Amino Acids (2011) 40:975–980
DOI 10.1007/s00726-010-0724-y

ORIGINAL ARTICLE

# PROlocalizer: integrated web service for protein subcellular localization prediction

**Kirsti Laurila · Mauno Vihinen**

Received: 16 June 2010 / Accepted: 10 August 2010 / Published online: 2 September 2010


**Abstract** Subcellular localization is an important protein property, which is related to function, interactions and other features. As experimental determination of the localization can be tedious, especially for large numbers of proteins, a number of prediction tools have been developed. We developed the PROlocalizer service that integrates 11 individual methods to predict altogether 12 localizations for animal proteins. The method allows the submission of a number of proteins and mutations and generates a detailed informative document of the prediction and obtained results. PROlocalizer is available at http://bioinf.uta.fi/PROlocalizer/.

**Keywords** Protein localization prediction · Cell compartments · Mutations · Disease-causing mutations · Prediction method

## Introduction

Cells have several membrane-enclosed compartments, which have different protein constituents and composition. The subcellular localization of a protein is important to know, as it is linked to protein function, interactions and activity in signaling pathways. Proteins are sorted to their destinations based on targeting signals that can appear anywhere in the protein sequence. During the protein translation in animals, proteins can enter the classical secretory pathway (SP) in which the protein is sorted via endoplasmic reticulum (ER), e.g., to the Golgi apparatus, plasma membrane, lysosome or to the extracellular matrix. Protein can also stay at ER. If the protein does not contain a signal peptide to the SP it can remain in the cytoplasm, enter non-classical secretory pathways (nSPs) or it can be sorted to the nucleus, peroxisomes or mitochondrion (Dönnes and Höglund 2004).

All details in protein sorting are not very well understood, but several SP targeting signals have been identified. These signals are usually N-terminal and their amino acid compositions are diverse as they are not strongly conserved (Stroud and Walter 1999). SP signal peptides have three part structure usually with polar and positively charged n-part, hydrophobic h-part and c-part, which often contains prolines and glycines to interfere with α-helix formation (Martoglio and Dobberstein 1998).

Mitochondrial proteins are targeted via N-terminal signal peptides as well. Another signal peptide can further sort the protein inside the mitochondrion. Nuclear localization signals are difficult to recognize as they are diverse and they can exist anywhere in the protein sequence. In addition, C-terminal signal peptides exist (Chou 2002). Of peroxisomal proteins, some are targeted with C-terminal sequences (Dönnes and Höglund 2004) and others with N-terminal signal peptide (Emanuelsson 2002).

Information on protein localization is scattered throughout publications and numerous databases and does not exist for many proteins. A protein can have several localizations, often depending on the state of the cell.

K. Laurila
Department of Signal Processing,
Tampere University of Technology,
P.O. Box 527, 33101 Tampere, Finland

K. Laurila · M. Vihinen (✉)
Institute of Medical Technology,
University of Tampere, 33014 Tampere, Finland
e-mail: mauno.vihinen@uta.fi

M. Vihinen
Science Center, Tampere University Hospital,
33520 Tampere, Finland

Experimental determination of protein subcellular localization can be time consuming and expensive even though couple of high-throughput approaches have been developed (Davis 2004; Falk et al. 2007). However, in these methods, protein localization can be interfered with reporter genes or cell fractionation. Consequently, numerous computational tools have been developed for localization prediction. The very first methods identified the existence of signal sequence, which was soon followed by methods for cleavage site prediction (Chou 2002). Dozens of prediction methods for subcellular localization have been developed, mainly for single compartments and in recent years some multicompartment predictors have also been released (for a review see Dönnes and Höglund 2004; Emanuelsson 2002; Schneider and Fechner 2004; Sprenger et al. 2006).

Present methods differ in many details, such as biological information used, algorithms and reliability. Many methods do predictions based on the amino acid sequence while others may need some additional information such as gene expression data. Several methods utilize amino acid composition because proteins in different environments are known to have different amino acid composition (Dönnes and Höglund 2004). Searches for predetermined signal peptide motifs and homology-driven approaches are also common. Several methods combine different rules to achieve better performance. The algorithms behind the methods also vary largely. The rules of prediction can be readily determined or they can be learnt from the training data set by different machine learning approaches. Most often used machine learning methods are Hidden Markov models, neural networks, self-organizing maps and support vector machines (Schneider and Fechner 2004).

The performance of the subcellular localization predictors varies. The localization of SP containing proteins can be predicted quite well while the knowledge of many non-SP signal peptides is insufficient as targeting sequences are usually just few residues long and as also the protein conformation affects the recognition of individual sites (Schneider and Fechner 2004).

For the single compartment predictors, the accuracy of prediction can be high, around 90% (Klee and Ellis 2005). Besides these binary predictors, some multipredictors have also been introduced. One of the pioneers in this field has been PSORT family of tools (Nakai and Horton 1999). In addition, some other methods have been launched including Hum-mPLoc (Shen and Chou 2007), HSLPred (Garg et al. 2005), LOCtree (Nair and Rost 2005), pTarget (Guda and Subramaniam 2005), Cello (Yu et al. 2006) Proteome analyst (Lu et al. 2004) and Euk-mPLoc 2.0 (Chou and Shen 2009). These predictors utilize machine learning methods such as support vector machine and k-Nearest Neighbours classification.

Based on individual predictors, a protocol was presented for a large number of subcellular localizations (Emanuelsson et al. 2007). The Scandinavian protocol is not actually a computer program, but a scheme by which individual predictions are manually run and interpreted. We evaluated the accuracy of WoLF PSORT (Horton et al. 2007) and Scandinavian protocol and predicted whether among 22,000 disease-related missense mutations are changes that could affect protein localization (Laurila and Vihinen 2009). According to the results, large number of diseases may arise due to mislocalization of proteins. Here we present an automated implementation of the Scandinavian protocol, which can be applied to predict the cellular localization of proteins and changes introduced by mutations.

## Materials and methods

PROlocalizer service implements the Scandinavian protocol of Emanuelsson et al. The method can be applied to animal proteins as the signals and localization machinery are well conserved in metazoans. The protocol can predict altogether 12 individual subcellular localizations, which are mitochondrial inner membrane, transmembrane; mitochondrial periplasmic space; mitochondrial matrix; Golgi, transmembrane; plasma membrane; secreted; ER lumen; nucleus; peroxisome; cytoplasmic; plasma membrane, GPI anchor; and plasma membrane, myristoylated. The prediction scheme is in Fig. 1. PROlocalizer is freely accessible for academic use at http://bioinf.uta.fi/PROlocalizer.

The protocol is based mainly on binary classifiers, which predict whether a protein is localized into a specific compartment or not. Of the 11 different programs in PROlocalizer, TargetP (Emanuelsson et al. 2000), SignalP (Bendtsen et al. 2004) and TMHMM (Krogh et al. 2001; Sonnhammer et al. 1998) were downloaded from http://www.cbs.dtu.dk/services/ and are run locally while programs Big-PI (Eisenhaber et al. 1998, 1999, 2000; Sunyaev et al. 1999) (http://mendel.imp.ac.at/sat/gpi/gpi_server.html), NMT (http://mendel.imp.ac.at/myristate/SUPLpredictor.htm), PeroxiP (Emanuelsson et al. 2003) (http://bioinfo.se/PeroxiP/), PredictNLS (Cokol et al. 2000) (http://cubic.bioc.columbia.edu/predictNLS/), PTS1 (Neuberger et al. 2003a, b (http://mendel.imp.ac.at/mendeljsp/sat/pts1/PTS1predictor.jsp), Golgi predictor (Yuan and Teasdale 2002) (http://ccb.imb.uq.edu.au/golgi/), Phobius (Käll et al. 2004) (http://phobius.sbc.su.se/), and Prosite (de Castro et al. 2006) (http://au.expasy.org/prosite/) are run over the Internet. If TargetP has a problem in sorting a protein to mitochondria with poor reliability coefficient (RC) (value 4 or 5) then also SignalP is used and the protein obtains two alternative localizations. Java and perl scripts were written

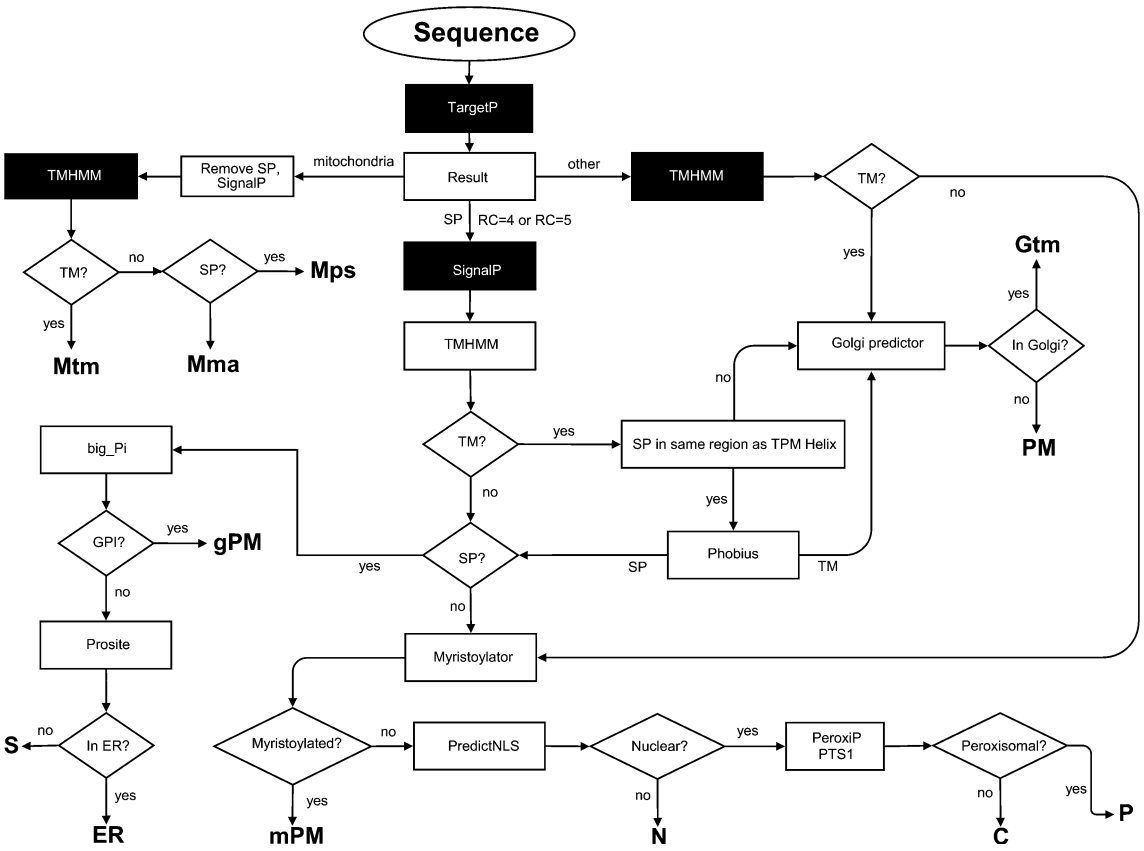



**Fig. 1** PROlocalizer prediction scheme. The service predicts proteins to 12 compartments: *C* cytosol; *ER* endoplasmic reticulum, lumen; *Gtm* Golgi, transmembrane; *Mtm* mitochondrial inner membrane (transmembrane); *Mps* mitochondrial periplasmic space; *Mma* mitochondrial matrix; *PM* plasma membrane; *S* secreted; *N* nucleus; *P* peroxisome; *gPM* plasma membrane, GPI anchor; and *mPM* plasma membrane, myristoylated. Programs run locally are indicated with *black box*. In some rare cases double localizations are predicted if the reliability coefficient (RC) in TargetP is high, i.e. poor

to submit the protein sequence to prediction programs and to automatically interpret the results and to generate the report for the user.

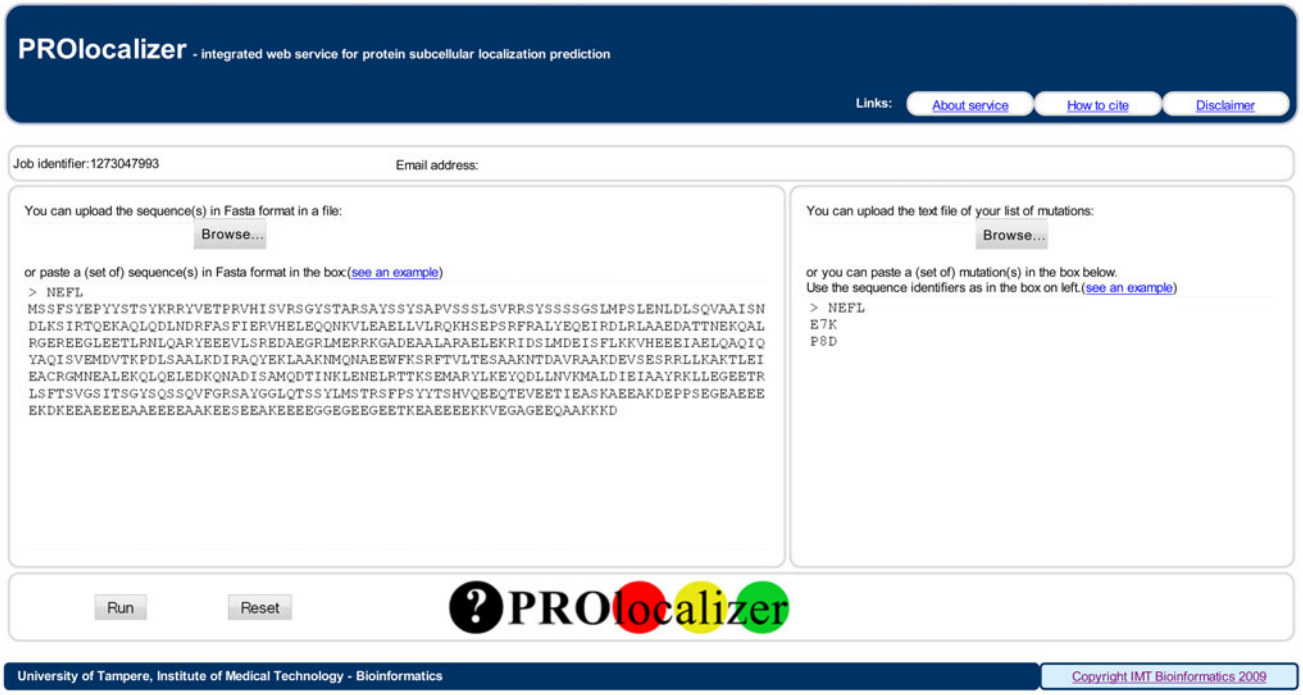
Briefly, the prediction protocol works as follows (Fig. 1): TargetP assigns whether the protein goes to secretory pathway or mitochondia or not. Then, mitochondrial proteins are further sorted to transmembrane, periplasmic space or matrix based on the analysis of transmembrane and signal peptide sequences. Transmembrane proteins have two prediction routes and are finally classified into those going to Golgi transmembrane or plasma membrane. Proteins with signal peptide(s) are predicted to their compartments based on the targeting motifs.

PROlocalizer has a user-friendly graphic interface (Fig. 2). The user needs to provide only protein sequence(s) in fasta format. Submission can be done via a file or pasting the information to the submission form. PROlocalizer can also efficiently predict the subcellular localization for relatively large protein datasets. Furthermore, effects of one or more missense mutations per protein can be predicted by PROlocalizer just by providing the wild type sequence and the original and mutated residues along with position information.

The length of a single protein sequence cannot be longer than 4,000 amino acids, because of the limitations in some of the programs. As many of the classifiers are run outside our system and can take some time, the submitter is provided with search ID. In case any of the servers running outside our system is down the analysis cannot be performed. The user should then return to the service to rerun the analysis. An e-mail is sent to the submitter once the analysis is finished. The e-mail contains the results and a link to a web page where the same data are available with a number of links. The web page will be available for a limited period of time. A pdf format report lists the details of the query and analysis, used programs and their versions and predicted localization(s) (Fig. 2).

## Results and discussion

Localization is a very important property for proteins, however, sometimes difficult to experimentally determine.



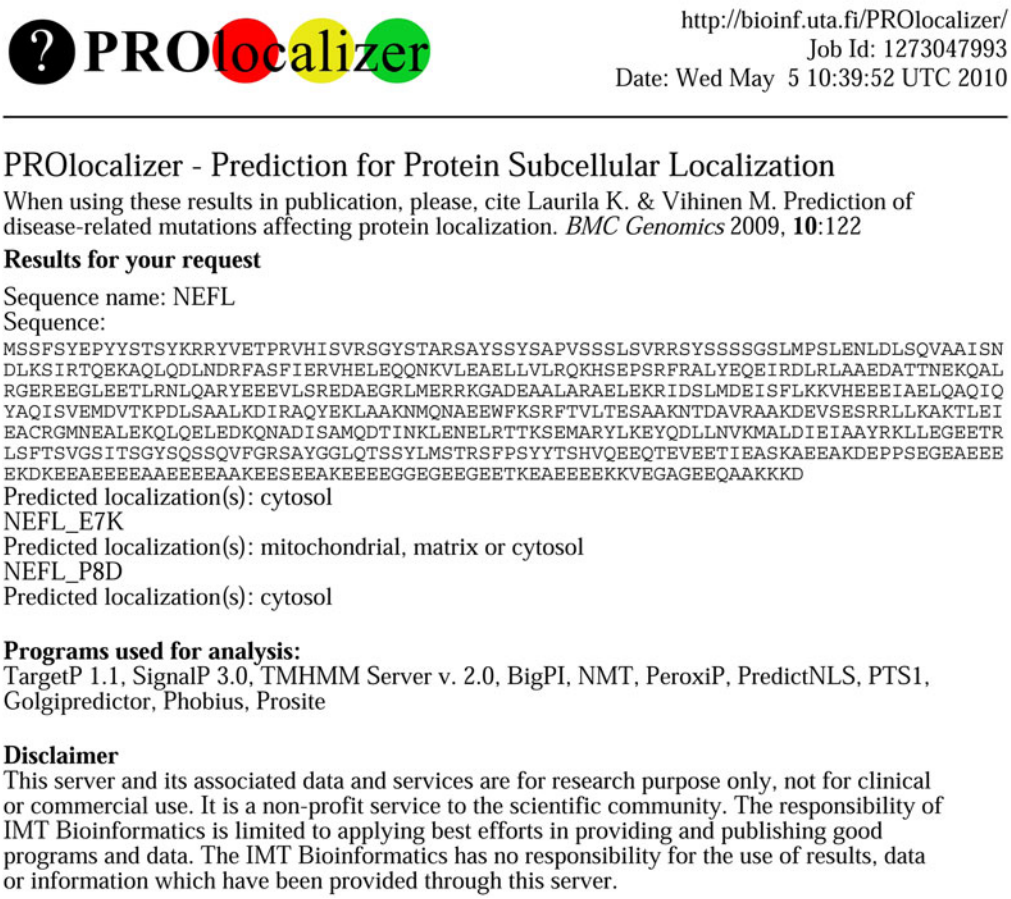




**Fig. 2** PROlocalizer home page and example of a prediction. *Top* PROlocalizer home page, *Bottom* example of the output for a prediction. Only part of the report is shown

A protein may have several localizations depending on the state of the cell and tissue. As novel protein sequences are identified at increasing pace it is beneficial to be able to predict properties of these molecules. PROlocalizer is based on the state-of-the-art binary predictors that have been combined to work as the Scandinavian protocol. We have automated the use, submission and interpretation of results in this protocol.

The performance of the Scandinavian protocol and the individual predictors that it is using has been previously evaluated with a dataset containing more than 1,500 proteins (Laurila and Vihinen 2009) and therefore we are not discussing the performance issue in detail here. The accuracy for different compartments varied from 0.55 to 0.97 being on average of 0.84. Additionally, the authors of the Scandinavian protocol discuss the performance in their article (Emanuelsson et al. 2007).

We have previously indicated that localization affecting mutations are likely involved in a number of disorders (Laurila and Vihinen. 2009). Due to the lack of sufficient number of known cases we were not able to provide statistical analysis for the performance when predicting mutation localization effects. As our previous study indicates, certain disease-causing mutations likely affect mutation signals just like other functionally and structurally crucial sites. Thus, analysis of variation effects may require localization predictions (Thusberg and Vihinen. 2009). PROlocalizer can also be used to analyze the mutation effects on protein subcellular localization. Compared with many other protein localization prediction services, the Scandinavian procedure behind the PROlocalizer does not utilize sequence homology searches or protein amino acid composition, instead the predictions are based on the identification of protein sorting signals. Thus, PROlocalizer is more likely to detect the point mutation effects on localization than several other prediction methods. Fully automated PROlocalizer has a user-friendly interface and it generates a detailed report of predictions. The service is freely available for academic, non-commercial use.

## Conclusions

PROlocalizer is a tool for prediction of altogether 12 protein subcellular localizations. It is implemented in a user-friendly, automated service. The user can submit at a time, a number of proteins and/or variations, if necessary. The service automatically interprets the result(s) based on the complex prediction scheme. PROlocalizer provides a detailed report in several formats, which also include in addition to the prediction results details for the used methods. In the future, we plan to extend the system to allow predictions for plants, fungi and bacteria.

**Acknowledgments** This work was supported by the Biocenter Finland, Finnish Academy (application number 213462, Finnish Programme for Centres of Excellence in Research 2006–2011), the Medical Research Fund of Tampere University Hospital, Sigrid Juselius Foundation, Tampere Graduate School in Information Science and Engineering (TISE), and Otto A. Malm Foundation.



## References

Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–795

Chou KC (2002) Prediction of protein signal sequences. Curr Protein Pept Sci 3:615–622

Chou KC, Shen HB (2009) A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. PLos One 5:e9931

Cokol M, Nair R, Rost B (2000) Finding nuclear localization signals. EMBO Rep 1:411–415

Davis TN (2004) Protein localization in proteomics. Curr Opin Chem Biol 8:49–53

de Castro E, Sigrist CJ, Gattiker A, Builliard V, Langendjik-Genevaux PS, Gasteiger E, Bairoch A, Hulo N (2006) ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res 34:W362–W365

Dönnes P, Höglund A (2004) Predicting protein subcellular localization: past, present, and future. Genomics Proteomics Bioinformatics 2:209–215

Eisenhaber B, Bork P, Eisenhaber F (1998) Sequence properties of GPI-anchored proteins near the omega-site: constraints for the polypeptide binding site of the putative transamidase. Protein Eng 11:1155–1161

Eisenhaber B, Bork P, Eisenhaber F (1999) Prediction of potential GPI-modification sites in proprotein sequences. J Mol Biol 292:741–758

Eisenhaber B, Bork P, Yuan Y, Löffler G, Eisenhaber F (2000) Automated annotation of GPI anchor sites: case study *C. elegans*. Trends Biochem Sci 25:340–341

Emanuelsson O (2002) Predicting protein subcellular localization from amino acid sequence information. Brief Bioinform 3:361–376

Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300:1005–1016

Emanuelsson O, Elofsson A, von Heijne G, Cristobal S (2003) In silico prediction of the peroxisomal proteome in fungi, plants and animals. J Mol Biol 330:443–456

Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2:953–971

Falk R, Ramstöm M, Ståhl S, Hober S (2007) Approaches for systematic proteome exploration. Biomol Eng 24:155–168

Garg A, Bhasin M, Raghava GP (2005) Support vector machine-based method for subcellular localization of human proteins

using amino acid compositions, their order, and similarity search. J Biol Chem 280:14427–14432

Guda C, Subramaniam S (2005) pTARGET: A new method for predicting protein sub-cellular localization in eucaryotes. Bioinformatics 21:3963–3969

Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K (2007) WoLF PSORT: protein localization predictor. Nucleic Acids Res 35:W585–W587

Käll L, Krogh A, Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338:1027–1036

Klee EW, Ellis LB (2005) Evaluating eukaryotic secreted protein prediction. BMC Bioinformatics 6:256

Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580

Laurila K, Vihinen M (2009) Prediction of disease-related mutations affecting protein localization. BMC Genomics 10:122

Lu Z, Szafron D, Greiner R, Lu P, Wishart DS, Poulin B, Anvik J, Macdonell C, Eisner R (2004) Predicting subcellular localization of proteins using machine-learned classifiers. Bioinformatics 20:547–556

Martoglio B, Dobberstein B (1998) Signal sequences: more than greasy peptides. Trends Cell Biol 8:410–415

Nair R, Rost B (2005) Mimicking cellular sorting improves prediction of subcellular localization. J Mol Biol 348:85–100

Nakai K, Horton P (1999) PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization. Trends Biochem Sci 24:34–35

Neuberger G, Maurer-Stroh S, Eisenhaber B, Hartig A, Eisenhaber F (2003a) Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. J Mol Biol 328:581–592

Neuberger G, Maurer-Stroh S, Eisenhaber B, Hartig A, Eisenhaber F (2003b) Motif refinement of the peroxisomal targeting signal 1 and evaluation of taxon-specific differences. J Mol Biol 328:567–579

Schneider GH, Fechner U (2004) Advances in the prediction of protein targeting signals. Proteomics 4:1571–1580

Shen HB, Chou KC (2007) Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem Biophys Res Commun 355:1006–1011

Sonnhammer EL, von Heijne G, Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. In: Proceedings of the 6th international conference on intelligent systems for molecular biology (ISMB-98), vol 6, 28 June–1 July, Montréal, pp 175–182

Sprenger J, Fink JL, Teasdale RD (2006) Evaluation and comparison of mammalian subcellular localization prediction methods. BMC Bioinformatics 7(Suppl 5):S3

Stroud RM, Walter P (1999) Signal sequence recognition and protein targeting. Curr Opin Struct Biol 9:754–759

Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, Kuznetsov EN (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng 12:387–394

Thusberg J, Vihinen M (2009) Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Hum Mutat 30:703–714

Yu CS, Chen YC, Lu CH, Hwang JK (2006) Prediction of protein subcellular localization. Proteins:Struct, Function Bioinformatics 64:643–651

Yuan Z, Teasdale RD (2002) Prediction of Golgi Type II membrane proteins based on their transmembrane domains. Bioinformatics 18:1109–1115