Protein Science (1999), *8*:978–984. Cambridge University Press. Printed in the USA.

ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites

OLOF EMANUELSSON,[1] HENRIK NIELSEN,[1,2] AND GUNNAR VON HEIJNE[1]

[1]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

(RECEIVED October 26, 1998; ACCEPTED January 15, 1999)

Abstract

We present a neural network based method (ChloroP) for identifying chloroplast transit peptides and their cleavage sites. Using cross-validation, 88% of the sequences in our homology reduced training set were correctly classified as transit peptides or nontransit peptides. This performance level is well above that of the publicly available chloroplast localization predictor PSORT. Cleavage sites are predicted using a scoring matrix derived by an automatic motif-finding algorithm. Approximately 60% of the known cleavage sites in our sequence collection were predicted to within ±2 residues from the cleavage sites given in SWISS-PROT. An analysis of 715 *Arabidopsis thaliana* sequences from SWISS-PROT suggests that the ChloroP method should be useful for the identification of putative transit peptides in genome-wide sequence data. The ChloroP predictor is available as a web-server at http://www.cbs.dtu.dk/services/ChloroP/.

Keywords: chloroplast; cleavage site; neural networks; protein sorting; transit peptide

The chloroplast is an organelle present in photosynthetic plants and algae. Even though chloroplasts have a genome of their own, most chloroplast proteins are encoded in the nuclear genome, translated in the cytosol, and post-translationally imported into the organelle. Most nuclearly encoded chloroplast proteins have an N-terminal presequence or transit peptide (cTP) that directs them to the chloroplast stroma (Soll & Tien, 1998). During or shortly after entry, the cTP is cleaved off by the stromal processing peptidase (SPP) (Robinson & Ellis, 1984). Proteins destined for the lumen of the intra-chloroplastic thylakoid compartment generally have a bi-partite targeting sequence composed of an N-terminal stroma-targeting cTP, followed by a thylakoid transfer domain that shares important features with signal sequences required for protein secretion in bacteria (von Heijne, 1990; Robinson et al., 1998).

cTPs from different proteins show a wide variation in length and sequence. They tend to be rich in hydroxylated residues and have a low content of acidic residues (von Heijne et al., 1989). A semi-conserved motif, (I/V)-X-(A/C)↓A, around the SPP cleavage site (arrow) has also been identified (Gavel & von Heijne, 1990). The only publicly available prediction method for automatic identification of cTPs is incorporated in the PSORT server (Nakai & Kanehisa, 1992) and is based mainly on discrimination according to amino acid content in certain sequence regions. PSORT does not provide a cleavage site prediction for cTPs.

Reprint requests to: Department of Biochemistry, Stockholm University, S-106 91 Stockholm, Sweden; e-mail: gunnar@biokemi.su.se.

In this paper we present a neural network-based predictor (ChloroP) that has been trained to discriminate N-terminal cTPs from other N-terminal sequences. Overall, ChloroP can discriminate cTPs from non-cTPs with high sensitivity and specificity. We have also constructed a scoring matrix-based method for prediction of cleavage sites. Interestingly, our analysis indicates that a majority of the cTP-containing proteins are cleaved first by the SPP, whereupon an additional 1–3 residues are removed from the mature protein by some other stromal proteolytic activity. The SPP cleavage site appears to be correctly predicted in about 60% of the sequences.

Results

Collection of cTP sequences

As described in Methods, cTP-containing proteins were extracted from SWISS-PROT. The initial set was screened for thylakoid transfer domains using the signal peptide predictor SignalP (Nielsen et al., 1997), and homology reduction was carried out using the Hobohm algorithm 2 (Hobohm et al., 1992). This left a collection of 80 sequences.

A literature check of this set revealed 11 cTPs where the cleavage site was predicted by homology to proteins with experimentally verified cleavage sites. For 10 of these entries, the experimentally verified sequence was used instead, and for the remaining case it turned out that the verified sequence was already present in the data set, so this entry was removed instead of re-

placed. The literature check also revealed that four proteins were erroneously included in the data set, two being chloroplast encoded, one having a thylakoid transfer domain that had been missed by SignalP, and the fourth being a chloroplast envelope protein. Altogether, five entries were thus removed from the initial set. Finally, for two proteins, the SWISS-PROT annotations of the cleavage sites were inconsistent with the literature, and the cleavage site assignment was changed in accordance with the information in the articles. The final set of positive examples thus included 75 sequences.

Analysis of total amino acid content of the homology reduced cTP set shows that the fraction of acidic residues (Asp and Glu) is very low and that Ser and Arg are over-represented in the cTPs, when compared to the mature part of the proteins (data not shown). This is well in accordance with previously reported cTP features (von Heijne et al., 1989).

The negative set also contained 75 nonhomologous sequences and was constructed from four subsets: mitochondrial, secretory, cytosolic, and nuclear proteins. The total data set of 150 sequences was divided into five subsets before network training, enabling cross-validation.

Neural network training

To discriminate between sequences having and not having a cTP, we proceeded in two steps. First, a neural network was trained to classify individual residues as either belonging or not belonging to a cTP. Second, the output from this first network (Fig. 1) was used as input for a second network that was trained to classify each sequence as either containing or lacking a cTP. A standard feed-forward network architecture with one hidden layer was chosen in both steps.

The final network chosen in the first step had a symmetric input window of 51 positions and 2 units in the hidden layer. Performance rose with the size of the input window, indicating that the network uses global information in its assignment. The number of nodes in the hidden layer played only a minor role, as long as there were at least two hidden nodes. After 200 training cycles, the performances, as measured by Mathews' correlation coefficient (Mathews, 1975), on the five different test sets varied between 0.62 and 0.77 for classification of residues. The rather small size of the homology reduced set could at least in part explain this variation, since the smaller a test set, the more variation in performance would be expected. Figure 1 shows the network output for a well-behaved test example. It also illustrates the need for a second step network, since the first network only gives a classification for each residue in a sequence, and not for the entire sequence.

In the second step, the chosen network architecture had 10 hidden units and the input window size was 100 residues. The correlation coefficient of the second step network for the five test sets varied between 0.68 and 0.93 for classification of sequences, resulting in an overall correlation coefficient of 0.76. Ninety-three percent of the cTPs were found, with 84% of the predicted cTPs being correct, i.e., sensitivity was 0.93 and specificity was 0.84. Eighty-eight percent of the sequences was correctly classified as cTP or non-cTP. Note that these performances are based only on test sequences; i.e., no sequence has been part of both training and test set in the same network run. The classification results in terms of the actual numbers of correctly/incorrectly predicted proteins are presented in Table 1.

Two further tests of ChloroP performance were done. First, among the 29 sequences that were initially removed from the training set because they were considered to represent bi-partite, stroma-thylakoid targeting sequences, 26 (90%) were predicted as having a cTP; this value is similar to the sensitivity obtained above with cross-validation.

Second, to get an idea of how well ChloroP will perform on genome-wide data where only a minority of all sequences represent chloroplast proteins, we analyzed the 715 sequences from *Arabidopsis thaliana* included in SWISS-PROT release 36. As shown in Table 2, 91 out of 95 sequences (96%) annotated as cTP-containing (and having an intact N-terminus) were correctly identified, while only 66 out of 620 sequences (11%) annotated as not having a cTP were falsely predicted to contain one. Plots of

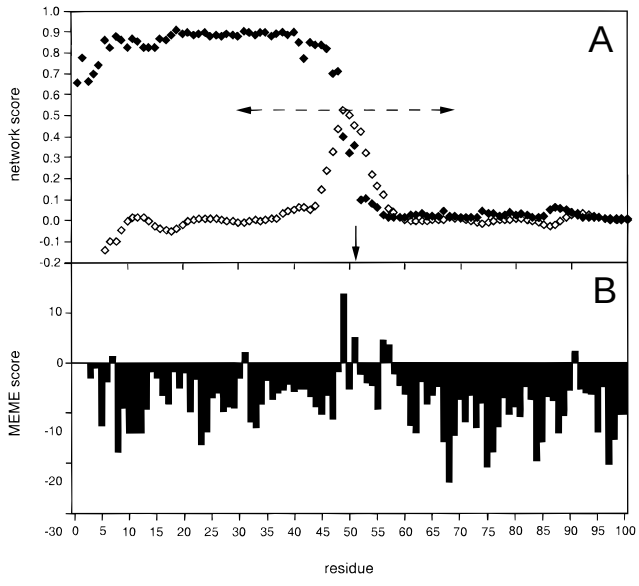

Fig. 1. First step network score (**A**, filled rhombs), "smoothed derivative" score (**A**, open rhombs), and MEME score (**B**) for the first 100 residues of SWISS-PROT entry P12372. The "smoothed derivative" score is calculated according to the formula in Methods. The horizontal, dashed arrow (**A**) corresponds to the 40 residues large window used for cleavage site motif searching. The SWISS-PROT annotated cleavage site is between residues 50 and 51 (arrow, **A**), while the cleavage site predicted by the MEME-based scoring matrix is between residues 48 and 49; the MEME score is defined so that the predicted cleavage site is directly N-terminal of the highest scoring residue (**B**).

Table 1. *ChloroP prediction results on the test set, showing the actual numbers of correctly/incorrectly classified proteins* [a]

Predicted localization	True localization				
	cTP	mTP	Signal	Cytosol	Nuclear
cTP	70	8	3	0	2
Not-cTP	5	12	17	20	13

[a] The positive set consists of chloroplast transit peptides (cTP). The negative set is divided into its four subsets representing different subcellular locations: mitochondrial (mTP), secretory (signal), cytosolic, and nuclear. Note that mTPs are more often falsely predicted as cTPs than the other negative categories.

Table 2. *ChloroP prediction results on 715 Arabidopsis thaliana sequences* [a]

Predicted localization	True localization			
	cTP	mTP	Signal	Other
cTP	91	11	4	51
Not-cTP	4	9	49	496

[a]The 715 sequences were retrieved from SWISS-PROT release 36. The "true localization" is based on the FT field of the SWISS-PROT entry, in the same manner as for the original training and test sets (see Methods). Both experimentally verified and otherwise annotated targeting peptides ("PROBABLE," "POTENTIAL," "BY SIMILARITY") were included in the respective classes.

sensitivity, specificity, and Mathews' correlation coefficient vs. the cutoff for the cTP/non-cTP prediction from the second network are shown in Fig. 2. The best correlation coefficient (0.78) is obtained for a cutoff = 0.52, close to the optimal cutoff (0.50) determined for the original test sets. For a cutoff of 0.52, the sensitivity is 0.93 and the specificity 0.72. As only 14 *A. thaliana* sequences were included in the training set (2 chloroplast and 12 nonchloroplast proteins) and only 11 in the initial collection of 237 cTP sequences before homology reduction, this should be a realistic estimate of ChloroP performance on genome-wide data.

PSORT comparisons

The performance of the final cTP predictor was compared with PSORT (Nakai & Kanehisa, 1992), a knowledge-based predictor that calculates the probability (or "certainty factors") that a protein belongs to any of a wide array of subcellular locations, among them the chloroplast. The four most probable locations for each protein are presented by PSORT. In our comparisons, we considered that a protein was assigned as a chloroplast protein if the chloroplast localization was among the four presented locations and had a certainty factor ≥0.200, since this turned out to give the best results among the interpretations tested. If these two demands were not fulfilled, the protein was considered to be predicted as a nonchloroplast protein. Using the same 150 sequence data set as was used in the neural network training, but now including the full amino acid sequences and interpreting the output as stated above, PSORT yielded a correlation coefficient of 0.47, with a sensitivity of 0.69 and specificity of 0.75. Another way of interpreting the PSORT output is to consider the first presented localization as the assigned location. This approach yielded a correlation coefficient of 0.43. Both ways of interpreting PSORT output are thus outscored by the ChloroP predictor.

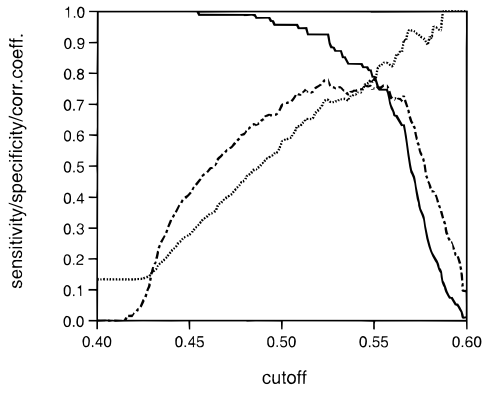

Fig. 2. Sensitivity (solid line), specificity (dotted line), and Mathews' correlation coefficient (dashed-dotted line) as a function of the cutoff in the second network for cTP/non-cTP determination for a set of 715 *A. thaliana* sequences. Lowering the cutoff yields a less restrictive prediction, suitable in a search for all possible cTP-candidates in genome-wide data, while raising the cutoff could be useful when priority is on minimizing the number of false positives in database annotation.

Cleavage site predictions

Our initial attempts to train a neural network to recognize cTP cleavage sites were unsuccessful. A recently published study of the in vitro cleavage specificity of SPP suggested a plausible explanation for this failure (Richter & Lamppa, 1998). While shown in Figure 3A, cTP cleavage sites listed in SWISS-PROT tend to have Arg in positions −2 and −3, five out of six precursor proteins tested in the in vitro cleavage assay were processed between Arg/Lys and Ala. In at least one of these cases, the mature protein isolated from chloroplasts lacks an additional residue from the N-terminus, suggesting that an uncharacterized stromal protease can remove one or a few N-terminal residues after the initial cleavage catalyzed by SPP. For this reason, most cleavage sites given in SWISS-PROT probably do not accurately represent the initial SPP cleavage site.

To circumvent this problem, we used the MEME motif-finding algorithm (Bailey & Elkan, 1994) to automatically detect the most conserved motif in the −20 to +6 region of a curated set of 62 cTPs (the cleavage site given in SWISS-PROT for 13 of the 75 cTPs used in the neural network training could not be confirmed in the literature and these entries were not used). Since MEME aligns its input sequences, it is capable of finding the most conserved motif in a set of unaligned sequences without prior knowledge of the exact position of the motif in each sequence. MEME also generates a log-odds scoring matrix for the motif, and this scoring matrix was used as a cleavage site predictor. The MEME scoring matrix generated for two different values of the MEME parameter "motif window length" ("short" and "8") yielded approximately the same prediction results. For the final prediction, we chose the value "8"; the consensus of the corresponding motif found by MEME was VR↓AAAVxx, where the SPP cleavage site suggested by the in vitro results is denoted by an arrow.

The output from the first neural network was used to approximately locate the cTP cleavage site by calculating a "smoothed derivative" of the network output curve and then requiring that the predicted cleavage site be within ±20 residues of the maximum of the smoothed derivative (Fig. 1). The position in this region with the maximum score calculated with the MEME log-odds scoring matrix was then taken as the predicted cleavage site.

As seen in Figure 4, only 3 of the 62 sequences had a predicted cleavage site located downstream of the cleavage site given in SWISS-PROT, 5 had a predicted cleavage site that coincided with the one given in SWISS-PROT, whereas one-half (31 sequences) had a predicted cleavage site located one or two residues upstream of the SWISS-PROT site. Assuming that all predictions that are at most two residues away from the site given in SWISS-PROT are the correct SPP sites, the simple combination of the neural network

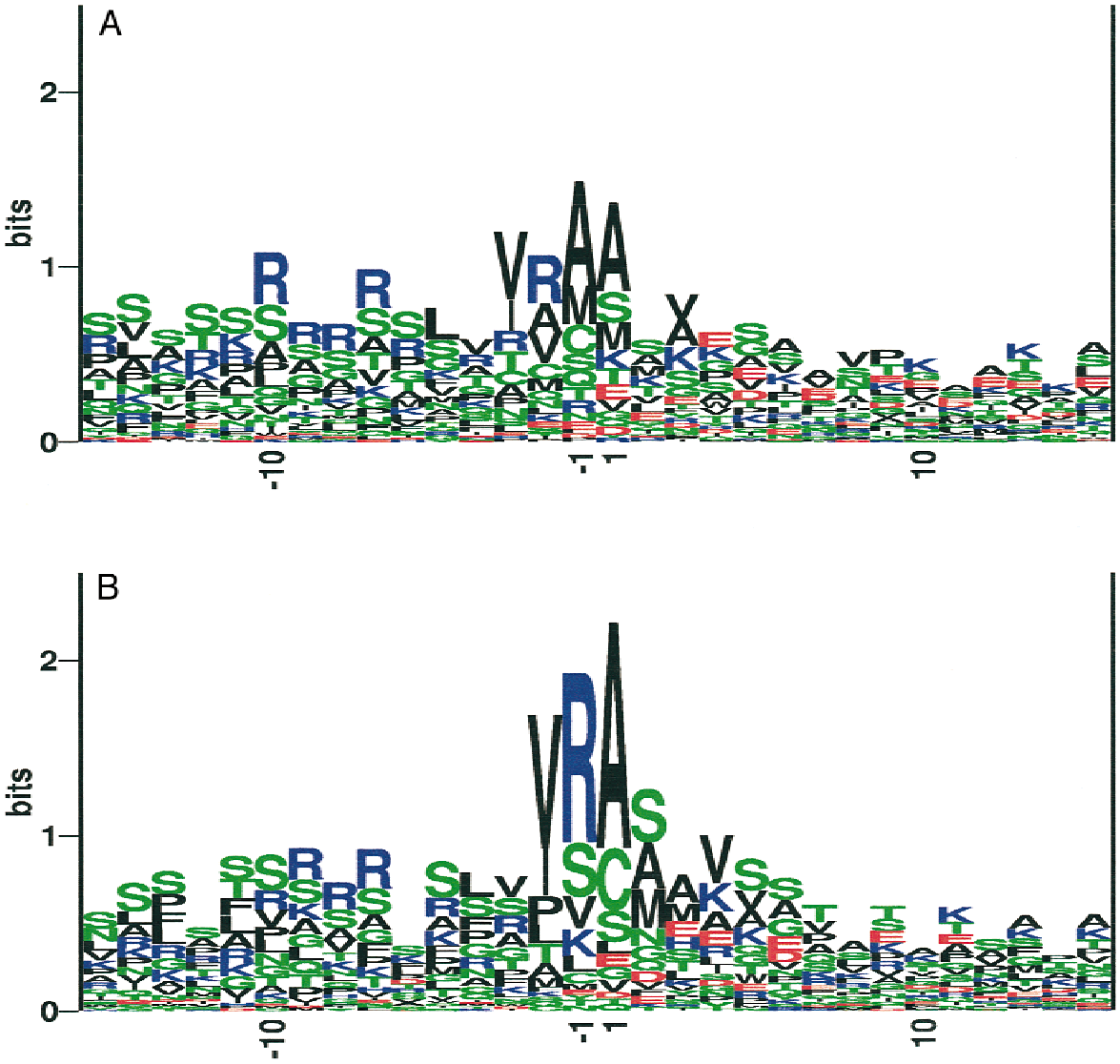

Fig. 3. Sequence logos constructed from the 62 sequences used in the cleavage site predictor development. (**A**) The sequences are aligned around their SWISS-PROT annotated cleavage site and (**B**) around the predicted cleavage site. The context of the predicted cleavage sites shows a higher degree of conservation than the annotated cleavage sites. Residue numbering is relative to the cleavage sites. Positively and negatively charged residues are shown in blue and red, uncharged, polar residues are shown in green, and hydrophobic residues in black.

smoothed derivative to approximately locate the cleavage site, and the MEME log-odds matrix thus gives around 60% correct predictions. Finally, it is worth noting that the alignment patterns, as shown in Figure 3, differ clearly between the SWISS-PROT annotated cleavage site (Fig. 3A) and the predicted cleavage site (Fig. 3B), showing that the context of the predicted cleavage sites is considerably more conserved than that of the annotated sites.

Discussion

Using a neural network approach, we have developed a predictor, ChloroP, that can discriminate between cTPs and non-cTPs with a sensitivity of 0.93 and a specificity of 0.84 on a homology-reduced test set. On a more demanding test set in which chloroplast proteins are a small minority (715 *A. thaliana* sequences), ChloroP yields a sensitivity of around 0.9 and a specificity of 0.6 (Table 2). Depending on the application, sensitivity/specificity may be balanced against each other by changing the threshold for the cTP/non-cTP discrimination (Fig. 2). Furthermore, by adding a simple scoring matrix, derived automatically from a set of approximately known SPP cleavage sites, on top of the neural network predictor, the correct SPP cleavage site can be predicted in approximately 60% of the cases. These performance levels are significantly better than those achieved prior to the PSORT predictor (Nakai & Kanehisa, 1992), and indicate that the ChloroP method should be useful for screening large sequence sets for putative chloroplast proteins.

A remaining problem is that the ChloroP predictor cannot discriminate very efficiently between cTPs and mitochondrial target-

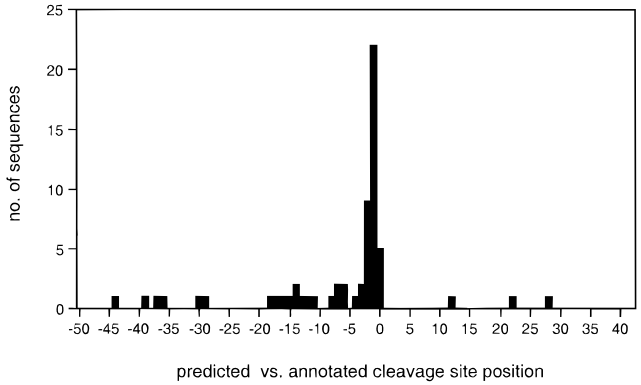

Fig. 4. Cleavage site prediction results using the MEME scoring matrix compared to SWISS-PROT annotations—54 out of 62 examined proteins are predicted to have shorter cTPs than what is annotated in SWISS-PROT (negative values), 5 are correctly predicted, and only 3 predicted to be longer (positive values).

ing sequences (mTPs) (see Tables 1, 2). In fact, these two kinds of sorting signals are sufficiently similar that a few exceptional cases are known, where one and the same sorting signal can route a passenger protein to both chloroplasts and mitochondria with similar efficiencies (Creissen et al., 1995; Chow et al., 1997; Akashi et al., 1998; Menand et al., 1998). Such dual targeting events are probably quite rare, however, and it may be possible to improve the discrimination between cTPs and mTPs by the simultaneous use of a cTP-specific and an mTP-specific neural network; such studies are underway.

It is striking that only 3 out of 62 sequences were predicted to have a longer cTP than what is stated in their SWISS-PROT entries, while 54 were assigned a shorter cTP and only 5 were "correctly" predicted (Fig. 4). The tendency to predict the cTPs as being shorter than their SWISS-PROT annotations does not seem to be an artifact of the asymmetric -20 to $+6$ window used to train MEME, since it was reproduced in MEME trainings using symmetric windows of ± 10 or ± 20 residues around the annotated cleavage site. The prediction performance was, however, slightly worse for the ± 10 window and significantly worse for the ± 20 window (data not shown). Given the recent experimental observation that cTPs are cleaved between Arg/Lys and Ala by purified SPP in vitro (Richter & Lamppa, 1998) and the fact that many cTP cleavage sites given in SWISS-PROT have Arg in positions -2 or -3 (Fig. 3A), it seems likely that most imported chloroplast proteins undergo additional N-terminal proteolysis after the initial SPP-catalyzed cleavage of the cTP. This unfortunately makes it impossible to accurately predict the final N-terminus of the mature protein, but our results nevertheless suggest that the ChloroP prediction will not be off by more than two residues from the mature N-terminus in approximately 60% of the cases.

Methods

Creation of training and test sets

Sequence data were obtained from SWISS-PROT release 35 (Bairoch & Apweiler, 1998). The positive (cTP) data set was extracted by collecting all entries containing an FT line with key name "TRANSIT" and description "CHLOROPLAST." Entries that showed any sign of ambiguity regarding the existence or length of the cTP ("PROBABLE," "POTENTIAL," "BY SIMILARITY") were excluded, leaving 266 sequences. The collected sequences were then analyzed with the signal sequence predictor SignalP (using the prokaryotic, gram-negative network) (Nielsen et al., 1997), and those that were assigned a cleavage site within ± 5 residues from the SWISS-PROT annotated cleavage site were excluded, since they most likely represent bi-partite stroma-thylakoid targeting sequences. The cTP set was thus reduced to 237 sequences.

The cTP parts (according to the SWISS-PROT annotation) of the remaining sequences were pairwise aligned using the Smith–Waterman algorithm and PAM250 scoring matrix; homology reduction was done using Hobohm algorithm 2 (Hobohm et al., 1992). The cutoff in the Hobohm algorithm was based on the extreme value distribution (Karlin & Altschul, 1990; Altschul et al., 1994), and chosen as the score at which the actual distribution of pairwise alignment scores deviates from the theoretical distribution of scores for pairwise alignments of randomly generated sequences (Pedersen & Nielsen, 1997). The total number of cTPs left after this procedure was 80, with an average cTP length of 50 amino acids. Careful sequence homology reduction is of great importance when the sequences are to be used for training a neural network, since too strong a similarity between sequences in the training and test sets could lead to an overestimation of prediction performance. Furthermore, the existence of too similar sequences in the training set may lead to a biased network, being better at recognizing cTPs of the over-represented type, but worse at recognizing all other types.

After a literature check of the 80 sequences, 5 were removed, and another 12 were replaced or corrected due to annotation inconsistencies or errors (see Results for discussion of these proteins). In addition to the cTP, 50 residues from the N-terminus of the mature protein were retained for each entry, yielding an average entry length of 100 amino acids. The final cTP collection of 75 proteins can be downloaded from the ChloroP web site.

The negative set was constructed in a similar way. Four homology-reduced subsets containing cytosolic, secretory, mitochondrial, and nuclear proteins, respectively, were collected and from each of them 20 sequences (15 from the nuclear subset) were included in the negative set so that the final positive and negative sets were of the same size. All proteins were taken from plants (as indicated by the node "PLANTA" in the OC line), except for the mitochondrial sequences, which were from all kinds of organisms since the homology reduced mitochondrial subset would otherwise have been too small. Such a nonrestrictive choice of mitochondrial sequences is justified by an earlier study showing that mitochondrial targeting peptides do not differ in any major way between organisms (Schneider et al., 1998). One hundred N-terminal residues were retained for each entry, equaling the average length in the positive set.

Before network training, the full dataset (150 sequences) was divided into five equally sized parts for cross-validation. Every network run was carried out with one part as test set and the remaining four as training sets, and this was repeated so that all five parts were used both for testing and training, but not both in the same run.

Neural network training

To construct the prediction method, two neural networks were used sequentially. Both were of the feed-forward type with sigmoidal

neurons (Minsky & Papert, 1968) and zero or one layer of hidden units, trained using error backpropagation (Rumelhart et al., 1986), but with different error functions.

The first neural network was based on the HOW package (S. Brunak, pers. comm.), used earlier (Brunak et al., 1991). It uses a logarithmic error function in the backpropagation algorithm and sparsely encoded sliding windows for encoding the input sequence data (Qian & Sejnowski, 1988; Brunak et al., 1991). Each position in the input sequence window occupies 20 input nodes (1 for each of the 20 amino acids). The node corresponding to the amino acid present at that position is switched on while the other 19 remain off. The network has one output unit that is trained to predict the state (cTP or non-cTP) of the central residue of the input window. The most N-terminal window has the most N-terminal residue in its central position. The input nodes belonging to the positions in the window thus not covered by the amino acid sequence remain off. The window is slid along the sequence, enabling the network to be trained on one residue and its environment at a time. Networks with window sizes from 7 to 51 positions and 0 to 8 nodes in the hidden layer were tested. The learning rate was set to 0.001, based on pilot studies (data not shown). The output of this first network is one score per amino acid in a sequence (Fig. 1).

The output from the first network was used as input to a second network that was based on the HOWLIN program (S. Brunak, pers. comm.). It uses real values as input and the standard error function in the backpropagation. The input was fixed to the output values from the first network corresponding to the N-terminal 100 positions in a sequence, a number chosen to be large enough to span the entire cTP for all entries in the cTP set used in this study. The network has one output unit that is trained to predict the presence or absence of a cTP in a sequence. The number of nodes in the hidden layer was varied between 0 and 20 and three different learning rates (0.001, 0.01, and 0.05) were tested; 0.001 was chosen for the final training. The output from the second network consists of one score per protein and the actual chloroplast/nonchloroplast localization assignment of the protein is based on whether this score is above or below a cutoff = 0.50, a value that yields the optimal Mathews' correlation coefficient over the test sets.

A common problem when using neural networks is to avoid overtraining, i.e., a decline in generalization ability, which often occurs after a certain point during training. In several neural network applications, this has been handled by monitoring test set performance during training and picking the network where performance on the test set was optimal (Qian & Sejnowski, 1988; Brunak et al., 1991; Nielsen et al., 1997). This approach has been criticized because it involves the test set for optimization of training length, so the performance might not reflect a true generalization ability. Although practical experience in a bioinformatics application has shown the performance on a new, independent test set to be as good as that found on the data set used to stop the training (Brunak et al., 1991), we have chosen to avoid optimization on the individual test sets by using a constant training length for all training sets in the cross-validation. A training length of 200 cycles was chosen for both types of network, based on a series of initial trial runs where the training and test sets performances were monitored to find a value where overtraining rarely occurred although near optimal test performance was reached. With the low learning rate we used, fluctuations in performances during training were small, and the exact choice of training length was not critical.

Test set performances were measured using Mathews' correlation coefficient (Mathews, 1975). The final network architectures were in both steps chosen as the best performing one, but with the complexity of the network taken into account, i.e., if two trained networks performed equally well, the least complex network architecture was chosen. Since there were five different training sets, five different networks were produced in both steps. In the final, web-accessible ChloroP predictor, a query sequence is processed through all five networks of the first step, and the resulting output values are averaged before they are presented as input to the five networks of the second step. The outputs from these five networks are also averaged, yielding a final score on which the cTP/non-cTP assignment is based. Both the actual prediction and the final network output score are presented to the user.

Cleavage site prediction

A scoring matrix based approach was used for cleavage site prediction. Since, as discussed in Results, the cleavage sites given in SWISS-PROT may not correspond exactly to the initial SPP cleavage sites, we used MEME (Bailey & Elkan, 1994), a web accessible tool for motif discovery, to identify the most conserved motif in a segment encompassing residues -20 to $+6$ relative to the SWISS-PROT cleavage sites and to generate a log-odds scoring matrix for the motif found. Only 62 sequences from the cTP set were used since the cleavage site given in SWISS-PROT could not be confirmed in the literature for 13 sequences, i.e., the mature part of these proteins was not N-terminally sequenced. Several runs were undertaken, varying the MEME "motif window length" parameter. Cleavage sites were predicted by first constructing a "smoothed derivative," ΔS, of the output profile from the first network and then searching for the maximum log-odds score in a window of ± 20 residues surrounding the position of the maximum in ΔS (Fig. 1). ΔS for position i was calculated according to the following formula, where S is the network output score:

$$\Delta S_i = \frac{1}{5}\left(\sum_{j=1}^{5} S_{i-j} - \sum_{j=0}^{5} S_{i+j}\right). \quad (1)$$

Logos

Sequence logos (Schneider & Stephens, 1990) were constructed for alignments of the 62 proteins around their annotated and predicted cleavage sites, respectively (Fig. 3). For each position in an alignment the frequencies of the amino acids present at that position are calculated, and the information content, as measured by the difference between maximum and actual Shannon entropy (Shannon, 1948), is represented by the height of the bars in the plot. The height of the letters within each bar represents the relative frequency of the corresponding amino acid at that position.

Acknowledgments

This work was supported by grants from the Swedish Natural and Technical Sciences Research Councils to GvH. HN was supported by a grant from the Danish National Research Foundation. We would like to thank Dr. Søren Brunak, The Technical University of Denmark, for help with the HOW package programs, Dr. Anders Krogh, The Technical University of Denmark, for valuable discussions concerning neural networks, and Drs. Kenneth Keegstra and John Froelich, Michigan State University, for help during the initial phase of the project.

References

Akashi K, Grandjean O, Small I. 1998. Potential dual targeting of an Arabidopsis archaebacterial-like histidyl-tRNA synthetase to mitochondria and chloroplasts. *FEBS Lett 431*:39–44.

Altschul S, Boguski MS, Gish W, Wootton JC. 1994. Issues in searching molecular sequence databases. *Nature Genetics 6*:119–129.

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *ISMB 2*:28–36.

Bairoch A, Apweiler R. 1998. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucleic Acids Res 26*:38–42.

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.

Chow KS, Singh P, Roper J, Smith A. 1997. A single precursor protein for ferrochelatase-I from Arabidopsis is imported in vitro into both chloroplasts and mitochondria. *J Biol Chem 272*:27565–27571.

Creissen G, Reynolds H, Xue YB, Mullineaux P. 1995. Simultaneous targeting of pea glutathione reductase and of a bacterial fusion protein to chloroplasts and mitochondria in transgenic tobacco. *Plant J 8*:167–175.

Gavel Y, von Heijne G. 1990. A conserved cleavage site motif in chloroplast transit peptides. *FEBS Lett 261*:455–458.

Hobohm U, Scharf M, Schneider R, Sander C. 1992. Selection of representative protein data sets. *Protein Sci 1*:409–417.

Karlin S, Altschul SF. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA 87*:2264–2268.

Mathews BW. 1975. Comparison of predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta 405*:442–451.

Menand B, Maréchal-Drouard L, Sakamoto W, Dietrich A, Wintz H. 1998. A single gene of chloroplast origin codes for mitochondrial and chloroplastic methionyl-tRNA synthetase in *Arabidopsis thaliana*. *Proc Natl Acad Sci USA 95*:11014–11019.

Minsky M, Papert S. 1968. *Perceptrons: An introduction to computational geometry*. Cambridge, Massachusetts: MIT Press.

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

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.

Pedersen AG, 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 TJ. 1988. Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol 202*:865–884.

Richter S, Lamppa GK. 1998. A chloroplast processing enzyme functions as the general stromal processing peptidase. *Proc Natl Acad Sci USA 95*:7463–7468.

Robinson C, Ellis RJ. 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 PJ, Robinson D, Mant A. 1998. Multiple pathways for the targeting of thylakoid proteins in chloroplasts. *Plant Mol Biol 38*:209–221.

Rumelhart DE, Hinton GE, Williams RJ. 1986. Learning internal representations by error backpropagation. In: Rumelhart D, McClelland J, Group PDP Research, eds. *Parallel distributed processing: Explorations in the microstructure of cognition, vol 1: Foundations*. Cambridge, Massachusetts: MIT Press. pp 318–362.

Schneider G, Sjöling S, Wallin E, Wrede P, Glazier E, von Heijne G. 1998. Feature-extraction from endopeptidase cleavage sites in mitochondrial targeting peptides. *Proteins 30*:49–60.

Schneider TD, Stephens RM. 1990. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res 18*:6097–6100.

Shannon CE. 1948. A mathematical theory of communication. *Bell System Tech J 27*:379–423, 623–656.

Soll J, Tien R. 1998. Protein translocation into and across the chloroplastic envelope membranes. *Plant Mol Biol 38*:191–207.

von Heijne G. 1990. The signal peptide. *J Membr Biol 115*:195–201.

von Heijne G, Steppuhn J, Herman SG. 1989. Domain structure of mitochondrial and chloroplast targeting peptides. *Eur J Biochem 180*:535–545.