining ones will be kept as a separated group.

4.5.3.2 Clustering criterions

The pull-down menu allows the selection of the criteria to be used for the computation. Several criteria can be selected at the same time.

The 'Life domain' criterion clusters the sequences into Eukaryota, Archaea, Prokaryota and Other groups. This criterion cannot be associated with another one.

4.5.3.3 Clustering methods

The 'Method' pull-down menu permits to choose the algorithm to be used for clustering computation.

- *Hierarchic clustering / secator,*
- *kmeans / DPC (Density of Points Clustering),*
- *mixture model / AIC criterion,*
- *mixture model / BIC criterion.*

The "Compute" button will launch the computation. The newly computed sequence clusters are directly displayed in the main Ordalie window.

4.5.3.4 Other buttons

The 'Reset' button will erase any clustering done so far and show the original clustering if any. The 'No Clusters' buttons removes all groups and leaves all the sequences as a single group.

The "Clusters Names" button will pop up a window allowing to give a name to each cluster. This cluster name may be used in subsequent analysis to identify the clusters, like in the "Tree" display, or the "Barcode" tool.

The 'Save' button will leave the clustering tool and the current clustering will be saved. The user is prompted whether to overwrite the current snapshot or to create a new one. The 'Return' button leaves the Clustering tool and displays the snapshot in its original state.

4.6 The Tree tool

The 'Tree' tool can be divided in two part. The first part consists in the tree building, which is done through the main Ordalie window. Once the tree is computed, its exploitation will be done in a dedicated new window.

The tree is computed using the FastME program [9] using default parameters. Ordalie computes first a distance matrix based on identity percentages calculated over the selected residue range. Although Bayesian based algorithms seem to produce more accurate trees, FastME is a good compromise between speed and accuracy.

4.6.1 Control panel for tree building

Figure 9: The Control panel of the Tree building tool

4.6.1.1 Selections

The left part of the Control panel deals with sequence and residue range selection.

- *Feature : displays the selected feature. The user can then select one or more feature items as a residue range.*
- *Clear : clears all residue ranges and all sequences previously selected.*
- *Select All : selects residues from the whole alignment.*

4.6.1.2 Options

The following buttons can be used to control the tree computation:

- *With PDB seq : includes PDB sequences in the tree computation. By default, Ordalie doesn't use PDB sequences as they are supposed to have their original sequence inside the snapshot.*
- *Pairwise / global : defines the type of gap removal algorithm to be used. 'Pairwise' (checkbutton on) means that, for each pairs of sequences, positions containing gap are excluded from the computation. 'Global' (checkbutton off) means that only complete columns of residues will be taken to compute the tree.*
- *Load Tree : it is possible to import a tree file into Ordalie. The tree file should be in a NEXUS format, and the tree leaves identifiers should match all or part of the sequence names present in the snapshot.*
- *Bootstrap : By setting the 'Bootstrap' checkbutton on, Ordalie will perform <N> bootstraps, N being the number entered in the text field located below the 'Bootstrap' checkbutton. Ordalie pre-computes an ad-hoc value for the bootstrap, the value being equal to 1.1 times the total number of sites used to compute the tree. A loaded tree can also be bootstrapped.*

4.6.1.3 Draw / Return

The 'Draw' button launches the computation, and draws the resulting tree in a separate and dedicated window. The 'Return' button leaves the Tree tool.

4.6.2 The Tree window

Each newly computed tree will appear in a new and dedicated window, that allows the exploration of the tree characteristics. Ordalie can to render two types of trees: dendrograms and radial trees. Some of the following options are specific to one or the other tree representation (see below).

The upper part of the tree rendering window is the drawing area, and the bottom part the control panel area.

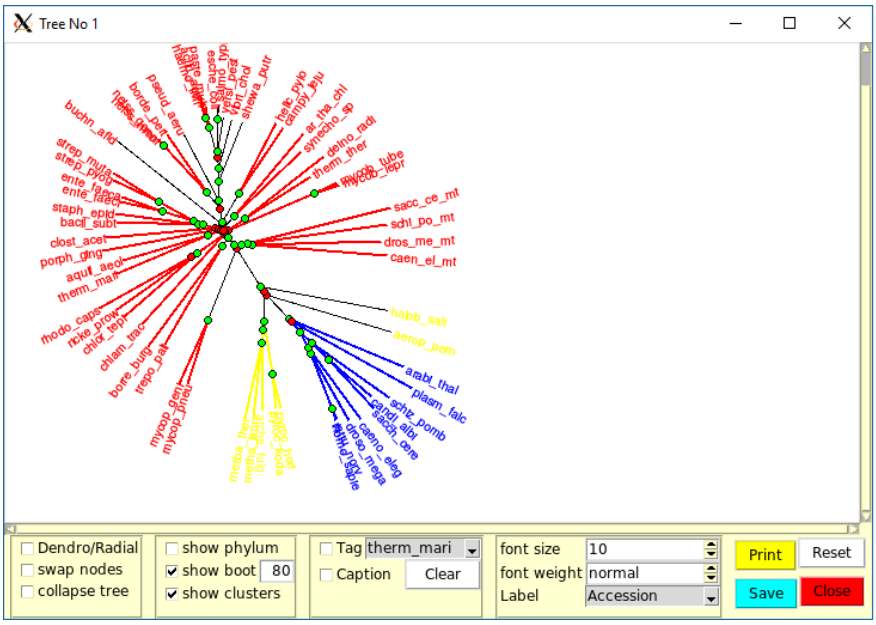

Figure 10: The Tree rendering window displaying a radial tree. The circles at each node indicate whether the bootstrap value for the node is higher (green) or lower (red) than the defined threshold. The sequences are coloured according to their cluster.

4.6.2.1 The Drawing area

The drawing area displays the current tree. The tree can be moved in all directions by simply dragging the mouse with <Button-1> down. A radial tree (see below) can be scaled by using the mouse wheel, while using the mouse wheel on a dendrogram will scroll up and down the tree. Finally, a right click <Button-3> will make a contextual menu appear which allows changing the dimensions of the tree. If the tree is a dendrogram, then the branch length and the height separating branches can be changed. If the tree is a radial tree, the tree can be rotated.

4.6.2.2 The Control panel

The control panel is divided in several parts. From left to right:
General tree options:

- *Dendrogram/radial : change the type of tree representation.*
- *Swap node : when activated, all the nodes of the tree present a little orange disc. By clicking on it, the branches going out from this node are swapped (rotated) around this node.*
- *Re-root tree (only for dendrogram) : when activated, all the nodes of the tree are marked with a little green disc. Although the tree is an unrooted tree, it is possible to define a new root (the two outmost branches) by clicking on a disc.*

Adding information to the tree representation:

- *Show Phylum : if the phylum information is available for the sequences, then the sequence names are coloured according to their life domain : eukaryots in red, archaea in blue, bacteria in yellow, and viruses in black.*
- *Show bootstrap : if the bootstrap computation has been done, each node will display a disc coloured in green or red depending on whether or not its bootstrap value is higher than the threshold (in %) defined in the text field next to the 'show bootstrap' checkbutton (default value 80%). Green circles correspond to boostrap values above the threshold, red circles correspond to values below. In the case of a dendrogram, the absolute value of the bootstrap is also written at each node. In radial tree, pointing the mouse over a disc will display the absolute value of the bootstrap.*
- *Show clusters : if the snapshot has been clustered, then the branches of the tree will be coloured according to the group they belong to.*

Tags and tree annotation:

- *Tag : the leaf specified in the combobox aside the checkbutton 'Tag' will be surrounded by a thick black box. This allows the quick identification of a sequence in case of a furnished tree.*
- *Caption : this will add a caption to the tree. The first time this option is invoked, or if the 'Clear' button is pressed, a window will appear to let the user customize the caption.*

Leaf labels:

- *Font size : changes the size of the font used to display labels (sequence names).*
- *Font weight : changes the font between normal and bold.*
- *Label : defines the 'text' to be displayed as leaf labels which is by default the sequence name. It may be useful to show the accession number, or the accession number with the species, etc ...*

Buttons:

- *Print : makes a PNG file of the current state of the tree.*
- *Reset : resets the tree to its original configuration.*
- *Close : closes the window.*

4.7 The Conservation tool

Traces of the evolution pressure that maintains the structure and function of a protein family can be found while examining the residue conservation along the snapshot. Both global and group conservations may help in deciphering functional sites like binding sites, interaction patches, or specialization coupled with intra-groups organization.

Ordalie offers several methods to compute conservation. Within this tool the user can try several methods to compute residue conservation. The results are temporarily kept until they are saved. A saved residue conservation computation becomes then a new feature attached to the current snapshot and can, as any other feature, be used in any tool allowing feature display.

4.7.1 Methods

Many methods exist to compute conservation, and they have been tested and compared extensively [17, 4]. Ordalie implements some of them, as well as two home-made conservation methods.

4.7.1.1 The 'Threshold' method

This method is essentially a counting method. Different levels of conservation are defined thanks to two pre-set thresholds. At a global level, a column with 100% conserved residues (the 'identity threshold') is assigned to 'identity conservation'. A column with at least 80% conservation (the 'global threshold') is considered a 'conserved' column. Within a group, however, only 'identity conservation' is considered. The thresholds can be changed through the 'Preferences' menu.

4.7.1.2 The automatic methods.

In these methods, only columns containing more than 5 residues are considered, and the computation proceeds through two steps. First, all the columns are scored with the chosen method. In a second step the columns with their associated scores are clustered, and the clusters are ranked according to their mean conservation scores. The two clusters containing the highest scores are considered to contain the columns corresponding to 'strictly conserved' and 'globally conserved' residues. The same computation is repeated for each group, but only the cluster with the highest scores is taken. The available automatic methods are:

- *BILD : Bayesian inferred score, see [2],*
- *Liu : taken from Liu* et al. *[10]*
- *Mean Distances : the algorithm implemented in ClustalX [13]*
- *Vector Norm : this method uses a physicochemical representation of amino acids based on their volume and polarity. The score is proportional to the most present amino acid in the column. The method is described in detail in the Appendix 6.3.*
- *Multi : each column is scored using several scores (here the 'Vector norm' and 'Mean Distance' scores) before being clustered.*

4.7.2 The Control panel

From left to right in the Control panel:

- *Method : the scoring method to be used.*

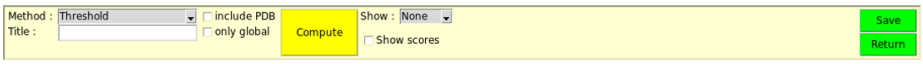

Figure 11: The control panel of the Conservation tool

- *Title : an optional title for the ongoing conservation calculation can be added.*
- *Include PDB : by default, the conservation calculation does not include PDB sequences as their genomic sequence is usually also included in the snapshot.*
- *Only global : if checked, the conservation will only be performed for the whole snapshot, conservation inside groups will be discarded.*
- *Compute : launches the computation.*
- *Show : Each computed score is saved temporarily and can be recalled using this combobox. By default, the score is called 'tmp<Method>-<x>' where <Method> is the method used and <x> an index. If the score is saved, its name will change to '<Method>-<x>'.*
- *Show scores : opens the Scores frame just below the alignment and displays the scores as a graph. The drawn scores correspond to the scores currently selected.*
- *Save : save the current score (the score indicated in the 'Show score' combobox) along with its 'Title'. The saved score name will be changed in that combobox, and the score will appear as a new feature in the 'Normal' mode.*
- *Return : returns to the 'Normal' mode.*

4.8 The Superposition tool

One of the strengths of Ordalie resides in its ability to link/map features to the 3D models (when available) of proteins. To exploit at best the feature mapping it is essential to proceed in the scope of the structural differences observed between proteins. To achieve that, Ordalie can superpose the structure according to feature, and/or user defined residue range.

4.8.1 The superposition algorithm

A protein structure can be made of several chains, which may be identical or not. A chain is usually composed of an amino acid polymer and ligands (in Ordalie, water molecules are considered as ligands). It is important here to understand that, although superposition computation are done using the polymer sequences, the entities that are moved (superposed) in Ordalie are the entire chains.

The chain superposition is done in three steps:

1. *Selection of the superposition zone. Depending on the structure, the zones may consist of an entire domain, or of selected secondary structures, for example.*

When applying a superposition to a chain, all residues of this chain (polymer AND ligands) are moved.

2. *Selection of the superposition zones. Depending on the structure, the zones may consist in a domain, or some selected seconddary structures for example.*
3. *Selection of the chains that would be superposed.*
4. *Selection, between the chains selected for superposition, of the reference chain. The reference chain will not move, all the other selected chains will be superposed onto it.*

The detailed superposition algorithm is presented in Appendix 6.5.

4.8.2 Control panel

Figure 12: The Control panel of the Superposition tool.

4.8.2.1 Selection

From left to right the superposition Control panel is made of:

- *Features : display the selected feature. The user can then select a feature item as a residue range.*
- *Clear : clears all residue ranges and all sequences previously selected.*
- *Select All : selects residue from the whole snapshot.*
- *All Helices : selects all helices present in the sequences.*
- *All Strands : selects all β-strands present in the sequences.*

The 'All Helices' and 'All Strands' selections will take, for each secondary structure type position, the minimal common part of all existing secondary structures present at that position.

4.8.2.2 Control

- *Superpose : launches the superposition. As mentioned previously, this will be done in two steps:*
  1. *Open the chain selection window where the user should select all the chains concerned by the current superposition.*
  2. *When done, the Reference chain window will open to choose the reference chain (the non-moving chain) among the previously selected chains.*
- *Display : opens the 3D Viewer (see the 3D Viewer section for details 4.9).*
- *Return : leaves the superposition tool.*

4.8.3 Example : dealing with a homo-dimer.

Suppose the loaded alignment concerns a protein known to be a homo-dimer (an α_2 structure) under biological conditions, and for which several 3D structures of some proteins coming from different organism have been solved. By investigating PDB ID (say 1abc and 1def), it is also known that all structures are made of two chains, A and B. The build alignment contains the corresponding sequences `PDB_1abc_A` *and* `PDB_1def_A`*.*

When loading the alignment, Ordalie will recognize the two PDB ID through their sequence names `PDB_1abc_A` *and* `PDB_1def_A` *and will then download from the PDB website the two structures with atomic coordinates, and store them inside a dedicated database. Note that Ordalie knows the coordinates for all the atoms of ALL chains of the structure, not only chain A.*

Several cases may be encountered when performing a superposition:

4.8.3.1 Only one sequence chain present in the snapshot

The snapshot contains the sequence named `PDB_1abc_A` *and* `PDB_1def_A`*. When superposing* `PDB_1def_A` *on* `PDB_1abc_A`*, only atoms of 1def chain A will change. Thus in the 3D Viewer, the whole structure of 1abc will be correct (its the non moving molecule), and 1def will have chain A on top of 1abc chain A, and 1def chain B somewhere in space. The symetry of the dimer is broken, as only chain A as moved.*

4.8.3.2 Moving a dimer.

Ordalie doesn't know anything about monomers, dimers, multimers in general. It is up to the user to provide the information, by giving Ordalie the sequences of the chains of interest.

To manipulate a multimer in Ordalie, all the sequences corresponding to the chains of the reference AND the sequences of the chains of the target structure should be present in the alignment.

If the alignment contains `PDB_1abc_A`, `PDB_1abc_B`, *and* `PDB_1def_A`, `PDB_1def_B`, *it is then possible to superpose the two dimers. A first superposition step where only* `PDB_1abc_A` *and* `PDB_1def_A` *are selected will bring* `PDB_1def_A` *on top of* `PDB_1abc_A`. *A second superposition step where only* `PDB_1abc_B` *and* `PDB_1def_B` *are selected will bring* `PDB_1def_B` *on top of* `PDB_1abc_B`.

4.9 The 3D Viewer tool

The 3D Viewer is one of the most useful tool in Ordalie. Although it does not offer all the features and functionalities that would a proper Molecular Visualization program like VMD or PyMol would do, it can be of great help in understanding protein features in the framework of protein structures.

4.9.1 Molecules and Objects

The Ordalie 3D Viewer is organized around the 'Molecule' and 'Object' notions. A 'Molecule' consists in all the chains (and consequently residues and atoms) that are present in a given PDB entry. An 'Object' belongs to a Molecule, and can be a composition of several elements (full chains, parts of chain, residues, ligands, etc...) belonging to that molecule. Objects can be painted with several colours and can contain several kinds of representation. Feature mapping only applies to objects.

By default, Ordalie will create 3 objects per molecule:

- *AL<molecule name><chain> which is a stick representation of all the atoms of the chain,*
- *CT<molecule name><chain> which is the Ca(or phosphate) trace of the chain,*
- *CA<molecule name><chain> which is the ribbon representation of the chain.*

At present, Ordalie does not handle hydrogen atoms.

4.9.2 Representation types

Ordalie is able to represent a structure in several ways.

- *Ribbon : the Ca/P smoothed ribbon. By default, the path atoms are the Ca and P for amino acids and nucleic acids respectively.*
- *Ca/P trace : a simple link between Ca or P atoms.*
- *CPK : each atom is represented as a sphere whose radius is the VDW radius of the atom.*
- *Pearl : the residue is simply represented by a solid sphere placed at the centre of mass of the residue.*
- *Atoms : a wireframe representation. Standard residues are drawn according to their topology. All other compounds (modified residues, ligands, ...) will be drawn according to a distances criteria. Depending on the quality of the structure, this may lead to chemically wrong atomic bonds.*

4.9.3 The 3D Viewer window

The 3D Viewer window can be divided in 4 parts. The top of the window is used to display information about picked atoms. Below is the 'Quick Mapping' panel. Below this panel, from left to right are the 'Molecular Objects' frame, the main 3D window, and the 'Actions' panel at the right. The 3D window can be maximized by hitting the <F1> on the keyboard, and hitting <F1> again gives the window its original geometry. All panels may be switched on or off by hitting the <F2> key.

4.9.3.1 Quick Mapping

The four comboboxes of this panel allow making a quick mapping of features on a given molecular object. The left outmost combobox selects the molecule, then the object onto which the feature will be mapped. There are then two features selectors. It is possible to map two features on a same object by selecting one feature with 'Feature 1' combobox, and a second feature with the 'Feature 2' combobox. The features are drawn in order, feature 1 before feature 2. Care should be taken when selecting Feature 1 and Feature 2 as Feature 2 can completely cover Feature 1. For example, if Feature 1 is set to conservation, which implies residues colouring, and Feature 2 is set to PFAM-A, a lot of conservation won't be seen as a PFAM domain extends to a large range of residues. In this case, Feature 1 should be set to PFAM, and Feature 2 to conservation.

4.9.3.2 Molecular Objects frame

Below the 'All On' and 'All Off' buttons that switch on and off all objects that have been defined in all Molecules is the list of all 3D molecules present in the snapshot. Aside the molecule name is the 'New' button that allows the definition of new objects for that molecule (see section 'Object Editor' 4.9.4 below). Clicking on a molecule name will open/close the list of the objects defined for that molecule. An object coloured in green is switched on and is displayed on the screen, a red object name means the object is switch off. Each object name is followed by the 'Edit' and 'Del' buttons, used to redefine and delete the object respectively.

4.9.3.3 The 3D window

This window contains the 3D objects themselves. The objects can be manipulated by the mouse through an arcball system, that is a virtual trackball. All the objects of the scene are enclosed in a sphere, and the objects are moved by dragging the sphere up and down and left to right using mouse Button-1, the mouse mimics the hand that would roll the sphere. The mouse wheel is used to zoom in and out the scene. A right drag with <Button-3> will translate the scene in the x-y plane. A <Control-B1> click will show the label of the atom being below the mouse pointer.

4.9.3.4 The Actions panel

Although the Ordalie 3D Viewer tool is not intended to be a complete Molecular graphics program, it still offers some functionalities which are, from top to bottom:

- *Reset : cancels all ongoing actions (for example, distance definition),*
- *Clear Ids : switches off all atom labels,*
- *Clear Distances : switches off all distances,*
- *Distances : computes the distance between two picked atoms,*
- *Centre On Atom : the picked atom becomes the rotation centre of the scene,*
- *Print : outputs a PS file of the scene,*
- *Full Screen : toggles the window into a full screen window,*
- *Stereo : not yet available,*
- *Close : closes the 3D Viewer window.*

4.9.4 The Object Editor

An object is an ensemble of residues and/or ligands belonging to one or several chains, and displayed in given styles with given colours. The Object Editor can be invoked to create a new object (the 'New' button) or to edit an existing object (the 'Edit' button).

Figure 13: The 3D object creation window.

4.9.4.1 Making / Editing an object

In case of a new object creation, the new object name should be entered in the top entry box. Two objects can not have the same name. The object edition can then be done in a five-step process:

1. *select the chain of interest and the type of residues in the chain : polymer residues (amino acids or nucleotides) or the ligands,*
2. *select a representation type,*
3. *select residues onto which to apply the selected style,*
4. *select a colour,*
5. *select residues onto which to apply the selected colour.*

This process is iterated until all pieces of the object are set up.

Finally, it is possible to add the molecular surface surrounding the object atoms by checking the checkbox at the bottom of the object window.

4.9.4.2 Residue selection

When applying a colour or a representation style, the user should specify the residues it should apply to. There are three ways to do so:

- *the 'All' button will select all the residues of the current chain. This may be useful to give all residues the same colour for example.*
- *The 'Selected' button : this refers to residues which have been selected with the mouse onto the residue frame. In this frame, clicking and dragging the mouse with <Button-1> down will select a range of residues.*
- *The 'Feature' combobox : this will select residues corresponding to the selected feature.*

The object is then finished by clicking on the 'OK' button. The new object will be added to the object list on the corresponding molecule.

4.10 The Features Editor

This tool is dedicated to feature management, feature creation, deletion and edition.

4.10.1 Control panel

The Control panel of the 'Features Editor' is really simple.

Figure 14: The Control panel of the Feature Editor tool

It consists, from left to right, in:

- *Features : displays the selected feature. The user can then select a feature item as a residue range.*
- *Clear Selection : clears all residue ranges and all sequences previously selected.*
- *Delete feature : This will delete the current feature for all sequences.*
- *Load Feat. File : loads a file containing user-defined features. The feature file format is described in Appendix 6.4.*
- *Features Summary : launches the Features summary tool which allows an overview of any feature within the snapshot. See section 4.11.*
- *Save : saves the feature changes and returns to the 'Normal' mode*
- *Return : returns to the 'Normal' mode and restores original features.*

4.10.2 Notation and Actions

It is important to understand the difference between a Feature and an Item of a Feature. Here, a Feature represents a set of instances of a given sequence characteristic that may be distributed over the whole snapshot. A Feature Item, or Item for short, is one instance of a Feature for a given sequence at a given place in the snapshot.

4.10.3 Contextual Menu

Contrary to all other tools, it is possible to interact directly with the features inside the snapshot window. A right click makes a contextual menu pop up, allowing several actions.

In the 'Feature Editor' tool, action of <Button-1> is changed through key combination :
<Button-1> alone : the action applies to the sequence under the mouse pointer.
<Control + B1> : the action applies to the group the sequence pointed by the mouse belongs to.
<Shift + B1> : the action applies to all the sequences present in the snapshot.

4.10.3.1 Select Item

Selects the Item just under the mouse pointer. If only <Button-1> is pressed, then the Item of the sequence will be selected, if <Control + B1> is pressed then all the Items at that position for sequences of the group will be selected, and if <Shift + B1> is pressed all Items appearing at that position for all sequences in the snapshot will be selected.

4.10.3.2 Select All Items

Selects all Items of a sequence, a group of sequence or the whole snapshot depending on the key pressed.

Selecting all Items for all sequences means that the whole Feature is selected. If it is subsequently deleted, then the whole feature will be deleted.

4.10.3.3 Select Region

A region (i.e., a residue range) can be selected by pressing and holding <Button-1> and then dragging the mouse along the sequence axis. Depending on whether no key, the <Control> key, or the <Shift> key is held down, the selected region

will cover the current sequence, the sequence group, or all sequences respectively. The selected region can then be used to define a new Item.

4.10.3.4 Clear Selection

Clears all selections currently set.

After having selected Items(s) or region, several options are then available.

4.10.3.5 Edit Item

If the selection refers to one or several already existing Items, it is possible to change some of their properties:

- *the residue range of the selected Items can only be changed if they refer to only one zone,*
- *the Item Colour,*
- *the Item Score,*
- *the Item Note.*

4.10.3.6 Define New ...

This option will make a window appear, allowing the description of the new item. If the 'Feature Name' entry is filled with an already existing feature, then the new item will be added to the item list of that feature. If the 'Feature Name' does not exist, a new feature is then created. In all cases the user is supposed to give to the item at least a Colour and optionally a Score and a Note.

4.10.3.7 Delete selected Items

This will delete the selected items from the current feature. Note that if all Items of a Feature have been selected, then this option will delete the Feature itself.

4.10.3.8 Propagate Items to this group

The Selected Items will be propagated to all the sequences of the group they belong to. If an Item to be propagated is already present in one or more sequence of the group, the Item will not be propagated.

4.10.3.9 Propagate Items to All

This will propagate the selected Items to all the sequences of the SNAPSHOT. If an Item to be propagated is already present in one or more sequence of the group, the Item will not be propagated.

4.11 The Features Summary

This representation can render several selected sequences and features on the same page. The sequences are not schematized as in the Snapshot Overview representation, but are shown as they appear in the alignment window.

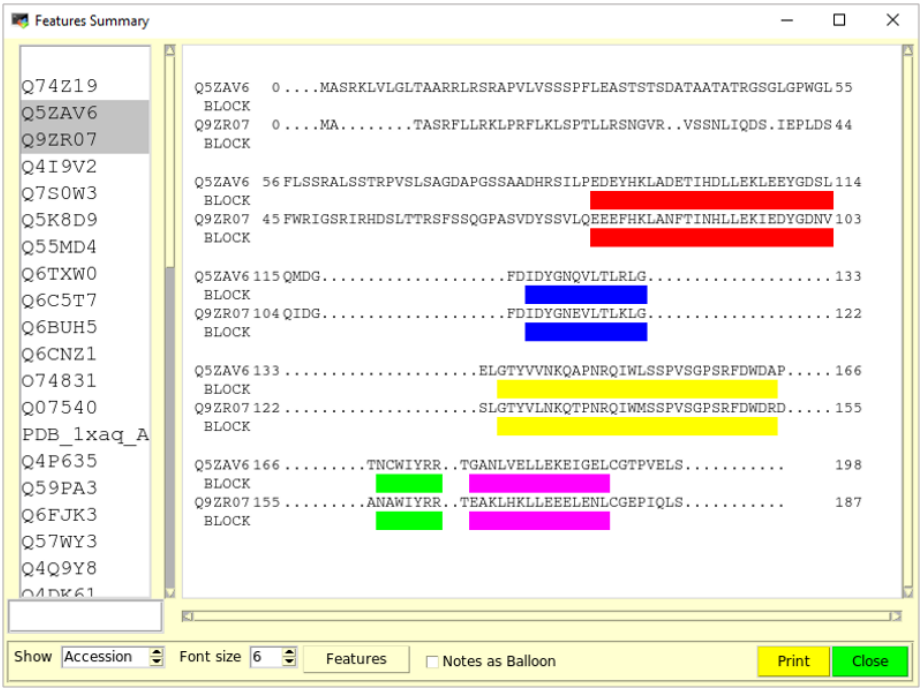

Figure 15: The Features Summary window, with the Drawing Area at the top and the Control panel at the bottom

4.11.1 Drawing Area

The top of the 'Features Summary' window is made of a listbox containing the sequence names on the left, and a drawing area on the right. Clicking on a name selects or unselects the corresponding sequence. Multiple sequences can be selected by holding the <Control> key down while clicking on their names with the mouse.

In the drawing area, for each sequence the sequence ID is written on the left, followed by the position of the first residue in the current sequence line, the amino acid sequence itself as present in the alignment, and the position of the last residue in the line. When a feature is selected, each sequence line is followed by a feature line, with the feature name beneath the sequence ID, and rectangles below sequence positions where the feature is present.

The Features Summary can be moved around by dragging the mouse while holding <Button-1>.

4.11.2 Control panel

Below the Drawing Area is the Control panel. On the left is a spinbox that selects the type of name the sequence should be referenced with, i.e. its sequence name, its accession number or its bank ID, when available. This choice applies in both the listbox and the drawing area. Follows the font size selector, and then the 'Features' selector. Any number of features can be selected by checking the button corresponding to the desired feature. The 'Notes as Balloon' checkbutton

renders or not the note attached to each feature as a flying balloon when the mouse pointer is over the feature. The 'Print' button will ask for a file name that will contain a PNG image of the current drawing area, and the window will disappear by clicking the 'Close' window.

4.12 The Annotation tool

This tool allows adding, modifying and removing "memo tags" anywhere on the snapshot. They can be displayed on or off at any time using the "Annotation" button on the snapshot bar.

A memo tag has three characteristics: its position and size, its colour, and the text note attached to it. The user can use them to annotate a piece of snapshot that will have to be manually re-aligned, to annotate a particular conserved zone of interest for example. The colour can be used as a classification mean, pink for structural annotation, cyan for sequence features, etc...

4.12.1 Control Panel

The Control Panel is fairly simple and is composed from left to right:

- *"Show annotation coloured in" can be used to show a given coloured annotation,*
- *"Show all" will display all available annotation*
- *"Hide all" will remove all annotation of the screen,*
- *"Dismiss" will leave the tool,*
- *"Return" will leave and save all changes made in the annotations.*

4.12.2 Keyboard Bindings

Annotation creation makes use of the mouse. Table 3 shows the actions associated with mouse button combined with keyboard.

Keys	**Action**
<Button-1>	Select the origin of a new annotation.
<Button-1 dragging>	Define the area of the current annotation
<Button-1 release>	Finish the annotation zone definition
<Control> + <Button-1>	Delete the annotation below the mouse pointer
<Shift> + <Button-1>	Edit the annotation below the mouse pointer

Table 3: Keys combination for annotation management

After creating an annotation zone (button-1 released), a window will pop up allowing the user to select the annotation colour and the associated text. The same window appears to edit a given annotation. Clicking <Return> will save all the changes made in annotations.

Figure 16: The barcode window

4.13 Barcode

The "Barcode" tool is another schematic representation of a snapshot. Figure 16 shows an example.

The upper frame of the tool is the drawing area. For each group of sequence of the snapshot a lane will be drawn

5 Menus Description

This chapter describes all the Ordalie menus. Only the entries which have not been already explained in this manual will be described in details here.

5.1 The File Menu

5.1.1 Open

Opens an alignment file. Several file formats are recognized:

- *TFA (Fasta) file format.*
- *MSF (Multiple Sequence File, GCG Wisconsin Package) format.*
- *Macsim/XML format. The Macsim format is an XML file output of the MACSIMS program or server (see [15] and http://www.lbgi.fr/ julie/MACSIMS/ for details).*
- *ALN (ClustalW) file format [14].*
- *RSF file format (Rich Sequence File), as output by the SeqLab program [16].*
- *ORD (Ordalie) file (ORD), Ordalie specific file format.*

A format checking is done using the file extension (. tfa/. fasta, .msf, .xml, .rsf, .aln or. ord recognized) against the file content. In case of inconsistency, the user will be asked for the correct file format.

5.1.2 Save

Save the entire current snapshot in the current file format with the current file name. The first time 'Save' is invoked, the user will be prompted to enter a file name. Subsequent call too "Save" will save the snapshot using the defined name and file format.

5.1.3 Save As ...

Ask for a file format and a file name to save the current snapshot. If there are selected sequences, the user is asked to save all sequences or the selected ones.

5.1.4 Save Window As

Save only the visible part of the snapshot in the specified file format and file name.

5.1.5 Close

Close all snapshots and all dependencies (Overview, 3D Viewer...), and set Ordalie ready to load a new file.

5.1.6 Print

When the 'Print' option is invoked, a 'Setup' properties window will pop up. Several printing parameters can be set:

- *image format : actual choices are PostScript, JPEG and PNG.*
- *Paper size : A4, A3 or US letter.*
- *Paper orientation : Landscape or Portrait*
- *Font size and weight : the font used to print alignments is always Courier New. By default, the font size is the same as the one used in the main Ordalie window. It can be changed from 6 to 14 by steps of 2 pts, adn the weight of the font can be 'normal' or 'bold',*
- *Print Area : By default, the whole alignment is printed. It is possible to print only the current window or the selected sequences can be printed too.*
- *Ruler : if desired, a ruler can be printed at the bottom of every page.*
- *Residue numbering : if desired, the residue numbering inside the sequence can be printed on every page for every sequences.*

5.1.7 Quit Ordalie

Guess...

5.2 The Edit Menu

5.2.1 Cut

Cut the selected sequences.

5.2.2 Copy

Copy the selected sequences into the memory buffer.

5.2.3 Paste

Paste the sequences present in the memory buffer. The sequences are pasted just below the current selected sequence. If no sequence is selected, then no sequences are pasted. Note that in case of pasting sequences that already exist in the snapshot, their names will be suffixed by '__<n>' where '<n>' is the copy number.

5.2.4 Preferences

Access the 'Preferences' panel. This gives access to Ordalie internals.

5.3 The View Menu

5.3.1 Bigger font

Increases the font size for sequence display. The Ordalie default font type and size can be defined through 'Edit' -> 'Preferences' -> 'General'.

5.3.2 Smaller font

Decreases the font size for sequence display.

5.3.3 Open Log Console

Most of Ordalie information and computations are logged. The Log window allows accessing the crude log text and to save it as a raw text file.

5.3.4 Output Log as ...

This will output the whole Ordalie lo in a file. A cascade menu allows selecting between an HTML or Text format. The output file name is the root name of the original alignment file suffixed by the format type.

5.3.5 Toggle Full Screen

Toggles the main window between its actual size and full screen size. This can also be achieved by hitting the <F1> key on the keyboard.

5.3.6 Show/Hide Icon Bar

Displays or hides the Icon bar buttons frame. This is useful when a maximal window size is required, for alignment editing for example.

5.3.7 Show/Hide Scores

This turns on and off the scores frame. When displaying a conservation score, either through the Features buttons or inside the 'Conservation' tool, the global score and groups scores for each column is plotted. By default, the Score frame is turned off.

5.3.8 Show/Hide Features frame

This turns on and off the Features frame. This is useful when a maximal window size is required, for alignment editing for example.

5.3.9 Show/Hide Phylum

If the "life domain" characteristic information is available for the sequences, then the sequence names are coloured according to their life domain : eukaryots in red, archaea in blue, bacteria in yellow, and viruses in black.

5.3.10 Show/Hide groups colours

Colours the sequences according to their groups.

5.3.11 Show/Hide sec. Str.

If the secondary structures are available, show or hide α-helices (red) and β-strands (green).

5.4 The Sequence Menu

5.4.1 Names as

In Ordalie, there are several ways to display sequences names:

- *by the sequence name, which is a name given by the user,*
- *by the accession number,*
- *by the bank Id,*
- *by the full organism name if available,*
- *by the Gene.Species abbreviation if available. The abbreviated names are of the form Homo.sapi for Homo sapiens, sacc.cere for Saccharomyces cerevisiae for example.*

By default, the Sequence Name is displayed.

In Macsims/Xml and Ordalie file formats, the Sequence Name, Accession Number and Bank ID are clearly identified for each sequence. For all other formats, there is only one ID per sequence, and this one will be taken as the sequence name, accession number and bank ID.

5.4.2 Identity tool

Launches the Identity tool, see 4.3.

5.4.3 Search motif

Launches the Search motif tool, see 4.4.

5.4.4 Retrieve Seq. Info.

This command is intended to retrieve information for some sequences by querying their original databases. When invoked a window appears asking if all sequences or only the selected ones should be treated. After pressing "OK" Ordalie will query the database to which the sequences belong to and retrieve corresponding information. If some amino acid sequences have changed, they will be immediately realigned against the other sequences with the MAFFT program. Once terminated a summary window appears showing the number of updates that occurred for Accession, Phylum, Taxa ID, Organism, Description and the amino acid sequence. Below the changes table the "Show details..." button launches a web page with the detailed changes that occurred for each sequence.

5.4.5 Edit Info Seq

Enters the sequence information edit tool. This tool gives access to all information available for a given sequence, and allows modifying some of them.
At the top of the "Edit Info. Seq." window a combobox allows selecting the desired sequence.

5.4.6 Features Editor

Launches the Features Editor tool, see 4.10.

5.4.7 Add feature from file

Opens a file dialogue asking for a feature file. The loaded features are immediately available in Ordalie. The feature file format is described in Appendix 6.4.

5.4.8 Features Summary

Launches the Feature Summary tool, see 4.11

5.4.9 Add in snapshot

This item triggers a menu allowing to add sequences or PDB from files or database. Choices are:

- *Sequences from database: download sequences from UniProt/RefSeq databases given the user specified accession list.*
- *Sequences from file: user supplied sequence file.*
- *PDB from sequence file:*
- *PDB from PDB ID: download a user supplied PDB ID.*

Depending on the choice a dedicated window pops up with corresponding fields to be filled. Figure 17 shows the window for adding sequences from file.

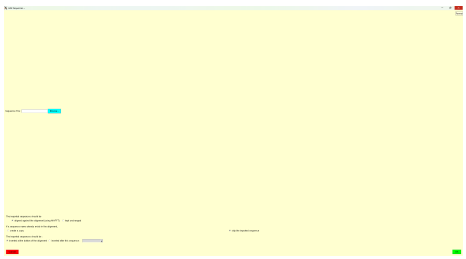

Figure 17: The window corresponding to adding sequences from file

After specifying a file, PDB ID or accession numbers, the user should select whether:

- *The added sequences should be aligned or not. Ordalie uses MAFFT to align the sequences against the snapshot sequences.*
- *To discard or to create a copy if there are duplicated sequences.*
- *To insert the newly added sequences at the bottom of the snapshot, just after a given sequence chosen by the dedicated combobox, or inserted into their corresponding cluster based upon sequence identity (only valid if added sequences are aligned).*

After sequences have been added, features positions are recomputed if needed, and basics sequences characteristics are computed for the new sequences (isoelectric point, molecular weight, identity, composition, ...) and are logged.

5.5 The Alignment Menu

5.5.1 Editor

Launches the Editor tool, see 4.2.

5.5.2 Re-Align sequences

Re-aligns the selected sequences by using the MAFFT program [7, 6, 5, 8]. The selected sequences are first degapped and then re-aligned against the remaining sequences. The re-aligned sequences stay at their original place.

5.5.3 Clustering

Launches the Clustering tool, see 4.5.

5.5.4 Add separator

In Ordalie a "separator" is simply a blank line between sequences. Adding separator allow the user to create his own clusters. This command will add a blank line separator just below the selected sequence. Be aware that only one sequence should be selected.

5.5.5 Remove Separator

Removes the separator just below the selected sequence. Be aware that only one sequence should be selected.

5.5.6 Remove All separators

Removes all separators, and unselects all sequences. This results in an unclustered snapshot.

5.5.7 Overview

Creates an instance of the Overview Window for the current snapshot, see "Snapshot Overview" section 4.1.

5.5.8 Conservation

Launches Conservation computation tool, see 4.7.

5.5.9 Tree building

Launches the Tree tool, see 4.6.

5.5.10 Annotate snapshot

Launches the Annotation tool, see 4.12.

5.5.11 Barcode

Launches the Barcode tool, see 4.13.

5.5.12 Features Editor

Launches the feature editor, see 4.10.

5.5.13 Features Summary

Launches the 'Features Summary' tool, see 4.11. Sequences selected in the main window will automatically be selected in the 'Features Summary' tool.

5.5.14 Toggle physicochem. col.

Toggle the colouring of residues with the same mapping as in the 'Editor'.

5.6 The Structure Menu

5.6.1 Superpose Structures

Launches the 'Superposition' tool if several PDB structures are available in the snapshot, see 4.8.

5.6.2 Display Structures

Launches the '3D Viewer' tool, see 4.9.

5.6.3 Colour Sec. Str. by identity

This tool allows colouring the secondary structures according to their sequence identity level in the snapshot. When invoked, a parameter window will pop up with the following parameters to select:

- *Compute similarities using mean sec. str. limits or by using a sequence reference,*
- *Type of color gradient : grey or color,*
- *Gradient limits : from min to max identity, or from 0 to 100%,*
- *Assign identity to B factor or not.*

5.6.4 Save PDB

As the 3D structures may be changed using the 'Superposition' tool, this option allows saving a given 3D structure present in the snapshot. It may be useful to output the superposed structures in order to render them in a more sophisticated drawing program.

When invoked, a window will pop up to ask the user for some parameters:

- *select the molecule,*
- *select all or a particular chain,*
- *select all residues or a specific residue range,*

5.7 The "?" Menu

5.7.1 About

This menu item pops up a window giving some information about Ordalie, such as the currently used version, and the URL for the Ordalie home page.

5.8 Help

5.8.1 Local Ordalie documentation

Launches this manual on the default web browser using local version.

5.8.2 On-line documentation

Launches the Ordalie documentation on a web browser using Ordalie website.

6 Appendix

6.1 The Ordalie database scheme

The core of Ordalie is build around an in-memory SQLite database [3] which scheme is given in figure XXX. Ordalie takes advantage of this underlaying database to store snapshots of alignments and their associated features. The "ordalie" table contains settings parameters saved at exit allowing the user to find the same state when launching Ordalie again. The "seqinfo" table contains sequence information that are not linked to amino acids positions (length, molecular weight, isoelectric point, ...) The "seqfeat" table is used to store features data mapped onto the residue sequence. Upon loading of a new alignment file, Ordalie creates a first snapshot as being a read-only copy of this alignment stored as the "original alignment" in the snapshot table. This table contains all snapshots created so far along with their name and description. The "seqali" table records the amino acid sequences as they appear in the snapshots. A link table "ln_snapshot_seqali" binds a given set of sequences to a given snapshot. Accordingly, the "featali" table stores features attached to aligned sequences in a given snapshot. A link table "ln_seqali_featali" couple these two tables. The "clustering" and "cluster" tables define a given clustering attached to a snapshot with its name, the ***method and residue zones*** *used to compute it, and the resulting clusters with their names respectively. The set of sequences defining a given cluster is available through the "ln_seqali_cluster" link table. The "colmeasure" and "colscore" tables correspond to conservation computations (column measurements) with their name and used method, and the conservation groups with name, value for each column of the group respectively. The conservation score for a given cluster is available through the link table "ln_cluster_colscore". Finally, the "annotation" table contains all information relative to annotation the user adds to a given snapshot. The Ordalie (. ord file extension) consists in a database dump.*

6.2 The FindPatterns syntax

6.2.1 Basic rules of syntax

The search pattern can include any legal sequence character, and also include several non-sequence characters, which are used to specify 'OR' matching, 'NOT' matching, 'begin' and 'end' constraints, and repeat counts. For instance, the pattern GASTE(X){20,30}FTG means searching GASTE, followed by 20 to 30 of any amino acid, followed by FTG. Following is an explanation of the syntax for pattern specification.

6.2.2 Implied Sets and Repeat Counts

Parentheses () enclose one or more symbols that can be repeated a certain number of times. Braces {} enclose numbers indicating how many times the symbols within the preceding parentheses must be found.

Sometimes, it is possible to leave out part of an expression. If braces appear without preceding parentheses, the numbers in the braces define the number of repeats for the immediately preceding symbol. One or both of the numbers within the braces may be missing. For instance, both the pattern GASG{2,}F

and the pattern GASG{2}F mean GAS, followed by G repeated from 2 to 350,000 times, followed by F; the pattern GASG{}F means GAS, followed by G repeated from 0 to 350,000 times, followed by F; the pattern GAS(TE){,2}F means GAS, followed by TE repeated from 0 to 2 times, followed by F; the pattern GAS(TE){2,2}F means GAS, followed by TE repeated exactly 2 times, followed by F (If the pattern in the parentheses is an OR expression (see below), it cannot be repeated more than 2,000 times).

6.2.3 OR Matching

Specifying several symbol choices can be easily done by enclosing the different choices between parentheses and separating the choices with commas. For instance, RGF(Q,A)S means RGF followed by either Q or A followed by S. The length of each choice need not be the same, and there can be up to 31 different choices within each set of parentheses. The pattern GAT(TG,T,G){1,4}A means GAT followed by any combination of TG, T, or G from 1 to 4 times followed by A. The sequence GATTGGA matches this pattern. There can be several parentheses in a pattern, but parentheses cannot be nested.

6.2.4 NOT Matching

The pattern GC~CAT means GC, followed by any symbol except C, followed by AT. The pattern GC~(A,T)CC means GC, followed by any symbol except A or T, followed by CC.

6.2.5 Begin and End Constraints

The pattern <GACCAT can only be found if it occurs at the beginning of the sequence range being searched. Likewise, the pattern GACCAT> would only be found if it occurs at the end of the sequence range.

6.3 The Vector Norm scoring method

This method is based on a vectorial representation of the 20 amino acids. This representation can be the same as the one used in the VRP representation, or can be for example, a volume/polarity couple. The score for a given column k can be computed then by:

*S(k) = nc/nt * /sum_ i=1^nV//sum_ 1=1^n/Vi/*

where nc is the number of residues in the column, nt is the total number of sequences. This function is bounded by 0. and N, where N is the number of sequences in the alignment.

6.4 The Feature File Format

It is possible to import features into Ordalie through a features file. It is also possible to add items to an existing feature, or to create completely new ones.

The feature file format looks like:

```
# This is an example of a feature file format
# A line starting by \# is a comment line that can be inserted everywhere
```

```
# Declare the feature
FEATURE MyFeat? PROPAGATE? ?all|group?

#
# Structure of the feature item:
# seq. name; coord. system; start; stop; color; score; note
Q65P3D; LOCAL; 23; 57; red; 0. 0; first item
Q65P3D; GLOBAL; 212; 345; blue; 0. 0; second one
FLK14Q; local; 123; 234; red; 0. 0; one more

# Then go to an other feature
FEATURE STRUCT
P12345; global; 2112; 2541; green; 0. 0; add one
```

To add some items to an existing feature, the feature name should be exactly identical to the one already present in the alignment as feature names are case-sensitive.

6.5 The superposition algorithm

Given two sets of atomic coordinates A and B, the following algorithm will try to minimize the RMS (Root Mean Squared) distance between A and B by moving B onto A. The algorithm can be separated in the following steps:

- *Compute the centre of mass (CDM) of the two sets and translate B atoms by the vector joining B CDM to A,*
- *Compute the 3 main inertia axis for the two sets, which correspond to the three eigen vectors of the Eigen decomposition of the whole atomic coordinates set,*
- *By turn, minimize the weighted RMS by rotating around the Eulerian angle associated to the current axis.*
- *Stops when converged, and issue superposition information.*

The output of this algorithm provides along with the RMS, the orientation matrix, translation vector, and rotations between the two molecules in different forms.

6.6 The command line options

Option	***Values***	***Description***
-convert	*<TFA/MSF/XML/ORD>*	*Converts the alignment into the format indicated by -convert. The converted output file name will have the form <alignemnt file>. <format>*
-precompute	*<0/1>*	*: precompute clustering and conservation for each PFAM domain*
-threshold	*<x>*	*Set conservation threshold level. <x> should be set between 51 and 100*
-batch	*<0/1>*	*Run Ordalie without windows and exit when finished*

References

[1] *Hirotugu Akaike. A new look at the statistical model identification.* IEEE Transactions on Automatic Control*, AC-19(6):716–723, December 1974.*

[2] *S. F. Altschul, J. C. Wootton, E. Zaslavsky, and Y. K. Yu. The construction and use of log-odds substitution scores for multiple sequence alignment.* PLoS Comput. Biol.*, 6(7):e1000852, Jul 2010.*

[3] *D. Richard Hipp. Sqlite home page.* `https://www.sqlite.org`*. Accessed: 2017-08-14.*

[4] *F. Johansson and H. Toh. A comparative study of conservation and variation scores.* BMC Bioinformatics*, 11:388, Jul 2010.*

[5] *K. Katoh, K. Kuma, T. Miyata, and H. Toh. Improvement in the accuracy of multiple sequence alignment program MAFFT.* Genome Inform*, 16(1):22–33, 2005.*

[6] *K. Katoh, K. Kuma, H. Toh, and T. Miyata. MAFFT version 5: improvement in accuracy of multiple sequence alignment.* Nucleic Acids Res.*, 33(2):511–518, 2005.*

[7] *K. Katoh, K. Misawa, K. Kuma, and T. Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.* Nucleic Acids Res.*, 30(14):3059–3066, Jul 2002.*

[8] *K. Katoh and H. Toh. Recent developments in the MAFFT multiple sequence alignment program.* Brief. Bioinformatics*, 9(4):286–298, Jul 2008.*

[9] *V. Lefort, R. Desper, and O. Gascuel. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program.* Mol. Biol. Evol.*, 32(10):2798–2800, Oct 2015.*

[10] *X. S. Liu and W. L. Guo. Robustness of the residue conservation score reflecting both frequencies and physicochemistries.* Amino Acids*, 34(4):643–652, May 2008.*

[11] *G. McLachlan and D. Peel.* Finite Mixture Models*. Wiley, 2000.*

[12] *E.G. Schwarz. Estimating the dimension of a model.* Annals of Statistics*, 6(2):461–464, 1978.*

[13] *J. D. Thompson, T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.* Nucleic Acids Res.*, 25(24):4876–4882, Dec 1997.*

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

[15] J. D. Thompson, A. Muller, A. Waterhouse, J. Procter, G. J. Barton, F. Plewniak, and O. Poch. MACSIMS: multiple alignment of complete sequences information management system. BMC Bioinformatics, *7:318, Jun 2006.*

[16] S. M. Thompson. Constructing and refining multiple sequence alignments with PileUp, SeqLab, and the GCG suite. Curr Protoc Bioinformatics, *Chapter 3:Unit 3.6, Feb 2003.*

[17] W. S. Valdar. Scoring residue conservation. Proteins, *48(2):227–241, Aug 2002.*

[18] N. Wicker, D. Dembele, W. Raffelsberger, and O. Poch. Density of points clustering, application to transcriptomic data analysis. Nucleic Acids Res., *30(18):3992–4000, Sep 2002.*

[19] N. Wicker, G. R. Perrin, J. C. Thierry, and O. Poch. Secator: a program for inferring protein subfamilies from phylogenetic trees. Mol. Biol. Evol., *18(8):1435–1441, Aug 2001.*

[20] D. D. Womble. GCG: The Wisconsin Package of sequence analysis programs. Methods Mol. Biol., *132:3–22, 2000.*