BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 7 2012, pages 1026–1027
doi:10.1093/bioinformatics/bts035

Sequence analysis
Advance Access publication February 12, 2012

SBAL: a practical tool to generate and edit structure-based amino acid sequence alignments

Conan K. Wang[1,*], Ursula Broder[1], Saroja K. Weeratunga[1], Robin B. Gasser[2], Alex Loukas[3,4] and Andreas Hofmann[1,2,4]

[1]Structural Chemistry Program, Eskitis Institute for Cell and Molecular Therapies, Griffith University, Brisbane, Qld 4111, [2]Department of Veterinary Science, The University of Melbourne, Vic 3052, [3]James Cook University, Cairns, Qld 4878 and [4]Queensland Tropical Health Alliance, Queensland, Australia

Associate Editor: Martin Bishop

ABSTRACT

Summary: Both alignment generation and visualization are important processes for producing biologically meaningful sequence alignments. Computational tools that combine reliable, automated and semi-automated approaches to produce secondary structure-based alignments with an appropriate visualization of the results are rare. We have developed SBAL, a tool to generate and edit secondary structure-based sequence alignments. It is easy to install and provides a user-friendly interface. Sequence alignments are displayed, with secondary structure assignments mapped to their corresponding regions in the sequence by using a simple colour scheme. The algorithm implemented for automated and semi-automated secondary structure-based alignment calculations shows a comparable performance to existing software.

Availability and implementation: SBAL has been implemented in Java to provide cross-platform compatibility. SBAL is freely available to academic users at http://www.structuralchemistry.org/pcsb/. Users will be asked for their name, institution and email address. A manual can also be downloaded from this site. The software, manual and test sets are also available as supplementary material.

Contact: conan.wang@griffith.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 14, 2011; revised on December 15, 2011; accepted on January 4, 2012

1 INTRODUCTION

Accurate amino acid sequence alignments are central to many bioinformatic, biochemical and structural studies. In structural biology, amino acid sequence alignments are used to not only identify conserved residues that may be functionally significant, but also regions of conserved secondary structure that may provide information on the overall fold of the molecule.

Of the alignment algorithms that are currently available, only a handful take into account secondary structure information; namely PRALINE (Simossis and Heringa, 2005), SPEM (Zhou and Zhou, 2005), PROMALS (Pei and Grishin, 2007), ISPAlign (Lu and Sze, 2008) and OPAL (Wheeler and Kececioglu, 2007). Despite using different approaches to implement secondary structure information into multiple sequence alignments, these programs have shown that significant improvements in accuracy can be achieved when secondary structure information is included.

However, a limitation of currently available programs is that they do not provide an easy means of visualizing alignments and secondary structure predictions after they have been generated. The visualization of alignments is an important part of the entire process of alignment generation, not only for quality control, but also for analysis and interpretation of the biological context of the alignments. Additionally, manual editing can frequently improve alignments because most algorithms apply heuristic approaches that may not generate an optimal solution.

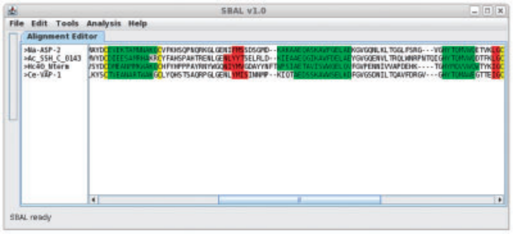
Despite the variety of alignment programs (Notredame, 2002) and viewers available (Anderson *et al.*, 2011; Goode and Rodrigo, 2007; Gouy *et al.*, 2010; Lord *et al.*, 2002; Waterhouse *et al.*, 2009), we found the process of generating secondary structure-based alignments and appropriate visualization difficult. Some tools are only available as web-based applications (Ginalski *et al.*, 2003; Pei *et al.*, 2008; Söding *et al.*, 2005) and others require complex installation and compilation. Some visualization tools are specialized for other purposes, such as for phylogenetic reconstructions. Additionally, visualization of secondary structure assignments on a multiple sequence alignment in most programs is complicated and does not allow for easy visual comparisons. We have implemented a simple-to-use and portable Java application that aids in the generation of structure-based sequence alignments by providing an automated alignment that may be edited and improved by the user through a user-friendly graphical user interface.

2 METHODS

SBAL is a Java application that builds on and extends fundamental Java classes developed within the Programme Collection for Structural Biology and Biophysical Chemistry (PCSB) (Hofmann and Wlodawer, 2002). In order to provide single sequence secondary structure prediction within SBAL, the neural network approach of PSIPRED (Bryson *et al.*, 2005) has been implemented in Java. In SBAL, secondary structure information is restricted to three elements: α-helix, β-strand and unstructured (random coil). We have also implemented a novel progressive alignment algorithm (see Supplementary Material for more details).

SBAL was tested on the BAliBASE (Thompson *et al.*, 1999) benchmark alongside ISPAlign (Lu and Sze, 2008), SPEM (Zhou and Zhou, 2005) and OPAL (Wheeler and Kececioglu, 2007). The performance of the programs was evaluated using the scoring software that accompanies the BAliBASE benchmark. The results are provided in the Supplementary Material.

[*]To whom correspondence should be addressed.



SBAL ready

Fig. 1. Main window display of SBAL. The layout design is simple to allow users to focus on the important features of the alignment. Regions of secondary structure are coloured (α-helical positions are coloured in green and β-strand positions are coloured in red). A variety of file, editing, alignment and analysis options can be accessed from the menu bar, and several shortcut keys have also been programmed. A dynamically updated status (bottom of the window) and progress bar (left of window) are provided for interactivity.

3 RESULTS

Currently, there is a need for a computational tool that facilitates the seamless transition from secondary structure-based alignments to the visualization of the results. In this context, we have developed SBAL to generate and edit secondary structure-based sequence alignments. SBAL has been implemented in Java to provide cross-platform compatibility. It is easy to install and provides an interactive user-friendly interface (Fig. 1). Users can choose to either generate secondary structure-based alignments using SBAL or import alignment results from other programs for visualization and editing, thus making SBAL a very versatile and practical tool for a wide range of scenarios.

A variety of established input formats (FASTA, MSF and ClustalW) are supported to ensure compatibility with commonly used alignment programs. Secondary structure information can be provided by different means: predicted secondary structure by PSIPRED (Bryson *et al.*, 2005) can be processed from the PSIPRED vertical (ss2-files) or horizontal (horiz-files) format, extracted from experimental 3D structures by DSSP (Kabsch and Sander, 1983), or directly from the header of Protein Data Bank files (HELIX and SHEET records). Secondary structure information is instantly mapped to the corresponding regions of the sequences by using a red–green colour scheme to facilitate visual comparisons.

We have implemented a progressive multiple alignment algorithm that aims to maximize secondary structure and amino acid matches using position-specific information. There are three fundamental steps: (i) an initial guide tree construction; followed by (ii) a dynamic programming-based alignment generation; and (iii) a final 'polishing step' that realigns the sequences. SBAL produces alignments of a comparable quality as those achieved using other structure-based alignment programs (see Supplementary Material for a more thorough analysis). Although some features that have been proposed to further improve accuracy have not been implemented in SBAL and are destined for inclusion in the near future.

Multiple amino acid sequence alignments can be modified by the user through insertion and deletion of gaps as well as horizontal movement of sequences with respect to the alignment. N- or C-terminal extensions of sequences that should not be included in the alignment can be truncated within SBAL. These functions can be executed for an individual sequence or a user-selected group of sequences. Scores representing the level of sequence and secondary structure conservation of the alignment are displayed in the status bar. The vertical order of sequences within the alignment can also be changed.

Alignments are saved in three file formats with the same root name: SBAL, HTML and FASTA. Reloading the SBAL file allows the user to continue to work on the alignment. The HTML output is considered the best format for dissemination, since it can be loaded into other editors or Word processing programs without losing the secondary structure mapping. The FASTA output file is provided for interfacing with other bioinformatic software.

Funding: Australian Research Council (to R.G., A.H. and A.L., in part, LP100100092); National Health and Medical Research Council (NHMRC) (to A.H., in part, APP1002898). C.K.W. is an NHMRC Early Career Research Fellow (536578), and A.L. is an NHMRC Senior Research Fellow.

Conflict of interest: none declared.

REFERENCES

Anderson,C.L. *et al*. (2011) SuiteMSA: visual tools for multiple sequence alignment comparison and molecular sequence simulation. *BMC Bioinformatics*, **12**, 184.

Bryson,K. *et al*. (2005) Protein structure prediction servers at University College London. *Nucleic Acids Res.*, **33**, W36–W38.

Ginalski,K. *et al*. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*, **22**, 1015–1018.

Goode,M.G. and Rodrigo,A.G. (2007) SQUINT: a multiple alignment program and editor. *Bioinformatics*, **23**, 1553–1555.

Gouy,M., *et al*. (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.*, **27**, 221–224.

Hofmann,A. and Wlodawer,A. (2002) PCSB - a program collection for structural biology and biophysical chemistry. *Bioinformatics*, **18**, 209–210.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Lord,P.W. *et al*. (2002) CINEMA-MX: a modular multiple alignment editor. *Bioinformatics*, **18**, 1402–1403.

Lu,Y. and Sze,S. (2008) Multiple sequence alignment based on profile alignment of intermediate sequences. *J. Comput. Biol.*, **15**, 767–777.

Notredame,C. (2002) Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics*, **3**, 131–144.

Pei,J. and Grishin,N.V. (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics*, **23**, 802–808.

Pei,J. *et al*. (2008) PROMALS3D: a tool for multiple sequence and structure alignment. *Nucleic Acids Res.*, **36**, 2295–2300.

Simossis,V.A. and Heringa,J. (2005) PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Res.*, **33**, W289–W294.

Söding,J. *et al*. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.*, **33**, W244–W248.

Thompson,J.D. *et al*. (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.*, **27**, 2682–2690.

Waterhouse,A.M. *et al*. (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.

Wheeler,T.J. and Kececioglu,J.D. (2007) Multiple alignment by aligning alignments. *Bioinformatics*, **23**, i559–i568.

Zhou,H. and Zhou,Y. (2005) SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures. *Bioinformatics*, **21**, 3615–3621.