W210–W213 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl093

pTARGET: a web server for predicting protein subcellular localization

Chittibabu Guda

Gen*NY*sis Center for Excellence in Cancer Genomics and Department of Epidemiology and Biostatistics,
University at Albany, State University of New York, 1 Discovery drive, Rensselaer, NY 12144-3456, USA

Received February 6, 2006; Revised February 18, 2006; Accepted March 7, 2006

ABSTRACT

The pTARGET web server enables prediction of nine distinct protein subcellular localizations in eukaryotic non-plant species. Predictions are made using a new algorithm [C. Guda and S. Subramaniam (2005) pTARGET [corrected] a new method for predicting protein subcellular localization in eukaryotes. *Bioinformatics*, 21, 3963–3969], which is primarily based on the occurrence patterns of location-specific protein functional domains in different subcellular locations. We have implemented a relational database, PreCalcDB, to store pre-computed prediction results for all eukaryotic non-plant protein sequences in the public domain that includes about 770 000 entries. Queries can be made by entering protein sequences or by uploading a file containing up to 5000 protein sequences in FASTA format. Prediction results for queries with matching entries in the PreCalcDB will be retrieved instantly; while for the missing ones new predictions will be computed and sent by email. Pre-computed predictions can also be downloaded for complete proteomes of *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila*, *Mus musculus* and *Homo sapiens*. The server, its documentation and the data are accessible from http://bioinformatics.albany.edu/~ptarget.

INTRODUCTION

Protein subcellular localization is a key functional characteristic of proteins. Subcellular localization of proteins in appropriate compartments is vital for the internal structure of the cell itself and for its functional integrity. Determination of the subcellular localization by experimental means is not practical for all proteins owing to time and cost constraints. Alternatively, several computational methods have been developed for the prediction of subcellular localization in eukaryotic proteins. These methods are broadly categorized into four classes. (i) Methods based on sorting signals relying on the presence of sorting signals that are recognized by location-specific transport machinery to enable their entry (1–3). (ii) Methods based on the differences in amino acid composition (AAC), pseudo-AAC (includes AAC plus sequence order) or amino acid properties of proteins from different subcellular locations (4–6). (iii) Methods based on lexical analysis of keywords (LOCkey) in the functional annotation of proteins (7). (iv) Methods using phylogenetic profiles (8) or domain projection method (9). Nevertheless, only a handful of methods are accessible online to the research community. We compiled a list of currently available web servers that offer access to web-based prediction of subcellular localizations for eukaryotic proteins (Table 1). Some of these methods can predict only a few types of locations owing to inherent limitations while some lack the robustness to handle the heterogeneity expected in eukaryotic proteomes. Moreover, some web servers are designed to process only one sequence or a limited number of sequences in batch queries, thus, limiting their use for proteome-wide predictions (Table 1).

Recently, we developed two prediction methods: MITOPRED, for predicting nucleus-encoded mitochondrial proteins (10,11) and pTARGET, for predicting nine distinct subcellular locations in eukaryotic proteomes (12) based on location-specific functional domains and AAC. pTARGET method is relatively robust for proteome-wide predictions since it does not rely on the presence of a signal or target peptides. Based on this method, here we present the pTARGET web server that can process proteome-scale queries, backed by a relational database, PreCalcDB, containing pre-computed predictions.

DESIGN AND IMPLEMENTATION

The pTARGET server has been designed using PERL-CGI interface to process user queries and display or email the

To whom correspondence should be addressed. Tel: +1 518 591 7155; Fax: +1 518 591 7151; Email: cguda@albany.edu

prediction results. A relational database, PreCalcDB containing pre-computed predictions has been developed to back the web server and to provide instant access to predictions for most of the eukaryotic protein sequences in the public domain. Perl DBD module was used to interface with the MySQL database. Query sequences are first searched against this database and predictions will be retrieved for matching entries; while for others, a new prediction process will be launched. The new prediction process includes searching the Protein family database (Pfam database, http://pfam.wustl.edu), which is the most time-consuming step in the prediction process. Pre-calculated prediction results are instantly displayed on the screen while those from new predictions are emailed to the user upon completion of the computation steps.

Algorithm

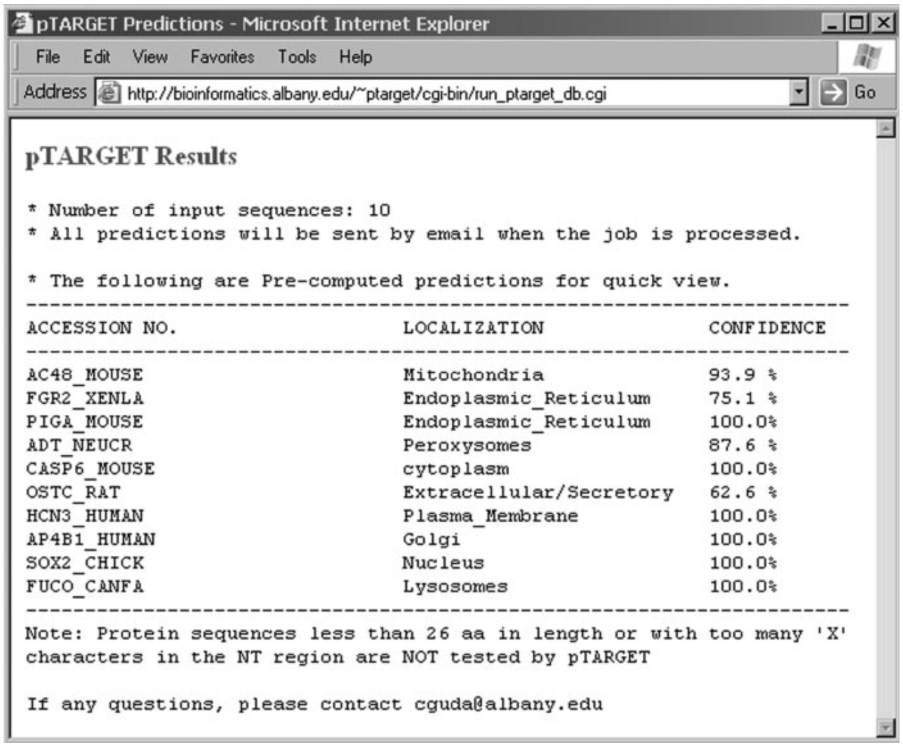
The pTARGET method (12) predicts proteins targeted to nine distinct subcellular locations in eukaryotic non-plant species. This prediction algorithm calculates two distinct scores, i.e. first, a score based on the presence or absence of location-specific Pfam domains (Pfam score) and second, a score based on the relative amino acid weights calculated from AAC (AAC score). The nine subcellular locations predicted by pTARGET include cytoplasm, endoplasmic reticulum, extracellular/secretory, golgi, lysosome, mitochondria, nucleus, plasma membrane and peroxisome.

Pre-computed prediction database

To expedite the response time, pre-computed predictions have been provided for all non-redundant eukaryotic protein sequences (excluding plant sequences) in the public domain (∼770 000 sequences). We have created a relational database, PreCalcDB, using the open source database MySQL 4.0 (downloaded from http://dev.mysql.com). PreCalcDB contains several relational tables to store protein sequences, headers and pTARGET prediction results. Protein sequence strings are treated as primary keys that are indexed to sequence accession IDs and to the prediction results. Programs for database development were written in SQL and supporting programs for accessing and manipulating the database were implemented in JAVA using the JDBC (Java Database Connectivity) API. Query sequences are searched against those in the PreCalcDB and for matching entries, predictions are retrieved and displayed instantly in the browser window. For the missing entries, the response time depends on the number of sequences requiring new predictions, which is ∼1 h for 60 sequences on the current server. If approved by the user, new query sequences and the prediction results will be automatically loaded back to the PreCalcDB to make them available to the next user. Since PreCalcDB stores comprehensive sets of protein sequences from major databases including Swiss-Prot and TrEMBL from EBI and the 'nr' database from GenBank, pre-computed predictions are available for the majority of user queries. Predictions for complete proteomes of important eukaryotic species including yeast (*Saccharomyces cerevisiae*), nematode (*Caenorhibditis elegans)*, fruit fly (*Drosophila melanogaster*), mouse (*Mus musculus*) and human (*Homo sapiens*) can also be downloaded from the web server.

Input and output formats

Users can enter protein sequences in the text box or upload a file containing up to 5000 protein sequences in FASTA format.


pTARGET Predictions - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://bioinformatics.albany.edu/~ptarget/cgi-bin/run_ptarget_db.cgi
Go

pTARGET Results

```
* Number of input sequences: 10
* All predictions will be sent by email when the job is processed.

* The following are Pre-computed predictions for quick view.
------------------------------------------------------------------------
ACCESSION NO.                   LOCALIZATION              CONFIDENCE
------------------------------------------------------------------------
AC48_MOUSE                      Mitochondria              93.9 %
FGR2_XENLA                      Endoplasmic_Reticulum     75.1 %
PIGA_MOUSE                      Endoplasmic_Reticulum     100.0%
ADT_NEUCR                       Peroxysomes               87.6 %
CASP6_MOUSE                     cytoplasm                 100.0%
OSTC_RAT                        Extracellular/Secretory   62.6 %
HCN3_HUMAN                      Plasma_Membrane           100.0%
AP4B1_HUMAN                     Golgi                     100.0%
SOX2_CHICK                      Nucleus                   100.0%
FUCO_CANFA                      Lysosomes                 100.0%
------------------------------------------------------------------------
Note: Protein sequences less than 26 aa in length or with too many 'X'
characters in the NT region are NOT tested by pTARGET

If any questions, please contact cguda@albany.edu
```

Figure 1. A screenshot showing pTARGET prediction results.

Table 1. Selected web servers for predicting protein subcellular localization online

Method	URL	Predicted location(s)	Scoring criteria	Batch option
LOCTarget	http://cubic.bioc.columbia.edu/services/LOCtarget/LOCtarget.html	11 subcellular Locations	Homology, Keywords, NLS	No
LOCTree	http://cubic.bioc.columbia.edu/services/loctree/	11 subcellular Locations	NLS, Prosite patterns, Homology, Keywords	Yes
MITOPRED	http://bioinformatics.albany.edu/~mitopred	Mit	Pfam domains, AAC	Yes
Mitoprot	http://ihg.gsf.de/ihg/mitoprot.html	Mit	Target Peptides	No
PredictNLS	http://cubic.bioc.columbia.edu/predictNLS	Nuc	NLS Patterns	No
ProSLP	http://www.ccbb.re.kr/proslp/	13 subcellular Locations	Homology	No
pSLIP	http://pslip.bii.a-star.edu.sg/	Cyt, Exc, Nuc, Mit, Pla	AAC physico-chemical properties	Yes
PSORT-II	http://psort.nibb.ac.jp/	12 subcellular Locations	TPs, SPs, AAC, rule-based, other	No
Sub-Loc	http://www.bioinfo.tsinghua.edu.cn/SubLoc	Cyt, Exc, Mit, Nuc	AAC	Yes
TargetP	http://www.cbs.dtu.dk/services/TargetP/	Exc, Mit	Target peptides	Yes
Wolf-PSORT	http://wolfpsort.seq.cbrc.jp/	12 subcellular Locations	TPs, SPs, AAC, rule-based, other	Yes

NLS-nuclear localization signals, AAC-amino acid composition, TPs-target peptides, SPs-signal peptides, Cyt-cytoplasmic, Exc-extracellular/secretory, Mit-mitochondrial, Nuc-nuclear, Pla-plasma membrane.

Results will be displayed on the screen and emailed in plain text format. Users should be aware that some spam filtering programs installed on the user's mail client could sort the emails from pTARGET server into a SPAM folder. As shown in the screen shot (Figure 1), each prediction is followed by a prediction confidence value. Prediction confidence value is calculated as the ratio of calculated score to the total score required to make a true prediction, and it is expressed in percentage. For example, a score of 50 is required to make a prediction with 100% confidence. If the calculated score for a query sequence is 45, the prediction confidence is 90%. All query sequences with calculated scores equal to or exceeding 50 are predicted with 100% confidence.

DISCUSSION

pTARGET web server is intended for performing genome-scale prediction of protein subcellular localizations in eukaryotic organisms excluding plant species. Since several metabolic pathways and organelles in plants are not the same as in animals, the distribution of protein functional domains in these two systems is different. Similarly, bacterial species do not possess all the subcellular locations as eukaryotic cells do. Hence, predictions with this web server should be used only in the context of eukaryotic non-plant proteomes. Prediction capabilities of other similar web servers are limited to a few types of locations in some cases while some can process only one query sequence at a time (Table 1), making them unsuitable for genome-scale predictions. Moreover, the underlying prediction algorithm (12) for this web server is primarily based on the Pfam domain occurrence patterns and hence is more robust than some other methods that require the presence of a signal or target peptide to make an accurate prediction. Since pTARGET predictions are compute-intense, we implemented the PreCalcDB as a back-end server to this web resource to significantly improve the response time. One limitation of pTARGET method is that proteins containing Pfam domains that exist in multiple subcellular locations or those without Pfam annotations are predicted solely based on their amino acid composition resulting in reduced prediction accuracy. However, over 70% of the protein sequences in the public domain are currently covered with Pfam annotations and this coverage is rapidly expanding. Hence, the prediction accuracy of pTARGET is expected to improve as more Pfam domains and more information on subcellular localization become available. We have implemented a self-enriching feature for the PreCalcDB which, with user's approval, can store the newly calculated predictions in the relational database to make them instantly accessible to another user. We will update the PreCalcDB every 6 months to provide the current and most accurate information to the research community.

ACKNOWLEDGEMENTS

This work has been supported by the startup funds to CG from the State University of New York at Albany (SUNY, Albany). The author is thankful to Research IT staff at SUNY, Albany and to Dr Russ Miller's group at SUNY, Buffalo for providing access to grid computing to run large-scale HMMER jobs. Funding to pay the Open Access publication charges for this article was provided by the SUNY startup funds.

Conflict of interest statement. None declared.

REFERENCES

1. Nielsen,H., Engelbrech,J., Brunak,S. and von Heijne,G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, **10**, 1–6.
2. Nakai,K. and Horton,P. (1999) PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, **24**, 34–36.
3. Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
4. Hua,S. and Sun,Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.
5. Chou,K.C. (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*, **43**, 246–255.
6. Sarda,D, Chua,G.H., Li,K.B. and Krishnan,A. (2005) pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. *BMC Bioinformatics*, **6**, 152.

7. Nair,R. and Rost,B. (2002) Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics*, **18**, S78–S86.
8. Marcotte,E.M., Xenarios,I., van Der Bliek,A.M. and Eisenberg,D. (2000) Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **97**, 12115–12120.
9. Mott,R., Schultz,J., Bork,P. and Ponting,C.P. (2002) Predicting protein cellular location using a domain projection method. *Genome Res.*, **12**, 1168–1174.
10. Guda,C., Fahy,E. and Subramaniam,S. (2004) MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics*, **20**, 1784–1794.
11. Guda,C., Guda,P., Fahy,E. and Subramaniam,S. (2004) MITOPRED: A web server for genome-scale prediction of mitochondrial proteins. *Nucleic Acids Res.*, **32**, W372–W374.
12. Guda,C. and Subramaniam,S. (2005) pTARGET: A new method for predicting protein sub-cellular localization in eukaryotes. *Bioinformatics*, **21**, 3963–3969.