Bioinformatics, 33(24), 2017, 3991–3992
doi: 10.1093/bioinformatics/btx521
Advance Access Publication Date: 16 August 2017
Application Note

OXFORD

Sequence analysis

ALIGNSEC: viewing protein secondary structure predictions within large multiple sequence alignments

Gilbert Deléage

Lyon University—CNRS, LBTI-UMR5305, Institute of Biology and Chemistry of Proteins, 7 passage du Vercors, 69367 Lyon cedex 07, France

Associate Editor: John Hancock

Received on June 12, 2017; revised on August 9, 2017; editorial decision on August 10, 2017; accepted on August 14, 2017

Abstract

Motivation: ALIGNSEC is a module within ANTHEPROT designed for the interactive display, edition and printing of large-scale multiple alignments integrating secondary structure predictions.
Availability and implementation: The ALIGNSEC module is part of the ANTHEPROT package (http://antheprot-pbil.ibcp.fr) which can be used freely for academic users. It is running on Windows Operating systems. For commercial use, please contact the author.
Contact: gilbert.deleage@ibcp.fr

Increasingly, biologists need computer tools to view their data. In the context of genomics, this is particularly the case for multiple alignments, and although many tools have been developed such as Jalview (Clamp *et al.*, 2004), Seaview (Galtier *et al.*, 1996), BioEdit, none offers the possibility of viewing multiple secondary structures predictions directly within the alignment with the gaps (insertions/deletions) created at the multiple sequence alignment step. In this context, I developed ALIGNSEC (a module integrated in ANTHEPROT) to respond to this lack. ALIGNSEC is interfaced with two powerful programs for multiple alignments of protein sequences such as Clustal Omega (Sievers *et al.*, 2011) and Muscle (Edgar, 2004).

ALIGNSEC allows interactive manipulation of large alignments (comprising, for example, 2000 sequences of 500 amino acids) of proteins. From the interface it is possible to eliminate sequences from the alignment and to restart the calculation on the remaining sequences. This allows the user to retain only the relevant sequences for further analysis. The alignment is editable by adding or removing gaps in one or more sequences. To improve readability, ALIGNSEC has two display modes: a colored box mode and an inverted video mode. In addition, the size of the font is adjustable (from 8 to 20) as well as the color mode according to (i) the identity level, (ii) the groups of amino acids established by Clustal Omega or (iii) the choice of the user. A mouse click onto the alignment returns the position in the sequence, the number and the name of the amino acid (see the small white box in the alignment part of the Fig. 1). To analyze the alignment, in addition to the standard display mode of the residues, it is possible to display only the conserved residues or the non-conserved residues. From ALIGNSEC, it is possible to select the sequences in order to display the UniprotKB entry (by right click of the mouse). The multiple selection also makes it possible to calculate the percentages of identity between the selected sequences and to display the corresponding matrix. A selected sequence (whose name is on a red background) can be analyzed by conventional tools such as BLAST or FASTA (against Uniprot or PDB), PROSITE signatures search, search for the most antigenic regions via the calculation of the hydrophobicity and antigenicity profiles or the prediction of the disordered regions thanks to IUPRED. Moreover, ALIGNSEC makes it possible to construct the repertoire of the amino acids used at each position in the alignment as well as the curve of entropy i.e. Shannon information content (Schneider and Stephens, 1990). Note that it is also possible to save in a text file the amino acid count (or frequency) of the whole alignment (or a part of it, which corresponds to the displayed window). Another functionality of ALIGNSEC is the display of a graphic 'logo' which is dynamically calculated according to the modifications performed in the alignment (editing, coloring). The movements in the alignment (in both directions) can be carried out in a conventional manner with scroll bars but also by means of the mouse wheel. A search function is very useful for locating specific fragments of sequence in the alignment. The strength and novelty of ALIGNSEC are the inclusion of predicted secondary structures directly into the alignment (see Fig. 1) using several methods (of a total of 8) including PHD (Rost *et al.*,

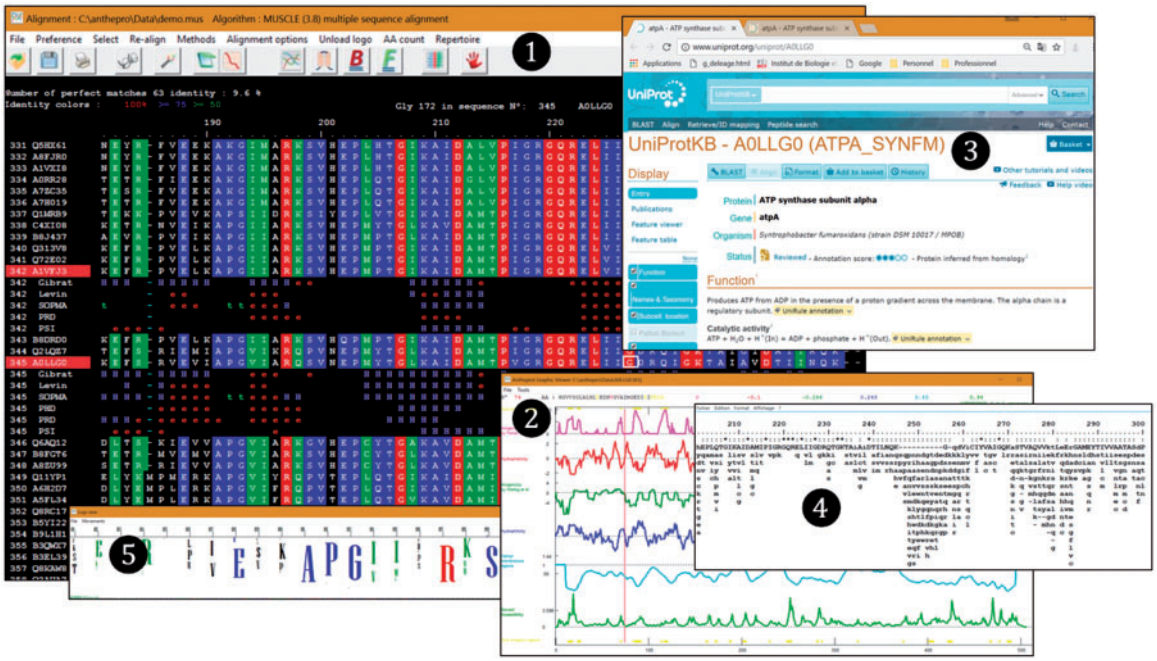

Fig. 1. General view of ALIGNSEC including predictions of secondary structures by GORII (Gibrat *et al.*, 1987), Levin (Levin *et al.*, 1986), SOPMA (Geourjon and Deléage 1995), Predator (Frishman and Argos, 1996) and PSIPRED (Jones, 1999) for A1VFJ3 and A0LLG0 (1). For A0LLG0 protein, the prediction obtained by PHD method (Rost *et al.*, 1994) is also shown. The physico–chemical profiles (2), the Uniprot entry of A0LLG0 accession number (3) are shown as satellite views. The repertoire text file (4) is also given with the 'sequence logo' graphic plot (5)

1994), PSIpred (Jones, 1999) and SOPMA (Geourjon and Deléage, 1995) thanks to the client/server capacity of ANTHEPROT (Deléage *et al.*, 2001). The agreement between the predictive methods is calculated using the SOV parameter (Zemla *et al.* 1999). In order to examine the agreement between the methods, it is possible to mask the sequences to focus on the secondary structures (it is also possible to mask all or part of the structure predictions). After user modifications, the optimized alignment can be saved in Clustal format (without secondary structures), in RTF or in ALIGNSEC format (with secondary structures). This format can be loaded again by ALIGNSEC. Note that the ALIGNSEC module of ANTHEPROT can be launched directly from a local file in CLUSTAL or ALIGNSEC format or from the NPS@ (Combet *et al.* 2000) web server for protein sequence analysis. At the printing level, a bitmap copy mode of the window is available even if the best results are obtained with the vector mode. In this latter case, full alignment is sent on the printer which can be a pdf virtual printer, which allows to manage formats from A4 to A0 allowing to draw very large posters containing quite large alignments.

ALIGNSEC is a powerful and versatile helper tool to correlate sequence structure with evolution function and can be considered as a computer tool to assist a wet laboratory in protein investigation. In the future, the addition of external programs to ALIGNSEC for automatic optimization of multiple alignments based on predicted secondary structure is feasible if the executable and the documentation files for such methods are available.

Funding

This work has been supported by the CNRS and University funds.

Conflict of interest: none declared.

References

Clamp,M. *et al.* (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.

Combet,C. *et al.* (2000) NPS@: network protein sequence analysis. *Trends Biochem. Sci.*, **25**, 147–150.

Deléage,G. *et al.* (2001) Antheprot: an integrated protein sequence snalysis software with client/server capabilities. *Comput. Biol. Med.*, **31**, 259–267.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1179.

Frishman,D., and Argos,P. (1996) Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Eng.*, **9**, 133–142.

Galtier,N. *et al.* (1996) SeaView and Phylo_win, two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.*, **12**, 543–548.

Geourjon,C., and Deléage,G. (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput. Appl. Biosci*, **11**, 681–684.

Gibrat,J.F. *et al.* (1987) Further developments of protein secondary structure prediction using information-theory-new-parameters and consideration of residue pairs. *J. Mol. Biol*, **198**, 424–443.

Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Levin,J.M. *et al.* (1986) An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Lett.*, **205**, 303–308.

Rost,B. *et al.* (1994) PHD: an automatic mail server for protein secondary structure prediction. *Proteins*, **10**, 53–60.

Schneider,T.D., and Stephens, R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Sievers, F. *et al.* (2011) DGFast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 359.

ZemlaA. *et al.* (1999) A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins-Structure Function and Bioinformatics*, **34**, 220–223.