doi:10.1006/jmbi.2000.3903 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2000) **300**, 1005–1016

JMB

Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence

Olof Emanuelsson[1], Henrik Nielsen[2], Søren Brunak[2] and Gunnar von Heijne[1]*

[1]*Stockholm Bioinformatics Center, Department of Biochemistry, Stockholm University, S-106 91, Stockholm, Sweden*

[2]*Center for Biological Sequence Analysis, The Technical University of Denmark, DK-2800, Lyngby, Denmark*

**Corresponding author*

A neural network-based tool, TargetP, for large-scale subcellular location prediction of newly identified proteins has been developed. Using N-terminal sequence information only, it discriminates between proteins destined for the mitochondrion, the chloroplast, the secretory pathway, and ''other'' localizations with a success rate of 85 % (plant) or 90 % (non-plant) on redundancy-reduced test sets. From a TargetP analysis of the recently sequenced *Arabidopsis thaliana* chromosomes 2 and 4 and the Ensembl *Homo sapiens* protein set, we estimate that 10 % of all plant proteins are mitochondrial and 14 % chloroplastic, and that the abundance of secretory proteins, in both *Arabidopsis* and *Homo*, is around 10 %. TargetP also predicts cleavage sites with levels of correctly predicted sites ranging from approximately 40 % to 50 % (chloroplastic and mitochondrial presequences) to above 70 % (secretory signal peptides). TargetP is available as a web-server at http://www.cbs.dtu.dk/services/TargetP/.

Keywords: protein sorting; genome annotation; neural networks; targeting sequences; cleavage sites

Introduction

Most proteins in a eukaryotic cell are encoded in the nuclear genome and synthesized in the cytosol, and many need to be further sorted to one or other subcellular compartment. When the final destination is the mitochondrion, the chloroplast, or the secretory pathway, sorting usually relies on the presence of an N-terminal targeting sequence that is recognized by a translocation machinery (Rusch & Kendall, 1995; Schatz & Dobberstein, 1996).

In most cases, the targeting sequence is proteolytically removed during or after the entry (Robinson & Ellis, 1984; Hawlitschek *et al*., 1988; Arretz *et al*., 1991). For further sorting within the organelle, additional targeting information may be located in a secondary targeting sequence, either placed adjacent to the original targeting sequence (this is the case for, e.g. thylakoid targeted chloroplast proteins, Figure 1), or in other regions of the protein.

Abbreviations used: SP, signal peptide; mTP, mitochondrial targeting peptide; MPP, mitochondrial processing peptidase; MIP, mitochondrial intermediate peptidase; IMS, intermembrane space; cTP, chloroplast transit peptide; SPP, stromal processing peptidase.

E-mail address of the corresponding author: gunnar@biokemi.su.se

Signal peptides (SPs) are responsible for targeting proteins to the ER for subsequent transport through the secretory pathway (Rapoport, 1992; von Heijne, 1990). SPs generally consist of three regions: a positively charged n-region, a hydrophobic h-region, and a polar c-region leading up to the signal peptidase cleavage site. The most well-conserved motif of SPs is the presence of a small and neutral amino acid at positions −3 and −1 relative to the cleavage site (von Heijne, 1983, 1985).

In mitochondrial targeting peptides (mTPs), Arg, Ala and Ser are over-represented while negatively charged amino acid residues (Asp and Glu) are rare. Only weak consensus sequences have been found, the most prominent being a conserved Arg in position −2 or −3 relative to the mitochondrial processing peptidase (MPP) cleavage site. Furthermore, mTPs are believed to form an amphiphilic α-helix that is of importance for import of the nascent protein into the mitochondrion (Gavel *et al*., 1988; Roise, 1997; Waltner & Weiner, 1996). Some matrix proteins are cleaved a second time by the mitochondrial intermediate peptidase (MIP), which removes an additional eight to nine residues from the mature protein (Kalousek *et al*., 1988; Isaya & Kalousek, 1994). A subset of the mitochondrial proteins are first imported into the matrix, where their

0022-2836/00/041005–12 $35.00/0

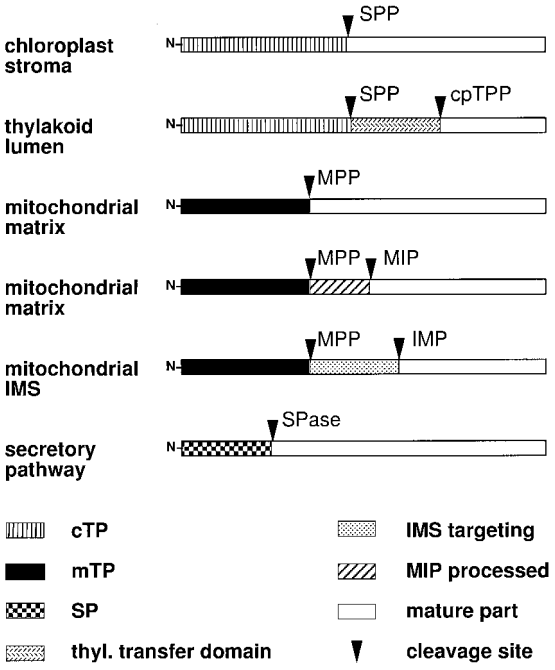

Figure 1. N-terminal targeting sequences, the corresponding final subcellular destinations, and the peptidases responsible for cleaving off the targeting sequences. The thylakoid transfer domain and the IMS targeting sequence are both SP-like. *cTP*, chloroplast transit peptide; *mTP*, mitochondrial targeting peptide; *SP*, signal peptide; *IMS*, intermembrane space (in mitochondria); *SPP*, stromal processing peptidase; *cpTPP*, chloroplast thylakoidal processing peptidase (homologous to SPase); *MPP*, mitochondrial processing peptidase; *MIP*, mitochondrial intermediate peptidase; *IMP*, mitochondrial inner membrane peptidase (homologous to SPase); *SPase*, signal peptidase.

mTP is cleaved off, and then re-exported to the intermembrane space (IMS) due to a second targeting signal (with some similarities to an SP) exposed after the removal of the mTP (Gasser *et al*., 1982; van Loon *et al*., 1987). Other variations on the theme exist, such as internal and C-terminal targeting information, and direct insertion into the outer membrane or IMS without first passing the matrix (Diekert *et al*., 1999; Lee *et al*., 1999).

The secondary structure of chloroplast transit peptides (cTPs) is not well characterized, and the sequence conservation around the stromal processing peptidase (SPP) cleavage site is not particularly strong (Gavel & von Heijne, 1990; Emanuelsson *et al*., 1999). Still, the cTP has a few distinguishing features such as low content of acidic residues and an over-representation of hydroxylated residues compared to the mature parts of chloroplast proteins (von Heijne *et al*., 1989). Thylakoid proteins have a bi-partite presequence structure (Figure 1) where the second signal is brought into action after the SPP cleavage of the N-terminal cTP. This thylakoidal transfer domain shares some important features with SPs (von Heijne, 1990; Robinson *et al*., 1998).

We have reported subcellular localization predictors designed to identify either SPs (SignalP) (Nielsen *et al*., 1997) or cTPs (ChloroP) (Emanuelsson *et al*., 1999) in a protein sequence. Here, we integrate and extend these efforts and present a novel subcellular localization predictor, TargetP, that assigns one of four different localizations (chloroplast, mitochondrion, ER/golgi/secreted, and ''other'') to a query sequence, and also predicts a potential cleavage site for presequence removal. A particularly important issue addressed in this work is the mutual discrimination between cTPs and mTPs, which was unsatisfying in ChloroP.

TargetP is built from two layers of neural networks, where the first layer contains one dedicated network for each type of presequence (cTP, mTP, SP), and the second is an integrating network that outputs the actual prediction (cTP, mTP, SP, other), Figure 2. A non-plant version of TargetP that distinguishes only between mTPs, SPs and other has also been constructed. All predictions are fully automatic and the expected performance profile can be customized to fit less restrictive searches for candidate proteins as well as highly conservative criteria for, e.g. database annotations. TargetP is able to discriminate between cTPs, mTPs, and SPs with sensitivities and specificities higher than what has been obtained with other available subcellular localization predictors, and has a relatively well-working cleavage site prediction capability for all involved target sequences.

Results

Data sets

As described in Methods, all sequences were extracted from SWISS-PROT and inappropriate sequences were removed before redundancy reduction, which was undertaken to avoid problems related to redundant data during neural network training and testing. To increase the size of the data sets as far as possible, also sequences annotated as ''POTENTIAL'', ''BY SIMILARITY''; or ''PROBABLE'' were included in their respective sets. These sequences lack experimental evidence for their cleavage sites, but since the networks in the first step are not trained to recognize cleavage sites specifically but instead whether or not a single residue is part of a targeting sequence, a misplaced cleavage site will only misclassify a few positions in each sequence. Therefore, we considered that a lower reliability in cleavage site position assignment would only marginally influence network performance. In the construction of mTP and cTP cleavage site predictors, though, only unambiguously annotated sequences were used (110 and 62, respectively).

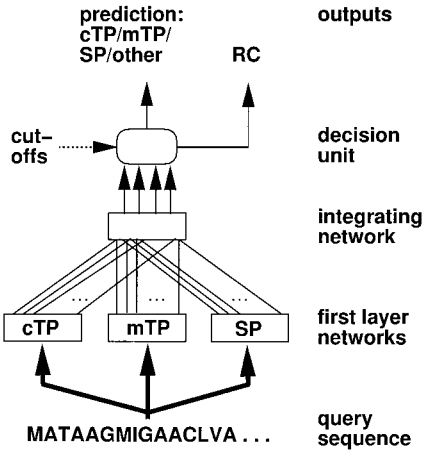

Figure 2. TargetP localization predictor architecture. TargetP is built from two layers of feed-forward neural networks, and on top a decision making unit, taking into account cutoff restrictions (if opted for) and outputting a prediction and a reliability class, RC, which is an indication of prediction certainty (see the text). The non-plant version lacks the cTP network unit in the first layer and does not have cTP as a prediction possibility.

Since the number of plant mTP sequences extracted from SWISS-PROT was too small to be useful, mTP sequences from all possible organisms were included when training both the plant and non-plant versions of TargetP. A cluster analysis of mTPs (Schneider *et al.*, 1998) using self-organized maps (Kohonen, 1982) did not reveal any significant species-specific features, suggesting that the use of non-plant mTPs in the training of the plant TargetP predictor and, conversely, the use of plant mTPs in the training of the non-plant TargetP predictor is reasonable. The redundancy reduced sets from which the training and test sets were built finally contained 141 cTP, 368 mTP, and 269 SP sequences (for the plant version of TargetP), and 371 mTP and 715 SP sequences (for the non-plant version of TargetP).

Neural network training

The three parallel networks in the first layer (cTP, mTP, SP) were trained using several combinations of network architecture parameters. The number of nodes in the hidden layer did not seem to be crucial, as long as there were at least two hidden nodes, while increased input window size in general improved performance (data not shown). For cTP networks, a window size of 55 positions was chosen, for mTP networks 35 positions (both plant and non-plant networks), and for SP networks, 31 (plant) or 27 (non-plant) positions. The chosen architectures of all first layer networks contained four hidden units. From these networks, the output scores corresponding to the 100 N-terminal positions in the sequence were fed into the integrating network. This network was trained on equal numbers of sequences from all the categories (size-equalized sets), and its performance was at its best when no hidden units were included. This indicates that once the first layer outputs have been obtained, the complexity of the sorting problem has been reduced, and is more or less linearly separable.

Localization prediction results on redundancy reduced test sets

Performance tests were done on several different data sets. First, TargetP was tested on the sequences that were used in the creation of the integrating layer networks, see Table 1. Cross-validation was applied, i.e. no sequence was tested on a network assembly whose training it had participated in. Since these sets were restricted (equalized) in size by the least abundant category of presequences (the cTP set for the plant version and the mTP set for the non-plant version of TargetP), testing was also undertaken with the full-size

Table 1. TargetP prediction performance, in actual numbers, on redundancy reduced non-equalized (size-equalized in parentheses) test sets

			Predicted category				
Set	True category	Number in category	cTP	mTP	SP	Other	Sensitivity
A. Plant	cTP	141 (140)	120 (119)	14 (14)	2 (2)	5 (5)	0.85 (0.85)
	mTP	368 (140)	41 (18)	300 (109)	9 (3)	18 (10)	0.82 (0.78)
	SP	269 (140)	2 (0)	7 (2)	245 (132)	15 (6)	0.91 (0.94)
	other	162 (135)	10 (5)	13 (9)	2 (5)	137 (116)	0.85 (0.86)
Specificity			0.69 (0.84)	0.90 (0.81)	0.96 (0.93)	0.78 (0.85)	
B. Non-plant	mTP	371 (370)	-	330 (330)	9 (8)	32 (32)	0.80 (0.89)
	SP	715 (370)	-	13 (6)	683 (354)	19 (10)	0.96 (0.96)
	other	1652 (370)	-	152 (47)	49 (8)	1451 (315)	0.88 (0.85)
Specificity			-	0.67 (0.86)	0.92 (0.96)	0.97 (0.88)	

In total 85.3(±3.5)% (85.8(±3.8)%) correct for plant protein predictor (940 (555) proteins) and 90.0(±1.1)% (90.0(±0.7)%) correct for non-plant protein predictor (2738 (1110) proteins), where standard deviations refer to the spread in performance of the five parallel networks.

redundancy-reduced sets, i.e. 940 sequences for plant version and 2738 for non-plant version. The prediction results are shown in Table 1. From this table, it can be seen that approximately 85 % of the 940 plant sequences were correctly predicted, with category-wise sensitivities in the interval 0.82-0.91, and for the non-plant protein predictor 90 % of the 2738 proteins were correctly predicted with category-wise sensitivities between 0.88 and 0.96. The specificities of plant and non-plant predictions are more scattered as they range from 0.67 (non-plant mTPs) to 0.97 (non-plant other) while the specificities when testing on the size-equalized sets were more uniform, essentially because the cTP and mTP specificities were higher. The poor discrimination between mTP and cTP which lowered the performance for the ChloroP predictor has been significantly improved, from almost 39 % of the mTP proteins falsely predicted as cTPs by ChloroP (data not shown) down to around 11 % for TargetP (calculated on ''full-size'' sets).

The full-size sets were also tested on PSORT (Nakai & Kanehisa, 1992; Horton & Nakai, 1997) and MitoProt (Claros, 1995; Claros & Vincens, 1996) as well as on TargetP's predecessors SignalP (Nielsen *et al*., 1997) and ChloroP (Emanuelsson *et al*., 1999). The results of these predictions are summarized in Table 2. TargetP performed better than PSORT in terms of sensitivity and specificity for almost all sets, with exceptions of non-plant ''other'' sensitivity where PSORT was better, and cTP and non-plant SP specificity where performance was equal. Each of the ''binary'' predictors MitoProt, ChloroP, and SignalP obtained better sensitivity for the category they were specialized on, but the specificities were in all cases lower than those of TargetP (for cTPs and plant-SPs much lower). In plant protein predictions, MitoProt was used in its three-state mode (see MitoProt instruction file), in which it distinguishes between mTPs and cTPs using Arg and Ser frequencies as proposed in (von Heijne *et al*., 1989). Although the obtained MitoProt cTP specificity was greater than that of ChloroP it was lower than those of PSORT and TargetP. Measured by the Matthews correlation coefficient, MCC (Matthews, 1975), TargetP performed better than both PSORT and MitoProt on all the tested categories.

To examine the possibilities to generate more reliable predictions for database annotation, where the focus is on specificity more than sensitivity, we tried two different ways of improving the certainty that query sequences really belong to the predicted category. The two approaches turned out to yield similar results when applied to the size-equalized test sets. The first approach was to demand the specificity (on the redundancy reduced test sets) to be above a certain level, which was implemented by imposing a cutoff restriction on the output scores in addition to the default winner-takes-all rule. Thus, for a specificity of 0.95, the corresponding sensitivities were found to be 0.63-0.96 (non-plant predictor) or 0.33-0.93 (plant predictor), see Table 3. The lower figure is in both cases for mTP

Table 2. Comparison of localization predictor performances on redundancy reduced non-equalized plant (940 proteins) and non-plant (2738 proteins) test sets

Predictor set	% Correct overall	Category	Specificity	Sensitivity	MCC	Reference
A. *TargetP*						
Plant	85.3	cTP	0.69	0.85	0.72	(This work)
		mTP	0.90	0.82	0.77	
		SP	0.95	0.91	0.90	
		other	0.78	0.85	0.77	
Non-plant	90.0	mTP	0.67	0.89	0.73	
		SP	0.92	0.96	0.92	
		other	0.97	0.88	0.82	
B. *PSORT*						
Plant	69.8	cTP	0.69	0.47	0.51	(Nakai & Kanehisa, 1992; Horton &)
		mTP	0.87	0.66	0.64	Nakai, 1997
		SP	0.74	0.82	0.69	
		other	0.47	0.78	0.50	
Non-plant	83.2	mTP	0 60	0 81	0 64	
		SP	0.93	0.64	0.71	
		other	0.87	0.92	0.73	
C. *MitoProt*						
Plant	80.3	cTP	0.59	0.45	0.44	(Claros, 1995; Claros & Vincens, 1996)
		mTP	0.77	0.84	0.67	
Non-plant	89.4	mTP	0.57	0.91	0 67	
D. *ChloroP*						
Plant	78.8	cTP	0.40	0.90	0.50	(Emanuelsson *et al*., 1999)
E. *SignalP*						
Plant	79.9	SP	0.59	0.94	0.62	(Nielsen *et al*., 1997)
Non-plant	92.4	SP	0.78	0.98	0.83	

MCC, Matthews correlation coefficient (calculated category-wise).

Table 3. Obtaining pre-defined specificities by imposing cutoff restrictions on TargetP predictions

				Predicted category											
				cTP			mTP			SP			other		
Set	Required minimum specificity	No. seqs with score > cutoff	Whereof correctly predicted	Spec.	Sens.	Cutoff	Spec.	Sens.	Cutoff	Spec.	Sens.	Cutoff	Spec.	Sens.	Cutoff
Plant	0.98	218 (39.3 %)	214 (98.2 %)	0.99	0.48	0.82	1.00	0.03	0.97	0.98	0.87	0.64	1.00	0.16	0.91
	0.95	333 (60.0 %)	318 (95.5 %)	0.95	0.59	0.73	0.96	0.33	0.86	0.96	0.93	0.43	0.95	0.43	0.84
	0.90	453 (81.6 %)	412 (90.0 %)	0.90	0.72	0.62	0.91	0.48	0.76	0.93	0.94	0.00	0.90	0.83	0.53
Non-plant	0.98	383 (34.5 %)	376 (98.1 %)	-	-	-	1.00	0.02	0.97	0.98	0.81	0.83	0.99	0.19	0.93
	0.95	881 (79.4 %)	841 (95.5 %)	-	-	-	0.95	0.63	0.78	0.96	0.96	0.00	0.95	0.69	0.73
	0.90	1033 (93.1 %)	955 (92.4 %)	-	-	-	0.91	0.79	0.65	0.96	0.96	0.00	0.90	0.84	0.52

Performed on the size-equalized plant (555 proteins) and non-plant (1110 proteins) sets. The actual specificity may be slightly higher than the required specificity, depending on prediction threshold effects between two adjacent cutoff levels. Spec., specificity; Sens., sensitivity.

prediction while the higher is for SP prediction. The non-equalized sets were also tested (data not shown), with similar resulting performances except that the limiting (in terms of size) categories in general scored lower sensitivities, while the largest categories of plant and non-plant proteins, mTP and ''other'', respectively, scored better sensitivities. This is not surprising given that the specificity requirements become harder to fulfill for the limiting categories as the number of potential false positives increases when going from size-equalized to non-equalized sets. In the publicly available predictor, predefined cutoffs corresponding to certain levels of specificity are provided.

Second, we tried the use of ''reliability classes'': a prediction is assigned a reliability class (*RC*) according to the difference, Δ, between highest and second-highest network output score. If $\Delta > 0.8$, then $RC = 1$; if $0.6 < \Delta < 0.8$, then $RC = 2$, etc. (five *RC*s in total). This feature is a useful indication of the level of certainty in the prediction for a particular sequence. Overall, 99 % of the sequences with $RC = 1$, and 93/95 % (plant/non-plant) of the sequences with $RC = 2$ were correctly predicted, Table 4. Besides the specificities within each particular *RC*, we also calculated the cumulative specificities and sensitivities. This is the values for all proteins predicted to a particular *RC* or better, so when calculating the cumulative performances for e.g. $RC = 2$, all proteins predicted either to $RC = 2$ or $RC = 1$ were included, and correspondingly for *RC*s 3, 4, and 5 (the latter of course equals the overall predictor performance without restriction cutoffs). While the specificity within an *RC* is an expression of the reliability of a specific prediction, the cumulative performance values show how much the sensitivity will be reduced, if the specificity is increased by only considering predictions at a particular *RC* or better.

Localization prediction results on *Arabidopsis thaliana* and *Homo sapiens* data sets from SWISS-PROT

All available *A. thaliana* and *H. sapiens* entries in SWISS-PROT (as of October 1999), with annotated subcellular location in the FT or CC fields, were collected and run through the predictors. Since quite a few of the entries were present in the training and test sets of TargetP, performance with the common sequences removed was also checked. Although this removal resulted in somewhat lower performances, the overall performance of TargetP was still the best among the tested predictors. A total of 84 % *A. thaliana* and 86 % *H. sapiens* sequences were correctly predicted by TargetP (considering only non-overlapping sequences), as compared to 68 and 69 %, respectively, for PSORT. In general, the performance patterns for all predictors were fairly similar to those on the full-size redundancy reduced sets except for cTP and mTP specificities. For mTP sets the specificities were clearly lower, with differences of 0.24 (TargetP, non-plant) to 0.55 (MitoProt, plant) units, while for cTP sets they were higher by 0.11 (PSORT) to 0.30 (ChloroP) units. For details on the tests of *A. thaliana* and *H. sapiens* SWISS-PROT sets, consult the supplementary material available at JMB Online.

Table 4. TargetP performance within the reliability classes (RCs)

				Predicted category											
				cTP			mTP			SP			other		
Set	RC	No. sequences predicted to RC	Whereof correctly predicted	Spec. cumulative	Sens. cumulative	Spec.	Spec. cumulative	Sens. cumulative	Spec.	Spec. cumulative	Sens. cumulative	Spec.	Spec. cumulative	Sens. cumulative	Spec.
Plant	1	173 (31.2 %)	172 (99.4 %)	1.00	0.24	1.00	1.00	0.22	1.00	0.99	0.66	0.99	1.00	0.12	1.00
	2	135 (24.3 %)	126 (93.3 %)	0.97	0.50	0.95	0.97	0.44	0.94	0.99	0.81	1.00	0.92	0.40	0.88
	3	98 (17.7 %)	83 (84.7 %)	0.93	0.62	0.81	0.92	0.59	0.82	0.99	0.88	1.00	0.89	0.65	0.85
	4	76 (13.7 %)	55 (72.4 %)	0.90	0.79	0.80	0.86	0.66	0.56	0.97	0.92	0.67	0.87	0.76	0.79
	5	73 (13.2 %)	40 (54.8 %)	0.81	0.85	0.42	0.81	0.78	0.62	0.93	0.94	0.33	0.85	0.86	0.68
Total		555 (100 %)	476 (85 8 %)												
Non-plant	1	432 (38.9 %)	426 (98.6 %)	-	-	-	0.97	0.28	0.97	1.00	0.62	1.00	0.98	0.25	0.98
	2	341 (30.7 %)	323 (94.7 %)	-	-	-	0.96	0.59	0.95	0.98	0.85	0.94	0.96	0.59	0.95
	3	153 (13.8 %)	127 (83.0 %)	-	-	-	0.93	0.75	0.82	0.97	0.90	0.87	0.93	0.72	0.83
	4	111 (10.0 %)	81 (73.0 %)	-	-	-	0.90	0.84	0.71	0.97	0.94	0.88	0.90	0.81	0.69
	5	73 (6.6 %)	42 (57.5 %)	-	-	-	0.86	0.89	0.53	0.96	0.96	0.55	0.88	0.85	0.67
In total:		1110 (100 %)	999 (90.0 %)												

Performed on the size-equalized plant (555 proteins) and non-plant (1110 proteins) sets. The *cumulative* specificities and sensitivities are the values for the sequences predicted to the particular reliability class, *RC*, or better. These values are comparable to the *sens.* and *spec.* values presented in Table 3 (see the text). Other values refer to the performance within each *RC*. Spec., specificity; Sens., sensitivity.

Cleavage site predictions

As mentioned in Methods, the TargetP cleavage site predictions of SPs and cTPs are the same as in the SignalP and ChloroP methods, while the mTP cleavage site prediction is a new feature. It consists of three competing scoring matrices, derived from sequences known to have an Arg in either −2, −3 or −10 relative to the annotated cleavage site. We tested the cleavage site prediction ability on the redundancy-reduced cTP/mTP/SP sets, using only the unambiguously annotated sequences (56 cTPs, 197 mTPs, 813 SPs), as well as on SWISS-PROT *A. thaliana* cTP and *H. sapiens* mTP sets (72 and 53 sequences, respectively), Table 5. The cleavage site prediction ability was in general good for SPs, 75% correct on the redundancy-reduced set, and around 65% correct on the *Arabidopsis* and *Homo* sets, while mTP and, in particular, cTP cleavage site prediction were not as reliable. On the human mTP set, TargetP predicted half, and MitoProt a quarter, of the cleavage sites correctly (the sequences that participated in the TargetP mTP cleavage site prediction development were removed from the TargetP results in this test). On the redundancy-reduced mTP set, TargetP again predicted approximately 50% of the cleavage sites correctly. Ninety-nine percent of the correctly predicted mTP sequences known to have an Arg residue in position −2, −3, or −10 relative to the annotated cleavage site, had their highest (prediction-determining) score from the matrix corresponding to their annotated cleavage site. It also turned out that the cleavage sites of proteins with an Arg residue in −10 were harder to predict correctly than those with an Arg residue in −2 or −3. For the two cTP sets (redundancy-reduced and *A. thaliana*) only around 10% of the cleavage sites were correctly predicted, while more than 40% were predicted to have their cleavage sites within ±2 residues from the annotated site. There was a strong bias towards predicting the cTP as shorter than annotated, and we have earlier suggested that this may depend on a to-date uncharacterized proteolytic activity in the stroma (Emanuelsson *et al.*, 1999). For prediction of mitochondrial cleavage site, though, no clear length bias was found.

Prediction of unannotated data sets

As a first application of TargetP, we analyzed the predicted protein coding regions of the newly sequenced *A. thaliana* chromosomes 2 and 4 (Lin *et al.*, 1999; The European Union Arabidopsis Genome Sequencing Consortia, 1999) and the *H. sapiens* Ensembl set (http://ensembl.ebi.ac.uk/), Table 6. Approximately 14% of the *Arabidopsis* sequences (both chromosome 2 and 4) were predicted as chloroplastic while the predicted mTP and SP abundances were around 10% and 15%, respectively, in all three sets. The results of these predictions are available on the TargetP web site. Not all sequences predicted to contain an SP are actually secreted, though; a subset of them are transmembrane (TM) proteins. To estimate the percentage of these, we used the TMHMM prediction method for TM helices (Sonnhammer *et al.*, 1998). We assigned all proteins predicted to contain one or more TM helices downstream of position 40 as TM proteins, while those with no TM helices were assigned as secreted proteins. Proteins predicted by TMHMM to contain only one TM helix within the N-terminal 40 residues were subjected to further analysis, since the predicted TM helix in these cases might actually be the hydrophobic part of a cleavable signal peptide. To this end, we used an experimental hidden Markov model-based version of SignalP (SignalP-HMM) (Nielsen & Krogh, 1998) which offers a better discrimination between cleaved signal peptides and uncleaved signal anchors than does the original neural network-based SignalP. If SignalP-HMM predicted a signal peptide (74%-83% of the group with one N-terminal predicted TM helix), the protein was assigned as secreted, otherwise as TM protein. This analysis lowered the estimates of secreted proteins to approximately 11% for the *Arabidopsis* sets and to 8% for the *Homo* set.

Table 5. Cleavage site predictions on redundancy-reduced, *Homo* and *Arabidopsis* sets

Predictor set	No. of seqs	% Correct	
		Exact	Within ±2 residues
TargetP			
cTP	58	6.9	44.8
mTP	197	50.8	59.9
SP	813	74.9	83.9
Homo-mTP	56	50.0	60.7
Homo-SP	1323	68.1	81.1
Arabidopsis-cTP	67	10.4	41.8
Arabidopsis-SP	53	62.3	77.4
MitoProt			
mTP	197	25.4	34.0
Homo-mTP	72	25.0	36.1

Note that TargetP cleavage site prediction performance is not cross-validated. However, the *Homo*-mTP set on which TargetP was tested contains no sequences that participated in the development of the TargetP mTP cleavage site predictor.

Table 6. TargetP predictions of unannotated *A. thaliana* and *H. sapiens* data sets

		Predicted abundance, %			
Data set	No. of seqs	cTP	mTP	SP	SP, no membrane proteins (*)
A. thaliana chr 2	4054	13.2	10.5	16.7	11.1
A. thaliana chr. 4	3744	13.9	10.1	17.2	11.6
H. sapiens (Ensembl)	10,228	-	9.3	12.8	8.0

The SP category marked (*) does not include predicted transmembrane proteins (see the text). The prediction for each sequence is available on the TargetP web site. The *A. thaliana* sets were downloaded from ftp://ftp.tigr.org/pub//data/a_thaliana/chromosomeII/(chromosome 2) and ftp://warthog.mips.biochem.mpg.de/pub/cress/chrIV/ESSAseq/(chromosome 4). The *H. sapiens* set was downloaded from ftp://ftp.sanger.ac.uk/pub/ensembl/data/pep/.

Discussion

The aim of this work was to provide improved subcellular localization predictions for proteins potentially sorted to the chloroplast, the mitochondrion, or the secretory pathway, and to generate a cleavage site prediction for mTPs as a complement to the already existing cTP and SP cleavage site predictions in ChloroP and SignalP. We have managed to increase (comparing to existing tools) the discrimination ability between the targeting sequences, especially in terms of specificity, and in particular the poor discrimination of cTPs and mTPs has been clearly improved when compared to ChloroP (Tables 1 and 2). In general, the one-category predictors MitoProt, SignalP, and ChloroP still yield a higher sensitivity on their particular presequences, but at the cost of reduced specificity. Letting the user choose cutoffs for the predictions is a means for fine-tuning the TargetP performance and biasing it towards more restrictive predictions (Table 3). To the same end, a classification of each prediction into one of five reliability classes has been developed as an indication of how certain a prediction is: the lower the class number, the safer the prediction (Table 4).

To test TargetP and its competitors on real-world applications, all *A. thaliana* and *H. sapiens* sequences available in SWISS-PROT were collected and processed through TargetP, PSORT, MitoProt, SignalP, and ChloroP. The performance of TargetP on these sets (86 and 84 % correctly predicted sequences for non-plant and plant sets) was only slightly lower compared to the cross-validated test set performance (Table 2), and was in almost all aspects superior to the other predictors. We conclude that the use of TargetP, for e.g. automatic annotation purposes, will yield significantly less false-positives at the cost of missing fairly modest numbers of true-positives compared to other available predictors.

The cleavage site predictions are not as reliable, but TargetP is still able to predict the correct cleavage site in approximately 40-50 % (cTPs and mTPs) or 70 % (SPs) of the tested proteins (Table 5). It is obvious that mTP and, in particular, cTP cleavage site predictions still are in great need of improvements. The scarce data is so far the biggest obstacle in this matter.

An analysis of three newly sequenced and unannotated data sets, *A. thaliana* chromosomes 2 and 4 and *H. sapiens* Ensembl set, suggests an abundance of roughly 10 % mTPs and 15 % SPs (including both secretory and membrane proteins) in all three sets, and 14 % cTPs in the plant set (Table 6).

In conclusion, the successful construction of the TargetP predictor demonstrates that protein sorting signals can be recognized with reasonable reliability from amino acid sequence data alone, thus to a certain extent mimicking the cellular recognition processes. It is likely that further improvements can be obtained by including, e.g. information from multiply aligned sequences or from analyses of the mature part of the proteins, downstream of the sorting signals (Andrade *et al.*, 1998; Chou & Elrod, 1999; Reinhardt & Hubbard, 1998). It should also be possible to extend the abilities of TargetP by searching for secondary targeting sequences such as thylakoid transfer domains immediately adjacent to the primary sorting signals.

Methods and Data Sets

General outline of data set creation

All data were extracted from SWISS-PROT (Bairoch & Apweiler, 2000). Release 36 was used for the plant data sets, and release 37 for the non-plant except for the mTP set in which the upgrades of release 38 also were included. Sequences were extracted by requiring the keyword EUKARYOTA in the OC (Organism Classification) field. Sequences exhibiting PLANTA as the second node in the OC field were extracted to the plant data sets. Targeting peptide entries marked as POTENTIAL, BY SIMILARITY, or PROBABLE in their FT field, but still with an explicitly annotated endpoint of the presequence, were also included in their respective sets (except non-plant SP set which was considered large enough without including such sequences). In the nuclear and cytosolic sets, sequences with any of these annotations as to their subcellular location annotations in their CC field were also accepted. Only sequences with an N-terminal Met residue were considered, and we also excluded the very few sequences containing B, Z, or X, in order to avoid possible noise from the ambigous positions in the training. Following the removal of these and other inadequate entries (see below), sequences with a high degree of similarity to other sequences were removed by redundancy reduction.

For redundancy reduction, Hobohm algorithm 2 was employed (Hobohm *et al*., 1992). Pairwise alignment was performed using the full Smith-Waterman algorithm and the PAM250 scoring matrix, as implemented in the search program of the FASTA package (Smith & Waterman, 1981; Pearson, 1990). The threshold score above which sequences were considered as too similar for network training was chosen as the score above which the actual distribution of scores deviated from the expected extreme value distribution of scores from a local alignment of random sequences (Karlin & Altschul, 1990; Altschul *et al*., 1994; Pedersen & Nielsen, 1997). Before comparison to the extreme value distribution, the score was corrected for the length difference of the aligned sequences by dividing the raw Smith-Waterman score with $\ln(m \times n)$ (Altschul & Gish, 1996), where m and n are the lengths of the two aligned sequences.

cTP data set for plant version of TargetP

cTP containing proteins were identified by requiring the annotation ''TRANSIT (...) CHLOROPLAST'' in the FT (Feature Table) field (566 proteins in total). From this set, entries with cleavage sites (CS), predicted by the SP-predictor SignalP (prokaryotic, gram-negative networks) (Nielsen *et al*., 1997) to lie within ±5 residues from annotated CS were removed since it could not be excluded that these cleavage sites resulted from the second cleavage of a bipartite stroma-thylakoidal targeting presequence. Sequences from algae were also removed since it has been shown that they are more similar to mTPs than to cTPs from higher plants (Franzén *et al*., 1990). Furthermore, eight chloroplast encoded proteins, one chloroplast envelope protein, and one having a thylakoidal transfer domain that had been missed by SignalP were excluded. The eight chloroplast encoded proteins were all annotated as of organellar origin (''Chloroplast'' in the OG, organelle, field) which is incompatible with their TRANSIT (...) CHLOROPLAST annotation in the FT field, since this annotation indicates that the protein has a transit peptide for import into chloroplasts. Since all these proteins were cytochrome f, known to be encoded in the chloroplast, the annotated presequence most likely is a thylakoidal transfer domain (the SWISS-PROT database curators have been informed). After these operations, a total of 432 entries were left in the cTP data set, with a mean cTP length of 56 amino acid residues. Redundancy reduction was then performed as described above. In the creation of the data set with positive training examples, the redundancy reduction was done including the annotated cTP and the first residue of the mature protein (resulting in 141 non-redundant entries) while the cTP entries to be used as negative examples in the training of mTP and SP networks were redundancy-reduced on the 68 N-terminal amino acid residues (corresponding to twice the length of the average mitochondrial transit peptide, mTP), leaving 123 sequences.

mTP data set for plant version of TargetP

Sequences annotated ''TRANSIT(...)MITOCHONDRION'' in their FT field (and with N-terminal mTP) were extracted to the mTP set. The mTP set consisted not only of plant sequences since the number of plant mTPs was too small to allow reliable network training. Previous studies have not been able to reveal significant species-correlated differences between mTPs (Schneider *et al*., 1998). Proteins annotated (according to the ''SUBCELLULAR LOCATION'' comment) as being located in the inter-membrane space were removed from the data set since they, in general, have a bi-partite presequence and the annotated cleavage site may thus stem from the cleavage of the second, IMS-targeting sequence (Figure 1). The 658 mTP containing sequences were left after these procedures. The redundancy reduction on this set for the use as positive training data was performed on the mTP plus the first mature residue and left 368 non-redundant proteins. For sequences to be used as negative examples in cTP and SP network training, redundancy reduction was performed on the 112 N-terminal amino acid residues (corresponding to twice the average length of cTP) resulting in 190 sequences.

SP data set for plant version of TargetP

The SP containing plant sequences were picked as those showing the keyword SIGNAL in their FT field (648 proteins in total). The data set was redundancy reduced on the actual SP plus the first residue of the mature protein for use as positive data in the SP network development, and on 112 residues (corresponding to twice the length of the average chloroplast transit peptide, cTP) for use as negative data in the training of the cTP and mTP networks. The resulting, redundancy reduced data sets, consisted of 269 and 82 sequences, respectively.

Cytosolic and nuclear data sets for plant version of TargetP

The cytosolic and nuclear plant sets were used as negative sequences in the training of all plant networks. The cytosolic and nuclear entries contained the string SUBCELLULAR LOCATION: CYTOPLASMIC/NUCLEAR in their CC field. The initial sets contained 537 cytosolic and 214 nuclear proteins. After redundancy reduction on the first 112 residues (to be used in the training of the cTP network), 87 cytosolic 48 nuclear proteins remained. For training of the mTP networks, redundancy reduction was applied on the 68 N-terminal residues, which left 108 cytosolic and 54 nuclear proteins.

mTP data sets for non-plant version of TargetP

The mTP set for the non-plant predictor was based on the mTP set for the plant protein predictor presented above, to which was added the mTP containing updates of SWISS-PROT releases 37 and 38. The redundancy reduction for the positive mTP set was performed on the mTP and the first three residues of the mature protein, since it had been shown that they potentially play a role in presequence recognition (Song *et al*., 1998). The initial set comprised 702 sequences. After redundancy reduction, the positive set contained 371 sequences, and the negative set 344 (reduced on the 44 N-terminal residues). The average mTP length was 34 amino acid residues.

SP data sets for non-plant version of TargetP

The non-plant signal peptide set contained 2292 sequences collected in the same way as the plant SP set, except that ambiguously annotated entries (see above) were not included. After redundancy reduction, performed on the SP and first residue of mature protein,

715 sequences were left. In the reduction procedure for the negative set (on the first 68 amino acid residues), the initial set was divided into two parts that were redundancy reduced separately for technical reasons. The two reduced sets were then concatenated and reduced once again, resulting in a set of 527 proteins. The average SP length was 22 residues.

Cytosolic and nuclear data sets for non-plant version of TargetP

These sets were collected from the non-plant sequences in SWISS-PROT release 37 annotated as SUBCELLULAR LOCATION: CYTOPLASMIC/NUCLEAR in their CC fields. There were 2274 cytoplasmic sequences and 4037 nuclear sequences initially. After redundancy reduction, 438 and 1214 sequences were left, respectively. Again, the redundancy reductions were performed (on the 68 N-terminal residues) by splitting each set into two separately reduced parts that were merged and reduced once again.

Training and test set construction

All training and test sets were truncated to a number divisible by five. Before network training, the data sets were divided into five equally sized parts for cross-validation. Each sequence participated either in the training or in the testing of a particular network, not both. For the first layer networks, the sets were constructed to contain equal numbers of positive and negative training sequences and the negative sequences consisted of approximately equal numbers of all the applicable non-positive categories. For the integrating layer network, the least abundant of the three (two in non-plant version) presequence categories determined the size of the other classes to assure training with equivalent numbers from each protein class (cTP, mTP, SP, other) (size-equalized sets). All sequence exclusions were random. The final plant TargetP training sets consisted of, for cTP networks 280 (half of which cTP-containing, positive examples), for mTP networks 730 (365 mTPs), for SP networks 530 (265 SPs), and for the integrating network 555 redundancy-reduced sequences (from four categories: cTP, mTP, SP, other). The final non-plant TargetP training sets consisted of, for mTP networks 740 (370 mTPs), for SP networks 1420 (710 SPs), and for the integrating network 1110 sequences (from three categories: mTP, SP, other). The cTP, mTP, and SP data sets can be downloaded from the TargetP web site.

Neural network architecture and training

The TargetP predictor has neural networks in two layers (Figure 2), roughly in the same manner as for ChloroP (Emanuelsson *et al*., 1999), with the first layer consisting of one network for each type of presequence (i.e. three in the plant version and two in the non-plant version), each assigning one score per residue. The outputs of the first layer networks are fed into the second (integrating) layer network, which outputs one score per query sequence and possible localization class (i.e. four in the plant version and three in the non-plant version). All neural networks in the predictor are of the feed-forward type with sigmoidal neurons (Minsky & Papert, 1968) and zero or one layer of hidden neurons, trained using error backpropagation (Rumelhart *et al*., 1986) but the implementations and chosen parameter values differ somewhat.

First layer networks are implemented using the HOW package, (Brunak *et al*., 1991) with a logarithmic error function and sparsely encoded sliding windows for input data encoding (Qian & Sejnowski, 1988; Brunak *et al*., 1991). Each position in the input sequence window occupies 20 input nodes (one for each of the 20 amino acid residues), and the node corresponding to the amino acid present at that position is switched on (i.e. set to one) while the others remain off (set to zero). The first layer networks are then trained to recognize whether or not the residue in the middle of the sliding window is part of a targeting sequence. Networks with different sizes of the sliding window and different numbers of nodes (0, 2, 4, 8) in the hidden layer were tested. Sliding window sizes ranged, for cTP networks, from 7 to 55 residues, for mTP networks from 7 to 35, and for SP networks from 7 to 31 residues (the upper limit roughly following the average presequence length). The learning rate was set to 0.001 based on earlier studies (Emanuelsson *et al*., 1999).

The second layer (integrating) network was implemented with the HOWLIN program from the HOW package, using a quadratic error function, and considers as input the outputs from the first layer cTP, mTP, and SP networks corresponding to the 100 N-terminal positions of the input sequence. For each residue in the query sequence there are thus three (plant version) or two (non-plant version) scores that are fed into the integrating layer network. The output from the top layer network is one score per type of targeting peptide, i.e. four for the plant version (cTP, mTP, SP, other) and three for the non-plant version (mTP, SP, other). In the default implementation the highest output score determines the prediction (winner takes all) but there is also a possibility to demand the output to be above a certain threshold to be valid as a prediction, thus altering the expected sensitivity/specificity balance (see Results). Due to the fivefold cross-validation, all previously mentioned networks (first and integrating layers) are not one single network but five parallel networks, each of which created from one training set and tested on one test set. In the final version, the prediction result is a combination of the outputs of all parallel networks.

Measuring prediction performance

Performances were in general measured as percentage correctly predicted sequences, and as sensitivity (fraction of positive examples predicted as positives):

$$sens = \frac{tp}{tp + fn}$$

and specificity (fraction of all positive predictions that are true positives):

$$spec = \frac{tp}{tp + fp}$$

where tp = true positives, fn = false negatives (under-prediction), and fp = false positives (over-prediction). The Matthews correlation coefficient, MCC (Matthews, 1975), defined as:

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fn)(tp + fp)(tn + fp)(tn + fn)}}$$

where tn = true negatives, was used in the comparison of performances of different predictors. MCC equals one for a perfect prediction, while it is zero for a completely random assignment.

Cleavage site predictions

cTP cleavage site prediction follows exactly the method described in (Emanuelsson *et al*., 1999), using the same networks and scoring matrix, and the same way of choosing the area within which the cleavage site is searched. All cTP cleavage site performances are given as the number of sequences correctly predicted within ±2 residues if not otherwise stated. SP cleavage is determined by processing the sequences through SignalP (Nielsen *et al*., 1997). The prediction of mTP cleavage sites is a new feature. The redundancy reduced mTP sequences, including the additions from SWISS-PROT releases 37 and 38 and excluding all with notations of uncertainty as to their cleavage sites (in total 197 sequences) were divided into 4 groups based on the Arg residue presence in positions −2, −3, and −10 relative to the annotated cleavage sites (Gavel *et al*., 1988). Sequences with an Arg residue in exactly one of these positions were kept in three separate groups and the rest (those with several or none Arg residues) were not used in the cleavage site prediction development. The three Arg groups (−2, −3, −10) contained 41, 39, and 30 sequences respectively. An automatic motif finding and score matrix generating program, MEME (Bailey & Elkan, 1994), was used to create one scoring matrix for each of the three groups. For each sequence, the 12 residues surrounding the Arg were included in the respective MEME training set. The final cleavage site was then predicted by simply letting the matrix generating the highest score on the sequence determine the site of cleavage. The search is restricted to the 120 N-terminal residues since mTPs longer than that only very rarely have been reported (only 2 out of 702 sequences in the unreduced mTP set are longer than 120 residues).

TargetP user instructions

The user is prompted to choose between the plant and non-plant versions of TargetP, and decides whether the default localization decision rule (winner-takes-all) is valid or should be completed with cut-off restrictions. There are also predefined cutoffs, corresponding to certain expected sensitivity/specificity combinations. Cleavage site prediction is another option. TargetP is available at http://www.cbs.dtu.dk/services/TargetP/.

Acknowledgments

This work was supported by grants from the Swedish Natural and Technical Sciences Research Councils and from the Foundation for Strategic Research to GvH. S.B. and H.N. are supported by a grant from the Danish National Research Foundation.

References

Altschul, S. F. & Gish, W. (1996). Local alignment statistics. *Methods Enzymol.* **266**, 460-480.

Altschul, S., Boguski, M. S., Gish, W. & Wootton, J. C. (1994). Issues in searching molecular sequence databases. *Nature Genet.* **6**, 119-129.

Andrade, M. A., O'Donoghue, S. I. & Rost, B. (1998). Adaption of protein surfaces to subcellular location. *J. Mol. Biol.* **276**, 517-525.

Arretz, M., Schneider, H., Wienhues, U. & Neupert, W. (1991). Processing of mitochondrial precursor proteins. *Biomed. Biochim. Acta,* **50**, 403-412.

Bailey, T. L. & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *ISMB,* **2**, 28-36.

Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45-48.

Brunak, S., Engelbrecht, J. & Knudsen, S. (1991). Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.* **220**, 49-65.

Chou, K. C. & Elrod, D. W. (1999). Protein subcellular location prediction. *Protein Eng.* **12**, 107-118.

Claros, M. G. (1995). MitoProt: a Macintosh application for studying mitochondrial proteins. *Comput. Appl. Biosci.* **11**, 441-447.

Claros, M. G. & Vincens, P. (1996). Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.* **241**, 779-786.

Diekert, K., Kispal, G., Guiard, B. & Lill, R. (1999). An internal targeting signal directing proteins into the mitochondrial intermembrane space. *Proc. Natl Acad. Sci. USA,* **96**, 11752-11757.

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

Franzén, L.-G., Rochaix, J.-D. & von Heijne, G. (1990). Chloroplast transit peptides from the green alga *Chlamydomonas reinhardtii* share features with both mitochondrial and higher chloroplast presequences. *FEBS Letters,* **260**, 165-168.

Gasser, S. M., Ohashi, A., Daum, G., Bohni, P. C., Gibson, J., Reid, G. A., Yonetani, T. & Schatz, G. (1982). Imported mitochondrial proteins cytochrome b2 and cytochrome c1 are processed in two steps. *Proc. Natl Acad. Sci. USA,* **79**, 267-271.

Gavel, Y. & von Heijne, G. (1990). A conserved cleavage-site motif in chloroplast transit peptides. *FEBS Letters,* **261**, 455-458.

Gavel, Y., Nilsson, L. & von Heijne, G. (1988). Mitochondrial targeting sequences. Why ''non-amphiphilic'' peptides may still be amphiphilic. *FEBS Letters,* **235**, 173-177.

Hawlitschek, G., Schneider, H., Schmidt, B., Tropschug, M., Hartl, F. U. & Neupert, W. (1988). Mitochondrial protein import: identification of processing peptidase and of PEP, a processing enhancing protein. *Cell,* **53**, 795-806.

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

Horton, P. & Nakai, K. (1997). Better prediction of protein cellular localization sites with the *k* nearest neighbors classifier. *ISMB,* **5**, 147-152.

Isaya, G. & Kalousek, F. (1994). Mitochondrial intermediate peptidase. In *Signal Peptidases* (von Heijne, G., ed.), pp. 87-103, R.G. Landes Company, Austin.

Kalousek, F., Hendrick, J. P. & Rosenberg, L. E. (1988). Two mitochondrial matrix proteases act sequentially

in the processing of mammalian matrix enzymes. *Proc. Natl Acad. Sci. USA,* **85**, 7536-7540.
Karlin, S. & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA,* **87**, 2264-2268.
Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **43**, 59-69.
Lee, C. M., Sedman, J., Neupert, W. & Stuart, R. A. (1999). The DNA helicase, Hmi1p, is transported into mitochondria by a C-terminal cleavable targeting signal. *J. Biol. Chem.* **274**, 20937-20942.
Lin, X., Kaul, S., Rounsley, S., Shea, T. P., Benito, M.-I., Town, C. D., Fujii, C. Y., Mason, T., Bowman, C. L., Barnstead, M., Feldblyum, T. V., Buell, C. R., Ketchum, K. A., Lee, J. & Ronning, C. M., *et al.* (1999). Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature,* **402**, 761-768.
Matthews, B. W. (1975). Comparison of predicted and observed secondary structure, of T4 phage lysozyme. *Biochim. Biophys. Acta,* **405**, 442-451.
Minsky, M. & Papert, S. (1968). *Perceptrons: An Introduction to Computational Geometry*, MIT Press, Cambridge, MA.
Nakai, K. & Kanehisa, M. (1992). A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics,* **14**, 897-911.
Nielsen, H. & Krogh, A. (1998). Prediction of signal peptides and signal anchors by a hidden markov model. *ISMB,* **6**, 122-130.
Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**, 1-6.
Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **183**, 63-98.
Pedersen, A. G. & Nielsen, H. (1997). Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. *ISMB,* **5**, 226-233.
Qian, N. & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *J. Mod. Biol.* **202**, 865-884.
Rapoport, T. A. (1992). Transport of proteins across the endoplasmic reticulum membrane. *Science,* **258**, 931-936.
Reinhardt, A. & Hubbard, T. (1998). Using neural networks for prediction of the subcellular location of proteins. *Nucl. Acids Res.* **26**, 2230-2236.
Robinson, C. & Ellis, R. J. (1984). Transport of proteins into chloroplasts. partial purification of a chloroplast protease involved in the processing of important precursor polypeptides. *Eur. J. Biochem.* **142**, 337-342.
Robinson, C., Hynds, P. J., Robinson, D. & Mant, A. (1998). Multiple pathways for the targeting of thylakoid proteins in chloroplasts. *Plant Mol. Biol.* **38**, 209-221.
Roise, D. (1997). Recognition and binding of mitochondrial presequences during the import of proteins into mitochondria. *J. Bioenerg. Biomembr.* **29**, 19-27.
Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning internal representations by error backpropagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol 1: Foundations* (Rumelhart, D., McClelland, J. & Group, P. R., eds), pp. 318-362, MIT Press Cambridge, MA.
Rusch, S. L. & Kendall, D. A. (1995). Protein transport *via* amino-terminal targeting sequences: common themes in diverse systems. *Mol. Membr. Biol.* **12**, 295-307.
Schatz, G. & Dobberstein, B. (1996). Common principles of protein translocation across membranes. *Science,* **271**, 1519-1526.
Schneider, G., Sjöling, S., Wallin, E., Wrede, P., Glaser, E. & von Heijne, G. (1998). Feature-extraction from endopeptidase cleavage sites in mitochondrial targeting peptides. *Proteins: Struct. Funct. Genet.* **30**, 49-60.
Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195-197.
Song, M.-C., Ogishima, T. & Ito, A. (1998). Importance of residues carboxyl terminal relative to the cleavage site in substrates of mitochondrial processing peptidase for their specific recognition and cleavage. *J. Biochem.* **124**, 1045-1049.
Sonnhammer, E. L. L., von Heijne, G. & Krogh, A. (1998). A hidden markov model for predicting transmembrane helices in protein sequences. *ISMB,* **6**, 175-182.
The European Union Arabidopsis Sequencing Consortium & The Cold Spring Harbor, Washington University in St Louis and PE Biosystems Arabidopsis Sequencing Consortium (1999). Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature,* **402**, 769-777.
von Heijne, G. (1983). Patterns of amino acids near signal sequence cleavage sites. *Eur. J. Biochem.* **133**, 17-21.
von Heijne, G. (1985). Signal sequences: the limits of variation. *J. Mol. Biol.* **184**, 99-105.
von Heijne, G. (1990). The signal peptide. *J. Membr. Biol.* **115**, 195-201.
von Heijne, G., Steppuhn, J. & Hermann, S. G. (1989). Domain structure of mitochondrial and chloroplast targeting peptides. *Eur. J. Biochem.* **180**, 535-545.
van Loon, A. P., Brandli, A. W., Pesold-Hurt, B., Blank, D. & Schatz, G. (1987). Transport of proteins to the mitochondrial intermembrane space: the ''matrix-targeting'' and the ''sorting'' domains in the cytochrome c1 presequence. *EMBO J.* **6**, 2433-2439.
Waltner, M. & Weiner, H. (1996). Conversion of a non-processed mitochondrial precursor protein into one that is processed by the mitochondrial processing peptidase. *J. Biol. Chem.* **271**, 21226-21230.

Edited by F. E. Cohen

(Received 17 February 2000; received in revised form 18 May 2000; accepted 22 May 2000)

http://www.academicpress.com/jmb

Supplementary material for this paper, comprising one Table is available from JMB Online