Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W277–W280***
doi:10.1093/nar/gki371

ProTeus: identifying signatures in protein termini

Iris Bahir and Michal Linial*

Department of Biological Chemistry, Institute of Life Sciences, Hebrew University, Jerusalem 91904, Israel

Received January 28, 2005; Revised and Accepted March 7, 2005

ABSTRACT

ProTeus (PROtein TErminUS) is a web-based tool for the identification of short linear signatures in protein termini. It is based on a position-based search method for revealing short signatures in termini of all proteins. The initial step in ProTeus development was to collect all *signature groups* (SIGs) based on their relative positions at the termini. The initial set of SIGs went through a sequential process of inspection and removal of SIGs, which did not meet the attributed statistical thresholds. The SIGs that were found significant represent protein sets with minimal or no overall sequence similarity besides the similarity found at the termini. These SIGs were archived and are presented at ProTeus. The SIGs are sorted by their strong correspondence to functional annotation from external databases such as GO. ProTeus provides rich search and visualization tools for evaluating the quality of different SIGs. A search option allows the identification of terminal signatures in new sequences. ProTeus (ver 1.2) is available at http://www.proteus.cs.huji.ac.il.

INTRODUCTION

Protein signatures are detected by a wide variety of methods. Most methods imply initial multiple sequence alignment to form a 'seed alignment' that is then generalized to build a consensus or a profile. These methods are the basis for most current knowledge on signatures in proteins. Due to an inadequate statistical significance score, very short signatures fail to be recognized by most search methods. A common property to all methods used for signature identification, i.e. InterPro (1), is that the relative position of the signature in the protein is not considered.

Biological examples are known in which the sequence of the protein terminus is critical for dictating protein cellular localization, sorting, stability or binding to a partner protein (2). The potential of ProTeus method to detect signatures in protein termini is illustrated by the known signature of KDEL at the C-terminal. This signature is known to tag endoplasmic reticulum (ER) resident proteins (3). Our method was able to detect 58 proteins with KDEL signature in the same position at the C-terminal, of which 54 were annotated by SwissProt (manually checked by experts) as localized to the ER, suggesting a false positive rate of only 7%. If the positional information is not taken into consideration, 1037 proteins that have KDEL in their sequence would be detected, resulting in a false positive rate of 94%. Thus, including positional information and an unbiased collection of proteins are crucial for the detection of short terminal signatures.

Herein, we present the ProTeus tool. It allows the search through a collection of preprocessed protein sets that share terminal signatures (referred to as SIGs). We expect many of the SIGs to account for previously overlooked functionally related groups. ProTeus (PROtein TErminUS) is presented as a website that supports inspection and new discovery of candidate SIGs. ProTeus is available at http://www.proteus.cs.huji.ac.il.

METHODOLOGY

ProTeus uses sequences which were taken from the SwissProt database. A pool of short signatures of 3–10 amino acids is collected from each terminal. All proteins were grouped according to the sequence signature and its relative position. Following removal of groups based on their size, we archived all groups that showed a high degree of correspondence to a functional annotation from SwissProt, InterPro and GO [Gene Ontology, (4)].

Resources and protein database

Protein sequences were taken from SwissProt version 40.28 (containing 114 053 proteins). Following removal of sequences annotated as 'fragments', a total of 106 920 proteins remained. To this set of protein, several external annotation sources were used: SwissProt keywords with 865 annotations (version: 40.28); InterPro (1) with 5551 annotations (version 5.2) and GO (4) with 5229 annotations (July 2002). ProtoNet version 4.0 (5,6) was used as a protein classification hierarchical scaffold.

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

Classifying proteins into SIGs

Datasets of N-terminal fragments and C-terminal fragments were created from the first ten and last ten amino acids of each protein, respectively.

All terminal fragments were grouped according to the appearance of signatures at a given location on the sequence. Proteins that shared a signature on their termini at the same position were grouped into the same Signature Group (SIG). Signatures span 3–10 amino acids in length, and contain either no or one undetermined amino acid (a gap), the latter referred to as *gapped SIG*. Often few continuous SIGs can be merged to create a unified gapped SIG.

Removal of irrelevant SIGs

We performed three sequential steps of SIG inspection and removal of SIGs; in order to reduce the number of proposed SIGs: (i) Groups with >10 proteins were removed. (ii) We tested each SIG for its correspondence with a biological annotation from GO, SwissProt or InterPro. We assigned to each SIG the most highly corresponding annotation using a score-based method. The score for a given annotation, k, and a set of proteins, P, in its data source is defined as:

$$\text{score}(P,k) = \frac{|P \cap K|}{|P \cup K|},$$

where K is the set of all SwissProt proteins that were assigned to annotation k. In order to identify the most significant signatures, we removed SIGs that received a low purity (<0.5). We defined Purity as the fraction of SIGs' proteins that intersect with the assigned annotation. If >9 proteins intersected with the annotation k, the SIG was not removed although the purity might be lower than 0.5. (iii) We removed all SIGs that contain proteins with a substantial overall sequence similarity. To this end, we took advantage of the scaffold of all proteins as reflected by ProtoNet (6). The hierarchical level of ProtoNet was applied in order to remove SIGs that share high level of sequence similarity. Following these steps, we composed a collection of functionally suggestive SIGs that are presented in http://www.proteus.cs.huji.ac.il/.

Quality measurement for the SIG assigned annotation

In order to test the significance of the annotation assigned to a SIG, we calculated the P-value for a group of proteins to have 'randomly' received the assigned annotation.

The P-value was calculated according to the hypergeometric distribution: the chance of getting x or more hits for an annotation when randomly picking a set of size g proteins out of a database of d proteins, given there are k proteins in the database with this annotation is:

$$P\text{-value}(x,g,d,k) = \sum_{i=x}^{\min(k,g)} \frac{\binom{k}{i}\binom{d-k}{g-i}}{\binom{d}{g}}.$$

We used an approximation for the binomial coefficient provided. This P-value is calculated for all available annotations.

SEARCH ProTeus

ProTeus offers five search options:

(i) *View all collected SIGs*. The user defines the N- or C-terminal, the source of data for searching; SwissProt (version 40.28) or a merge of SwissProt (41.21 and TrEMBL 24.8) and the annotation source for searching; SwissProt, GO or InterPro.

(ii) *Search by an annotation*. The search for significant SIGs covers all annotations (a complete term or a partial one) derived from external annotation sources including GO, SwissProt or InterPro.

(iii) *Scan a protein*. Any protein, whether external or part of the protein database, may be tested for a match with the collected SIGs.

(iv) *Search for a signature*. The user may provide any suggested signature (continuous or gapped) from 3 to 10 amino acids in length. The SIGs that correspond to that signature are presented.

(v) *BLAST your protein*. This is a specialized BLAST (7) version tuned for short sequences that uses the BLAST algorithm to search for signatures on the termini of the user's protein sequence. This BLAST option allows detection of signatures that are degenerate or include multiple gaps.

REPRESENTATION OF A SIG

For each of the five search modes, the final result is a summary line with detailed information on the properties of the selected SIG. Table 1 summarizes the information presented in the summary line for this SIG. The example presented is a result of browsing the C-terminal collection of ProTeus based on the SwissProt annotations.

The Signature KDEL (mentioned in the Introduction) is specified by its position, −4 (note that position −1 refers to the most C-terminal amino acid): 58 proteins share the KDEL signature, 54 of them are annotated by SwissProt as 'Endoplasmic reticulum'. Thus, the calculated purity is 93%.

Table 1. A summary table for a selected SIG of KDEL signature

No. of motifs	13
Signature	KDEL
Relative position	−4
SIG size	58
Annotation source	SwissProt
No. of proteins intersecting with annotation	54
Purity	0.93
ProtoNet cluster	227 261
Cluster size	89 264
General annotation frequency	1.04%
Protolevel	99.97
Connection ratio	17.05
No. of kingdoms	3
No. of taxons	72
No. of taxa per kingdom	24
Average length	529.67
SD	175.81

The SIG is defined by the identity of the N- or C-terminal, the underlying annotation source, the signature, the position of the signature in the protein and the corresponding dominant annotation associated with it.

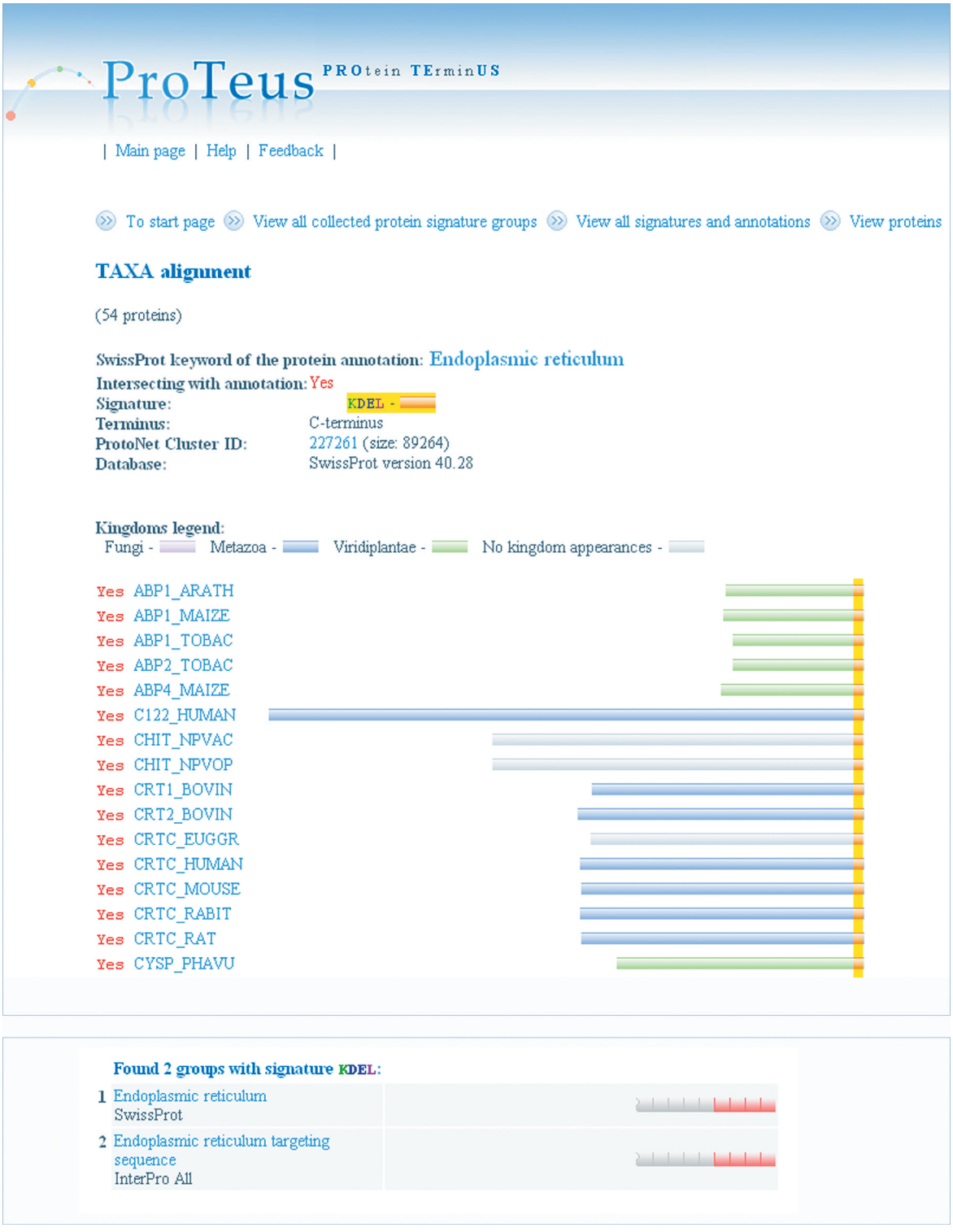

Figure 1. A sample of ProTeus website search results. Upper screen: shows a section from the taxonomical view on the proteins within specific SIG of the KDEL signature. Lower screen: the result of searching SIG with a KDEL signature revealed 2 SIGS based on annotations from external sources. The list of proteins and the source of annotations are available through an active link.

The abundance of this annotation in the database is 1.04%. Additional information refers to the number of taxonomical kingdoms that are covered by this SIG (72 different taxa in 3 different kingdoms). The average length and SD of the set of proteins that accounts for the best annotation ('Endoplasmic reticulum') is reported. The degree of sequence similarity is shown by the 'Protolevel' and the 'Connection ratio' (see 'help' in the ProTeus website for explanations).

From the summary line, the user may request a full description of the proteins in this SIG according to several modes of visualizations:

(i) A full list of proteins for downloading in several routinely used formats.
(ii) A graphical view for the proteins that share the best corresponding functional annotation (marked as YES); the proteins that have the signature in an identical position, however, the annotation term that specifies the SIGs is missing (marked as NO) and the combined set (marked as ALL). For each of the protein sets, the user can activate a local ClustalW tool (8) for creating a multiple alignment consensus. The terminal signature is marked as a yellow patch on the protein sequence.
(iii) A PANDORA (9) based visualization on a set of proteins marked with YES, NO or ALL (see above). PANDORA presents an integrated biological view of protein sets based on knowledge-based functional annotation sources, and offers statistical evaluation of these sets.
(iv) A local BLAST search for any pairs of proteins within the SIG.

Screen shots from ProTeus website are shown in Figure 1. The upper screen represents a graphical view of the taxonomy diversity of a selected SIG with the KDEL signature, the relevant signature is marked by a yellow bar. The different kingdoms are color-coded. The lower screen shows the results following the search for a SIG with the KDEL signature at its C-terminal.

At any step, the user can search and inspect a SIG that is either continuous or gapped. Furthermore, a link to the ProtoNet database is available. This link allows the user to access the rest of the proteins in the database that share the annotation specified by the SIG of interest.

MAINTENANCE AND FUTURE DIRECTIONS

The analysis described above is based on SwissProt (version 40.28). An identical scheme was applied to a larger database that combines SwissProt (version 41.21) and TrEMBL (version 24.8) with over one million proteins. Additional external sources of annotations such as protein–protein interactions will be included in future versions. A 'feedback' option for experimental biologists is presented to allow an input from the community on the validity of the proposed SIGs. The ProTeus website will be updated twice a year in conjugation with ProtoNet updates (10).

CONCLUSIONS

ProTeus provides a collection of few hundreds of SIGs covering the SwissProt database and much larger number when combining proteins from SwissProt and TrEMBL. ProTeus focuses on signatures that are often undetected by routine search programs. The set of proteins within a SIG often span a broad phylogenic diversity and a large variation in protein size. An interesting case is represented by those SIGs in which some of the proteins are marked as hypothetical. In such instances, it is appealing to suggest functional inference with other annotated proteins within the same SIG.

ProTeus provides an online interactive tool that allows detecting previously known and potentially overlooked signatures in protein termini.

ACKNOWLEDGEMENTS

We thank Noam Kaplan for his valuable suggestions and support. We thank the ProtoNet team for developing maintenance and support throughout. Special thanks to Alex Savenok for ProTeus web site design. Grant support is by the NoE European BioSapiens consortium. I.B. is supported by a fellowship of the SCCB, The Sudarsky Center for Computational Biology. 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 HG02602-01.

Conflict of interest statement. None declared.

REFERENCES

1. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D. *et al*. (2000) *InterPro*—an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics*, **16**, 1145–1150.
2. Chung,J.J., Shikano,S., Hanyu,Y. and Li,M. (2002) Functional diversity of protein C-termini: more than zipcoding? *Trends Cell Biol*., **12**, 146–150.
3. Pelham,H.R. (1990) The retention signal for soluble proteins of the endoplasmic reticulum. *Trends Biochem. Sci*., **15**, 483–486.
4. Camon,E., Magrane,M., Barrell,D., Binns,D., Fleischmann,W., Kersey,P., Mulder,N., Oinn,T., Maslen,J., Cox,A. *et al*. (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res*., **13**, 662–672.
5. Kaplan,N., Friedlich,M., Fromer,M. and Linial,M. (2004) A functional hierarchical organization of the protein sequence space. *BMC Bioinformatics*, **5**, 196.
6. Sasson,O., Vaaknin,A., Fleischer,H., Portugaly,E., Bilu,Y., Linial,N. and Linial,M (2003) ProtoNet: hierarchical classification of the protein space. *Nucleic Acids Res*., **31**, 348–352.
7. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
8. Higgins,D.G. and Sharp,P.M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*, **73**, 237–244.
9. Kaplan,N., Vaaknin,A. and Linial,M. (2003) PANDORA: a keyword-based analysis of protein sets by integration of annotation sources. *Nucleic Acids Res*., **31**, 5617–5626.
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.