BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 3 2005, pages 393–395
doi:10.1093/bioinformatics/bti013

The SSEA server for protein secondary structure alignment

Paolo Fontana[1], Eckart Bindewald[2,†], Stefano Toppo[3], Riccardo Velasco[1], Giorgio Valle[4] and Silvio C. E. Tosatto[4,*]

[1]Istituto Agrario di San Michele all'Adige, via E. Mach 1, 38010 S.Michele all'Adige (TN), Italy, [2]Center of Excellence in Bioinformatics, State University of New York at Buffalo, USA, [3]Department of Biological Chemistry and [4]Department of Biology and CRIBI Biotechnology Centre, University of Padova, V.le G. Colombo 3, 35121 Padova, Italy

Received on March 24, 2004; revised on August 11, 2004; accepted on August 29, 2004
Advance Access publication September 3, 2004

ABSTRACT

Summary: We present a web server that computes alignments of protein secondary structures. The server supports both performing pairwise alignments and searching a secondary structure against a library of domain folds. It can calculate global and local secondary structure element alignments. A combination of local and global alignment steps can be used to search for domains inside the query sequence or help in the discrimination of novel folds. Both the SCOP and PDB fold libraries, clustered at 95 and 40% sequence identity, are available for alignment.

Availability: The web server interface is freely accessible to academic users at http://protein.cribi.unipd.it/ssea/. The executable version and benchmarking data are available from the same web page.

Contact: silvio@cribi.unipd.it

INTRODUCTION

The secondary structure of proteins is a useful feature for both three-dimensional (3D) structure prediction and classification. Russell *et al.* (1996) were among the first to demonstrate how a simple mapping of secondary structures can be used to predict the fold of a protein. All major protein 3D structure classification schemes distinguish primarily between secondary structure classes on the first level, e.g. Structural Classification of Proteins (SCOP) (Andreeva *et al.*, 2004). Common protein folds were shown to be distinguishable depending on the arrangement of secondary structure elements using an automated taxonomy (Przytycka *et al.*, 1999). This concept was further developed into a secondary structure element alignment (SSEA) method (McGuffin *et al.*, 2001; Bindewald *et al.*, 2003).

The structural genomics initiatives aim at the experimental solution of novel protein folds. However, after experimental elucidation many proteins turn out to share known folds. The reliable identification of novel folds is therefore becoming increasingly relevant. As secondary structure prediction is quite reliable (Albrecht *et al.*, 2003), a secondary structure similarity search was proposed recently as a possible way to help discriminate novel folds in structural genomics (McGuffin and Jones, 2002). We have implemented the SSEA server for SSEA as a way to help in this process. It captures the intuitive concept of aligning entire elements of secondary structure, which was shown to perform significantly better than residue-based secondary structure alignments (McGuffin *et al.*, 2001). On a difficult test set of 248 distantly related protein folds with little sequence similarity, it yielded 30% correct first hits compared to sequence-based methods such as PSI-BLAST and GenTHREADER with 13 and 14%, respectively (Bindewald *et al.*, 2003).

SSEA allows the user to search with a secondary structure against a library of protein folds and establish the most similar ones. Low alignment scores imply a low similarity between two secondary structures, and a higher probability of a novel fold.

PROGRAM OVERVIEW

The SSEA scheme was implemented in the fold-recognition program MANIFOLD (Bindewald *et al.*, 2003), which uses a slightly simplified version of the previously described algorithm (Przytycka *et al.*, 1999). Each secondary structure element is represented by a letter for the state (H, E, C), and a number corresponding to the length of the region. Matches (H → H, E → E, C → C) are scored by the length of the shorter fragment. Mismatches (H → E, E → H) do not contribute to the score. Structure to coil matches ({H, E} → C)

*To whom correspondence should be addressed.

†Present address: Basic Research Program, SAIC-Frederick, Inc., Laboratory of Experimental and Computational Biology, National Cancer Institute, Building 469, Frederick, MD 21702, USA.

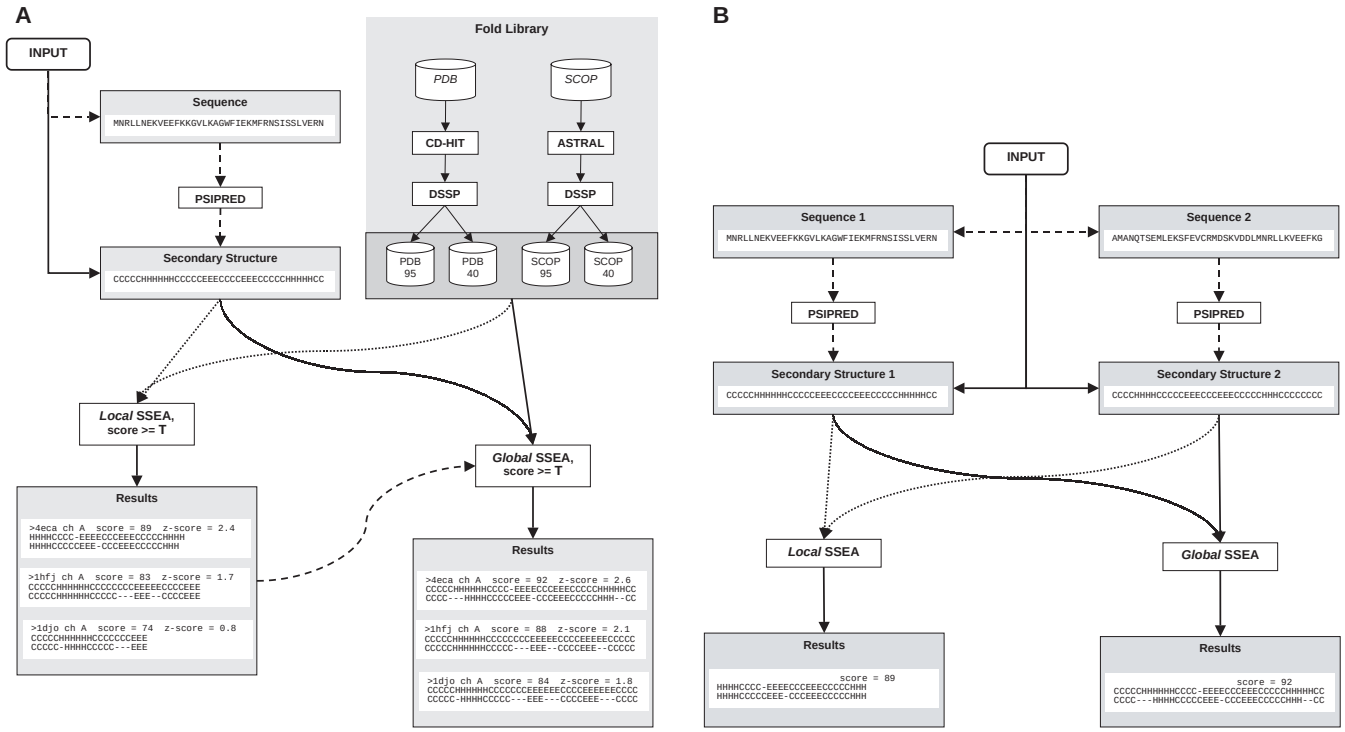

Fig. 1. Flow chart of the SSEA server. (**A**) Database alignment mode and (**B**) one versus one alignment mode.

are weighted half the length of the shorter segment. Unlike the original method (Przytycka *et al*., 1999), we do not split secondary structure elements to obtain better matches. Both global and local alignments are supported. The similarity score is normalized in the range 0–100. A *Z*-score is also calculated over all predictions in the database alignment mode to estimate the statistical significance.

The web server computes pairwise alignments of secondary structures as shown in Figure 1. It can either scan a single secondary structure against a representative fold library (Figure 1A) or two secondary structures against each other (Figure 1B). In both cases, the user can choose between submitting the secondary structure directly or having the server predict it from the amino acid sequence with PSIPRED (Jones, 1999).

Four different fold libraries, based on the Protein Data Bank (PDB) (Bourne *et al*., 2004) and SCOP (Andreeva *et al*., 2004) clustered at different similarity thresholds are available. The PDB fold libraries were created by clustering the February 2004 PDB release at 95 and 40% sequence identity with CD-HIT (Li *et al*., 2002), and running DSSP (Kabsch and Sander, 1983) to extract the secondary structure signatures. These fold libraries represent entire protein chains. A different representation is given by the SCOP fold library. The ASTRAL (Chandonia *et al*., 2004) sequences for SCOP 1.65 at 95 and 40% sequence identity were selected and the corresponding secondary structure extracted as before. The SCOP fold libraries are especially useful for determining the similarity to known protein domains, although not every PDB structure is classified in SCOP (e.g. recently released structures).

The server can compute either local or global alignments, with the latter usually aligning longer stretches at the cost of lower overall similarity. Since the local alignment may produce more reliable estimates of segments to be aligned (e.g. similar domains), but poor scoring, it is possible to follow a two-step protocol. Once the local alignments are computed, the user may select a subset of proteins to re-align using the global algorithm.

In order to assess the reliability of the results with respect to the returned *Z*-score, we used a test set consisting of 98 proteins searched against the SCOP-95 fold library. Using global alignment, the server produced no false positives at a *Z*-score threshold of 4.6. The full benchmarking data are available on the website.

The results of the alignment procedure are presented to the user as a dynamic HTML page. It contains either the alignment (for pairwise alignments) or a list of aligned protein structures above a user-supplied similarity score threshold (fold library search). The actual alignment is shown together with the normalized similarity score. In case of fold library searches, the output contains a brief summary at the top of the page, including the description and SCOP code of hits (where available), followed by each alignment in decreasing order of similarity. Links to the relevant

databases (PDB and SCOP) are provided for matching proteins.

ACKNOWLEDGEMENTS

The authors are grateful to Mario Albrecht for insightful discussions. S.T. is funded by a 'Rientro dei cervelli' grant from the Italian Ministry for Education, University and Research (MIUR). This work was supported by the project 'Advanced Biology' funded by the Fondazione delle Casse di Risparmio di Trento e Rovereto. Part of the research was also funded by Telethon (Italy), grant number B057-I.

REFERENCES

Albrecht,M., Tosatto,S.C.E., Lengauer,T. and Valle,G. (2003) Simple consensus procedures are effective and sufficient in secondary structure prediction. *Protein Eng.*, **16**, 459–462.

Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32** (Database issue), D226–D229.

Bindewald,E., Cestaro,A., Hesser,J., Heiler,M. and Tosatto,S.C.E. (2003) MANIFOLD: protein fold recognition based on secondary structure, sequence similarity and enzyme classification. *Protein Eng.*, **16**, 785–789.

Bourne,P.E., Addess,K.J., Bluhm,W.F., Chen,L., Deshpande,N., Feng,Z., Fleri,W., Green,R., Merino-Ott,J.C., Townsend-Merino,W. *et al*. (2004) The distribution and query systems of the RCSB Protein Data Bank. *Nucleic Acids Res.*, **32** (Database issue), D223–D225.

Chandonia,J.M., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, **32** (Database issue), D189–D192.

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

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

Li,W., Jaroszewski,L. and Godzik,A. (2002) Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*, **18**, 77–82.

McGuffin,L.J. and Jones,D.T. (2002) Targeting novel folds for structural genomics. *Proteins*, **48**, 44–52.

McGuffin,L.J., Bryson,K. and Jones,D.T. (2001) What are the baselines for protein fold recognition? *Bioinformatics*, **17**, 63–72.

Przytycka,T., Aurora,R. and Rose,G.D. (1999) A protein taxonomy based on secondary structure. *Nat. Struct. Biol.*, **6**, 672–682.

Russell,R.B., Copley,R.R. and Barton,G.J. (1996) Protein fold recognition by mapping predicted secondary structures. *J. Mol. Biol.*, **259**, 349–365.