Downloaded from orbit.dtu.dk on: Aug 18, 2022

DTU Library

Signal peptides and protein localization prediction

Nielsen, Henrik

Published in:
Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics

Publication date:
2005

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Nielsen, H. (2005). Signal peptides and protein localization prediction. In *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics* John Wiley and Sons Ltd.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Signal peptides and protein localization prediction

Henrik Nielsen
Center for Biological Sequence Analysis, The Technical University of Denmark, Lyngby, Denmark

1. Introduction

In 1999, the Nobel prize in Physiology or Medicine was awarded to Günther Blobel "for the discovery that proteins have intrinsic signals that govern their transport and localization in the cell". Since the subcellular localization of a protein is an important clue to its function, the characterization and prediction of these intrinsic signals – the "zip codes" of proteins – has become a major task in bioinformatics.

Here, I will review the most important methods for the prediction of subcellular localization, also known as protein sorting. Owing to the limited space, this review is far from complete; especially, applications that are not publicly available on-line are ignored. Generally, there are two approaches to protein localization prediction: signal detection, that is, prediction of the sorting signals themselves, and prediction based on global properties (amino acid composition and/or physicochemical variables) that are characteristic of different subcellular compartments.

2. Secretory signal peptides

The best known "zip code" is the secretory signal peptide, which targets a protein for translocation across the plasma membrane in prokaryotes and across the endoplasmic reticulum (ER) membrane in eukaryotes (von Heijne, 1990). It is an N-terminal peptide, typically 15–30 amino acids long, which is cleaved off during translocation of the protein across the membrane. There is no simple consensus sequence for signal peptides, but they typically show three distinct compositional zones: an N-terminal region that often contains positively charged residues, a hydrophobic region of at least six residues, and a C-terminal region of polar uncharged residues with some conservation at the −3 and −1 positions relative to the cleavage site.

In a very early bioinformatics application, von Heijne (1986) developed a weight matrix for recognition of the signal peptide cleavage site. This weight matrix has found an extremely wide usage. It does not exist as a WWW-server, but it is included in PSORT (see below).

A newer method is SignalP (**http://www.cbs.dtu.dk/services/SignalP**), which is based on neural networks (NNs) and hidden Markov models (HMMs) (Bendtsen *et al*., 2004). See Article 98, **Hidden Markov models and neural networks**, Volume 8 for an introduction to these machine-learning methods. A comparison of signal peptide prediction methods showed that both NNs and HMMs outperform the weight matrix (Menne *et al*., 2000). This still seems to be the case, even though a newer application using weight matrices has become available (Hiller *et al*., 2004) (**http://www.predisi.de/**).

The HMM included in SignalP is a complex architecture that does not adhere to the well-known profile HMMs. However, Zhang and Wood (2003) showed that the task can be done with an only slightly lower performance using a profile HMM implemented in the standard HMMER package (the model can be downloaded from **http://share.gene.com/**).

It should be noted that far from all proteins with secretory signal peptides are actually secreted to the outside of the cell. In gram-negative bacteria, they by default end up in the periplasmic compartment, and a separate mechanism is needed to secrete them to the growth medium (Pugsley *et al*., 1997). In eukaryotes, proteins translocated across the ER membrane are by default transported through the golgi apparatus and exported by secretory vesicles, but some proteins have specific retention signals that hold them back in the ER, the golgi or the lysosomes. In general, these retention signals are poorly characterized, one exception being the ER retention signal that has the consensus sequence KDEL or HDEL (van Vliet *et al*., 2003).

Some transmembrane proteins also have a cleavable secretory signal peptide that initiates translocation, whereafter the translocation is halted by a transmembrane α-helix that acts as a stop-transfer signal, leaving the protein integrated in the membrane. For a comparison of various publicly available methods for predicting transmembrane helices, see Chen *et al*. (2002) (*see* Article 38, **Transmembrane topology prediction**, Volume 7 and Article 65, **Analysis and prediction of membrane protein structure**, Volume 7).

Transmembrane helices often lead to false-positives in signal peptide prediction and vice versa. Recently, a combined HMM that deals with this problem by modeling both these signals, Phobius, (**http://phobius.cgb.ki.se**) has become available (Käll *et al*., 2004).

Other membrane proteins do not have transmembrane domains, but are linked to the membrane by a covalently attached lipid group. Prokaryotic lipoproteins have signal peptides that are cleaved by a special signal peptidase, and their cleavage site has a characteristic consensus signal with a 100% conserved cysteine in position +1. Two publicly available signal peptide prediction methods are designed to recognize prokaryotic lipoprotein signal peptides: LipoP, (**http://www.cbs.dtu.dk/services/LipoP/**), which is based on a combination of NNs and HMMs (Juncker *et al*., 2003); and SPEPlip, (**http://gpcr.biocomp.unibo.it/predictors/**), which is based on NNs combined with a simple pattern matching (Fariselli *et al*., 2003). In eukaryotes, some proteins are linked to the membrane by a glycosylphosphatidylinositol (GPI) anchor at the C-terminus or a myristoyl anchor at the N-terminus. These can be predicted with the big-Π and NMT tools (**http://mendel.imp.univie.ac.at/mendeljsp/sat/index.jsp**) (Eisenhaber *et al*.,

2003), and the myristoylation also with Myristoylator (**http://www.expasy.org/tools/myristoylator/**) (Bologna *et al*., 2004).

3. Other localization signals

The target peptides of chloroplasts and mitochondria are also N-terminal cleavable peptides (Schatz and Dobberstein, 1996). They are less well characterized than the secretory signal peptide, but they are both rare in negatively charged residues and able to form amphiphilic α-helices (Bannai *et al*., 2002; Bruce, 2000).

A widely used method to predict mitochondrial transit peptides (mTPs) is Mitoprot (**http://websvr.mips.biochem.mpg.de/cgi-bin/proj/medgen/mitofilter**) (Claros and Vincens, 1996). It is a feature-based method, using a linear combination of a number of sequence characteristics such as amino acid abundance, maximum hydrophobicity, and maximum hydrophobic moment (α-helix amphiphilicity) that are combined into an overall score.

A newer method, MITOPRED (**http://mitopred.sdsc.edu**), does not rely on mitochondrial targeting signals, but is based on Pfam domain occurrence patterns and the amino acid compositional differences between mitochondrial and non-mitochondrial proteins (Guda *et al*., 2004).

For chloroplast transit peptides, there is a NN-based method available, ChloroP (**http://www.cbs.dtu.dk/services/ChloroP/**) (Emanuelsson *et al*., 1999). A successor of ChloroP is TargetP (**http://www.cbs.dtu.dk/services/TargetP/**), which provides prediction of both chloroplast transit peptides, mitochondrial transit peptides, and secretory signal peptides (Emanuelsson *et al*., 2000). Both ChloroP and TargetP use a combination of NNs to calculate a transit or signal peptide score, and a weight matrix to locate the transit peptide cleavage sites.

Another NN-based method is Predotar (**http://genoplante-info.infobiogen.fr/predotar/**) (Small *et al*., 2004). In contrast to TargetP that uses moving windows to calculate the transit peptide score, Predotar uses a fixed window comprising the first 40–60 amino acids of the sorting signal. Like TargetP, it predicts mitochondrial, chloroplast, and secretory signals.

Bannai *et al*. (2002) tested a large number of physicochemical features of N-terminal parts of proteins with signal or transit peptides and obtained a combination of simple rules that yielded a discriminative performance fairly close to that of TargetP. Interestingly, a simple hydrophobicity scale even outperformed the NN-based TargetP on plant signal peptides. The resulting method is called iPSORT (**http://hypothesiscreator.net/iPSORT/**).

Not all localization signals are N-terminal and cleavable. Nuclear localization signals can occur internally in the sequence and are not cleaved. The method PredictNLS (**http://cubic.bioc.columbia.edu/predictNLS/**) (Cokol *et al*., 2000) predicts nuclear localization by comparing the query sequence to a database of experimentally verified NLS sequences and derived signals.

Peroxisomes also have their own protein import machinery. Two uncleaved signals are known: the C-terminal PTS1 and the N-terminal PTS2. PTS1 with the consensus sequence -SKL is best known, and a predictor is available (**http://mendel.imp.univie.ac.at/PTS1/**) (Neuberger *et al*., 2003).

4. Global property methods

In addition to the recognition of the sorting signals, prediction of protein sorting can exploit the fact that proteins of different subcellular compartments differ in global properties, reflected in the amino acid composition. Andrade *et al*. (1998) found that the signal in the total amino acid composition, which makes it possible to identify the subcellular location, is due almost entirely to surface residues. While the signal-prediction methods are probably closer to mimicking the information processing in the cell, methods based on global properties can work also for genomic or EST sequences where the N-terminus of the protein has not been correctly predicted. One drawback is that such methods will not be able to distinguish between very closely related proteins or isoforms that differ in the presence or absence of a sorting signal.

The NNPSL method (**http://predict.sanger.ac.uk/nnpsl/**) (Reinhardt and Hubbard, 1998) uses NNs trained on overall amino acid composition to predict localization. The method distinguishes between three bacterial compartments (cytoplasmic, periplasmic, and extracellular) and four eukaryotic compartments (cytoplasmic, extracellular, mitochondrial, and nuclear). Interestingly, plant proteins were found to be very poorly predicted, and are not included in the present method.

Nair and Rost (2003), also working with NNs, found that prediction could be improved by using information from protein structure. Specifically, they calculated amino acid composition separately for three categories of secondary structure (helix, sheet, and coil) and for surface-accessible residues. Naturally, the improvement was most pronounced when applied to proteins of known structure, but even a predicted secondary structure (according to an NN) was able to enhance prediction. The resulting method is implemented in a database, LOC3D (**http://cubic.bioc.columbia.edu/db/LOC3d/**), and a web server, LOCtarget (**http://cubic.bioc.columbia.edu/services/LOCtarget/**) (Nair and Rost, 2004).

There is a rapidly growing number of subcellular localization prediction methods based on amino acid composition and related features. The SubLoc method (**http://www.bioinfo.tsinghua.edu.cn/SubLoc/**) (Hua and Sun, 2001) is based on support vector machines (SVMs, *see* Article 110, **Support vector machine software**, Volume 8). The data set used to train SubLoc is that of Reinhardt and Hubbard (1998), but the predictive performance is significantly better than the NN. Three newer SVM applications are Esub8, PLOC, and ESLpred. Esub8 (**http://bioinfo.tsinghua.edu.cn/CoupleLoc/eu8.html**) (Cui *et al*., 2004) uses the amino acid composition of the first and last half of each sequence and distinguishes between eight subcellular locations. PLOC (**http://www.genome.ad.jp/SIT/ploc.html**) (Park and Kanehisa, 2003) uses, in addition to amino acid composition, amino acid pairs (adjacent or separated by one to three positions) to enhance prediction. It distinguishes between as many as twelve subcellular locations. ESLpred (**http://www.imtech.res.in/raghava/eslpred/**) (Bhasin and Raghava, 2004) is a four-location predictor based on the NNPSL data, which uses both amino acid composition, adjacent amino acid pairs, PSI-BLAST output, and physicochemical properties.

It should be stressed that no prediction method is better than the data set used to train it. One problem that is rarely properly addressed in global property methods is homology in the data. If the data used to test a method has sequences

that are significantly homologous to sequences in the training data, the apparent performance of the method is an overestimate. To compute a true generalizable performance, the data set should be reduced so that no homologous pairs remain (Hobohm *et al*., 1992). Reinhardt and Hubbard (1998) reduced their data set, but only removed sequences with more than 90% identity, which is clearly much higher than the homology threshold. Newer methods, with the exception of PLOC, are trained with the Reinhardt and Hubbard (1998) data or without homology reduction at all. Therefore, care should be taken when comparing performance measures.

5. Integrated methods

PSORT (**http://psort.nibb.ac.jp** or **http://www.psort.org/**) (Nakai and Kanehisa, 1991, 1992; Nakai and Horton, 1999) is an integrated system of several prediction methods, using both sorting signals and global properties. Some of the components are developed within the PSORT group, others are implementations of methods published elsewhere, including selected PROSITE patterns. PSORT is the only publicly available system that shows this degree of integration. In addition to localization (up to 16 different possible locations in plant cells), it also predicts motifs for posttranslational modifications such as lipid attachment.

All the constituent predictors provide feature values, which are then integrated to produce a final prediction. In the original version, PSORT I, the integration was done in the style of a conventional knowledge base using a collection of "if-then" rules, while the newer PSORT II version uses quantitative machine-learning techniques, such as probabilistic decision trees and the k nearest neighbors classifier to integrate scores from all the features. PSORT II is available for animal and yeast proteins (11 locations), while plant proteins still have to rely on PSORT I.

For gram-negative bacteria, there is a recently improved version named PSORT-B (**http://www.psort.org/psortb/index.html**) (Gardy *et al*., 2003) discriminating between five possible locations (cytoplasm, inner membrane, periplasm, outer membrane, and extracellular). It uses a combination of homology searches to proteins of known localization, PROSITE motifs, signal peptide, and transmembrane helix predictors based on HMMs, and a SVM-based predictor using amino acid composition.

Drawid and Gerstein (2000) developed a different integrated system for localizing all the proteins in the yeast genome to one of the five possible compartments (cytoplasm, nucleus, mitochondria, membrane, or secretory pathway). It is a Bayesian system integrating 30 features comprising both specific motifs (e.g., signal sequences or the HDEL motif), overall properties of a sequence (e.g., surface composition or isoelectric point), and whole-genome data (e.g., absolute mRNA expression levels or their fluctuations). The method is not available for submission of new sequences, but predictions for all known yeast genes can be retrieved (**http://bioinfo.mbb.yale.edu/genome/localize/**).

Acknowledgments

I thank Gunnar von Heijne and Jacob Engelbrecht for comments on the manuscript.

References

Andrade MA, O'Donoghue SI and Rost B (1998) Adaptation of protein surfaces to subcellular location. *Journal of Molecular Biology*, **276**, 517–528.

Bannai H, Tamada Y, Maruyama O, Nakai K and Miyano S (2002) Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, **18**, 298–305.

Bendtsen JD, Nielsen H, von Heijne G and Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology*, **340**, 783–795.

Bhasin M and Raghava GPS (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Research*, **32**, W414–W419.

Bologna G, Yvon C, Duvaud S and Veuthey A-L (2004) N-Terminal myristoylation predictions by ensembles of neural networks. *Proteomics*, **4**, 1626–1632.

Bruce BD (2000) Chloroplast transit peptides: structure, function and evolution. *Trends in Cell Biology*, **10**, 440–447.

Chen CP, Kernytsky A and Rost B (2002) Transmembrane helix predictions revisited. *Protein Science*, **11**, 2774–2791.

Claros MG and Vincens P (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. *European Journal of Biochemistry*, **241**, 779–786.

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

Cui Q, Jiang T, Liu B and Ma S (2004) Esub8: a novel tool to predict protein subcellular localizations in eukaryotic organisms. *BMC Bioinformatics*, **5**, 66.

Drawid A and Gerstein M (2000) A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *Journal of Molecular Biology*, **301**, 1059–1075.

Eisenhaber F, Eisenhaber B, Kubina W, Maurer-Stroh S, Neuberger G, Schneider G and Wildpaner M (2003) Prediction of lipid posttranslational modifications and localization signals from protein sequences: big-Π, NMT and PTS1. *Nucleic Acids Research*, **31**, 3631–3634.

Emanuelsson O, Nielsen H, Brunak S and von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology*, **300**, 1005–1016.

Emanuelsson O, Nielsen H and von Heijne G (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Science*, **8**, 978–984.

Fariselli P, Finocchiaro G and Casadio R (2003) SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. *Bioinformatics*, **19**, 2498–2499.

Gardy JL, Spencer C, Wang K, Ester M, Tusnády GE, Simon I, Hua S, deFays K, Lambert C, Nakai K, *et al*. (2003) PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Research*, **31**, 3613–3617.

Guda C, Fahy E and Subramaniam S (2004) MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics*, **20**, 1785–1794.

Hiller K, Grote A, Scheer M, Munch R and Jahn D (2004) PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Research*, **32**, W375–W379.

Hobohm U, Scharf M, Schneider R and Sander C (1992) Selection of representative protein data sets. *Protein Science*, **1**, 409–417.

Hua S and Sun Z (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.

Juncker AS, Willenbrock H, von Heijne G, Brunak S, Nielsen H and Krogh A (2003) Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Science*, **12**, 1652–1662.

Käll L, Krogh A and Sonnhammer ELL (2004) A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology*, **338**, 1027–1036.

Menne KML, Hermjakob H and Apweiler R (2000) A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics*, **16**, 741–742.

Nair R and Rost B (2003) Better prediction of sub-cellular localization by combining evolutionary and structural information. *Proteins*, **53**, 917–930.

Nair R and Rost B (2004) LOCnet and LOCtarget: sub-cellular localization for structural genomics targets. *Nucleic Acids Research*, **32**, W517–W521.

Nakai K and Horton P (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends in Biochemical Sciences*, **24**, 34–35.

Nakai K and Kanehisa M (1991) Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins*, **11**, 95–110.

Nakai K and Kanehisa M (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, **14**, 897–911.

Neuberger G, Maurer-Stroh S, Eisenhaber B, Hartig A and Eisenhaber F (2003) Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. *Journal of Molecular Biology*, **328**, 581–592.

Park K-J and Kanehisa M (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, **19**, 1656–1663.

Pugsley AP, Francetic O, Possot OM, Sauvonnet N and Hardie KR (1997) Recent progress and future directions in studies of the main terminal branch of the general secretory pathway in Gram-negative bacteria – a review. *Gene*, **192**, 13–19.

Reinhardt A and Hubbard T (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Research*, **26**, 2230–2236.

Schatz G and Dobberstein B (1996) Common principles of protein translocation across membranes. *Science*, **271**, 1519–1526.

Small I, Peeters N, Legeai F and Lurin C (2004) Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, **4**, 1581–1590.

van Vliet C, Thomas EC, Merino-Trigo A, Teasdale RD and Gleeson PA (2003) Intracellular sorting and transport of proteins. *Progress in Biophysics and Molecular Biology*, **83**, 1–45.

von Heijne G (1986) A new method for predicting signal sequence cleavage sites. *Nucleic Acids Research*, **14**, 4683–4690.

von Heijne G (1990) The signal peptide. *The Journal of Membrane Biology*, **115**, 195–201.

Zhang Z and Wood WI (2003) A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics*, **19**, 307–308.