BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 15 2012, pages 2072–2073
doi:10.1093/bioinformatics/bts302

Advance Access publication May 21, 2012

SALIGN: a web server for alignment of multiple protein sequences and structures

Hannes Braberg[1,2,3], Benjamin M. Webb[2,3], Elina Tjioe[2,3], Ursula Pieper[2,3], Andrej Sali[2,3,*] and M.S. Madhusudhan[4,5,6,*]

[1]Department of Cellular and Molecular Pharmacology, [2]Department of Bioengineering and Therapeutic Sciences, [3]Department of Pharmaceutical Chemistry, and California Institute of Quantitative Biosciences (QB3), University of California, San Francisco, CA 94158, USA, [4]Bioinformatics Institute, 30 Biopolis Street, #07-01, Matrix, 138671 Singapore, [5]Department of Biological Sciences, National University of Singapore, Singapore and [6]School of Biological Sciences, Nanyang Technological University, Singapore

Associate Editor: Burkhard Rost

ABSTRACT

Summary: Accurate alignment of protein sequences and/or structures is crucial for many biological analyses, including functional annotation of proteins, classifying protein sequences into families, and comparative protein structure modeling. Described here is a web interface to SALIGN, the versatile protein multiple sequence/structure alignment module of MODELLER. The web server automatically determines the best alignment procedure based on the inputs, while allowing the user to override default parameter values. Multiple alignments are guided by a dendrogram computed from a matrix of all pairwise alignment scores. When aligning sequences to structures, SALIGN uses structural environment information to place gaps optimally. If two multiple sequence alignments of related proteins are input to the server, a profile–profile alignment is performed. All features of the server have been previously optimized for accuracy, especially in the contexts of comparative modeling and identification of interacting protein partners.

Availability: The SALIGN web server is freely accessible to the academic community at http://salilab.org/salign. SALIGN is a module of the MODELLER software, also freely available to academic users (http://salilab.org/modeller).

Contact: sali@salilab.org; madhusudhan@bii.a-star.edu.sg

Received and revised on April 2, 2012; accepted on April 27, 2012

1 INTRODUCTION

As part of our effort to better understand cellular processes, we annotate the functions of proteins, classify them into families, determine or model their 3D structures, categorize these structures, etc. To achieve these aims, robust and accurate methods for aligning protein sequences and structures with one another are essential. Alignment methods can be divided into three categories based on the information used to align proteins: sequence–sequence, sequence–structure, and structure–structure alignments. The SALIGN web server is designed to provide a user-friendly interface to producing robust and accurate solutions for the alignment problems in these different categories. Although several programs and servers (e.g. Di Tommaso *et al*., 2011; Shatsky *et al*., 2004) align protein sequences or protein structures with one another, SALIGN can align sequences, structures, or a combination of sequences and structures.

2 PROGRAM OVERVIEW

The multi-purpose alignment module of MODELLER (Sali and Blundell, 1993), SALIGN (Madhusudhan, *et al*., 2006; 2009; Marti-Renom *et al*., 2004; Pieper *et al*., 2011), can produce alignments of two or more protein sequences, based on sequence and/or structure information. Given a set of input protein sequences and/or structures, all possible pairwise alignments are calculated using dynamic programming.

SALIGN provides two methods for constructing multiple alignments from pairwise alignments, 'tree alignment' and 'progressive alignment' (Madhusudhan *et al*., 2009). The tree algorithm first creates a dendrogram of the input proteins from a matrix of all pairwise alignment scores. This guides the multiple alignment by serially aligning, in pairwise fashion, the closest linked branches to each other. The progressive alignment algorithm is computationally less expensive, where two arbitrary sequences are first aligned to one another, followed by the addition of a third; in n-1 steps, a multiple alignment of n sequences is created.

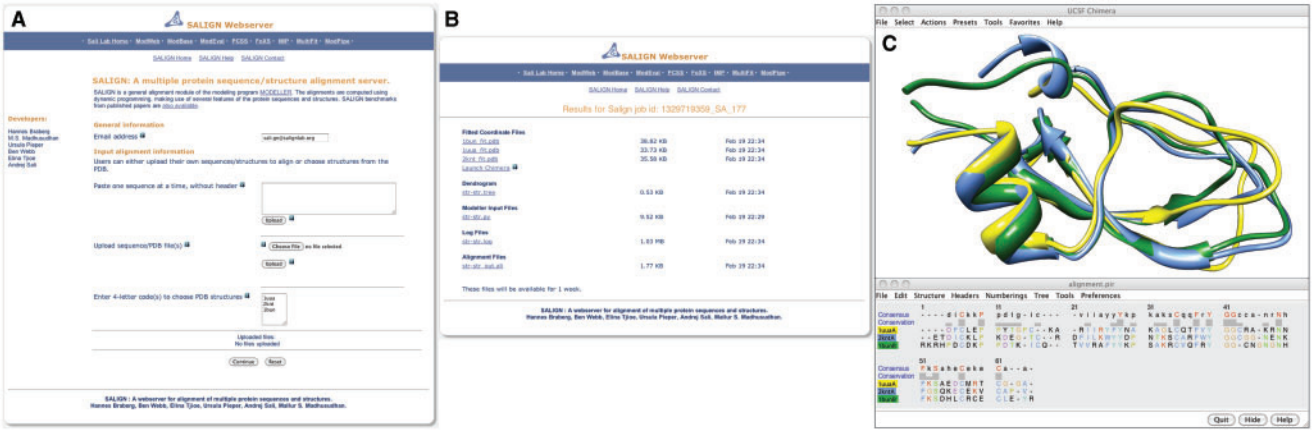
If two pre-aligned blocks of sequences are to be aligned, the profile–profile alignment method is used (Marti-Renom *et al*., 2004). To align a block of sequences to a block of structures, the ALIGN2D algorithm is used, where affine gap penalties are replaced by an environment-dependent gap penalty function (Madhusudhan *et al*., 2006).

SALIGN's default settings suffice for many applications. It has been fine tuned and extensively tested for alignment accuracy (Davis *et al*., 2006; Madhusudhan *et al*., 2006; 2009; Marti-Renom *et al*., 2004; 2007; Pieper *et al*., 2011). Nevertheless, the interface allows the user to manipulate many options if so desired.

3 SERVER DESCRIPTION

The main interface to the server allows the user to input protein structures and/or sequences. Structures can be uploaded in protein data bank (PDB) format or specified by their four-letter PDB codes. Sequences can be uploaded as files in the FASTA or MODELLER

[*]To whom correspondence should be addressed.


A
B
C
SALIGN Webserver
Sali Lab Home · ModWeb · ModBase · ModEval · PCSS · FoXS · IMP · MultiFit · ModPipe
SALIGN Home SALIGN Help SALIGN Contact
Developers:
Continue
Reset
UCSF Chimera
File Select Actions Presets Tools Favorites Help
alignment.pir
File Edit Structure Headers Numberings Tree Tools Preferences
Consensus
Conservation
1uuaA
2kntA
1bunB
Quit
Hide
Help

SALIGN: A multiple protein sequence/structure alignment server.

SALIGN is a general alignment module of the modeling program MODELLER. The alignments are computed using dynamic programming, making use of several features of the protein sequences and structures. SALIGN benchmarks from published papers are also available.

General information

Email address

Input alignment information

Users can either upload their own sequences/structures to align or choose structures from the PDB.

Paste one sequence at a time, without header

Upload sequence/PDB file(s)

Choose File no file selected

Enter 4-letter code(s) to choose PDB structures

Uploaded files:
No files uploaded

Results for Salign job id: 1329719359_SA_177

Fitted Coordinate Files		
1bun_fit.pdb	36.62 KB	Feb 19 22:34
1uua_fit.pdb	33.73 KB	Feb 19 22:34
2knt_fit.pdb	35.58 KB	Feb 19 22:34
Launch Chimera		
Dendrogram		
str-str.tree	0.53 KB	Feb 19 22:34
Modeller Input Files		
str-str.py	9.52 KB	Feb 19 22:29
Log Files		
str-str.log	1.03 MB	Feb 19 22:34
Alignment Files		
str-str_out.ali	1.77 KB	Feb 19 22:34

These files will be available for 1 week.

SALIGN : A webserver for alignment of multiple protein sequences and structures.
Hannes Braberg, Ben Webb, Elina Tjioe, Ursula Pieper, Andrej Sali, Mallur S. Madhusudhan.

Fig. 1. A snapshot of the web server input (**A**) and results page (**B**). A link from the results page for structure–structure alignments launches the Chimera software (Pettersen *et al.*, 2004) for visual display (**C**). In this illustration, the proteins identified by the PDB codes 1uuaA, 2kntA, and 1bunB are structurally aligned.

Protein Information Resource (PIR) format, or pasted into the sequence window. Alignment files in the MODELLER PIR format can also be used to specify the input structures; the server searches the PDB for the structures denoted in the alignment file. In the event no structure is found matching an alignment entry, it is treated as a sequence.

Given the input, the server decides upon the optimal alignment protocol. If structures were input, the user has the option of specifying a structure segment (e.g. a single chain or domain). Further, on the advanced view page, the user can override the default alignment protocol and parameter values. The parameters that can be tweaked depend on the selected alignment protocol. For instance, if a structure–sequence alignment is chosen, parameters related to the environment-dependent gap penalty function can be changed. An explanation for each of the tunable parameters is provided in the help pages or via links to the MODELLER manual.

After the successful completion of an alignment task, the user is e-mailed a link to an archive of output files. The archive contains the final alignment file, superimposed coordinates if structures were aligned, a dendrogram file if a tree was constructed, and the MODELLER input and log files. The input file(s) can be used with a standalone version of MODELLER (version 9 or higher). The log file contains information about the calculation, including root mean square deviation values, numbers of equivalent positions, and any warnings and errors. If structures have been aligned, the output page also includes a link that displays the aligned structures in the molecular graphics viewer Chimera (Pettersen *et al.*, 2004). This functionality provides an instant visualization of the alignment (Fig. 1).

ACKNOWLEDGEMENTS

We thank Eric Pettersen for helpful discussions.

Funding: National Institutes of Health (NIH), U.S.A [U54 GM094662 and R01 GM54762 to A.S.]

Conflict of Interest: none declared.

REFERENCES

Davis,F.P. *et al*. (2006) Protein complex compositions predicted by structural similarity, *Nucleic Acids Res*., **34**, 2943–2952.

Di Tommaso,P. *et al*. (2011) T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension, *Nucleic Acids Res*., **39**, W13–W17.

Madhusudhan,M.S. *et al*. (2006) Variable gap penalty for protein sequence-structure alignment, *Protein Eng. Des Sel.*, **19**, 129–133.

Madhusudhan,M.S. *et al*. (2009) Alignment of multiple protein structures based on sequence and structure features, *Protein Eng. Des. Sel.*, **22**, 569–574.

Marti-Renom,M.A. *et al*. (2004) Alignment of protein sequences by their profiles, *Protein Sci.*, **13**, 1071–1087.

Marti-Renom,M.A. *et al*. (2007) DBAli tools: mining the protein structure space, *Nucleic Acids Res*., **35**, W393–W397.

Pettersen,E.F. *et al*. (2004) UCSF Chimera–a visualization system for exploratory research and analysis, *J. Comput. Chem*., **25**, 1605–1612.

Pieper,U. *et al*. (2011) ModBase, a database of annotated comparative protein structure models, and associated resources, *Nucleic Acids Res*., **39**, D465–D474.

Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints, *J. Mol. Biol*., **234**, 779–815.

Shatsky,M. *et al*. (2004) A method for simultaneous alignment of multiple protein structures, *Proteins*, **56**, 143–156.