# 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 System Requirements

Ordalie is a cross-platform application available for Windows, MacOS, and Linux (32/64-bit architectures).

### 1.1.1 Graphics

To render 3D structures, Ordalie leverages the **OpenGL** library. While this is standard on most modern graphics cards, specific attention is needed for certain systems:

On MacOS (version *Mojave* and later), OpenGL is no longer bundled with the OS by default. Users may need to install it manually (e.g., via the *App Store* or developer tools) to enable 3D visualization.

### 1.1.2 Connectivity

While Ordalie functions offline, an active internet connection is highly recommended to access remote features, including:

- Direct querying of biological databases (PDB, UniProt, NCBI).
- On-the-fly sequence alignments.
- Web service integration.

## 1.2 Installation Procedure

The latest version of Ordalie can be downloaded from: `http://www.lbgi.fr/ordalie`. Select the installer corresponding to your platform (Windows, MacOS, or Linux 64-bit).

### 1.2.1 Windows Installation

Launch the `.exe` installer and follow the on-screen instructions. Upon completion, Ordalie will launch automatically, and a shortcut will be created on your desktop.

### 1.2.2 Linux and MacOS Installation

For Unix-based systems, Ordalie is provided as a compressed archive. Follow these steps in your terminal to complete the installation:

1. **Extract the archive:**
   ```
   % tar -zxf setup-linux-x86_64.tar.gz
   ```
2. **Run the installation script:**
   ```
   % cd setup-linux-x86_64
   % ./install.tcl
   ```

**1.2.2.1 Directory Selection:** By default, Ordalie installs to your `/home/<user>/` directory. If you wish to install it system-wide (e.g., in `/usr/local`), ensure you have the necessary `sudo` or administrative privileges.

**1.2.2.2 Cleanup:** Once the installation is confirmed, you may safely delete the setup directory:

```
% cd ..
% rm -rf setup-linux-x86_64
```

# 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. Since the 1970s, the massive increase of protein 3D structures and sequences has led to more subtle definitions, such as super-family or sub-family organizations.

- **Super-family:** The broadest level of classification. Members share distant evolutionary ancestry and often a common structural fold, performing a wide range of biological roles.
- **Family:** Proteins within a super-family that are more closely related in terms of sequence, structure, and function, typically sharing a common evolutionary origin.
- **Sub-family:** The most specific level. Members are highly related, often performing near-identical functions and having evolved 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 functional divergence.

## 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) in a hierarchical manner, and within different contexts, such as phylogeny or 3D structure.

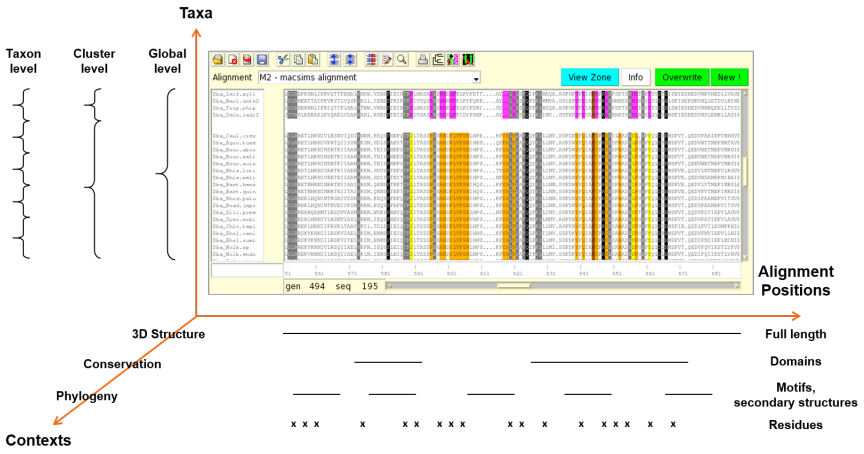


Figure 1: Diagram of the Ordalie philosophy

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

The information distributed along the sequence (horizontal axis) is analyzed across several scales:

- **Large-scale features:** Overall domain architecture and broad evolutionary conserved regions.
- **Intermediate-scale features:** Secondary structure elements, functional motifs (SLiMs), and molecular recognition patches.
- **Local features:** Post-translational modification (PTM) sites, catalytic residues, and site-specific conservation scores.

The taxonomic depth constitutes the second axis:

- **Global scale:** To define family-wide hallmarks and conserved motifs.
- **Intermediate scale:** To distinguish sub-families by identifying residues specifically conserved within a subgroup.
- **Individual scale:** To pinpoint mutations or unique variations specific to a particular taxon.

The third analytical axis focuses on computational contexts, unified within a structural framework. This allows sequence-based features to be spatially mapped and compared directly onto 3D structures.

## 2.3 Database and Snapshot Management

At the core of Ordalie lies a dedicated **SQLite database** engine. This architecture ensures high-performance data handling and persistent storage for both system-wide configurations and alignment-specific data.

A major strength of this architecture is the ability to perform non-linear, exploratory analyses through a system of **Snapshots**. Rather than a linear versioning system, snapshots in Ordalie represent parallel analytical hypotheses. They allow the user to freeze and compare different interpretations of the same dataset:

**The Database:** Acts as a central, relational repository. By using the native `.ord` file format, all sequences, metadata, and analytical branches are bundled into a single, portable file, ensuring data integrity across sessions.

**Snapshots:** A *snapshot* captures a specific state of the exploratory process. It is not merely a backup, but a record of a particular scientific choice, such as:

- A specific **clustering strategy** (e.g., comparing different phylogenetic groupings or functional sub-families).
- A unique **alignment topology**: This allows for the comparison of different ways to register sequence zones. For instance, a specific domain can be anchored to one position in one snapshot and realigned to a different evolutionary landmark in another.

- A unique set of curated **sequence features** or annotations mapped onto a given alignment solution.

Snapshots are managed via the dedicated *Snapshot Bar* (see section 3.6.2), providing a seamless way to switch between different analytical scenarios without data loss.

# 3 Ordalie Basics

This section presents the fundamental operations and interface logic of Ordalie. While these concepts are introduced here, they are explored in greater detail in the specialized sections of this manual.

## 3.1 Alignment Handling

Ordalie supports a wide range of MSAs file formats, including FASTA, MSF, RSF, ClustalW ALN, Macsim/XML, and its native ORD format. Upon loading, the software performs a consistency check between the file extension and its internal structure, raising an error in case of a mismatch. Basic statistical information is automatically logged in the *Log Console* for both the global alignment and individual sequences (e.g., length, isoelectric point). Pairwise identity scores are also computed and logged for the global set and any defined groups.

## 3.2 Snapshots

As already mentioned in the introduction (see 2.3) a single Multiple Sequence Alignment can be the subject of various analytical interpretations. In Ordalie, these distinct states—such as different clustering configurations or specific conservation models—are called **Snapshots**. This system allows users to explore multiple hypotheses in parallel, with the ability to annotate, save, and switch between states without data loss (see 3.6.2). This non-linear workflow is made possible by the integrated SQLite database.

## 3.3 Sequence Names and Databases

Ordalie automatically parses sequence names to identify UniProt, RefSeq, or PDB accessions. By default, common database prefixes (e.g., `sw|` for SwissProt, `gi|` for NCBI Gene Identifier, `PDB|` for Protein Data Bank) are stripped for clarity; for example, `>sw|P12345` will be displayed as `P12345`. This behavior and the list of recognized separators can be customized via *Preferences*. If sequence names follow standard accession formats, Ordalie can fetch additional metadata from remote databases (see 5.4.4). *Note: While DNA/RNA alignments can be viewed or edited, Ordalie is primarily optimized for protein sequences; some advanced analytical tools may be disabled for nucleotide data.*

## 3.4 Conventions and Selections

### 3.4.1 Mouse and Keyboard terminology

In this manual, the mouse buttons are designated as <**B1**> (Left), <**B2**> (Middle), and <**B3**> (Right). Keyboard keys are represented between brackets (e.g., `<Ctrl>`).

### 3.4.2 Selection Mechanisms

Ordalie uses a standardized selection logic for both sequence names and residue ranges.

#### 3.4.2.1 Sequence Names Selection

Sequence names selection is done using a combinaison of <B1>, <Shift> and <Control> keys ins a way obseerved in other programs.

| Action | Result |
|---|---|
| `<B1>` | Selects the sequence under the pointer. |
| `<Ctrl> + <B1>` | **Toggle selection:** Adds an unselected sequence to the current selection or removes a selected one. |
| `<Shift> + <B1>` | **Range selection:** Selects all sequences between the previous selection and the current click. |
| `<Ctrl> + <a>` | Selects all sequences in the snapshot. |

Table 1: Shortcuts for sequence name manipulation.

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

#### 3.4.2.2 Residue Range Selection

By default, residue range selection is disabled in Ordalie. It is enabled within specific modules (Cluster, Tree, Identity, etc.).

| Action | Result |
|---|---|
| `<B1>` | Sets the start of the selection zone. |
| `<B3>` | Sets the end of the selection zone (highlighted in grey). |
| `<Ctrl> + <B3>` | Unselects the zone under the mouse pointer. |
| `<Ctrl> + <B1>` | Selects the **feature** (e.g., PFAM domain) under the pointer. |

Table 2: Shortcuts for residue and feature selection.

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

## 3.5 The SQLite Database and .ord Format

Ordalie uses an embedded SQLite database to manage complex data structures (system data, snapshots, features). The **.ord** file format is a portable SQLite database containing:

- **Session State:** All variables and view settings.
- **Sequence Data:** Metadata not linked to the primary sequence (pI, descriptions).
- **Analytical Branches:** All snapshots and their specific clustering.

Using the `.ord` format as your primary working file is highly recommended to ensure all analytical context is preserved. The scheme of the database can be found in Appendix 6.1.

## 3.6 The Main Window

The interface is organized vertically into several functional areas (see fig. 2).

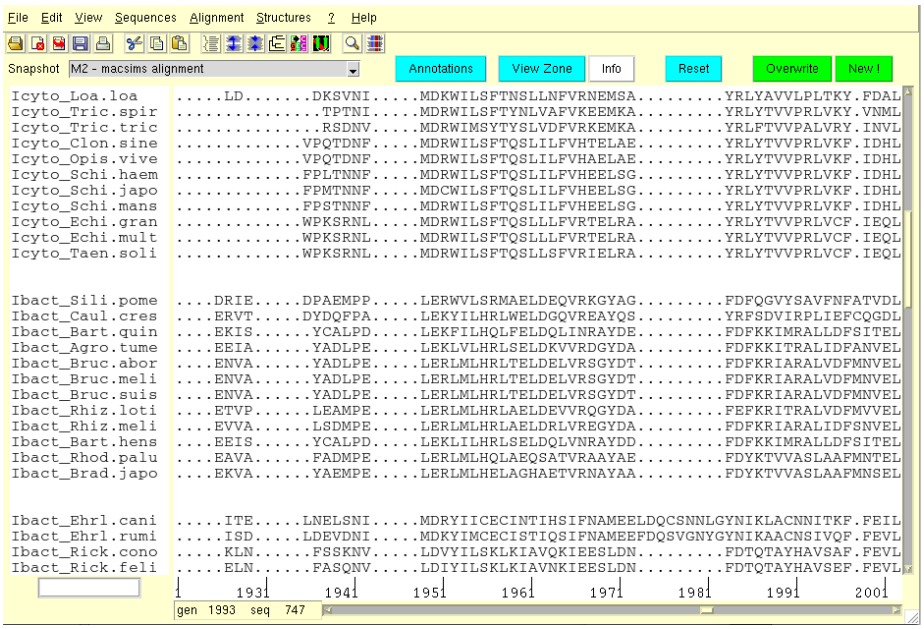


Figure 2: Ordalie main window

### 3.6.1 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 "Edit" gives access to the common cut/copy/paste options and to the Preferences. 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, add sequences or structures coming either from databases or from a file to the current sncapshot. 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 files. Finally, the "Help" 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, with a tooltip appearing on mouse hover.

### 3.6.2 The Snapshot Bar

Located above the alignment, this bar allows you to:

- Switch between snapshots (i.e. the different analytical scenarios) via a dropdown menu.
- View the information attached to the current snapshot (*Info*).
- View the zones that lead to the current clustering (*View Zone* button).
- Reset the current snapshot(*Reset*).
- Save changes to the current snapshot (*Overwrite*) or create a new one (*New*).

### 3.6.3 The Snapshot Frame

This is the primary workspace, divided into the **Sequence Names** (left) and the **Alignment View** (right).

`<Mouse-wheel>` scrolls the view vertically.
`<Ctrl> + <Mouse-wheel>` scrolls the alignment horizontally.

#### 3.6.3.1 Sequence Names Pane

Sequence names associated with a PDB structure are highlighted in red. When the mouse pointer hovers over a sequence name, a "tooltip" (yellow message window) displays available metadata—whether imported from Macsim/XML and ORD files or retrieved online (see 5.4.4). For a more comprehensive view, a Right-click (<B3>) on any name opens a detailed information window containing the accession number, database ID, organism, sequence length, and full description. Directly below the names list, a dedicated search box allows you to find sequences by name or fragment. Typing a search string and pressing <Return> will automatically scroll the alignment to position the first match at the top of the window.

#### 3.6.3.2 Alignment View Panel

The right side of the frame displays the amino acid sequences, accompanied by a ruler at the bottom to indicate column positions. Navigation is managed through horizontal and vertical scrollbars. As you move the mouse over the alignment, the **position counter** updates in real-time to show two distinct coordinates for the residue currently under the pointer:

- **The 'seq' (Local) position:** This indicates the specific position of the residue within its own individual sequence (excluding gaps).
- **The 'gen' (Global) position:** This represents the absolute coordinate of the residue within the current snapshot grid.

In this manual, the position within the sequence is referred to as the *local position*, while the position within the snapshot is referred to as the *global position*.

**Navigation tip:** You can "jump" along the alignment by typing a number on the keypad followed by the `<Left>` or `<Right>` arrows (e.g., `200 + <Right>`).

### 3.6.4 The Scores Frame

The Scores frame is hidden by default but can be toggled via the `View -> Show/Hide scores` menu. Once residue conservation has been computed, a score is assigned to every column of the snapshot.

These scores are normalized between 0 and 100 and displayed graphically:

- **Global level:** Represented by a **black line**, reflecting the conservation across the entire snapshot.
- **Group level:** If sequence groups are defined, their specific conservation scores are displayed using lines that match their respective **group colors**.

This overlay allows for an immediate visual comparison between family-wide hallmarks and group-specific constraints.

### 3.6.5 The Control Panel

The Control Panel is located at the bottom of the main window. In its default state, the frame displays a welcome message. However, as the analysis progresses, this area becomes a dynamic interface populated with buttons representing the available **features** of the current snapshot.

**3.6.5.1 Feature Visibility:** Each button acts as a toggle for a specific feature:

- **Activation (Red):** Pressing a button turns it **red** and displays the corresponding feature on the alignment.
- **Deactivation (Green):** Pressing the button again turns it **green** and removes the feature from the view.

The features are rendered in the order their buttons are activated. To change the layering (e.g., to place a specific domain on top of a conservation score), simply toggle the buttons to reorder the display stack!

**3.6.5.2 Contextual Adaptation:** The content of the Control Panel is context-sensitive. When switching between different Ordalie tools (e.g., from the Editor to the Clustering tool), the panel will automatically update to show the controls specific to that module. Detailed descriptions of these specific controls are provided in their respective sections of this manual.

## 3.7 Features

**Features** are a central pillar of the Ordalie architecture. A feature is defined as a biological or structural characteristic mapped onto a specific coordinate range. These can be categorized by their scope:

- **Sequence Features:** Specific to a single sequence (e.g., a PTM site or a single PFAM domain).
- **Group Features:** Characteristics shared by a specific cluster of sequences.
- **Global Features:** Hallmarks applicable to the entire snapshot or protein family.

A single feature can encompass multiple *items*; for example, a "Domains" feature may contain several distinct PFAM occurrences along the sequence. One of the primary strengths of Ordalie is its ability to investigate these features across multiple dimensions, such as projecting sequence-based motifs directly onto 3D structures.

### 3.7.1 Data Provenance

Features can be integrated into Ordalie through three main channels:

1. **External software:** Parsing XML output from the Macsims program [18].
2. **Dedicated files:** Using the Ordalie feature file format (see section 6.4).
3. **Manual curation:** User-defined annotations created via the *Features Editor* tool (see 4.10).

Ordalie does not automatically generate biological features. The only exception is the **Conservation Tool**, which creates new feature items based on each calculated conservation score.

# 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* provides a condensed, schematic representation of the entire alignment. This tool is designed to facilitate navigation and global feature inspection. Multiple overview windows can be opened simultaneously for the same snapshot, allowing for different levels of zoom or feature mapping.


Figure 3: The Snapshot Overview tool window, showing the alignment scheme and control panel.

### 4.1.1 The Snapshot Frame

In this view, the alignment is abstracted: each residue is represented by a single grey pixel on a white background. This miniaturized "map" provides two main interactive functions:

- **Feature Mapping:** Any sequence or global feature can be overlaid onto this scheme, highlighting motifs or domains across the whole family.
- **Navigation & Synchronization:** A **stippled rectangle** (bounding box) indicates the region currently visible in the main alignment window. Clicking anywhere on the scheme will instantly center the main window on that specific position.

### 4.1.2 Control Panel

The control panel at the bottom of the window manages the display and interaction:

- **Feature Selector (Combobox):** By default, this is set to **'automatic'**, meaning the Overview reflects exactly what is displayed in the main window. Selecting a specific feature here will force its display on the Overview independently of the main window settings.
- Zoom (+ / -): These buttons adjust the scale of the schematic pixels, allowing the user to switch between a dense family-wide view and a more detailed motif-level view.

- **Print:** Exports the current schematic state as a PNG image. The software will prompt for a destination file name.
- **Close:** Terminates the Overview instance.

## 4.2 Editor

The quality of downstream analyses (phylogeny, conservation, or feature mapping) is strictly dependent on the accuracy of the Multiple Sequence Alignment (MSA). While automated algorithms are highly advanced, manual curation remains an essential step. Ordalie integrates a high-performance sequence editor inspired by the *SeqLab* interface (GCG Wisconsin Package), designed for intuitive and precise alignment refinement.

Upon entering the Editor, sequences are automatically colored according to their physico-chemical properties.


Figure 4: Default physico-chemical amino acid coloring scheme in the Editor.

*Note: The default coloring scheme can be customized in the* `Edit -> Preferences` *menu. Active features remain visible on the snapshot to guide the editing process.*

To ensure data consistency, some analytical functionalities are temporarily disabled while the Editor tool is active.

### 4.2.1 Control Panel

The Editor control panel at the bottom of the window provides the following management tools:

**Clear selections:** Deselects all currently selected sequences.

**Group / Ungroup:** Groups the selected sequences. Grouped sequences are assigned a unique name color and, crucially, **behave as a single entity** during editing (moving one moves all members of the group).

**Lock / Unlock:** By default, the editor is **Locked**, meaning only gap characters (`.`) can be inserted or deleted. Unlocking the sequences allows for the direct insertion or deletion of amino acid residues.

**Map/Unmap residues:** A toggle to enable or disable the physico-chemical color scheme.

**Remove columns of gaps:** Scans the snapshot and deletes any columns consisting entirely of gaps. This operation is also performed automatically upon exiting the Editor.

**Show/Hide features:** Provides a menu to toggle specific features on or off. Features are layered in the order they are activated.

**Temp. Save:** Exports the current state to a `.tfa` file. This acts as a manual checkpoint during long editing sessions.

**Search Motif:** Switch to the search mtif tool, see 4.4

**Save & Return:** Exits the editor and prompts the user to either overwrite the current snapshot or create a new one with the changes.

**Cancel:** Exits the editor and discards all changes, restoring the snapshot to its original state.

### 4.2.2 Edition Actions

Clicking within the alignment frame activates a **yellow blinking cursor**. The following keyboard interactions are then available:

- **Navigation:** Use the `Arrow` keys to move the cursor through the alignment grid.
- **Deletion:** `<Backspace>` removes the character preceding the cursor. *Reminder: In Lock mode, this only works on gaps.*
- **Shifting (The "Push" mechanism):**
  - `<Shift-Right>`: Pushes the block of contiguous residues to the right of the cursor until the next common gap.
  - `<Shift-Left>`: Pushes the block of contiguous residues to the left of the cursor until the previous common gap.

#### 4.2.2.1 Numeric Buffer (Rapid Navigation):

To speed up editing, you can type a number (0-9) before a command. The digits are stored in a buffer (e.g., typing `1` then `2` stores `12`). Pressing a navigation or shift key will then apply that action `12` times and clear the buffer.

## 4.3 The Identity Tool

The Identity tool provides quantitative data on the sequence identity percentages within the snapshot. It allows for precise comparisons between specific sequences or defined residue ranges (features), making it essential for characterizing sub-family divergence or domain-specific conservation.

Figure 5: The control panel of the Identity tool.

### 4.3.1 Control Panel

The interface is divided into selection management on the left and computation/results on the right.

#### 4.3.1.1 Selection Management

Before computing identity, the user must define the scope of the analysis:

- **Feature:** A dropdown menu to select specificfeatures (e.g., a specific domain). You can select one or multiple item of different features as residue ranges.
- **Clear:** Resets all current sequence and residue range selections.
- **Select All:** Automatically selects the entire length of all sequences in the snapshot.

#### 4.3.1.2 Computation and Results

Once the scope is defined, the following actions are available:

**Compute:** Triggers the identity calculation. A detailed summary is automatically sent to the **Log Console**, see 5.3.3.

**Summary:** Opens a new window providing aggregate statistics for the entire selection and for each individual group:

- The **average identity percentage**.
- The **standard deviation**.
- Identification of the pairs with the **maximum** and **minimum** identity percentages.

**Pairwise Comparison:** Using the two dedicated comboboxes, the user can select any two sequences to display their specific identity percentage and the length of the ungapped sequences compared.

**Details :** Opens a new window providing detailed information for the selected sequence pair. The identity percentage and sequence lengths are displayed. For each sequence, its most identical (closest) and most distant sequences are specified, along with their respective identity percentages. If the snapshot is clustered, the window also highlights the closest and most distant sequences both within and outside the sequence's cluster. 6

**Return:** Exits the Identity tool and returns to the main window.

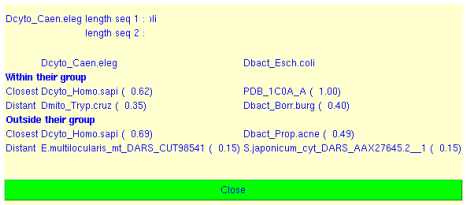


Figure 6: The "Details" window

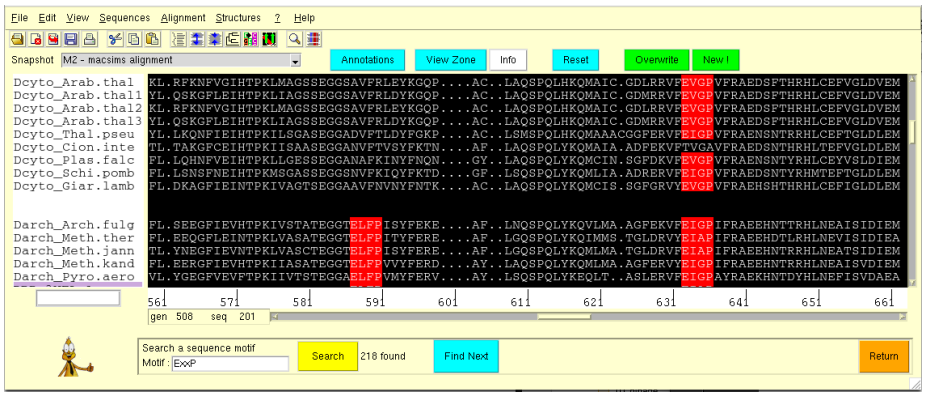


Figure 7: The Search Motif tool in action. Detected occurrences are highlighted in red against a high-contrast black background.

## 4.4 The Search Motif Tool

The Search Motif tool allows for rapid identification of specific sequence patterns within the current snapshot.

### 4.4.1 Search Pattern Syntax

Ordalie adopts the robust `FindPatterns` syntax from the GCG Wisconsin Package [23]. This allows for complex queries, including ambiguous residues and variable-length gaps. A comprehensive guide to this syntax is provided in Appendix 6.2.

### 4.4.2 Control Panel

The Search Motif interface is minimalist and designed for speed:

**Motif Entry Box:** The text field where the search pattern is defined.

**Search:** Launches the initial scan across all sequences in the snapshot.

**Find Next:** Navigates the view to the next occurrence of the motif.

**Return:** Exits the search tool and restores the standard display.

**4.4.2.1 Visual Feedback:** When a search is successful, Ordalie switches to a high-contrast mode: the snapshot background turns **black**, and every instance of the found motif is highlighted in **red**. This visual "negative" ensures that even single-residue motifs are easily spotted in large alignments.

The “Search Motif” is the only tool that remains accessible while the **Editor** tool is active, facilitating real-time curation based on specific motif locations.

## 4.5 The Clustering Tool

The **Clustering** tool enables the grouping of sequences based on numerical criteria derived from sequence characteristics. The computation can integrate all or part of the sequence, as well as specific snapshot columns. Users can combine multiple numerical criteria and select a specific clustering algorithm.

Once launched, the newly created sequence clusters are automatically displayed in the main window. While clustering trials are stored temporarily during the session, any specific clustering result can be permanently saved as a new snapshot.

### 4.5.1 Criteria

The tool currently supports the following numerical criteria:

- **Identity percentage**: Calculated for each sequence over the selected residue range against all other selected sequences.
- **Length**: The total length of the sequence.
- **Hydrophobicity**: (Note: Only available for Macsim alignments).
- **Isoelectric point (pI)**: Computed using the EMBOSS pKa values for amino acids.
- **Amino acid composition**: The relative percentage of the 20 standard amino acids within a given sequence.

### 4.5.2 Algorithms

Ordalie automatically determines the optimal number of groups. The clustering algorithms and the methods for defining cluster counts are integrated from the **Cluspack** package. Available methods include:

- **K-means / DPC (Density of Points Clustering)** [21]: Cluster identification is based on point density criteria, while the final selection is performed via the k-means algorithm.
- **Hierarchic / Secator** [22]: Groups are identified through ascending hierarchical classification. Cluster selection is determined by an inertia loss criterion.
- **Mixture Model** [12] / **AIC** [1] or **BIC** [13]: Data distribution is modeled using Gaussian distributions, with clustering optimized according to the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC).

### 4.5.3 Control Panel

#### 4.5.3.1 Selections

The left section of the Control Panel manages sequence and residue range selection:


Figure 8: The Clustering tool control panel interface.

- **Feature**: Displays the currently selected feature. Users can select one or more items from this feature to define a residue range.
- **Clear**: Resets all previously selected residue ranges and sequences.
- **Select All**: Selects all columns in the current snapshot.

*Note on sequence selection:* If no sequences are manually selected, the tool processes **all** sequences. If a subset is selected (minimum 4 sequences required), the clustering applies only to that subset; unselected sequences are grouped into a separate "Others" category.

#### 4.5.3.2 Clustering criteria

The pull-down menu allows for the selection of one or more criteria.

The special **'Life domain'** criterion (Eukaryota, Archaea, Prokaryota, etc.) is categorical and cannot be combined with other numerical criteria.

#### 4.5.3.3 Clustering methods

The 'Method' menu provides access to the algorithms:

- Hierarchic clustering / Secator
- K-means / DPC
- Mixture model / AIC criterion
- Mixture model / BIC criterion

Clicking **"Compute"** updates the main Ordalie window with the resulting groups.

#### 4.5.3.4 Other buttons

- **Reset**: Reverts to the original clustering state of the snapshot.
- **No Clusters**: Removes all existing groups, merging all sequences into a single block.
- **Clusters Names**: Opens a dialog to rename clusters. These names are preserved in other modules, such as the **Tree** display or the **Barcode** tool.

- **Save**: Ends the session and saves the results. You will be prompted to either overwrite the current snapshot or create a new one.
- **Return**: Exits the tool without saving changes, restoring the snapshot to its initial state.

## 4.6 The Tree Tool

The **Tree** tool is divided into two main phases: tree construction, performed within the main Ordalie window, and tree exploration, which takes place in a dedicated visualization window.

The tree is computed using the **FastME** program [10] with default parameters. Ordalie first generates a distance matrix based on identity percentages calculated over the selected residue range. While Bayesian-based algorithms can offer higher precision, FastME provides an excellent balance between computational speed and phylogenetic accuracy.

### 4.6.1 Control Panel for Tree Building

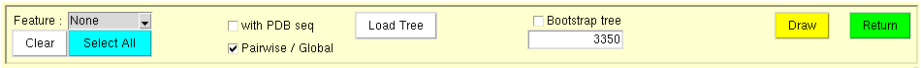


Figure 9: The Tree building tool control panel.

#### 4.6.1.1 Selections

The left section of the Control Panel manages sequence and residue range selection:

- **Feature**: Displays the currently selected feature. Users can select one or more items from this feature to define the residue range for computation.
- **Clear**: Resets all previously selected residue ranges and sequences.
- **Select All**: Selects all residue columns from the entire alignment.

#### 4.6.1.2 Options

The following settings control the tree computation process:

- **With PDB seq**: Includes PDB sequences in the computation. By default, Ordalie excludes them as they are typically redundant with existing sequences in the snapshot.
- **Pairwise / Global**: Defines the gap removal strategy. *Pairwise* (checked) excludes gaps on a per-pair basis during distance calculation. *Global* (unchecked) uses only complete columns (without any gaps) for the entire computation.
- **Load Tree file**: Imports an external tree file in **NEXUS** format. Leaf identifiers must match the sequence names present in the snapshot.

- **Bootstrap**: When enabled, Ordalie performs $N$ bootstrap replicates. The default value is set to 1000, standard value for publication purposes. Imported trees can also be bootstrapped.

#### 4.6.1.3 Draw / Return

The **"Draw"** button launches the computation and opens the tree in a new window. The **"Return"** button exits the Tree tool.

### 4.6.2 The Tree Visualization Window

Each newly computed tree is rendered in a dedicated window designed for deep exploration of phylogenetic characteristics. Ordalie supports two primary visualization modes: **Dendrograms** (rectangular layouts) and **Radial Trees** (circular layouts). While most options are universal, some are tailored to a specific representation.

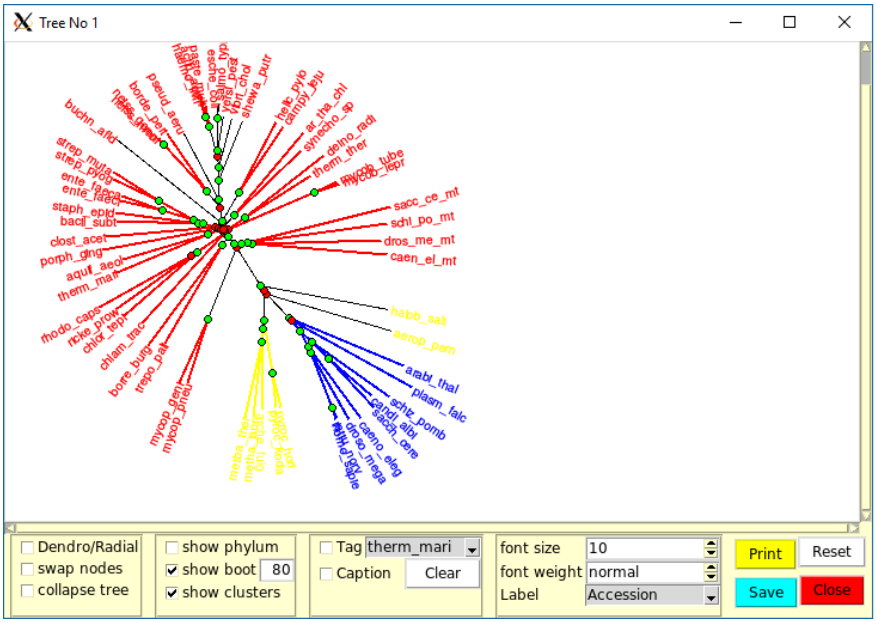


Figure 10: The Tree rendering window. In this radial view, nodes are marked with circles indicating if bootstrap values exceed (green) or fall below (red) the user-defined threshold. Sequences are color-coded by cluster.

The interface is divided into a large **Drawing Area** (top) and a comprehensive **Control Panel** (bottom).

#### 4.6.2.1 The Drawing Area

This interactive area allows for direct manipulation of the tree:

- **Navigation**: Move the entire tree by dragging it with the mouse (Button-1).

- **Scaling & Scrolling**: In a *Radial Tree*, the mouse wheel scales the diagram. In a *Dendrogram*, the wheel scrolls vertically through the sequences.
- **Customization**: A right-click (Button-3) opens a contextual menu. For dendrograms, you can adjust branch lengths and vertical spacing between taxa. For radial trees, this menu allows for precise rotation of the entire structure.

#### 4.6.2.2 The Control Panel

The panel is organized into four functional groups to streamline the analysis:

**1. General Tree Options**

- **Dendrogram/Radial**: Switches the graphical representation of the tree.
- **Swap Node**: Enables orange markers at each node; clicking a marker rotates the descendant branches around that node.
- **Re-root Tree** (Dendrogram only): Enables green markers at nodes. Clicking a marker sets a new root, effectively redefining the outgroup branches.

**2. Visual Annotations**

- **Show Phylum**: Automatically colors sequence names based on their taxonomic domain: by default, Red (Eukaryota), Blue (Archaea), Yellow (Bacteria), and Black (Viruses).
- **Show Bootstrap**: Displays colored discs at nodes based on a confidence threshold (default 80%). Green circles represent high confidence, while red indicates lower support. In dendrograms, absolute bootstrap values are printed; in radial trees, they appear as tooltips when hovering over a node.
- **Show Clusters**: If a clustering analysis was previously performed, branches are color-coded to match their respective clusters.

The colors assigned to life dommains can be changed using "Edit -> Preferences" menu item.

**3. Labels and Metadata**

- **Tag**: Highlights a specific leaf (chosen via the combobox) with a thick black frame for rapid identification.
- **Caption**: Adds a title or explanatory text to the tree. The tool allows for full customization of the caption text and style.
- **Font**: Adjusts the font size and weight (Normal/Bold).

- **Labels**: The **Label** menu selects which metadata to display (Sequence Name, Accession Number, Species, etc.).

**4. Global Actions**

- **Print**: Captures the current tree view and exports it as a high-quality PNG file.
- **Reset**: Restores the tree to its default parameters and orientation.
- **Close**: Exits the visualization window.

## 4.7 The Conservation Tool

Evolutionary pressure, which maintains the structure and function of protein families, leaves detectable traces in residue conservation patterns. Analyzing both global and group-specific conservation helps in identifying key functional sites, such as ligand-binding pockets, protein-protein interaction patches, or residues linked to functional specialization within sub-groups.

Ordalie provides a comprehensive suite of methods to quantify this conservation. Users can experiment with different algorithms, and results are stored temporarily during the session. Once a computation is saved, it is integrated as a new **Feature** within the snapshot, making it available for visualization and analysis across all other Ordalie tools.

### 4.7.1 Conservation Methods

Residue conservation can be measured using various established metrics [20, 5]. Ordalie implements several standard algorithms alongside two proprietary methods.

**4.7.1.1 The 'Threshold' Method** This is a simple frequency-based counting method. Two pre-set thresholds define the levels of conservation:

- **Identity Conservation Threshold**: Set by default to 100%. Columns reaching this level are flagged as 'Identity conservation'.
- **Global Conservation Threshold**: Set by default to 80%. Columns meeting this criteria are considered 'Globally conserved'.

At the group level, only the Identity threshold is applied. These values can be customized in the 'Preferences' menu.

**4.7.1.2 Automatic Methods** These methods follow a two-step statistical approach and only process columns containing more than five residues:

1. **Scoring**: Each column is assigned a raw score based on the chosen algorithm.
2. **Clustering**: Scores are clustered using a distribution analysis. Clusters are then ranked by their mean score. The two highest-ranking clusters are designated as 'Strictly conserved' and 'Globally conserved'. For group-specific analyses, only the top-ranking cluster is retained.

Available automatic methods include:

- **BILD (Bayesian Inferred Log-Likelihood Divergence)**: A Bayesian scoring method that assesses column conservation by comparing a posterior distribution, derived from Dirichlet mixtures, against a background null model [2]. It is particularly robust for small alignments as it uses prior biological knowledge to stabilize frequency estimations, making it highly effective for identifying functional motifs with limited data.
- **Shannon Entropy**: A classic information-theoretic metric that quantifies the uncertainty or randomness at a specific alignment position [14, 15]. The score is calculated based on the frequencies of the observed residues in a column; a lower entropy signifies a highly conserved position, while a higher entropy reflects high variability. It serves as the baseline for most evolutionary conservation analyses.
- **Liu**: An entropy-based approach proposed by Liu *et al.* [11] that quantifies conservation by measuring the reduction in uncertainty at a specific position. Unlike Shannon's classic entropy, this method incorporates sequence weighting or stereochemical properties to better distinguish between purely invariant columns and those exhibiting conservative substitutions.
- **Mean Distances**: The matrix-based algorithm popularized by ClustalX [16]. It calculates the average of all pairwise scores between residues in a column using a substitution matrix (e.g., BLOSUM or PAM). This method effectively captures the "average" biochemical similarity.
- **JSD (Jensen-Shannon Divergence)**: A powerful information-theoretic metric used to measure the distance between the observed residue distribution in a column and a reference background distribution (e.g., BLOSUM62 frequencies) [3]. The JSD is symmetric and bounded, providing a normalized score.
- **Vector Norm**: A proprietary method that projects amino acids into a multidimensional physicochemical space, primarily defined by properties such as Van der Waals volume and polarity. The score is calculated as the magnitude of the mean property vector in this space; a high norm indicates a strong consensus where residues share similar physical constraints. See Appendix 6.3 for the detailed geometric derivation.
- **Multi**: A meta-scoring hybrid approach designed to increase the signal-to-noise ratio. By combining orthogonal metrics chosen by the user through a dedicated window, it aggregates complementary methods that do not measure exactly the same properties.

#### 4.7.2 The Control Panel

The Control Panel provides the following options:

- **Method**: Selection of the scoring algorithm.
- **Title**: An optional label to identify the specific computation.

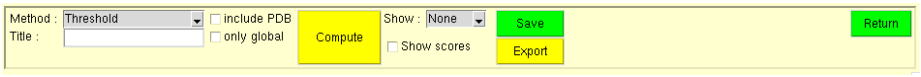


Figure 11: The Conservation tool control panel interface.

- **Include PDB**: If enabled, includes PDB sequences in the calculation (disabled by default to avoid redundancy with the corresponding PDB genomic sequences).
- **Only Global**: If checked, group-specific conservation is ignored, focusing only on the entire alignment.
- **Compute**: Launches the conservation analysis.
- **Show**: A dropdown menu to navigate between temporary scores (labeled *tmp*<*Method*>-$x$).
- **Show Scores**: Opens a graphical frame below the alignment to display the selected conservation scores as a histogram.
- **Save**: Permanently saves the selected score as a new Feature, making it accessible in the main 'Normal' mode. The name of the feature becomes <*method*>-$x$ in the dropdown menu.
- **Return**: Exits the tool, close the 'Scores' frame if opened, and returns to the main interface.

## 4.8 The Superposition Tool

One of the key strengths of Ordalie is its ability to map sequence-based features onto 3D protein models. To fully exploit this mapping, it is essential to account for structural differences between proteins. The Superposition tool allows users to align 3D structures based on specific features or user-defined residue ranges.

### 4.8.1 The Superposition Algorithm

A protein structure may consist of several chains, which can be identical or distinct. Each chain typically comprises an amino acid polymer and various ligands (in Ordalie, water molecules are treated as ligands).

While superposition calculations are performed using polymer sequences, the entities physically moved in the 3D space are the entire chains. Applying a superposition to a chain moves all its components, including both the polymer residues AND the ligands.

The chain superposition process follows three steps:

1. **Zone Selection**: Identification of the structural regions to be used for the alignment (e.g., an entire domain or specific secondary structures).

2. **Chain Selection**: Selection of all chains that need to be moved or used as a reference.

3. **Reference Selection**: Choosing one chain among the selected ones to act as the fixed reference. All other chains will be aligned onto this stationary target.

The detailed mathematical superposition algorithm is presented in Appendix 6.5.

### 4.8.2 Control Panel


Figure 12: The Superposition tool control panel.

#### 4.8.2.1 Selection Tools

The left part of the panel manages the residues used for the alignment calculation:

- **Features**: Displays the active feature. Users can select specific feature items to define the residue range.
- **Clear**: Resets all current residue and sequence selections.
- **Select All**: Selects all residues in the current snapshot.
- **All Helices**: Selects all helical regions across the sequences.
- **All Strands**: Selects all β-strands across the sequences.

The 'All Helices' and 'All Strands' options identify the minimal common intersection of secondary structures at each position across all selected sequences.

#### 4.8.2.2 Control Tools

- **Superpose**: Launches the alignment process. This triggers two sequential windows:
  1. The **Chain Selection** window to pick all chains involved.
  2. The **Reference Chain** window to designate the fixed structure.
- **Display**: Opens the **3D Viewer** for visual inspection (see Section 4.9).
- **Return**: Exits the Superposition tool.

### 4.8.3 Example: Handling a Homo-dimer

Consider a protein known to be a homo-dimer (an $\alpha_2$ quaternary structure). Suppose we have two structures from different organisms, PDB 1abc and 1def, both composed of chains A and B. If the alignment contains only the sequences `PDB_1abc_A` and `PDB_1def_A`:

**4.8.3.1 Scenario 1: Single chain in the snapshot** Ordalie identifies the PDB IDs and downloads the full atomic coordinates for both structures (including chains B). However, if you superpose `PDB_1def_A` onto `PDB_1abc_A`, only the atoms of 1def Chain A will be aligned. In the 3D Viewer, the 1abc dimer remains intact, but for 1def, Chain A will be correctly placed while Chain B will remain in its original coordinate space, effectively "breaking" the 1def dimer symmetry.

**4.8.3.2 Scenario 2: Moving a complete multimer** Ordalie does not automatically infer multimeric states; the user must define the chains of interest.

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

If the alignment contains `PDB_1abc_A`, `PDB_1abc_B`, `PDB_1def_A`, and `PDB_1def_B`, you can superpose the dimers in two steps:

1. Superpose `PDB_1def_A` onto `PDB_1abc_A`.
2. Superpose `PDB_1def_B` onto `PDB_1abc_B`.

This ensures the entire biological assembly is correctly aligned in the 3D space.

## 4.9 The 3D Viewer Tool

The **3D Viewer** is one of the most powerful tools in Ordalie. While it is not intended to replace dedicated molecular visualization software like VMD or PyMol, it is specifically designed to bridge the gap between sequence features and 3D structural analysis.

### 4.9.1 Molecules and Objects

The viewer is organized around two fundamental concepts:

- **Molecule**: Represents all atoms, residues, and chains associated with a specific PDB entry.
- **Object**: A custom-defined selection of elements (entire chains, specific domains, residues, or ligands) belonging to a Molecule.

*Note:* Feature mapping is applied exclusively to **Objects**. By default, Ordalie automatically generates three objects for each chain:

- `AL_<molecule>_<chain>`: A stick representation of all atoms.
- `CT_<molecule>_<chain>`: A Cα(or phosphate) backbone trace.
- `CA_<molecule>_<chain>`: A ribbon (cartoon) representation.

At present, Ordalie does not handle hydrogen atoms.

#### 4.9.2 Representation Types

Multiple styles are available to visualize structural data:

- **Ribbon**: A smoothed path following Cα(amino acids) or P (nucleic acids) atoms.
- **Cα/P Trace**: A simplified line connecting successive Cαor P atoms.
- **CPK**: Space-filling spheres based on Van der Waals (VDW) radii.
- **Pearl**: A simplified view where each residue is represented by a single sphere at its center of mass.
- **Atoms**: A wireframe (line) representation. Standard residues follow internal topology rules, while other compounds (ligands, modifications) use distance-based bonding, which may vary depending on structure quality.

#### 4.9.3 The 3D Viewer Interface

The window is divided into four main functional areas:

1. **Atom Information** (Top): Displays details about picked atoms.
2. **Quick Mapping Panel** (Below information): For rapid feature visualization.
3. **Molecular Objects Frame** (Center-left): Lists defined molecules and their objects.
4. **3D Window** (Center-right): The main interactive viewport.
5. **Actions Panel** (Far right): Toolset for analysis and export.

##### 4.9.3.1 Keyboard Shortcuts

- <**F1**>: Toggles maximizing/restoring the 3D window geometry.
- <**F2**>: Toggles the visibility of all side and top panels.

#### 4.9.3.2 Quick Mapping

This panel uses four comboboxes to map features onto structures. You can map up to two features simultaneously on a single object. *Important:* Features are rendered sequentially (Feature 1 then Feature 2). If Feature 2 covers a larger area (like a PFAM domain) than Feature 1 (like Conservation), Feature 1 might be hidden. In such cases, reverse the order: set PFAM as Feature 1 and Conservation as Feature 2.

#### 4.9.3.3 Molecular Objects Frame

This frame manages the hierarchy of structures:

- **Status**: Object names in **green** are visible; names in **red** are hidden.
- **All On / All Off**: Switches visibility for every defined object.
- **New / Edit / Del**: Used to create, modify, or remove objects for a specific molecule.

#### 4.9.3.4 Interaction in the 3D Window

Ordalie uses a virtual **Arcball** (trackball) system for navigation:

- **Rotate**: Drag with mouse **Button-1** (Left click).
- **Zoom**: Use the **Mouse Wheel**.
- **Translate (Pan)**: Drag with mouse **Button-3** (Right click).
- **Identify**: <**Control-Button-1**> click on an atom to display its label.

#### 4.9.3.5 The Actions Panel

Key functionalities include:

- **Distances**: Calculates the distance between two sequentially picked atoms.
- **Centre On Atom**: Centers the rotation of the scene on the last picked atom.
- **Print**: Exports the current scene as a PostScript (PS) file.
- **Reset / Clear**: Resets distances, labels, or ongoing picking actions.

### 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).

The name of a new object is enterred in the entry box at the top of the interface. Be aware that objects names should be unique. Then the creation or edition of an object is an iterative five-step process:

1. Select the **Chain** and components (Polymer residues or Ligands).

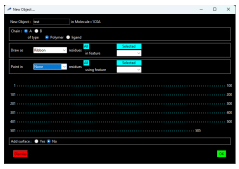

Figure 13: The 3D object creation and edition interface.

2. Select a **Representation Type**.
3. Define the **Residue Range** for that style.
4. Select a **Colour**.
5. Define the **Residue Range** for that 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.

#### 4.9.4.1 Residue Selection Methods

Three modes of residue selection are available for applying styles or colors:

- **All**: Applies to the entire chain.
- **Selected**: Applies to residues highlighted in the residue frame (via click-and-drag with Button-1).
- **Feature**: Selects residues based on a predefined sequence feature (e.g., specific domains or conserved sites).

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 Tool

This tool is dedicated to feature management, including creation, deletion, and edition of sequence characteristics.

### 4.10.1 Control Panel

The Control panel of the **Features Editor** provides quick access to global actions.


Figure 14: The Control panel of the Feature Editor tool

The panel includes the following options (from left to right):

- **Features:** Displays the currently selected feature. The user can select a specific feature item as a residue range.
- **Clear Selection:** Resets all residue ranges and sequences previously selected.
- **Delete Feature:** Permanently removes the current feature across all sequences.
- **Load Feat. File:** Imports a file containing user-defined features (see Appendix 6.4 for format details).
- **Features Summary:** Launches the summary tool for an overview of all features within the snapshot (see section 4.11).
- **Save:** Commits changes and returns to 'Normal' mode.
- **Return:** Cancels changes, restores original features, and returns to 'Normal' mode.

### 4.10.2 Notation and Actions

It is essential to distinguish between a **Feature** and a **Feature Item**:

- **Feature:** A general set of instances representing a specific sequence characteristic (e.g., "Alpha Helix").
- **Feature Item (or Item):** A single instance of a Feature occurring at a specific position for a specific sequence.

### 4.10.3 Contextual Menu

Unlike other tools, the Feature Editor allows direct interaction within the snapshot window via a right-click contextual menu.

In this tool, the behavior of `<Button-1>` (Left Click) is modified by modifier keys:

- `<Button-1>` alone: Applies action to the sequence under the pointer.
- `<Control + B1>`: Applies action to the entire group containing the sequence.
- `<Shift + B1>`: Applies action to all sequences in the snapshot.

**4.10.3.1 Select Item** Selects the Item directly under the mouse pointer. The scope (sequence, group, or all) depends on the modifier keys mentioned above. If <Control> is pressed with <Button-1> then the item present in the sequences of the current group will be selected. If <Shift> is pressed then the item present in all sequences of the snapshot will be selected.

**4.10.3.2 Select All Items** Selects all Items of a sequence, a group, or the whole snapshot depending on the modifier key.

Selecting all Items for all sequences effectively selects the entire Feature. If deleted, the Feature itself is removed from the project.

**4.10.3.3 Select Region** Define a residue range by clicking and dragging `<Button-1>`. This region can be used to define a new Item. The scope follows the standard `Control/Shift` modifiers.

**4.10.3.4 Clear Selection** Clears all active selections.

#### 4.10.4 Edition Options

After selecting Items or a region, the following options become available in the menu:

**4.10.4.1 Edit Item** Modifies properties of existing Items. If multiple Items are selected, some properties may be locked unless they refer to a single zone:**Residue range** (Only editable for single-zone selections), **Item Colour**, **Score**, and **Note**.

**4.10.4.2 Define New...** Opens a configuration window:

- If the **Feature Name** matches an existing one, the item is added to that feature.
- If the name is new, a new Feature is created.
- Users must assign a **Colour**; **Score** and **Note** are optional.

**4.10.4.3 Delete Selected Items** Removes the selected items. Deleting all items of a feature will result in the deletion of the feature itself.

**4.10.4.4 Propagate Items (Group/All)** **Propagate to Group** will copy selected Items to all sequences in the current group while **Propagate to All** will copy selected Items to every sequence in the SNAPSHOT. Note that propagation will skip sequences where the Item is already present.

### 4.11 The Features Summary

The **Features Summary** tool allows the simultaneous rendering of multiple selected sequences and their associated features on a single page. Unlike the *Snapshot Overview*, sequences in this view maintain their appearance as seen in the main alignment window.

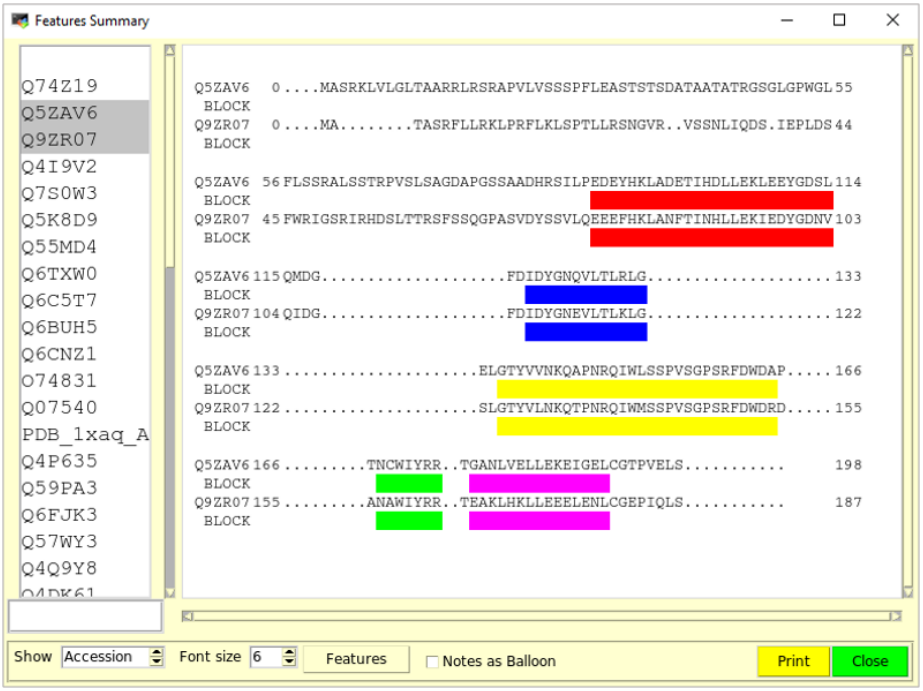


Figure 15: The Features Summary window, showing the Drawing Area (top) and the Control Panel (bottom).

### 4.11.1 Drawing Area

The upper portion of the window consists of a listbox for sequence selection and a main drawing area. Clicking a sequence name in the listbox toggles its selection. Multiple sequences can be selected by holding the `<Control>` key while clicking. For each selected sequence, the display includes:

- The **Sequence ID** on the left.
- The **Start and End positions** of the residues for the current line.
- The **Amino acid sequence** as it appears in the alignment.

When a feature is active, a dedicated line appears below the sequence. The feature name is displayed under the sequence ID, and colored rectangles mark the positions where the feature is present. The entire view can be panned by clicking and dragging with `<Button-1>`.

### 4.11.2 Control Panel

Located below the drawing area, the Control Panel provides several configuration options:

**ID Type Selector (Spinbox):** Choose how sequences are referenced (Sequence Name, Accession Number, or Bank ID). This change updates both the listbox and the drawing area.

**Font Size Selector:** Adjusts the text scale for better readability.

**Features Selector:** A list of checkboxes allowing the user to overlay any number of features onto the sequences.

**Notes as Balloon:** When enabled, the notes attached to a feature appear as a tooltip (balloon) when hovering over the feature with the mouse.

**Print:** Prompts for a filename to export the current drawing area as a **PNG image**.

**Close:** Exits the tool and closes the window.

## 4.12 The Annotation Tool

This tool allows users to add, modify, and remove "memo tags" anywhere on the snapshot. These annotations can be toggled on or off at any time using the **Annotation** button located on the snapshot bar.

A memo tag is defined by three main characteristics: its position and size, its color, and its attached text note. These tags are particularly useful for marking zones of interest, such as conserved regions or areas requiring manual realignment. Furthermore, color-coding can serve as a classification system—for instance, using pink for structural annotations and cyan for sequence features.

### 4.12.1 Control Panel

The Control Panel is designed for straightforward management of visibility and tool states. From left to right, it includes: The **Show annotation coloured in** option filters the display to a specific color, while **Show all** and **Hide all** manage the global visibility of all annotations on the screen. To exit the tool, the **Dismiss** button leaves without saving further changes, whereas the **Return** button exits and saves all modifications made to the annotations.

### 4.12.2 Keyboard Bindings

Annotation creation and editing are primarily mouse-driven. The interaction logic is summarized in Table 3.

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

Table 3: Keys combination for annotation management

Upon releasing the mouse button after defining a zone, a configuration window appears. This interface allows the user to specify the annotation's color and enter the associated text. This same window is invoked when editing an existing annotation. Finally, clicking **Return** ensures all changes are committed.

## 4.13 Barcode

The **Barcode** tool provides an alternative schematic representation of a snapshot, offering a condensed bird's-eye view of the alignment. A typical example of this visualization is shown in Figure 16.


Figure 16: The Barcode window.

The upper frame of the tool contains the primary drawing area. In this representation, a dedicated lane is drawn for each sequence group within the snapshot. These lanes serve as a visual summary, where conserved regions, gaps, or specific features are rendered as vertical bars of varying thickness and color, effectively creating a "barcode" signature for each group. This allows for rapid comparison of structural or sequence patterns across different clusters of sequences.

# 5 Menus Description

This chapter describes all the Ordalie menus. Only the entries which have not been already explained in the 'Tools' section 4 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, extensions .tfa .fa.tfasta.fasta.
- MSF (Multiple Sequence File, GCG Wisconsin Package) format, extension .msf.
- Macsim/XML format. The Macsim format is an XML file output of the MACSIMS program or server (see [18] and `http://www.lbgi.fr/~julie/MACSIMS/` for details). Accepted extensions are .xml, .macsim, .macsims.
- ALN (ClustalW) file format [17], extension .aln.
- RSF file format (Rich Sequence File), as output by the SeqLab program [19], extension .rsf.
- ORD (Ordalie) file, Ordalie specific file format, extensions .ord, .ordalie

A format checking is done using the file extension 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

This command pastes the sequences currently stored in the clipboard. Sequences are inserted immediately following the currently selected sequence. If no sequence is selected, no sequences are pasted. When pasting sequences that already exist in the snapshot, their names will be automatically suffixed with `__<n>`, where `<n>` denotes the copy number.

#### 5.2.4 Preferences

Open the *Preferences* panel to access the default internal parameters of Ordalie.

Figure17 shows the *Preferences* panel. Clicking an item in the list box on the left displays the corresponding settings.

- **General**

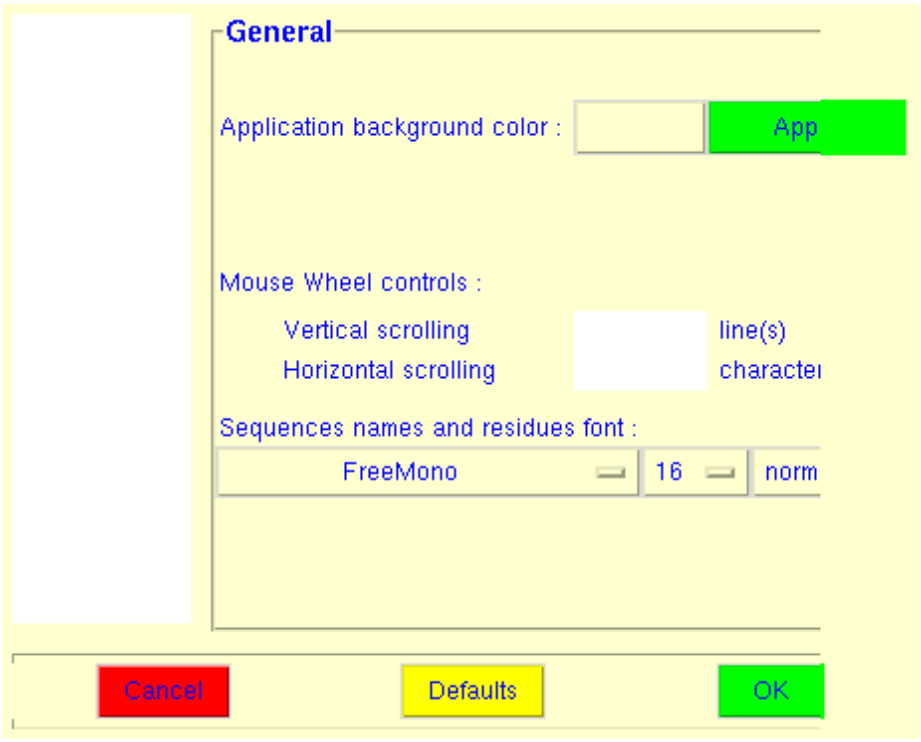


Figure 17: The Preferences panel.

- **Application background:** Click the colored button to open a color palette and select the application background color.
- **Mouse wheel controls:** Defines the number of lines or columns moved for each mouse wheel increment.
- **Sequence names and residues font:** Menu buttons allow selection of font family, size, and weight (normal or bold). Changes are applied immediately in the main window.

- **Sequences**
  - **Remove bank prefix when loading sequences:** Sequence names may be prefixed by their database identifier. By default, Ordalie removes this prefix.
  - **Add a new bank prefix:** Ordalie recognizes the following prefixes by default: SPT, SW, PDB, VARSPLIC, UniRef90, UniRef100, PQS, ENS, SP, UNP, TR, REF, GB, DBJ, ENA. A new prefix can be entered in the text field.
  - **Life domain colors:** Click the *Eukaryota*, *Archaea*, *Bacteria*, or *Viruses* buttons to change the associated colors using a color palette.

- **Conservation**
  - **Conservation type colors:** Defines foreground and background colors for the three conservation levels used in Ordalie.
  - **Threshold:** Threshold above which a column is considered conserved.
  - **Use similarity:** Includes physicochemical similarity when computing conservation. Default is *No*.

- **Editor**
  - **Lock sequences at start-up:** By default, only gaps can be added or removed. Selecting *Yes* allows amino acid insertion or deletion in the editor.
  - **Residue mapping colors:** Defines how residues are displayed in the editor. The default scheme is based on the SeqLab editor [19]. A colorblind-friendly scheme is also available. Clicking the *Default* button restores the original colors.

- **3D Display**
  - **Default ribbon color:** Color used for ribbon representation when no feature is applied.
  - **Precompute surface:** Determines whether the solvent-accessible surface is computed when creating a new object. Default is *No*.
  - **Color ramps:** Clicking a button changes the corresponding color.

- **Trees**
  - **Gap removal:** Selects the gap removal strategy (see Section 4.6.1). Default is *Pairwise*.

- **Type of tree:** Default is dendrogram.
- **Tree font:** Defines size and weight (normal or bold) for tree labels.
- **Bootstrap:** Enables tree bootstrapping. Default is *No*.

## 5.3 The View Menu

### 5.3.1 Bigger Font

Increases the font size for the sequence display. The default font type and size for Ordalie can be configured via `Edit` → `Preferences` under the **General** tab.

### 5.3.2 Smaller Font

Decreases the font size for the sequence display.

### 5.3.3 View Log

Most of Ordalie's operations and computations are recorded in a log. Selecting this item allows the log to be displayed either in a dedicated window as raw text or in a web browser as an HTML document. The log window also provides an option to save the content as a raw text file.

### 5.3.4 Save Log File

Exports the entire Ordalie log to a file. A sub-menu allows the user to choose between HTML and Text formats. The output file name is derived from the root name of the original alignment file, appended with the corresponding extension (`.txt` or `.html`).

### 5.3.5 Toggle Full Screen

Toggles the main window between its current size and full-screen mode. This action can also be triggered by pressing the `F1` key.

### 5.3.6 Show/Hide Icon Bar

Displays or hides the shortcut icon bar. Hiding this bar is useful when maximum vertical space is required, such as during intensive alignment editing.

### 5.3.7 Show/Hide Scores

Toggles the visibility of the scores panel. When a conservation score is active (enabled via the Features buttons or the **Conservation** tool), the global and group-specific scores for each column are plotted. By default, the scores panel is hidden.

### 5.3.8 Show/Hide Features Panel

Toggles the visibility of the Features panel. This is useful for maximizing the alignment display area during manual editing.

### 5.3.9 Show/Hide Phylum

If taxonomic information (Life Domain) is available for the sequences, the sequence names are color-coded accordingly: **Eukaryota** in red, **Archaea** in blue, **Bacteria** in yellow, and **Viruses** in black. These default colors can be customized in the **Preferences** panel.

### 5.3.10 Show/Hide Group Colors

Applies color-coding to sequences based on their assigned groups.

### 5.3.11 Show/Hide Secondary Structures

If secondary structure information is available, this option displays or hides α-helices (represented in red) and β-strands (represented in green).

## 5.4 The Sequence Menu

### 5.4.1 Display Names

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

- **Sequence Name**: A user-defined name.
- **Accession Number**: The unique identifier from the source database.
- **Database ID**: The database entry identifier.
- **Full Organism Name**: The complete scientific name (if available).
- **Genus.Species Abbreviation**: A shortened version of the organism name (if available). For example, *Homo sapiens* becomes *Homo.sapi* and *Saccharomyces cerevisiae* becomes *sacc.cere*.

By default, the **Sequence Name** is displayed.

In Macsims/XML and Ordalie file formats, the Sequence Name, Accession Number, and Bank ID are explicitly identified for each sequence. For all other formats, where only one ID per sequence is typically provided, this identifier will be used as the Sequence Name, Accession Number, and Bank ID simultaneously.

### 5.4.2 Identity Tool

Launches the Identity tool; for further details, see Section 4.3.

### 5.4.3 Search Motif

Launches the Search Motif tool; for further details, see Section 4.4.

### 5.4.4 Retrieve Sequences Information

This command retrieves up-to-date information for specific sequences by querying their original databases. Upon invocation, a dialog box appears asking whether to process all sequences or only the selected ones. Once confirmed, Ordalie queries the corresponding databases to fetch the latest metadata.

If any amino acid sequences have been updated in the database, they will be automatically realigned with the existing sequences using the MAFFT program. After completion, a summary window displays the number of updates performed for Accession, Phylum, Taxa ID, Organism, Description, and the amino acid sequence itself. Below the summary table, the "Show details..." button opens a web page providing a detailed log of the changes for each sequence.

### 5.4.5 Edit Sequence Information

Opens the Sequence Information Editor. This tool provides access to all metadata available for a specific sequence and allows the user to modify individual fields. A dropdown menu (combobox) at the top of the window enables the selection of the desired sequence to be edited.

### 5.4.6 Add Feature from File

Opens a file dialog to import a feature file. Once loaded, these features are immediately available within Ordalie. The supported feature file format is described in Appendix 6.4.

### 5.4.7 Export Features to File

This function allows exporting features associated with a selected set of sequences into a feature file (see 6.4).

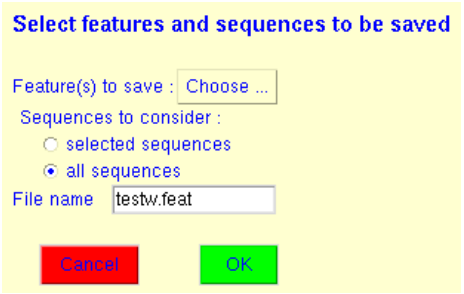


Figure 18: The Export Features window

Figure 18 shows the "Export Features" window.

Clicking the *Features* menu button displays the list of available features as a set of checkboxes. The user can select one or more features to export.

The sequences from which the features are extracted can be either all sequences or only the currently selected ones.

Finally, the user must specify the name of the output file.

### 5.4.8 Features Editor

Launches the Features Editor tool; see Section 4.10.

### 5.4.9 Features Summary

Launches the Feature Summary tool; see Section 4.11.

### 5.4.10 Add Sequences to Snapshot

This item opens a menu for adding sequences from databases or local files.

- **From Database**: Ordalie downloads sequences from UniProt or RefSeq based on a user-provided list of accession numbers.
- **From File**: The user provides a sequence file in TFA, MSF, or ALN format for Ordalie to parse.

The sequences are then added to the current snapshot.

A dedicated window appears based on the selection to collect the necessary parameters. Figure 19 illustrates the interface for adding sequences from a database.


Figure 19: The interface for adding sequences from a database.

After specifying the source, the user should define the following settings:

- **Alignment**: Choose whether the added sequences should be automatically aligned using MAFFT against the existing snapshot sequences.
- **Duplicates**: Choose whether to discard duplicate sequences or create copies. In the case of duplicates, the new sequence names will be suffixed with `__<n>`, where `<n>` represents the copy number.
- **Insertion Point**: Specify whether to append sequences at the bottom of the snapshot, insert them immediately after a specific sequence (selected via a dropdown menu), or cluster them based on sequence identity (available only if the sequences are being aligned).

Once the sequences are added, feature positions are recomputed if necessary. Basic sequence characteristics (isoelectric point, molecular weight, identity, composition, etc.) are then calculated for the new entries and recorded in the log.

### 5.4.11 Add PDB to Snapshot

A PDB structure can be added by specifying a PDB ID (for automatic download) or by providing a local PDB file.

## 5.5 The Alignment Menu

### 5.5.1 Editor

Launches the Alignment Editor tool; see Section 4.2.

### 5.5.2 Re-Align Sequences

Re-aligns the selected sequences using the MAFFT program [8, 7, 6, 9]. The selected sequences are first degapped and then re-aligned against the remaining sequences. The re-aligned sequences maintain their original vertical positions within the snapshot.

### 5.5.3 Clustering

Launches the Clustering tool; see Section 4.5.

### 5.5.4 Add Separator

In Ordalie, a "separator" is a blank line inserted between sequences. Sequences separated by these lines constitute a distinct group or cluster. Adding separators allows users to define their own custom clusters. This command inserts a blank separator line immediately below the selected sequence. Note that exactly one sequence must be selected for this operation.

### 5.5.5 Remove Separator

Removes the separator located immediately below the selected sequence. Note that exactly one sequence must be selected.

### 5.5.6 Remove All Separators

Removes all separators and clears all sequence selections, resulting in an unclustered snapshot.

### 5.5.7 Overview

Creates an instance of the Overview Window for the current snapshot; see the "Snapshot Overview" in Section 4.1.

### 5.5.8 Conservation

Launches the Conservation computation tool; see Section 4.7.

### 5.5.9 Tree Building

Launches the Phylogenetic Tree tool; see Section 4.6.

### 5.5.10 Annotate Snapshot

Launches the Annotation tool; see Section 4.12.

### 5.5.11 Barcode

Launches the Barcode tool; see Section 4.13.

### 5.5.12 Toggle Physicochemical Coloring

Toggles residue coloring using the same mapping as in the **Editor**. The residue color scheme can be customized in `Edit` → `Preferences` under the **Editor** tab.

## 5.6 The Structure Menu

### 5.6.1 Superpose Structures

Launches the **Superposition** tool, provided that multiple PDB structures are available in the current snapshot; see Section 4.8.

### 5.6.2 Display Structures

Launches the **3D Viewer** tool; see Section 4.9.

### 5.6.3 Color Secondary Structures by Identity

This tool allows for the coloring of secondary structures based on their sequence identity levels within the snapshot. Upon invocation, a dialog box appears with the following configuration options:

- **Similarity Computation**: Calculate similarities using either the average secondary structure limits or a specific reference sequence.
- **Color Gradient**: Select between a grayscale or a multi-color gradient.
- **Gradient Scale**: Define the limits from the minimum to maximum identity, or across a fixed 0% to 100% range.
- **B-factor Assignment**: Choose whether to map the identity scores to the B-factor column of the PDB file.

### 5.6.4 Save PDB

Since 3D structures may be modified (e.g., coordinates transformed) using the **Superposition** tool, this option allows the user to save any 3D structure present in the snapshot to a file. This is particularly useful for exporting superposed structures to more advanced molecular rendering software.

When selected, a window prompts the user to define the export parameters:

- Select the target molecule.
- Select either all chains or a specific chain.
- Select all residues or a specific residue range.

To complete the process, the user will be prompted to specify the output PDB filename.

## 5.7 The Help Menu

### 5.7.1 About Ordalie

This menu item opens a window providing essential information about Ordalie, including the current version number and the link to the official Ordalie home page.

### 5.7.2 Local Documentation

Launches this manual locally using the built-in Tcl HTML viewer.

### 5.7.3 Online Documentation

Launches the Ordalie documentation in a web browser, redirecting to the official Ordalie website.

# 6 Appendix

## 6.1 The Ordalie database scheme

The core of Ordalie is build around an in-memory SQLite database [4] which scheme is given in figure 20.

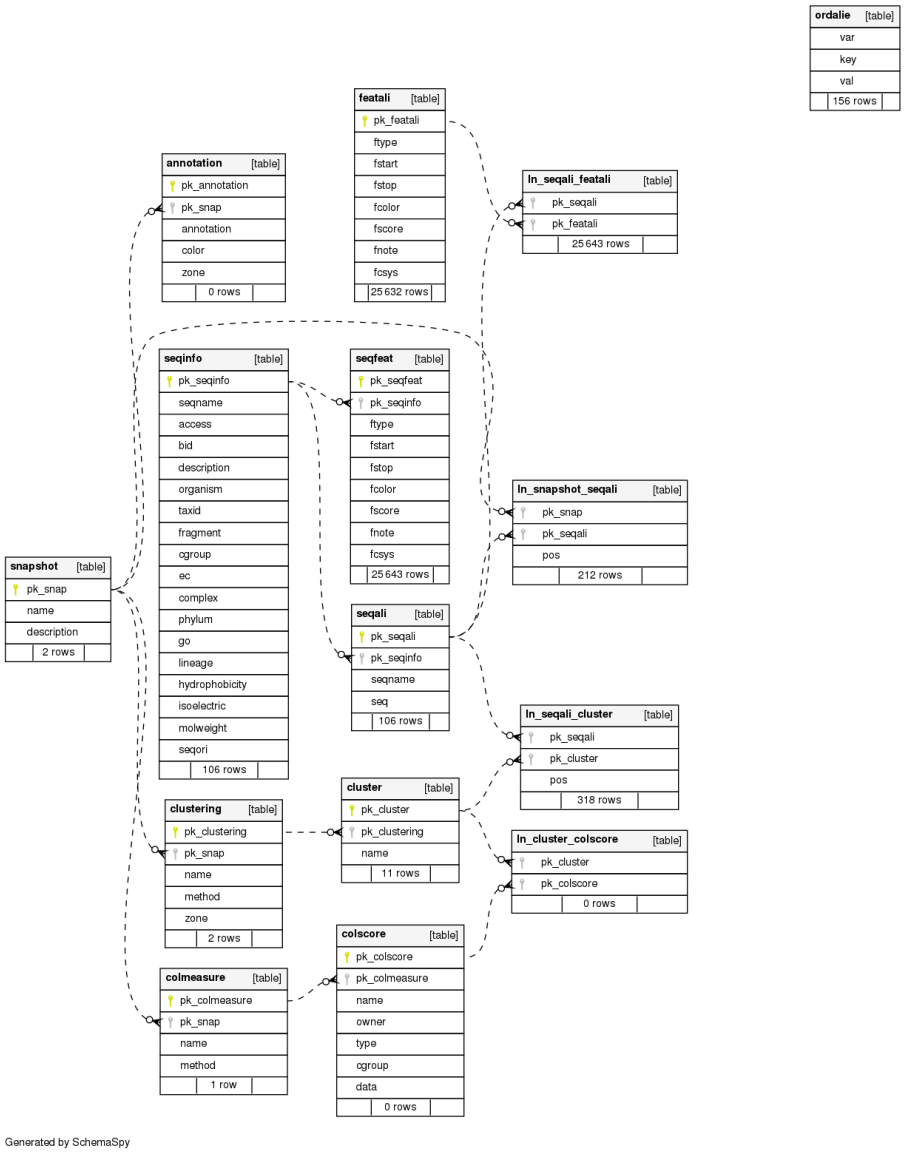


Figure 20: Scheme of the ordalie database.

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] John A Capra and Mona Singh. Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23(15):1875–1882, 2007.

[4] Richard D. Hipp. *SQLite: An SQL Database Engine in a C Library.* Hwaci, Charlotte, NC, 2024.

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

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

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

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

[14] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[15] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(4):623–656, 1948.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

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