ELSEVIER

Biochimica et Biophysica Acta 1541 (2001) 114–119

BIOCHIMICA ET BIOPHYSICA ACTA
BBA

www.bba-direct.com

Review

Prediction of organellar targeting signals

Olof Emanuelsson, Gunnar von Heijne *

Stockholm Bioinformatics Center, Stockholm University, S-10691 Stockholm, Sweden

Received 25 July 2001; accepted 13 August 2001

Abstract

The subcellular location of a protein is an important characteristic with functional implications, and hence the problem of predicting subcellular localization from the amino acid sequence has received a fair amount of attention from the bioinformatics community. This review attempts to summarize the present state of the art in the field.

Keywords: Chloroplast; Mitochondria; Presequence; Protein import; Sorting

1. Introduction

The general problem to predict the subcellular location of a protein from its amino acid sequence has long been a central one in bioinformatics. To date, three conceptually different approaches have been proposed: to look for the targeting signals that the cell uses as 'address labels', to base the prediction on the observation that proteins from different cellular compartments tend to differ in subtle ways in their overall amino acid composition, and to use evolutionary relationships (based on the endosymbiotic origin of organelles) to infer the subcellular localization. There are even one or two 'meta-methods' in which outputs from a range of 'primary' prediction/analysis methods are combined in an optimal way. Each approach has its strengths and weaknesses, and since no across-the-board benchmarking tests have been performed, it is not yet possible to make a fair comparison between all the different methods proposed by different authors.

In this review, we have chosen to first discuss the most commonly used methods for predicting individual subcellular localizations – the secretory pathway, mitochondria, chloroplasts, and the nucleus – and then describe a couple of attempts to construct integrated predictors that try to 'sort' proteins between multiple compartments. The reader is also referred to a recent (and somewhat more ambitious) review by Nakai [1] for further details.

2. Prediction of signal peptides for secretion

N-Terminal signal peptides target proteins to the secretory pathway in eukaryotic cells, and for translocation across the cytoplasmic membrane in bacteria. It has long been known that they have a tripartite design with a short positively charged amino-terminal segment (n-region), a central hydrophobic segment (c-region), and a more polar C-terminal seg-

* Corresponding author. Fax: +46-8-153679.
E-mail address: gunnar@dbb.su.se (G. von Heijne).

0167-4889/01/$ – see front matter
PII: S0167-4889(01)00145-8

ment that is recognized by the signal peptidase enzyme. The first methods to identify signal peptide sequences were published already in the mid 1980s [2–4], but the currently most widely used method is the neural network-based SignalP predictor [5]. SignalP combines two different neural networks: one that is trained to discriminate between residues that belong and do not belong to a signal peptide (the S-score), and one that is trained only to recognize signal peptidase cleavage sites (the C-score). The cleavage site is predicted by multiplying together the C-score and the negative 'derivative' of the S-score (this serves to focus the prediction on the region where the S-score changes from high to low), while the discrimination between proteins that have and do not have a signal peptide is based on the mean S-score evaluated from the N-terminus to the predicted cleavage site. The current version of SignalP was trained on three different signal peptide data sets – one with eukaryotic signal peptides, one with signal peptides from Gram-negative bacteria, and one from Gram-positive bacteria – and hence is to some extent optimized for different organisms.

SignalP-HMM is a new version of SignalP that is based on a hidden Markov model formalism [6]. This predictor was developed in order to improve the discrimination between signal peptides and N-terminal transmembrane anchor segments, but is in other respects comparable to the original SignalP predictor.

According to a recent benchmarking study [7], SignalP and SignalP-HMM perform equally well when it comes to discriminating between proteins with and without signal peptides, although the neural network version seems to be slightly better in predicting signal peptidase cleavage sites, Table 1. SignalP-HMM is however clearly superior for discriminating between cleavable signal peptides and N-terminal anchors. The two SignalP versions clearly outperformed the other programs tested, and thus seem to be the best signal peptide predictors available at the moment.

It should be mentioned that Chou recently reported a method similar in spirit to a weight-matrix method but including statistics on pairwise correlations between the positions closest to the signal peptidase cleavage site [8]; however, this method was not included in the benchmarking study.

Table 1
Performance of the two versions of SignalP: hidden Markov model version (HMM) and neural network version (NN)

SingalP version	Cleavage site location, % correct			Discrimination, MCC			
	Euk	G−	G+	SP/non-SP			SP/SA
				Euk	G−	G+	Euk
HMM	69.5	81.4	64.5	0.94	0.93	0.96	0.74
NN	72.4	83.4	67.5	0.97	0.89	0.96	0.39

The cleavage site location is measured in the percentage correctly assigned cleavage sites. The discrimination is measured in Mathews' correlation coefficient (MCC) which is one (1) for a perfect prediction and zero (0) for a totally random assignment [33]. The discrimination is given both between signal peptide-containing (SP) and signal peptide-lacking (non-SP) proteins and between secreted proteins (SP) and proteins anchored in the membrane (SA). The table is adapted from [6].

3. Prediction of mitochondrial targeting peptides

Mitochondrial targeting peptides are enriched in positively charged residues (Arg in particular), lack negatively charged residues, and have the ability to form amphiphilic α-helices [9]. The amphiphilic structure is important for binding to receptors in the outer mitochondrial membrane [10,11], and the net positive charge may be needed during the $\Delta\Psi$-driven import across the inner mitochondrial membrane [12].

Three popular methods for predicting mitochondrial targeting peptides are TargetP [13], MitoProt [14], and Predotar (see Table 3). Both Predotar and TargetP are neural network predictors and are conceptually similar to SignalP. They are not clear-cut single location predictors since they also deal with other presequences; both investigates chloroplast transit peptide presence and in addition to this TargetP handles signal peptide prediction. Predotar is essentially aimed for plant sequences. The performance of Predotar and TargetP is discussed in Section 6 and summarized in Table 2.

MitoProt predicts localization of a protein by calculating a number of physicochemical parameters from its amino acid sequence, and then computing a linear discriminant function (LDF) which is compared to a cutoff for mitochondrial/non-mitochondrial localization prediction. Both MitoProt and Tar-

Table 2
Comparison of localization prediction for three multi-category predictors

Predictor	Location	Plant set			Non-plant set		
		% Correct	Sensitivity	Specificity	% Correct	Sensitivity	Specificity
TargetP	Chloro.	85.3	0.85	0.69	90.0	–	–
	Mito.		0.82	0.90		0.80	0.67
	Secr.		0.91	0.95		0.96	0.92
PSORT	Chloro.	69.8	0.47	0.69	82.5	–	–
	Mito.		0.66	0.87		0.81	0.60
	Secr.		0.82	0.74		0.64	0.93
Predotar	Chloro.	84.8	0.82	0.77	76.3	–	–
	Mito.		0.86	0.87		0.86	0.50
	Secr.		(0.80)	n/a		(0.65)	n/a

The plant data set contains 940 proteins (from Swiss-Prot release 36) and the non-plant set contains 2738 proteins (from Swiss-Prot release 37) with annotated localization. Note that Predotar is not intended for use on non-plant set, hence its partly poor performance on this set. Sensitivity is the fraction of true positive predictions relative to the set of proteins known to be localized in respective compartment. Specificity is the fraction of true positive predictions relative to the set of proteins predicted to respective compartment. 'Percent correct' refers to the fraction of all proteins in a set for which the correct location is predicted.

getP suggest a potential cleavage site of the predicted mitochondrial targeting peptides.

Predotar, TargetP and MitoProt only predict N-terminal mitochondrial targeting sequences, and no method exists that will identify import signals present elsewhere in the protein, although such signals are known to exist [15,16].

4. Prediction of chloroplast transit peptides

N-Terminal chloroplast transit peptides have highly variable lengths, contain very few negatively charged residues, and are highly enriched for hydroxylated amino acids. Two neural-network based predictors are available: ChloroP [17] and Predotar (see Table 3). ChloroP also includes a separate module (based on a weight matrix) for predicting the transit peptide cleavage site. A comparison of localization prediction performance between Predotar and TargetP (of which ChloroP is a part) using 940 plant sequences from Swiss-Prot can be found in Table 2.

Many thylakoid proteins have composite targeting signals with a typical transit peptide followed by a thylakoid targeting signal. The latter is usually very similar to the signal peptides found on secretory proteins, and can be identified by SignalP or SignalP-HMM (our unpublished data). A specialized weight matrix for predicting the cleavage site is available for thylakoid signal peptides [18].

5. Prediction of nuclear localization signals

Nuclear localization signals are composed of one (monopartite) or a pair of (bipartite) short positively

Table 3
Web addresses of predictors

Predictor	Web address (URL)
ChloroP	http://www.cbs.dtu.dk/services/ChloroP/
MitoProt	http://www.mips.biochem.mpg.de/cgi-bin/proj/medgen/mitofilter
predictNLS	http://maple.bioc.columbia.edu/predictNLS/
Predotar	http://www.inra.fr/Internet/Produits/Predotar/
PSORT	http://psort.nibb.ac.jp/
SignalP	http://www.cbs.dtu.dk/services/SignalP/
TargetP	http://www.cbs.dtu.dk/services/TargetP/
TMHMM	http://www.cbs.dtu.dk/services/TMHMM/

charged stretches in the protein chain. The monopartite nuclear localization signal has a short consensus sequence, K(K/R)X(K/R), and binds to a pocket on the surface of the importin α receptor [19]. In bipartite nuclear localization signals, the monopartite motif is combined with a second small cluster of basic residues, 10–12 residues N-terminal to the first.

The basic clusters can be found anywhere within the protein chain, and are exposed on the surface of the folded protein. Since the entire chain has to be searched for nuclear localization signals, it is difficult to avoid false positive predictions. The best predictor available at the moment is based on a large collection of mono- and bipartite motifs [20]. It is capable of finding 43% of the known nuclear proteins with no false positive predictions on the set of Swiss-Prot entries (release 38) with unambiguously annotated localization. This is achieved through collection of known NLSs and their homologues, and applying an 'in silico mutagenesis' to extend the motifs as far as possible without matching any non-nuclear proteins.

6. Integrated methods for predicting subcellular localization

In these days of whole-genome sequencing, what is obviously needed are integrated prediction methods that somehow represent the entire protein sorting potential of the cell and assign the most likely subcellular localization to a protein based on its amino acid sequence. This also includes sorting within an organelle or a pathway: between, e.g., the mitochondrial outer membrane, intermembrane space, inner membrane, and matrix, or between the different compartments along the secretory pathway. In eukaryotic cells, the number of distinct compartments is thus very large.

The pioneering work in this area is due to Nakai and Kanehisa [21,22]. His PSORT program now distinguishes between 17 different subcellular localizations (10 for a newer, retrained version called PSORT 2 that uses a slightly different decision algorithm), and integrates a number of pre-existing prediction programs as well as calculated characteristics such as overall amino acid composition within a unified framework [23,24]. Drawid and Gerstein [25] have recently presented a system that is similar in spirit to PSORT but uses a different formalism (Bayesian statistics) for integrating multiple kinds of information (everything from SignalP predictions to microarray expression profiles). The method was applied to the full *Saccharomyces cerevisiae* proteome, and thus provides estimates of the fraction of all yeast proteins found in different compartments. A predictor based only on overall amino acid composition and pairwise residue correlations has been developed by Chou [26].

The TargetP predictor [13] has a more limited scope than PSORT, and only differentiates between secretory proteins, mitochondrial proteins, chloroplast proteins, and everything else. The method looks for N-terminal sorting signals by feeding the outputs from SignalP, ChloroP, and an analogous mitochondrial predictor (not available as a standalone predictor) into a 'decision neural network' that makes the final choice between the different compartments. Although not yet integrated into TargetP, membrane proteins can be predicted with high reliability by programs such as TMHMM [27,28]. TargetP predicts signal peptides with high sensitivity and specificity but performs less well on mitochondrial targeting peptides and chloroplast transit peptides, Table 2. Modules for predicting cleavage sites in the different targeting signals are also included in TargetP; again, performance is much better on the signal peptides than on the other two classes of peptides.

Predotar is primarily aimed at predicting the chloroplast/mitochondrion sorting problem (thus dealing with plant sequences), and can also predict dual localization – both chloroplastic and mitochondrial – which is an existing reality for some proteins [29]. The level of overall prediction accuracy is around 85% on a plant test set, the same as for TargetP, Table 2. The two predictors differ however somewhat in their performance on the subsets and trying both predictors on sequences of interest could prove useful.

Finally, an interesting approach to subcellular localization prediction has been presented by Eisenberg and co-workers [30]. They use a protein's 'phylogenetic profile' (i.e., a list of the presence or absence of orthologs to the query protein in all fully sequenced genomes) to predict its localization,

based on the assumption that the endosymbiont origin of different compartments will be reflected in the phylogenetic profiles of their respective proteomes. Thus, mitochondrial proteins (even the nuclearly encoded ones) will be most highly related to proteins from bacteria such as *Rickettsia prowasekii* [31], whereas chloroplast proteins will be most highly related to those found in photosynthetic bacteria.

Unfortunately, the different methods discussed in this section have not been evaluated together using a common benchmark (since the different methods do not distinguish between the same set of compartments, such an evaluation is not trivial). TargetP has the conceptual advantage that it tries to identify biologically well-characterized sorting signals and hence allows a certain amount of 'critical evaluation by eye' after the prediction has been made. The phylogenetic-profile approach also has a clear biological foundation, and again a human user may critically evaluate the results (i.e., the list of orthologs) against his or her biological knowledge. The purely statistical methods are at a disadvantage in this respect since they are based on sequence characteristics that are not easily evaluated by eye and, insofar as they incorporate amino acid composition measures, only correlate with subcellular localization indirectly (e.g., as a result of surface-exposed residues being adapted to a low-pH environment [32]).

7. Conclusions

The complex compartmentalization of a biological cell cannot yet be accurately captured by bioinformatics. For compartments where the sorting signals can to a good approximation be regarded as short stretches of amino acids with little interaction with the rest of the protein, the sequence analysis tools now available do a decent job. In cases where the sorting signals are presented in the context of a folded protein, however, they are very difficult to identify and one often has to resort to purely statistical approaches (amino acid composition) or methods based on sequence similarity. With improved fold recognition and three-dimensional structure prediction algorithms, it may eventually become possible both to detect these more complex sorting signals and to predict the location of a protein based on its general surface characteristics. In any event, the prediction of subcellular protein localization will most likely remain an important problem area for bioinformatics for some time to come.

References

[1] K. Nakai, Adv. Protein Chem. 54 (2000) 277–344.
[2] G. von Heijne, Eur. J. Biochem. 133 (1983) 17–21.
[3] G. von Heijne, Nucleic Acids Res. 14 (1986) 4683–4690.
[4] D.J. McGeoch, Virus Res. 3 (1985) 271–286.
[5] H. Nielsen, J. Engelbrecht, S. Brunak, G. von Heijne, Protein Eng. 10 (1997) 1–6.
[6] H. Nielsen, A. Krogh, Intell. Syst. Mol. Biol. 6 (1998) 122–130.
[7] K.M.L. Menne, H. Hermjakob, R. Apweiler, Bioinformatics 16 (2000) 741–742.
[8] K.-C. Chou, Protein Eng. 14 (2001) 75–79.
[9] G. von Heijne, EMBO J. 5 (1986) 1335–1342.
[10] Y. Abe, T. Shodai, T. Muto, K. Mihara, H. Torii, S. Nishikawa, T. Endo, D. Kohda, Cell 100 (2000) 551–560.
[11] T. Muto, T. Obita, Y. Abe, T. Shodai, T. Endo, D. Kohda, J. Mol. Biol. 306 (2001) 137–143.
[12] W. Voos, H. Martin, T. Krimmer, N. Pfanner, Biochim. Biophys. Acta 1422 (1999) 235–254.
[13] O. Emanuelsson, H. Nielsen, S. Brunak, G. von Heijne, J. Mol. Biol. 300 (2000) 1005–1016.
[14] M.G. Claros, P. Vincens, Eur. J. Biochem. 241 (1996) 779–786.
[15] N. Wiedemann, N. Pfanner, M.T. Ryan, EMBO J. 20 (2001) 951–960.
[16] C.M. Lee, J. Sedman, W. Neupert, R.A. Stuart, J. Biol. Chem. 274 (1999) 20937–20942.
[17] O. Emanuelsson, H. Nielsen, G. von Heijne, Protein Sci. 8 (1999) 978–984.
[18] C.J. Howe, T.P. Wallace, Nucleic Acids Res. 18 (1990) 3417.
[19] M. Hodel, A. Corbett, A. Hodel, J. Biol. Chem. 276 (2001) 1317–1325.
[20] M. Cokol, R. Nair, B. Rost, EMBO Rep. 1 (2000) 411–415.
[21] K. Nakai, M. Kanehisa, Proteins Struct. Funct. Genet. 11 (1991) 95–110.
[22] K. Nakai, M. Kanehisa, Genomics 14 (1992) 897–911.
[23] P. Horton, K. Nakai, in: Fifth International Conference on Intelligent Systems for Molecular Biology (ISMB-97) Halkidiki, Greece,, 1997, pp. 147–152.
[24] K. Nakai, P. Horton, Trends Biochem. Sci. 24 (1999) 34–35.
[25] A. Drawid, M. Gerstein, J. Mol. Biol. 301 (2000) 1059–1075.
[26] K.-C. Chou, Proteins Struct. Funct. Genet. 43 (2001) 246–255.
[27] A. Krogh, B. Larsson, G. von Heijne, E. Sonnhammer, J. Mol. Biol. 305 (2001) 567–580.
[28] S. Möller, M. Croning, R. Apweiler, Bioinformatics 17 (2001) 646–653.

[29] I. Small, H. Wintz, K. Akashi, H. Mireau, Plant Mol. Biol. 38 (1998) 265–277.
[30] E. Marcotte, I. Xenarios, A. van Der Bliek, D. Eisenberg, Proc. Natl. Acad. Sci. USA 97 (2000) 12115–12120.
[31] S.G.E. Andersson, A. Zomorodipour, J.O. Andersson, T. SicheritzPonten, U.C.M. Alsmark, R.M. Podowski, A.K. Näslund, A.S. Eriksson, H.H. Winkler, C.G. Kurland, Nature 396 (1998) 133–140.
[32] M. Andrade, S. O'Donoghue, B. Rost, J. Mol. Biol. 276 (1998) 517–525.
[33] B. Matthews, Biochim. Biophys. Acta 405 (1975) 442–451.