RESEARCH

ORDALIE: digging and mapping knowledge on multiple sequence alignments

Luc Moulinier[1]*
, Raymond Ripp[1]
, Anne Friedrich[2]
, Marie Sissler[3]
, Audrey Defosset[1]
, Jean Muller[4,5]
and Olivier Poch[1]

*Correspondence: luc.moulinier@unistra.fr
[1]Complex Systems and Translational Bioinformatics, ICube, UMR 7357, CNRS, Université de Strasbourg, Fédération de Médecine Translationnelle, 11, rue Humann, 67000 Strasbourg, France
Full list of author information is available at the end of the article

Abstract

Background: Text for this section.

Findings: Text for this section.

Conclusion: Text for this section.

Keywords: multiple sequence alignment; visualisation; editing; transversal analysis

Introduction

Multiple Sequences Alignment (MSA) plays a key role in modern bioinformatics as cornerstone tool for diverse biological studies such as sequence analysis, evolutionary studies, comparative genomics... All these approaches have in common to use MSAs as primary source of integration allowing the retrieval of several kind of data (structural, functional and even cellular, tissular localisation or developmental stage) in the framework of the closeness of the related aligned sequences. Recently, MSAs became a major support for the study of inter-individual genomic variation pathogenicity providing the best evaluation scoring systems of potential disease-causing insertion/deletion or amino acids replacements [1, 2]. All this constatations yield to new requirements for programs handling MSAs.

Being considered as a main data source or a data mapping environment, a fruittful MSA exploitation is directly dependent on the accuracy of the alignments. Although the research on algorithms dedicated to align sequences is still intensive and the outcoming softwares are more and more accurate [3, 4, 5, 6], the need of manual MSA inspection, curation and editing is still necessary. Furthermore MSA as data source are usually customized depending on the targeted study, the sequences can be clustered differently for example. and the embedded featuresmay also . This appeals to maintain several instances of a same alignment between which the user can swap.

Numerous programs for MSA visualisation, edition and manipulation already exist [7, 8] and are still being developped [9, 10] displaying all different strengths and weaknesses. // We present here ORDALIE (ORDered ALignment Information Explorer), a workbench dedicated to the analysis and exploration of the informational content of a protein MSA.

ORDALIE has been designed around three main ideas : a powerfull set of editing capabilities, the ability to annotate sequences with characteristic features, and a mechanism to handle several instances of a given alignment (hereafter called "snapshots"). The ORDALIE ability to manage, manipulate and store several snapshots and their associated features data combined the associated tools provided give the biologist a frame to perform a transversal analysis of the alignment. Finally, as bioinformatics become a tool used in all laboratories, the graphical user intece of ORDALIE benefits a special care.

The ORDALIE core components

The database

The core of ORDALIE is build around a in-memory SQLite database [11] which scheme is given in figure ??. ORDALIE takes advantage of this underlaying database to store snapshots and their associated features.

The "ordalie" table contains settings parameters saved on exit allowing the user to find the same state when launching ORDALIE again. The "seqinfo" table contains sequence information that are not linked to aminoacids positions (sequence length, molecular weight, isoelectric point, ...). The "seqfeat" table is used to store features data mapped onto the aminoacids sequence. Upon loading of a new alignment file, ORDALIE creates a first snapshot as being a read-only copy of this alignment and it is stored as "original alignment". 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" couples this two tables. The "clustering" table defines a clustering with its name, residue zones used to compute it and the algorithm used, and the "cluster" table stores the resulting clusters with their names. The "ln_seqali_cllluster" link table defines the set of sequences belonging to a given cluster. The parameters for the computation of the residue conservation along the columns (columns scores measurements), i.e. name and used method used, are stored in the "colmeasure" table while the column scores themselves for each cluster are stored in the "colscore" table. The table "ln_cluster_colscore" makes the cluster ans columns scores linkage. Finally, the "annotation" table contains all information relative to annotation the user added to a given snapshot.

The ORDALIE type file (.ord file extension) consists in a database dump.

Snapshot Management

ORDALIE reads and writes alignments in MSF, Fasta, Cl ustal, MACSIMS/XML file formats and the specific ORDALIE file format. Once the alignment is loaded, the user can create several instances of thsi alignment, hereafter called "snapshot". The user can navigate between snapshots and modify or delete them.

The user can cut, copy, paste sequence(s) inside a snapshot. Empty sequences names (called "separators") can be inserted or removed in order to create user-defined clusters outside the "Clustering" tool (see below). New equences can be imported as FASTA type sequences. Sequences can be aligned or re-aligned against the current set of sequences of the snapshot using the DDBJ MAFFT web service [12]. The whole or parts of a snapshot can be printed for publication issues.

Snapshot Editing

Manual curation remains an important task in order to imrpove the global alignment quality. We develop a dedicated Tk widget written in C for performance issues allowing fast snapshot edition. The ”Editor” functionnalities are based but extend the old but famous SeqLab editor [13] that was part of the GCG Wisconsin package [14]. By default, the user can only insert/remove gaps inside one or several sequences. Options for grouping/ungrouping sequences and removing columns of gaps are provided. Amino acids may be edited after unlocking sequence. The residues coloring scheme is adapted from SeqLab and can be customized. After edition, the user can save or discard the modified snapshot.

Features

In ORDALIE, a feature is defined as a position dependent sequence characteristic like the extend of a PFAM domain, a transmenbrane regions, PROSITE sites, exons limits for example. A feature is specified by the sequence(s) it applies to, a start and stop position, a color, an associate score, a note, a coordinates system (”global” for snapshot position or ”local” for sequence relative position). Features are imported into ORDALIE through the Macsims program XML output file [15], by using a dedicated feature file format, or manually defined using the ”Feature tool”. With this tool, features can be created, modified or deleted.

The ORDALIE user interface

The ORDALIE interface has been designed to be intuitive and easy to use for non-bioinformatician users. The ORDALIE emainwindow is divided in three parts. The top part contains the menu bar, the icons bar which are shortcuts to most usefull menu items, and the snapshot management bar, allowing snapshot selection, creation and deletion. The middle part of the window contains the sequences names list on the left and the amino acid sequences on the right. Sequences names can be searched through a text box placed below the sequences names. The first matching name is then positionned at the top of the names list. Below the amino acid sequences, the current position of the mouse pointer are shown both in global alignment and local sequence coordinates systems. A ruler gives the columns positions attached to a horizontal slider. The bottom part of the window hosts by default the on/off buttons for the available features. There is one button per available feature. The user can display several features at the same time, the features being stacked in the order they are switched on. When calling a tool, the features buttons are replaced by the control panel of the tool. Each tool has its own control panel (see below for a brief description), and leaving the tool restores the features buttons.

The ORDALIE associated Tools

Trees

ORDALIE builds phylogenetic trees based on all or parts of the aligned sequences, and all or parts of the snapshot columns. Based upon the user selection, ORDALIE computes a distance matrix using identity percentage, and then processes the distance matrix by the FastME program [16] using default parameters. Although likelihood-based trees are more reliable, the speed and reliability of FastME is enough to have a first insight into the protein phylogeny. The robustness of the tree nodes can be assessed through bootstrap statistics. The user can also import a pre-computed tree in Newick format. The computed or imported tree is then displayed in a dedicated window. The tree can be viewed as a dendrogram or as a radial tree. The user can re-root the tree, swap nodes, display bootstrap values, show nodes above a bootstrap threshold, change leaves labelling, color leave according to their cluster or their life domain, print the tree and more. In the drawing area a contextual menu allows scaling of the tree and rotation of a radial tree. The ORDALIE ability to build tree on parts of the snapshot allows the user to estimate the evolution differences between domains of the protein for example.

Clustering

Sequence clustering is a tool that may enlight differences between sequences according to a given criteria. The ORDALIE clustering tool can be applied to all or parts of the columns. The user should choose one or several numerical criterias and one clustering method. The available criterias are identity percentage, isoelectric point, sequence length, hydrophobicity and aminoacid composition. The clustering algorithms along with criterias to automatically define the number of clusters are taken from the Cluspack package. The available methods are: hierarchical clustering/secator [17], k-means clustering with DPC (Density Point Clustering) [18], and mixture model clustering with AIC or BIC criteria [19, 20, 21]. The special ”Life Domain” criterion clusters sequences into ”Eukaryota”, ”Archaea”, ”Bacteria”, ”Viruses” or ”unknown” according to the sequence lineage. This lineage is part of the Macsims/XML information file or can be retrieved by ORDALIE.

All computed clustering can then be saved and retrieved later.

Conservation

The sequence conservation constitutes a fingerprint of the evolution pressure effects and reveals zones of functional interest in the sequence. When narrowing the focus of the conservation analysis at the cluster level, the outcoming information may enlight the clustering choice, for example, sequences clustered according to mesohiles, psychrophile and therophiles will display a different pattern than sequences clustered according to phylogenetic ranks.

Many algorithms exist to compute residues conservation or dispersion [22, 23]. One of the interesting feature of ORDALIE is that it not only computes the conservation scores along the snapshot columns, but it automatically defines and displays the two most conserved groups of columns in the snapshot. Indeed, the conservation scores are clustered using a hierarchic clustering and the “secator” programs. The two groups having the highest mean conservation score corresponding to the two most conserved groups of residues are then displayed with a black and dark gray background respectively.

ORDALIE implements six methods to compute residue conservation: ”Threshold”, ”Liu”, ”Mean Distances”, ”Vector Norm”, ”Multi” and ”BILD” [24, 25, 26, 27]. The “Vector Norm” is a home-made method. The ”Threshold” method makes a simple counting of conserved residues inside a column. The columns are then partitionned into three groups: 100% conserved, ¿80% conserved and ¿60% physicochemical conserved using the aminoacids groups PAGST, DEQN, KRH, FYW, ILMV, and C.

Conservation computation can be done either considering the whole set of sequences as one group (”Global” option) or taking the sequence clustering into account. In the later case, ORDALIE outputs the first two global conservation groups and also the most conserved columns in each seuence clusters. The conservation scores can be plotted under the snapshot columns by setting the ”Show Score” ckeckbox. ORDALIE stores temporarily all computation trials until the user saves the most relevant ones.

3D viewer

The main interest of the “3D Viewer” tool resides in its ability to map snapshot features to a 3D structure. Contextualizing a feature within 3D structure aspects like molecular surface accessibility, spatial closeness to other features, etc ... is greatly valuable to understand a feature function.

In ORDALIE, if a sequence name refers to a PDB ID [28], the corresponding 3D structure coordinates file is automatically downloaded, processed and can be displayed in the “3D viewer” tool. By default, ORDALIE builds three 3D models representations: a full atoms model, a Cα trace model and a flat ribbon model. Models rotation and translation follow a mouse virtual trackball. All available features can be mapped on a model. After selecting a structure and a model, the user can map two features onto the model. A model builder allows to create customized model representation and coloring.

Other tools

The “Overview” tool creates a schematic representation of the current snapshot as a white and grey pixel map onto which one or more features can be drawn. This gives a an overview of the features distribution along the snapshot. The overview can be scaled and printed.

The “Search” tool allows to find motifs inside the aligned sequences. The search motif syntax follows the FindPattern program syntax [14]. The pattern may be degenerated. Occurences of the motif appear in red in the sequence window.

The “Fetch Information” tool requests the UniProt and RefSeq databases [29, 30] using sequence IDs to retrieve relevant information, like organism name, description, lineage, etc...

Conclusion

By embedding in a same program a database system along with alignment manipulation and exploration tools as well as a complete feature definition and mapping, ORDALIE provides a framework to explore several alignment analysis hypothesis by considering several contexts (thermoophilicity, protein interaction networks, protein sub-families, ...). The ORDALIE file type (a dump of the internal database) ensures a.backup and restore session capability. The interconnected tools (clustering, phylogenetic tree, conservation computation, 3D structure mapping) allow a broad variety of information mining for alignment exploitation. Finally, the ORDALIE interface has been designed to be intuitive and user-friendly making ORDALIE accessible to non bioinformaticians users.

Availability and Requirements

ORDALIE is written in Tcl/Tk [31], version 8.6.6. The alignment editor Tk widget (Biotext) is written in C for performances issues and is included in the ORDALIE distribution. Installers and binaries are provided for Windows, Mac OS X and Linux as well as source code and documentation at http://www.lbgi.fr/ordalie. ORDALIE is an open-source program and is distributed under the LGPL licence.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
Not applicable.

Competing interest
Not applicable.

Funding
Author's contributions
All authors read and approved the final manuscript.

Acknowledgements
Text for this section . . .

Author details
[1]Complex Systems and Translational Bioinformatics, ICube, UMR 7357, CNRS, Université de Strasbourg, Fédération de Médecine Translationnelle, 11, rue Humann, 67000 Strasbourg, France. [2]Université de Strasbourg, CNRS, Génétique Moléculaire, Génomique, Microbiologie, UMR 7156, , 67000 Strasbourg, France. [3]Architecture et Réactivité de l'ARN, CNRS, Université de Strasbourg, IBMC, 15, rue René Descartes, 67084 Strasbourg Cedex, France. [4]Laboratoire de diagnostic génétique, Institut de Génétique Médicale d'Alsace, Hôpitaux Universitaires de Strasbourg, , 67000 Strasbourg, France. [5] Laboratoire de Génétique Médicale, Institut de Génétique Médicale d'Alsace, INSERM U1112, Fédération de Médecine Translationnelle de Strasbourg (FMTS), Université de Strasbourg, , 67000 Strasbourg, France.

References

1. Adzhubei, I., Jordan, D.M., Sunyaev, S.R.: Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet **Chapter 7**, 7–20 (2013)
2. Kumar, P., Henikoff, S., Ng, P.C.: Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc **4**(7), 1073–1081 (2009)
3. Thompson, J.D.: Statistics in Bioinformatics: Methods for Multiple Sequence Alignment. Elsevier, ??? (2017)
4. Bawono, P., Dijkstra, M., Pirovano, W., Feenstra, A., Abeln, S., Heringa, J.: Multiple Sequence Alignment. Methods Mol. Biol. **1525**, 167–189 (2017)
5. Chatzou, M., Magis, C., Chang, J.M., Kemena, C., Bussotti, G., Erb, I., Notredame, C.: Multiple sequence alignment modeling: methods and applications. Brief. Bioinformatics **17**(6), 1009–1023 (2016)
6. Zambrano-Vega, C., Nebro, A.J., Garcia-Nieto, J., Aldana-Montes, J.F.: M2Align: parallel multiple sequence alignment with a multi-objective metaheuristic. Bioinformatics **33**(19), 3011–3017 (2017)
7. List of alignments visualisation software. https://en.wikipedia.org/wiki/List_of_alignment_visualization_software. Accessed: 2017-08-14
8. Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., Barton, G.J.: Jalview Version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics **25**(9), 1189–1191 (2009)
9. Barson, G., Griffiths, E.: SeqTools: visual tools for manual analysis of sequence alignments. BMC Res Notes **9**, 39 (2016)
10. Larsson, A.: AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics **30**(22), 3276–3278 (2014)
11. Hipp, D.R.: SQLite home page. https://www.sqlite.org. Accessed: 2017-08-14
12. Katoh, K., Misawa, K., Kuma, K., Miyata, T.: MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. **30**(14), 3059–3066 (2002)
13. Thompson, S.M.: Constructing and refining multiple sequence alignments with PileUp, SeqLab, and the GCG suite. Curr Protoc Bioinformatics **Chapter 3**, 3–6 (2003)
14. Womble, D.D.: GCG: The Wisconsin Package of sequence analysis programs. Methods Mol. Biol. **132**, 3–22 (2000)

15. Thompson, J.D., Muller, A., Waterhouse, A., Procter, J., Barton, G.J., Plewniak, F., Poch, O.: MACSIMS: multiple alignment of complete sequences information management system. BMC Bioinformatics **7**, 318 (2006)
16. Lefort, V., Desper, R., Gascuel, O.: FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. Mol. Biol. Evol. **32**(10), 2798–2800 (2015)
17. Wicker, N., Perrin, G.R., Thierry, J.C., Poch, O.: Secator: a program for inferring protein subfamilies from phylogenetic trees. Mol. Biol. Evol. **18**(8), 1435–1441 (2001)
18. Wicker, N., Dembele, D., Raffelsberger, W., Poch, O.: Density of points clustering, application to transcriptomic data analysis. Nucleic Acids Res. **30**(18), 3992–4000 (2002)
19. McLachlan, G., Peel, D.: Finite Mixture Models. Wiley, ??? (2000)
20. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control **AC-19**(6), 716–723 (1974)
21. Schwarz, E.G.: Estimating the dimension of a model. Annals of Statistics **6**(2), 461–464 (1978)
22. Valdar, W.S.: Scoring residue conservation. Proteins **48**(2), 227–241 (2002)
23. Johansson, F., Toh, H.: A comparative study of conservation and variation scores. BMC Bioinformatics **11**, 388 (2010)
24. Liu, X., Li, J., Guo, W., Wang, W.: A new method for quantifying residue conservation and its applications to the protein folding nucleus. Biochem. Biophys. Res. Commun. **351**(4), 1031–1036 (2006)
25. Liu, X.S., Guo, W.L.: Robustness of the residue conservation score reflecting both frequencies and physicochemistries. Amino Acids **34**(4), 643–652 (2008)
26. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G.: The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25**(24), 4876–4882 (1997)
27. Altschul, S.F., Wootton, J.C., Zaslavsky, E., Yu, Y.K.: The construction and use of log-odds substitution scores for multiple sequence alignment. PLoS Comput. Biol. **6**(7), 1000852 (2010)
28. The RCSB Protein Data Bank. https://www.rcsb.org. Accessed: 2017-08-14
29. Pundir, S., Martin, M.J., O'Donovan, C.: UniProt Protein Knowledgebase. Methods Mol. Biol. **1558**, 41–55 (2017)
30. Pundir, S., Martin, M.J., O'Donovan, C.: UniProt Protein Knowledgebase. Methods Mol. Biol. **1558**, 41–55 (2017)
31. Tcl Developer's XChange. https://www.tcl.tk. Accessed: 2017-08-14

Figures

Figure 1 Sample figure title. A short description of the figure content should go here.

Figure 2 Sample figure title. Figure legend text.

Tables

Table 1 Sample table title. This is where the description of the table should go.

	B1	B2	B3
A1	0.1	0.2	0.3
A2	...	..	.
A3	..	.	.

Additional Files

Additional file 1 — Sample additional file title
Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title
Additional file descriptions text.