W372–W374 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh374

MITOPRED: a web server for the prediction of mitochondrial proteins

Chittibabu Guda[1], Purnima Guda[1], Eoin Fahy[1] and Shankar Subramaniam[1,2,*]

[1]San Diego Supercomputer Center and [2]Departments of Bioengineering, Chemistry and Biochemistry, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

Received February 13, 2004; Revised and Accepted March 15, 2004

ABSTRACT

MITOPRED web server enables prediction of nucleus-encoded mitochondrial proteins in all eukaryotic species. Predictions are made using a new algorithm based primarily on Pfam domain occurrence patterns in mitochondrial and non-mitochondrial locations. Pre-calculated predictions are instantly accessible for proteomes of *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila*, *Homo sapiens*, *Mus musculus* and *Arabidopsis* species as well as all the eukaryotic sequences in the Swiss-Prot and TrEMBL databases. Queries, at different confidence levels, can be made through four distinct options: (i) entering Swiss-Prot/TrEMBL accession numbers; (ii) uploading a local file with such accession numbers; (iii) entering protein sequences; (iv) uploading a local file containing protein sequences in FASTA format. Automated updates are scheduled for the pre-calculated prediction database so as to provide access to the most current data. The server, its documentation and the data are available from http://mitopred.sdsc.edu.

INTRODUCTION

Finding the subcellular location of a protein is an important step towards finding its function, as well as the pathway it is associated with, in the cell. We are especially interested in predicting mitochondrial proteins owing to their role in a variety of complex biochemical processes and their association with over 100 known human diseases (http://www.neuro.wustl.edu/neuromuscular/mitosyn.html). The Swiss-Prot database provides annotation for subcellular location based on the experimental evidence, but such reliable information is available only for a small number of proteins. In the case of *Saccharomyces cerevisiae*, subcellular locations have been experimentally determined for the entire proteome using immunodetection methods (1,2); however, such approaches are not feasible for all genomes. Over the past decade, several *in silico* prediction methods have been developed for determining the sub-cellular location of proteins (3–8). However, none of these methods is suitable for genome-scale prediction of mitochondrial proteins due to inherent limitations in the prediction protocols such as dependence on the presence of signal sequences or cleavage sites. Recently, we developed a new method (MITOPRED) for genome-scale prediction of mitochondrial proteins based primarily on Pfam domain occurrence patterns (9). Here, we present a web server to make genome-scale predictions using the MITOPRED algorithm.

DESIGN AND IMPLEMENTATION

This web server has been designed using a PERL-CGI interface to access user queries. Depending on the input data, the program either retrieves pre-calculated predictions stored on the server database or launches a MITOPRED process as shown in Figure 1. To expedite the prediction process, the interface provides built-in mapping facilities to match either the Swiss-Prot or TrEMBL (jointly known as 'SPTr') accession number or the input sequence with corresponding values in the pre-calculated entries. Input sequences are matched using hexadecimal hashing methods from the MD5 Perl module, and those without matches are separated. For matching entries, predictions are retrieved from the pre-calculated database, and for others a new prediction process is launched. A new prediction processes includes searching the protein family database (Pfam database, http://pfam.wustl.edu), which is the most time-consuming step, depending on the number of sequences. Predictions can be done at different confidence cutoffs such as '99%', '85%' and '60%'. Intuitively, at higher confidence levels, the number of predictions is lower; however, the prediction accuracy is high. Pre-calculated results are instantly displayed on the screen while those from new predictions are emailed to the user upon completion of the computation steps.

*To whom correspondence should be addressed. Tel: +1 858 822 0986; Fax: +1 858 822 3782; Email: shankar@ucsd.edu

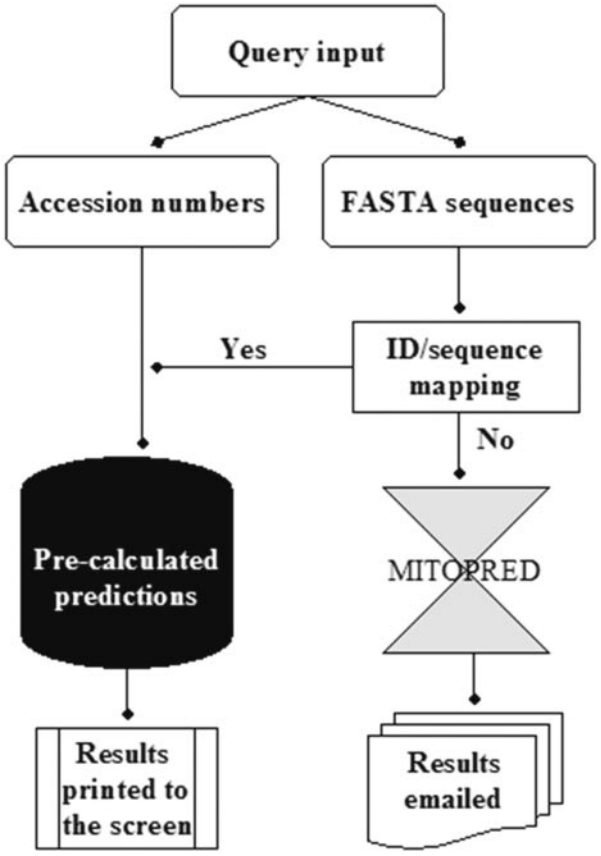

Figure 1. Flow diagram of implementation of the MITOPRED server.

Algorithm

The algorithm is based primarily on the occurrence of mitochondria-specific Pfam domains and the differences in the amino acid compositional values between mitochondrial and non-mitochondrial protein sequences (9). A query sequence is scored based on its N-terminal and C-terminal amino acid composition and the presence or absence of mitochondria-specific or non-mitochondria-specific Pfam domains. Pfam score is calculated using only Pfam-A annotations, since Pfam-B annotations are not very reliable.

Pre-calculated predictions

To expedite the response time, pre-calculated predictions have been provided for the entire eukaryotic sequence set in the SPTr database release 42.0 (~500 000 sequences) at different confidence levels. For example, it takes only 20 s to retrieve predictions for the entire proteome of yeast when a local file containing yeast SPTr accession numbers is uploaded. Predictions for complete proteomes of important eukaryotic species such as yeast (*S.cerevisiae*), nematode (*C.elegans*), fruit fly (*Drosophila melanogaster*), mouse (*Mus musculus*), human (*Homo sapiens*) and *Arabidopsis thaliana* can also be downloaded from the web server.

Query interface

Users can enter the input data in four different ways: by (i) entering accession numbers; (ii) uploading a local file with accession numbers; (iii) entering protein sequences; (iv) uploading a local file containing protein sequences in FASTA format. Since new prediction processes are very time consuming, we limit the number of sequences per search to 500. In the queries using sequences, users are required to select the source of the sequences as 'yeast/animal' or 'plant' species. This is because the program used for predicting plant sequences is a slightly different variant of that used for animal sequences due to the presence of chloroplasts in plant species.

Input and output formats

SPTr accession numbers can be entered in the text box as space-delimited or comma-delimited or one accession number per line format; however, an uploaded file should be in one accession number per line format. Input protein sequences should be entered or uploaded in FASTA format only. Results are displayed on the screen or emailed in plain text format as one line per accession number or sequence format. As shown in the screen shot (Figure 2), each prediction is followed by a prediction confidence value. Prediction confidence value is calculated as the ratio of calculated score to the total required score (for being a mitochondrial sequence) expressed as a percentage.

DISCUSSION

This server is intended primarily for making genome-scale predictions of mitochondrial proteins in eukaryotic organisms. Other popular servers such as TargetP (http://www.cbs.dtu.dk/services/TargetP), PSORT (http://psort.nibb.ac.jp), MITOPROT (http://ihg.gsf.de/ihg/mitoprot.html) and Predotar (http:// www.inra.fr/predotar/) serve the same purpose, however, these servers are not suitable for genome-scale predictions since their methods rely primarily on the presence of target peptides or cleavage sites in the sequences that are lacking in the majority of mitochondrial sequences. Our method (MITOPRED) does not require such signal peptides but relies primarily on the Pfam domain occurrence patterns and hence could be used for predicting the entire proteome of a genome. Also, this method is sufficiently robust to use against all eukaryotic species without having to be trained on species-specific data. One limitation of our method is that proteins containing Pfam domains that exist in both mitochondrial and non-mitochondrial locations or those without Pfam annotations are predicted solely based on their amino acid composition, resulting in reduced prediction accuracy. However, about 70% of the Swiss-Prot and TrEMBL sequences are currently covered by Pfam annotations, and this coverage is rapidly expanding. The prediction accuracy of our program is expected to improve as more Pfam domains and more information on subcellular localization become available. Since this web server is scheduled to update automatically for each SPTr and Pfam database update, it will provide current and the most accurate information on mitochondrial proteins to the community. We are currently working on expanding this methodology to predict proteins destined for other subcellular locations and on creating a web server to access such predictions.

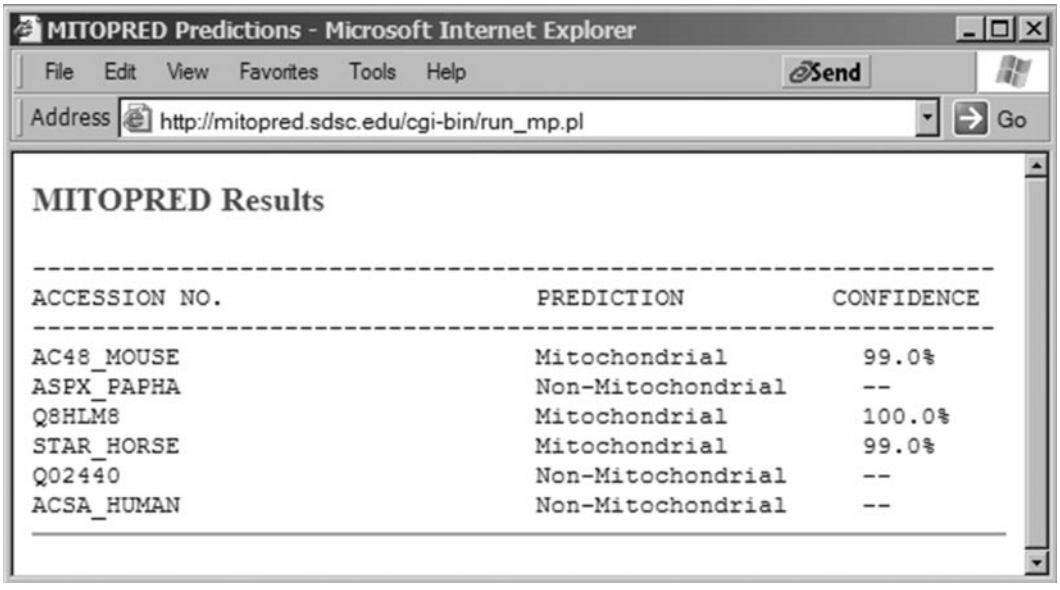

Figure 2. A screen shot of prediction results from MITOPRED.

ACKNOWLEDGEMENTS

The authors are thankful to Dr Giridhar Chukkapalli (San Diego Supercomputer Center) for assistance in running HMM jobs. This project is supported by the University of California Life Sciences Informatics (LSI) Program/Mitokor grant (L99-10077).

REFERENCES

1. Kumar,A., Agarwal,S., Heyman,J.A., Matson,S., Heidtman,M., Piccirillo,S., Umansky,L., Drawid,A., Jansen,R., Liu,Y. *et al.* (2002) Subcellular localization of the yeast proteome. *Genes Dev.*, **16,** 707–719.
2. Huh,W.-K., Falvo,J.V., Gerke,L.C., Carroll,A.S., Howson,R.W., Weissman,J.S. and O'Shea,E.K. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.
3. 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.
4. Fujiwara,Y., Asogawa,M. and Nakai,K. (1997) Prediction of mitochondrial targeting signals using hidden Markov models. In Miyano,S. and Takagi,T. (eds), *Genome Informatics*. Universal Academy Press, Inc., Tokyo, Japan, pp. 53–60.
5. 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.
6. 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.
7. Claros,M.G. and Vincens,P. (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.*, **241**, 779–786.
8. Drawid,A. and Gerstein,M. (2000) A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J. Mol. Biol.*, **301**, 1059–1075.
9. Guda,C., Fahy,E. and Subramaniam,S. (2004) A genome-scale method for the prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics*, 10.1093/bioinformatics/bth171.