Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W81–W84***
doi:10.1093/nar/gki389

ProTarget: automatic prediction of protein structure novelty

Ori Sasson and Michal Linial[1,*]

School of Computer Science and Engineering and [1]Department of Biological Chemistry, Institute of Life Sciences, Hebrew University of Jerusalem, 91904 Israel

Received January 28, 2005; Revised and Accepted March 10, 2005

ABSTRACT

ProTarget is a Web-based tool for the automatic prediction of fold novelty. It offers the structural genomics community a method for target selection by providing an online analysis of any new or pre-existing sequence for its relationship to any previously solved three-dimensional structure. ProTarget takes as input an amino acid sequence. Regions of this sequence that exhibit high similarity to an existing PDB (Protein Data Bank) sequence are removed, leaving one or more subsequences. Each of these subsequences is then analyzed against a clustering of the protein space to determine the likelihood of its representing a new structural superfamily. This likelihood is derived from the distance in the clustering between the (sub)sequence and sequences that have known structures. The output of ProTarget is a graphical visualization of the protein of interest together with the likelihood that a protein sequence represents a novel structural superfamily. ProTarget is updated regularly and currently covers over 160 000 protein sequences from the SwissProt and PDB databases. ProTarget is available at http://www.protarget.cs.huji.ac.il.

INTRODUCTION

More than a million protein sequences are publicly available, yet the number of proteins for which the three-dimensional (3D) structure has been determined by X-ray and NMR technologies is significantly smaller. The goal of structural genomics (SG) is coverage of the protein fold space, and in particular completion of the structural representations of all proteins in selected model organisms (1). The number of newly discovered structures archived in the Protein Data Bank (PDB) (2) that originate from SG projects is constantly increasing. One of the most critical tasks for SG is target selection (3), the process of choosing a relatively small set of protein sequences such that, once their structures are solved, the impact in terms of covering the fold space is maximized (4). However, the actual number of proteins required to achieve the goal of full coverage of the protein structural space remains unknown (5).

Sequence-based protein classification offers a way to reduce redundancy in terms of structural representatives. Namely, all members in a family are assumed to share a structure similar to the family's representative. ProtoNet, an agglomerative hierarchal clustering of all proteins (6), provides a scaffold containing protein sequences organized as a condensed family tree (7).

Herein, we present a Web-based tool for selecting target proteins and provide a sorted list to be used in SG projects. ProTarget makes use of the tree of all sequence-based clusters presented by ProtoNet (7). It is based on a global statistical–computational learning procedure, as presented in (8). The prediction is that clusters that are distal from other solved structures within the graphical protein tree will have the highest probability of representing a new structural superfamily. ProTarget provides a significant measurement for such prediction and indicates those domains within the protein sequence that are already solved. ProTarget predictions can be used to choose protein targets with higher chances of representing new superfamilies among all proteins with unsolved structures.

PROTARGET ALGORITHM—AN OUTLINE

ProTarget is based on applying the ProtoNet clustering algorithm (6) to create Proto3D, which is a ProtoNet-like tree in which all sequences of domains from the PDB are included. Proto3D includes all 114 033 SwissProt sequences (release 40.28) and all domains from PDB (as of September 2002). A total of ~36 700 sequences of structural domains from the PDB were retained after excluding those that are based on theoretical modeling. The hierarchical clustering is used to

*To whom correspondence should be addressed. Tel: +972 2 6585425; Fax: +972 2 6586448; Email: michall@cc.huji.ac.il
Present address:
Michal Linial, School of Computer Science and Engineering, University of Washington, Seattle, WA 98195, USA

predict whether a given sequence is associated with a novel structure. The hierarchy of the ProtoNet clustering can be accessed at http://www.protonet.cs.huji.ac.il.

Technically, this is achieved using the notion of the lowest solved ancestor (LSA), that is, the lowest ancestor of a sequence that is also an ancestor of a solved sequence (i.e. a sequence that is in the PDB archive). In other words, the LSA is the lowest common ancestor for any sequence in the Proto3D or any query sequence. The notion of LSA provides a powerful method for predicting structural novelty (8). The predicted likelihood of a sequence representing a novel family is determined by comparing the actual level of the LSA in the clustering hierarchy with a predetermined unified threshold. The threshold was determined as the best separator between sequences with an already known superfamily and those that represent a new, presently unknown superfamily. The choice of this threshold is determined by a statistical learning technique described by Kifer and colleagues (8).

The ProTarget algorithm employs a precalculation step to filter structurally solved sequences. It performs a BLAST search (9) of the query protein against the database of all solved structures (i.e. the PDB). If the query protein is similar enough (BLAST *E*-score $\leqslant 1 \times 10^{-5}$) to one or more solved domains, it is broken down into several fragments (a procedure referred to as 'cropping'). All fragments that overlap with a known domain are filtered out, and each of the remaining fragments that is long enough (>30 amino acids) is subjected to the following steps.

We insert a new protein or a partial sequence into the ProtoNet clustering by associating it with the most suitable ProtoNet cluster based on its BLAST similarity to other sequences. In some cases, there is no apparent similarity to any protein in ProtoNet, and in such cases the protein is marked as 'isolated' and is not treated any further ('isolated' segments are associated with an undefined confidence score). The next step involves calculating the level of the LSA of the

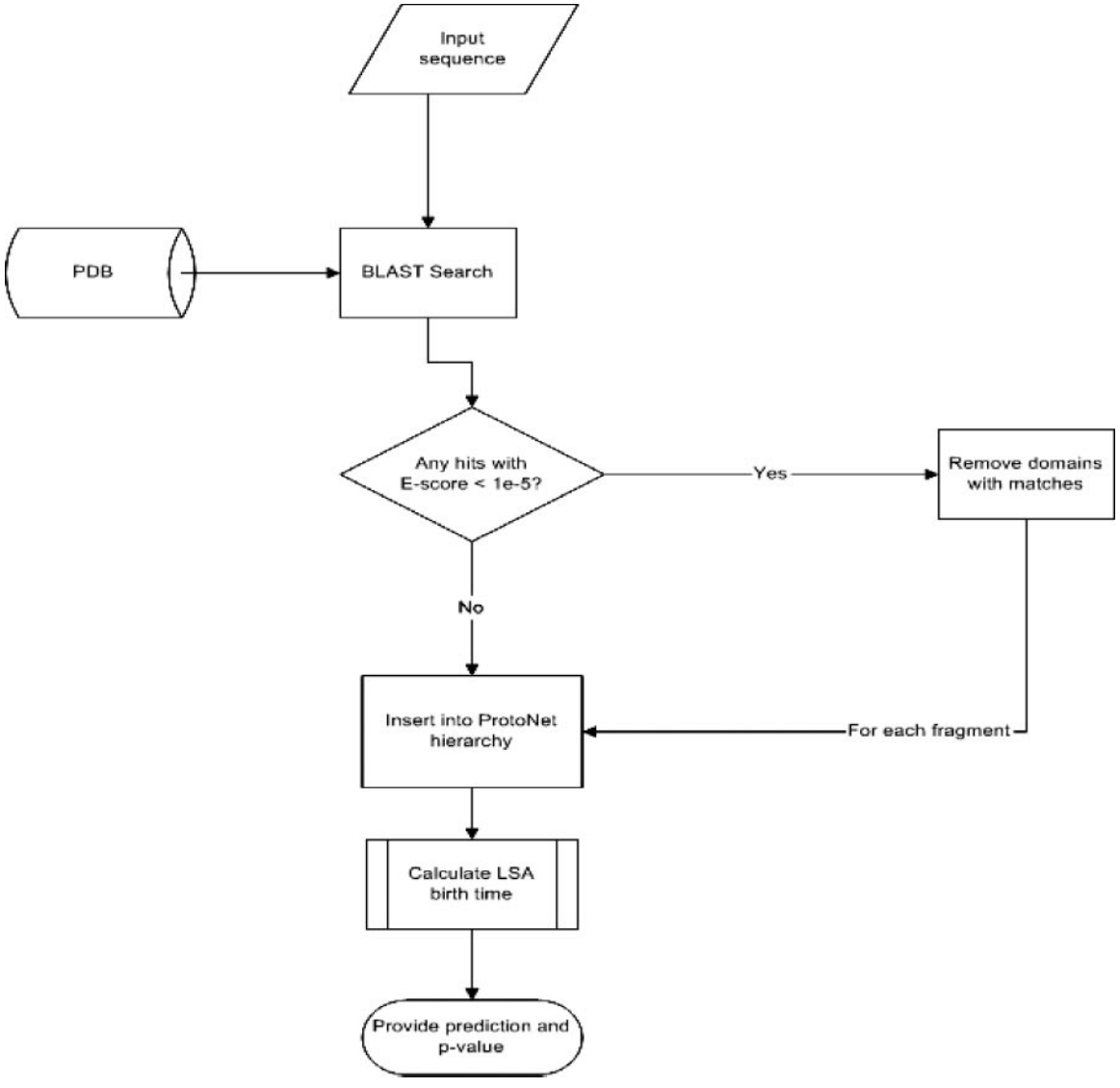

Figure 1. A schematic illustration of the ProTarget algorithm. Sequences are filtered using a BLAST similarity search to remove subsequences similar to those in the PDB, and inserted into the ProtoNet clustering hierarchy, where a prediction is provided. For details see (8).

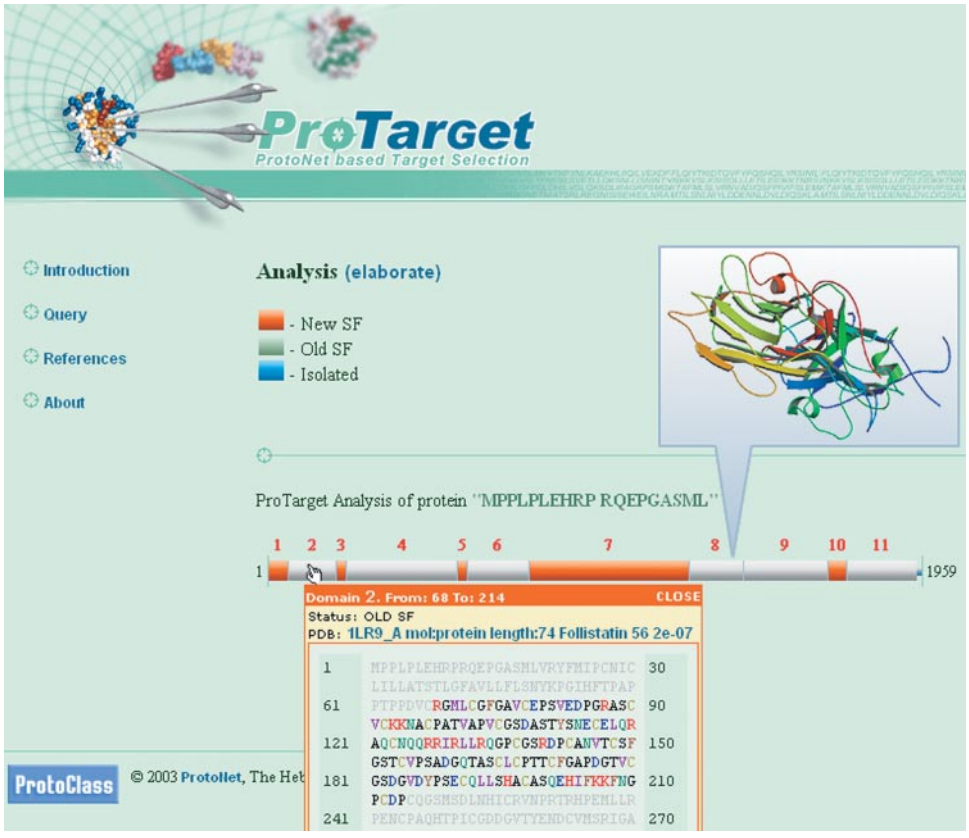

Figure 2. Graphical view of the ProTarget output for a query protein. The query protein (AGRN_RAT) of 1959 amino acids is fragmented by the cropping method to 11 segments. Segments labeled 'New SF' (new superfamily) are colored orange; those labeled 'Old SF' are in white. Each segment is associated with its sequence, the confidence score for its being new (score =1 indicates highest confidence) or a link to the best hit in the PDB. The inset shows the structure from the PDB that corresponds to segment 8. This is PDB 1PZ9, a 201 amino acid domain from the tail of AGRN-CHICK. The *E*-value of the similarity score between the sequence of any segment and the best correspondence structure from the PDB is indicated in the header of the interactive opened frame. An example for such frame for segment 2 is shown.

sequence, and finally a prediction is given based on whether this value is above or below the predetermined threshold, together with a confidence score. The confidence score depends on the distance of the level of the LSA from the threshold. The further the LSA of the sequence from the threshold, the higher (most significant) is our prediction. A scheme of the ProTarget algorithm is shown in Figure 1.

PROTARGET WEB SERVER

The ProTarget Web server at http://www.protarget.cs.huji.ac.il allows user-supplied targets to be ranked according to their likelihood of belonging to new superfamilies. The ProTarget user interface is straightforward and requires users to input a protein sequence and its name. The sequence may be a new or a pre-existing sequence. The output from the ProTarget prediction algorithm is shown graphically in Figure 2.

A query protein is labeled according to three color-coded categories based on the ProTarget predictions: 'old superfamily', 'new superfamily' and 'isolated'. If the protein is cropped into fragments, each fragment may have a different label. It is possible to move the cursor over the color-coded areas to view the respective subsequence as well as the confidence level of the prediction (ranging from 0 to 1). The example of the rat agrin protein following cropping into 11 segments is shown in Figure 2. Note that the very short segments are probably too short to be autonomous domains, and thus they may serve as linkers.

VERSIONS AND UPDATES

Proto3D will be updated once a year following expansion of the ProtoNet database. Currently, ProtoNet 4.0 includes over one million proteins from SwissProt and TrEMBL (10). The next version of ProTarget will include the 1 072 911 protein sequences from ProtoNet 4.0 combined with the 54 745 domains from the SCOP database (version 1.65) (11). Note that the online search against the PDB and the cropping procedure are performed using the most recently updated version of the PDB. In a future release we will include the option to submit larger numbers of sequences and will provide a mode for saving the results for further analysis.

ACKNOWLEDGEMENTS

Ilona Kifer and O.S. jointly developed the algorithm and the validation tests underlying ProTarget prediction. We thank Ilona for setting up the ProtoMap-based version of ProTarget. We thank Nati Linial and Elon Portugaly for fruitful discussions throughout this study. The authors wish to thank

the outstanding ProtoNet team and especially Alex Savenok for the ProTarget web design. M.L. is a member of the Sudarsky Center for Computational Biology (SCCB) at the Hebrew University of Jerusalem. This study was partially supported by the CESG consortium (NIMSG, NIH) and the European SPINE consortium. Funding to pay the Open Access publication charges for this article was provided by the National Science Foundation under grant DBI-0218798 and the National Institutes of Health under grant HG 02602-01.

Conflict of interest statement. None declared.

REFERENCES

1. Burley,S.K. and Bonanno,J.B. (2002) Structural genomics of proteins from conserved biochemical pathways and processes. *Curr. Opin. Struct. Biol.*, **12**, 383–391.
2. Westbrook,J., Feng,Z., Chen,L., Yang,H. and Berman,H.M. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.*, **31**, 489–491.
3. Watson,J.D., Todd,A.E., Bray,J., Laskowski,R.A., Edwards,A., Joachimiak,A., Orengo,C.A. and Thornton,J.M. (2003) Target selection and determination of function in structural genomics. *IUBMB Life*, **55**, 249–255.
4. Sali,A. (1998) 100,000 protein structures for the biologist. *Nature Struct. Biol.*, **5**, 1029–1032.
5. Vitkup,D., Melamud,E., Moult,J. and Sander,C. (2001) Completeness in structural genomics.[comment]. *Nature Struct. Biol.*, **8**, 559–566.
6. Sasson,O., Linial,N. and Linial,M. (2002) The metric space of proteins-comparative study of clustering algorithms. *Bioinformatics*, **18**, S14–S21.
7. Kaplan,N., Friedlich,M., Fromer,M. and Linial,M. (2004) A functional hierarchical organization of the protein sequence space. *BMC Bioinformatics*, **5**, 196.
8. Kifer,I., Sasson,O. and Linial,M. (2005) Predicting fold novelty based on ProtoNet hierarchical classification. *Bioinformatics*, **21**, 1020–1027.
9. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
10. Kaplan,N., Sasson,O., Inbar,U., Friedlich,M., Fromer,M., Fleischer,H., Portugaly,E., Linial,N. and Linial,M. (2005) ProtoNet 4.0: a hierarchical classification of one million protein sequences. *Nucleic Acids Res.*, **33**, D216–D218.
11. Lo Conte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.