FOX (FOld eXtractor): A novel protein fold recognition method using iterative PSI-BLAST searches and structural alignments

Stefano Toppo[(1)], Paolo Fontana[(2)], Riccardo Velasco[(2)], Giorgio Valle[(3)] and Silvio C.E. Tosatto[(3)]

(1) Dip. di Chimica Biologica, Universita' di Padova
(2) Istituto Agrario di San Michele all'Adige
(3) Dip. di Biologia & CRIBI Biotech Centre, Universita' di Padova

Keywords. Structural Genomics, Protein Fold Recognition, Secondary Structure, Sequence Alignment

We present a novel fold recognition method based on the combination of detailed sequence searches and structural information. Presently the protocol implements two different approaches to assign the correct fold to the target protein sequence: the first is based on database secondary structure search and the second is based on iterative database sequence search.
In the first phase a secondary structure prediction of the target is performed and based on the ConSSPred [1] protocol. This prediction is used to search for hits against a database of known secondary structures extracted from PDB (using DSSP). The search is based on a two-step strategy: the first step is based on a Smith-Waterman local secondary structure similarity search with a specific substitution matrix optimized for secondary structure alignment [2]. The second is based on a global alignment based on SSEA [3] (Secondary Structure Element Alignment), as implemented in our program MANIFOLD [4], to refine the score and the alignment itself in the region extracted from the first step. At the end of the first phase a list of hits that share a similar secondary structure topology with the target sequence is extracted.
The second phase is based on a modified protocol for scanning the sequence database called SENSER [5]. In the beginning of the second phase, BLASTP [6] is used to scan the target sequence against the NR database. These initial hits are clustered to reduce sequence bias and a seed alignment with 20 or fewer sequences generated. This step ensures that PSI-BLAST [7] can be jump-started with a more sensitive initial profile, increasing its sequence diversity. PSI-BLAST is run for four iterations (e-value inclusion threshold 10e-3) on the NR60 database of known sequences. NR60 is produced by applying the CD-HIT [8] algorithm to cluster the NR database at 60% sequence identity. Sequences producing NR60 hits with the query are assigned either to the significant sequence space (e-value <= 10e-3) or the trailing end (e-value <= 10) for further use. The profile is used to search the PDBAA database of sequences with known structure. If a significant PDBAA hit (e-value <= 10) is found, the protocol proceeds to the back-validation step (see below). If no significant hit is found, or the hit does not back-validate, a new PSI-BLAST search, using the above "4+1" protocol on NR and PDBAA, is started for the highest ranking sequence (i.e. lowest e-value) in the significant sequence space. Sequences from NR60 matching the query are also assigned to either the significant sequence space or the trailing end. Significant PDBAA hits are again submitted to back-validation. If no significant PDBAA hit is recorded and the significant sequence space has been exhausted, then the protocol uses the trailing end sequences as additional starting points for PSI-BLAST searches. In contrast to previous sequences, which were assumed to be similar enough to the target to imply homology, these sequences are submitted to back-validation before proceeding to the "4+1" PSI-BLAST protocol. The back-validation step consists in using PSI-BLAST to find the target starting from a different query sequence, found as described above. I.e. due to the asymmetric nature of PSI-BLAST, if sequence A finds sequence B it is not always the case that B also finds A. Sequences that back-validate are more likely to be correct hits. Once a sequence from PDBAA back-validates and its secondary structures is compatible with the one of the target sequence as found in the first phase, the protocol builds a target to template alignment and stops. The procedure described so far serves to identify a template structure for the target sequence. In order to produce an accurate alignment, HMMER [9] is used to build a hidden Markov model (HMM) based on the HOMSTRAD [10] sequence alignment. The target is then aligned to the template using this HMM. Preliminary results for the method indicate a clear increase in both detection rate and alignment accuracy for distantly homologous sequences. Presently FOX has been tested on Fischer-68 [11] test set to compare its performance with standard PSI-BLAST searches, GenTHREADER [12] and the original SENSER protocol. As expected the introduction of the secondary structure prediction of the protein target and the database secondary structure searches in the first phase have increased detection sensitivity and sensibility of the method compared to profile based searches as PSI-BLAST and SENSER protocol (Fig. 1). The performance is comparable to GenTHREADER showing that right template structure is always found in the top 50 hits as shown in Fig. 1. Further score optimization and development are required to definitely test the entire protocol and make the program available as a web-based server from our group's web site (http://protein.cribi.unipd.it/).

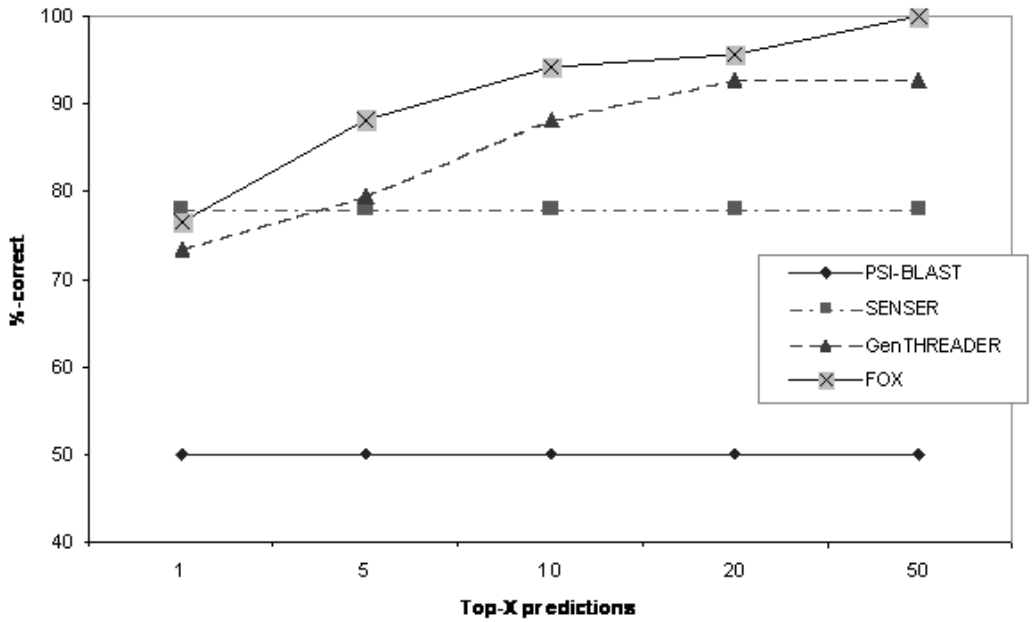

Fig. 1 percentage of template hits (y-axis) relative to the Top-x predictions (x-axis) detected by PSI-BLAST, SENSER, GenTHREADER and FOX. The results are reported for Fischer-68 test set.

References:

[1] M. Albrecht, S.C.E. Tosatto, T. Lengauer and G. Valle, Simple consensus procedures are effective and sufficient in secondary structure prediction. *Protein Engineering*, 16: 459-462, 2003.

[2] A. Wallqvist, Y. Fukunishi, L.R. Murphy, A. Fadel, R.M. Levy. Iterative sequence/secondary structure search for protein homologs: comparison with amino acid sequence alignments and application to fold recognition in genome databases. *Bioinformatics* 16(11):988-1002, 2000.

[3] T. Przytycka, R. Aurora and G.D. Rose. A protein taxonomy based on secondary structure. *Nature Struct. Biol* 6(7):672-682, 1999.

[4] E. Bindewald, A. Cestaro, J. Hesser, M. Heiler, S.C.E. Tosatto. MANIFOLD: Protein fold recognition based on secondary structure, sequence similarity and enzyme classification *Protein Engineering* 16(11): 785-789, 2003.

[5] K.K. Koretke, R.B. Russell and A.N. Lupas, Fold recognition without folds. *Protein Science*, 11:1575-1579, 2002

[6] S.F. Altschul, W. Gish, W. Miller, E.W. Myers and D.J. Lipman. Basic Local Alignment Search Tool. *J. Mol. Biol.*, 215:403-410, 1990.

[7] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acid. Res.* 25:3389-3402, 1997.

[8] W. Li, L. Jaroszewski and A. Godzik. Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics* 18:77-82, 2002.

[9] E.R. Eddy. Profile Hidden Markov Models. *Bioinformatics* 14:755-763, 1998.

[10] P.I.W. de Bakker, A. Bateman, D.F. Burke, R.N. Miguel, K. Mizuguchi, J. Shi, H. Shirai and T.L. Blundell. HOMSTRAD: adding sequence information to structure-based alignments of homologous protein families. *Bioinformatics*, 17:748-749, 2001.

[11] D. Fischer, A. Elofsson, D. Rice and D. Eisenberg. Assessing the performance of fold recognition methods by means of a comprehensive benchmark *Pac. Symp. Biocomput*. 300-318, 1996.

[12] D.T. Jones. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol Biol.* 287(4):797-815, 1999.