LumenP—A neural network predictor for protein localization in the thylakoid lumen

ISABELLE WESTERLUND, GUNNAR VON HEIJNE, AND OLOF EMANUELSSON

Stockholm Bioinformatics Center, and Department of Biochemistry and Biophysics, Stockholm University,
SE-106 91 Stockholm, Sweden

(RECEIVED February 13, 2003; FINAL REVISION June 19, 2003; ACCEPTED July 11, 2003)

Abstract

We report the development of LumenP, a new neural network-based predictor for the identification of proteins targeted to the thylakoid lumen of plant chloroplasts and prediction of their cleavage sites. When used together with the previously developed TargetP predictor, LumenP reaches a significantly better performance than what has been recorded for previous attempts at predicting thylakoid lumen location, mostly due to a lower false positive rate. The combination of TargetP and LumenP predicts around 1.5%–3% of all proteins encoded in the genomes of *Arabidopsis thaliana* and *Oryza sativa* to be located in the lumen of the thylakoid.

Keywords: Thylakoid lumen; protein sorting; chloroplast; prediction

Supplemental material: See www.proteinscience.org

From the point of view of protein sorting, the chloroplast of higher plants presents a complex problem where nuclearly encoded proteins must not only be targeted to the organelle, but must further be sorted between at least six different suborganellar compartments: the outer envelope membrane, the interenvelope space, the inner envelope membrane, the stroma, the thylakoid membrane, and the thylakoid lumen. Targeting to the stromal compartment depends on an N-terminal chloroplast transit peptide (cTP). Further transport to the thylakoid lumen depends on a lumenal targeting peptide (lTP) that is located immediately downstream of the cTP in the nascent polypeptide (Robinson et al. 1998), and that is in most cases cleaved off from the nascent chain by the thylakoid processing peptidase (Halpin et al. 1989). Sorting signals for the three envelope compartments are not well understood.

Prediction of the subcellular localization of a protein from its amino acid sequence is an important area in bioinformatics (Emanuelsson and von Heijne 2001). One approach to this problem is to try to emulate the cellular process of sorting signal recognition. One of the more widely used methods of this kind is TargetP (Emanuelsson et al. 2000), a neural network-based predictor that assigns proteins to four different locations: the secretory pathway, mitochondria, chloroplasts, and "all other compartments." An attempt at a fully comprehensive prediction scheme is PSORT1 (Nakai and Kanehisa 1992) that distinguishes between no less than 17 compartments when applied to plant proteins.

The recognition of cTPs is already a part of TargetP, but the program does not yet include a routine for predicting lTPs. Because lTPs are quite similar to the signal peptides that target proteins for secretion in bacteria, one way to identify lTPs is to use TargetP to first search for cTPs, followed by a search for signal peptides using the SignalP predictor (Nielsen et al. 1997; Nielsen and Krogh 1998). Such an approach has been used with some success, for example, by Peltier et al. (2002) and Schubert et al. (2002). However, we assumed that performance would be even better with a dedicated lTP predictor trained on a proper lTP data set.

Here, we report such a predictor—LumenP—that has been constructed in a similar way as the existing components of TargetP. When coupled with TargetP, LumenP al-

Reprint requests to: Gunnar von Heijne, Stockholm Bioinformatics Center, and Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden; e-mail: gunnar@dbb.su.se; fax: 46-8-15 36 79.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.0306003.

lows proteins of the thylakoid lumen to be identified with high confidence.

Results

Training sets and neural network architecture

As described in Materials and Methods, an initial data set of 259 lTP-containing proteins was collected from Swiss-Prot, from the literature, and from experimental analyses of the lumenal proteomes of *Pisum sativum* (Peltier et al. 2000; Schubert et al. 2002) and *Arabidopsis thaliana* (Peltier et al. 2002; Schubert et al. 2002) thylakoids. These sequences were further classified according to whether or not the lTP contained the diagnostic "twin-arginine" signal that is found in proteins that are imported into the thylakoid via the TAT-pathway (twin-arginine translocation pathway; Berks et al. 2000). One hundred thirty eight such proteins were identified, and the remaining 121 proteins were assumed to be imported using the Sec pathway.

Based on the observation that only 4% of the combined cTP+lTP signals were longer than 130 residues, only the N-terminal 130 residues were analyzed for each protein. The 138 proteins in the TAT group and the 121 proteins in the Sec group were treated as separate positive training sets, and each set was redundancy-reduced as described in Materials and Methods. After this step, 50 nonhomologous sequences were left in the TAT group and 43 in the Sec group. Because the precise extent of the lTP signal is in general not known for these sequences, a stretch of 35 residues upstream of the lTP cleavage site (determined by experiments if available, otherwise by similarity) was annotated as belonging to the lTP for the neural network training procedure. A redundancy-reduced negative training set of 50 130-residues long nonthylakoid sequences (10 stromal proteins, 10 mitochondrial proteins, 10 nuclear proteins, 10 secreted proteins, and 10 cytosolic proteins) was also collected.

For both the TAT and Sec training sets, two networks—one on top of the other—were trained in the same way as has been done previously for the SignalP (Nielsen et al. 1997) and ChloroP (Emanuelsson et al. 1999) predictors. The first network was trained with the amino acid sequence as input and a sliding window of size 35 residues. The output of the first networks is one score per residue, giving the probability that this residue is part of an lTP. The output values for residues 21 to 110 for each protein were then used as input to a second network with 90 input nodes. The output of the second network is one score per protein, giving the probability that the protein has an lTP. The networks were trained using fivefold cross-validation.

Prediction of lTP cleavage site

An important part of the predictor is a scoring-matrix-based method for predicting the cleavage site location of the lumenal targeting peptide. Focusing exclusively on the region around the cleavage site we pooled the TAT and Sec datasets because these signals are assumed to be cleaved by the same protease. Thus, the entire set of 93 redundancy reduced lumenal sequences were used in the construction of the scoring matrix. First, the proteins were aligned (without gaps) around their cleavage sites, and eight alignment positions were extracted—six from the lTP and two from the mature part of the protein, thus covering the c-region of the signal sequence (von Heijne 1983). Then, the cleavage site scoring matrix was constructed by recording for each position i in the alignment, the frequencies $f_{i,j}$ of each amino acid j, and contrasting this frequency with the frequency p_j of that amino acid in a background set (see Materials and Methods). The resulting scoring matrix can be used to scan candidate sequences for the most probable cleavage site. The search is limited to the sequence region comprising residues 50–150 (this includes all known cTP+lTP lengths).

The SignalP predictor can also be used to predict lTP cleavage sites. In this case, we employed a truncation scheme (described below) resulting in 26 suggested cleavage sites per sequence, and the one with the highest cleavage site score was chosen as the final prediction. We also compared our results with those obtained using an older scoring-matrix method that was specifically designed to predict lTP cleavage sites (Howe and Wallace 1990).

Performance tests and comparisons

The results of different combinations of test sets analyzed by one or both of the Sec- and TAT-trained neural networks are shown in Table 1. Not surprisingly, the Sec and TAT networks perform best on their respective set of sequences. For a cutoff score of 0.67, both reach a Matthews' correlation coefficient (MCC) of 0.74, with the corresponding sensitivities and specificities in the range 0.8–0.9.

Because it is not known beforehand whether a particular protein uses the TAT or the Sec pathway, we also analyzed

Table 1. *Scores from different neural network combinations for the redundancy reduced TAT and Sec test sets of thylakoid lumen proteins and the common negative training set (cleavage site scores not taken into account)*

Test set	Neural network used	MCC	Sensitivity	Specificity
Sec + negative	Sec	0.74	0.84	0.88
	Sec+TAT	0.64	0.84	0.78
TAT + negative	TAT	0.74	0.90	0.85
	Sec+TAT	0.70	0.90	0.82
Sec + TAT + negative	Sec+TAT	0.67	0.87	0.89

When the Sec- and TAT-trained networks were used in parallel (Sec+TAT), the highest of the two output scores was chosen. The cutoff score used for discriminating between sequences with and without an lTP was 0.67 (see Fig. 1A).

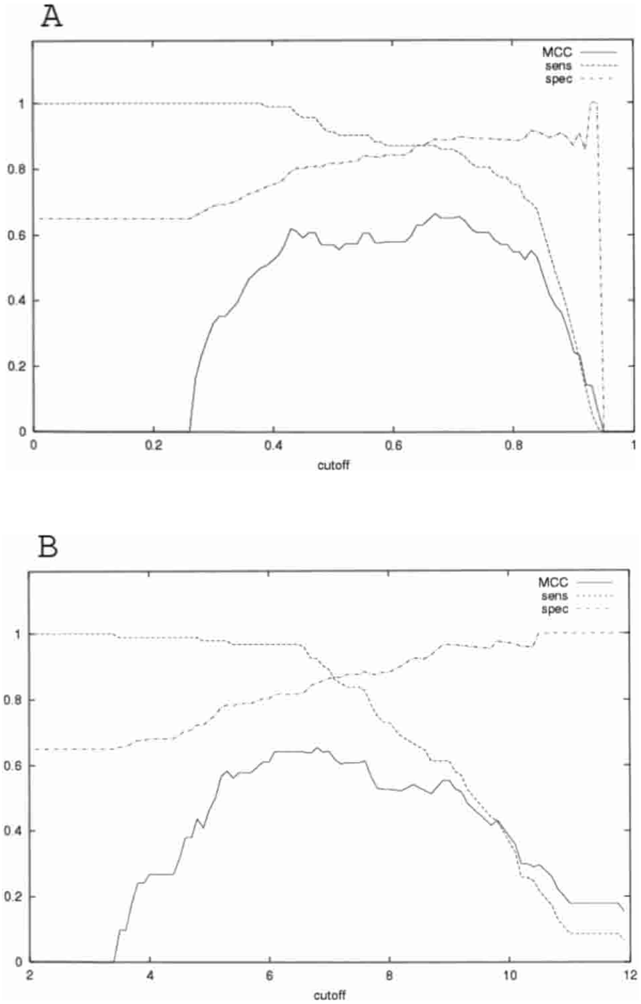

Figure 1. Performance characteristics for different cutoff values using the Sec+TAT-positive dataset and the mixed negative data set together with the Sec+TAT networks (*A*) and using the scoring-matrix constructed for prediction of cleavage sites as a discriminator lumenal/nonlumenal (*B*). Matthews' correlation coefficient, sensitivity, and specificity are plotted against the score cutoff used to discriminate between lTP and non-lTP sequences. *MCC*, Matthews' correlation coefficient; *sens*, sensitivity; *spec*, specificity.

each sequence by both the Sec- and TAT-trained networks, choosing the highest output score from the two networks. For the cutoff score of 0.67, the MCC value is 0.67, while the specificity and sensitivity are still mainly in the range 0.8–0.9 (Fig. 1A). In Figure 1B, the corresponding values when using only the cleavage site score for discrimination are shown. Clearly, there is a prediction power also in the cleavage site scores. Thus, for the LumenP predictions, a combination of the network score and the cleavage site score was used for the discrimination lumenal/nonlumenal. From an investigation of all possible cutoff combinations, we concluded that a sequence should be required to have a network score above 0.47 and a cleavage site score above 6.80 to be predicted as lumenal (data not shown). If not stated otherwise, these cutoffs were used throughout this report. Choosing other cutoffs is, of course, possible, and would result in a changed sensitivity/specificity balance. The exact choice of cutoffs was not critical to the performance on the test set.

In Table 2, the results from a comparison between LumenP, different versions of SignalP (Nielsen et al. 1997; Nielsen and Krogh 1998), and PSORT1 (Nakai and Kanehisa 1992) on the LumenP redundancy reduced test set is shown. TargetP (Emanuelsson et al. 2000) was used to preselect proteins with a predicted cTP. When combined with TargetP, LumenP has an MCC value of 0.72, a sensitivity of 0.82, and a specificity of 0.96. For the TargetP+SignalP analysis, we used the same truncation scheme as in Peltier et al. (2002), that is, for each protein with a predicted cTP, 20–80 residues were removed (in steps of five residues) from the N-terminus, and the protein was predicted as lumenal if at least one of these truncated forms was predicted as lumenal (using both Gram-negative and Gram-positive versions of SignalP, resulting in 26 predictions per protein). The combination of TargetP and the HMM version of SignalP performed slightly better than the crossvalidated TargetP+LumenP predictor (Table 2), while using the NN version of SignalP resulted in performance levels on par with TargetP+LumenP. PSORT1 had a much lower sensitivity than both TargetP+LumenP and TargetP+SignalP.

Because the redundancy reduced test set used in the construction of the predictor is relatively limited in size, there was a need to further test the prediction accuracies of the various predictors. Specifically, because the ratio of the number of positive (93) and negative (50) sequences in the test set is far from the ratio expected in a genome-wide scan, we paid special attention to the tendency of falsely predicting nonthylakoidal sequences as thylakoidal (overprediction), because such a feature would not be captured very well by the test set.

All *Plantae* sequences from Swiss-Prot annotated as containing a secretory signal peptide were analyzed (745 se-

Table 2. *LumenP, SignalP, and PSORT1 predictions on the redundancy reduced test set*

Predictor	MCC	Sensitivity	Specificity
LumenP	0.67	0.87	0.89
TargetP + LumenP	0.72	0.82	0.96
TargetP + SignalP (NN)	0.71	0.89	0.90
TargetP + SignalP (HMM)	0.76	0.89	0.93
PSORT1	0.37	0.39	0.95

For LumenP, fivefold crossvalidation was used, and for the TargetP+LumenP results, the effects of cleavage site prediction were also incorporated, as described in the text. For SignalP, the truncation scheme (see text) was used. PSORT1 performance was evaluated based on highest ranked prediction.

Table 3. *Estimates of the false positive rates for thylakoid lumenal protein prediction using four sets collected from Swiss-Prot: all* Plantate *sequences annotated as containing a signal peptide (SP); all* Plantae *sequences annotated as being stromal; all proteins (with or without localization annotation) not annotated as thylakoid lumen from* Arabidopsis thaliana; *and all proteins (with or without localization annotation) not annotated as thylakoid lumen from* Oryza sativa

Set	Number of sequences	TargetP+LumenP	TargetP+SignalP (NN)	TargetP+SignalP (HMM)	PSORT1
Plantae SP	745	3 (0.4%)	8 (1.1%)	5 (0.7%)	0 (0%)
Plantae chloroplast stroma	878	165 (18.8%)	515 (58.7%)	253 (28.8%)	43 (4.9%)
Arabidopsis not thylakoid	1034	61 (5.9%)	169 (16.3%)	81 (7.8%)	12 (1.2%)
Oryza not thylakoid	402	5 (1.2%)	40 (10.0%)	17 (4.2%)	5 (1.2%)

For SignalP, the truncation scheme (see text) was used.

quences) as well as a set of 878 plant sequences annotated as being located in the chloroplast stroma. As shown in Table 3, LumenP (with the TargetP preselection step) predicted only 0.4% of the signal peptides as being lTPs, while the number of false positives from the stromal set was 18.8%. The combination of TargetP and SignalP, however, predicted a much higher fraction of the stromal proteins as lumenal, 28.8% (HMM) or 58.7% (NN). Table 3 also shows the results of analyses of all *Arabidopsis* and *Oryza* proteins present in Swiss-Prot that were not annotated as being thylakoid lumenal. Again, LumenP performed better than SignalP-HMM, and they both performed significantly better than the SignalP-NN. The PSORT1 results support the observation that PSORT1 is very conservative in its prediction of lumenal proteins.

From these tests, we conclude that the major advantage of the LumenP predictor is the reduced number of chloroplast stromal proteins falsely predicted as being lumenal, while still retaining a high level of sensitivity compared to the SignalP-based approaches. A way to further decrease the number of false positives is to focus on the subset of proteins with the TAT pathway motif and require the presence of the TAT motif (Arg-Arg) in a specified region upstream of the predicted cleavage site. We looked for the motif in the sequence region −18 to −32 relative to the cleavage site (Peltier et al. 2002) and found that the number of false positives were reduced for LumenP and SignalP in approximately equal amounts: 21% and 25% (for LumenP and SignalP, respectively) of the true stromal proteins predicted as false positives were still predicted as lumenal when in addition to the results from the predictors also demanding the presence of the TAT motif.

Investigating the TargetP+LumenP performance on the *Arabidopsis* and *Oryza* Swiss-Prot subsets with known subcellular localization, we found that almost all (88% and 100% for *Arabidopsis* and *Oryza,* respectively, data not shown) of the proteins with a Swiss-Prot subcellular localization annotation that were incorrectly assigned as lumenal were annotated as stromal, which is in accordance with the previous tests.

The results of the LumenP and SignalP approaches for prediction of cleavage sites were rather similar to each other; 72 of the 93 (77.4%) cleavage sites in the redundancy-reduced test set were correctly predicted by LumenP, compared to 70 (75.3%) using the computationally more costly SignalP-NN analysis (Table 4). Testing the performance on the entire set of 259 lumenal proteins revealed again similar performance levels, 54.8% of the cleavage sites were correctly predicted by LumenP and 55.6% by SignalP. It is surprising that both LumenP and SignalP performed much better on the redundancy reduced set of 93 proteins than on the remaining 166 lumenal proteins; we have no good explanation for this at present. In accordance with previous findings (Nielsen and Krogh 1998), we found that the NN version of SignalP is clearly better than the HMM version in predicting cleavage sites, even though the HMM outperforms the NN on the lumenal/nonlumenal prediction (Tables 2 and 3). The Howe-Wallace scoring matrix, which is based on only 12 sequences, performed worse than all the other methods, but again performed significantly better on the test set (93 sequences) than on the entire lumenal set (259 sequences).

Conclusions

We have devised a new neural network predictor, LumenP, that is intended to be used together with TargetP to identify

Table 4. *Cleavage site prediction*

Cleavage site predictor	Data sets	
	Test set (93 Sec+TAT)	Lumenal set (259 Sect+TAT)
LumenP	72 (77.4%)	142 (54.8%)
SignalP (NN)	70 (75.3%)	144 (55.6%)
SignalP (HMM)	61 (65.6%)	131 (50.6%)
Howe-Wallace	45 (48.4%)	102 (39.4%)

Comparison of cleavage site (CS) prediction performance between the new LumenP method, SignalP (both NN and HMM versions), and the Howe-Wallace scoring matrix. The number and percent of correctly predicted cleavage sites is shown. For the SignalP analysis, we used both Gram-negative and Gram-positive versions and applied the truncation scheme described in the text. From the resulting 26 suggested cleavage sites (13 each from Gram-negative and Gram-positive versions, respectively), the one with the highest cleavage site score (NN: max Y-score, HMM: C-max score) was chosen as the final prediction.

nuclearly encoded plant proteins destined for the lumen of the thylakoid. As judged by fivefold crossvalidation using a positive set of 93 nonhomologous lumenal proteins (further divided into a TAT and a Sec group) that all contain both a cTP and an lTP, and a mixed negative set of 50 proteins not present in the thylakoid lumen, we find that the combination of TargetP (to identify proteins with a cTP) and LumenP (to identify lTPs in the proteins passed on from TargetP) reaches a Matthews' correlation coefficient of 0.72 with a sensitivity of 0.82 and a specificity of 0.96 (Table 2) when including LumenP cleavage site scores. The lTP cleavage site prediction performance was on par with an approach based on a truncation scheme requiring 26 runs of SignalP to arrive at a prediction. Thus, the LumenP method for predicting cleavage sites is simpler to use and as accurate as the SignalP approach (Table 4).

To evaluate the frequency of false-positive predictions on a realistic test set, we applied LumenP and TargetP+LumenP to all plant proteins found in Swiss-Prot that were annotated as either containing a signal peptide (i.e., nuclearly encoded secretory proteins) or as being located in the chloroplast stroma. Although LumenP by itself identifies many signal peptides as being lTPs (data not shown), almost no secretory proteins survive through the combined TargetP+LumenP predictor. In contrast, 19% of the proteins annotated as chloroplast stromal (i.e., having a cTP but not an lTP) are predicted as lumenal by TargetP+LumenP (Table 3). This rather high value suggests that some of these proteins may be misannotated and are, in fact, lumenal, which is further supported by the observation that on the Swiss-Prot *A. thaliana* and *Oryza sativa* subsets with annotated subcellular location, almost all of the incorrectly assigned lumen proteins were annotated as stromal. A comparison with the TargetP+SignalP approach (including the sequence truncation scheme described above) for predicting lumenal sequences revealed that the most significant contribution of LumenP is to reduce the number of false positives.

A final application was to scan all the *Arabidopsis* and *Oryza* ORFs predicted from the complete genome sequences. Using prescreening by TargetP and then LumenP (cutoff pair 0.67/6.80), 417 out of 25,826 (1.6%) *Arabidopsis*, and 1200 out of 41,915 (2.9%) *Oryza* proteins were predicted as being located in the lumen of the thylakoid. A full listing of the predicted lumenal proteins is provided as Supplemental Material.

Materials and methods

Datasets used in training and testing of LumenP

Positive set

Two hundred fifty-four sequences for thylakoidal lumenal proteins were generously provided by Jean-Benoît Peltier (Cornell University). This set was expanded by searching Swiss-Prot release 40 (O'Donovan et al. 2002) for lumenal sequences not present in the Peltier data set. Sequences were extracted by searching for "THYLAKOID LUMEN" in the FT field and "SUBCELLULAR COMPARTMENT: THYLAKOID LUMEN" in the CC field. Targeting peptide entries marked as POTENTIAL, BY SIMILARITY, or PROBABLE were excluded. By this approach, five more sequences were found. Including these sequences and their orthologs resulted in a final positive set of 259 sequences. By searching for the so-called twin-arginine motif Arg-Arg upstream of the hydrophobic region in the lTPs, 138 of these were classified as belonging to the TAT pathway and 121 as belonging to the Sec or other pathways. The data set is available as Supplemental Material.

Because lTPs have been shown to have much less sequence conservation than the mature part of the protein (Peltier et al. 2002), only the cTP+lTP part of the proteins was used for the training of LumenP. The cTP part was not removed because the exact cTP cleavage sites are generally not known. Also, TargetP prediction of cTP cleavage sites is not very reliable (Emanuelsson et al. 1999). Instead, a stretch of 35 residues upstream of the lTP cleavage site (determined by experiments if available, otherwise by similarity) roughly corresponding to the average length of lTPs, were annotated as belonging to the lTP in the neural network training procedure.

Negative set

A mixed negative set of roughly the same size as the positive TAT and Sec sets was constructed. This set contained proteins destined for the chloroplast (but not the thylakoid lumen), mitochondrion, cytoplasm, nucleus, and secretory pathway, in equal numbers.

The chloroplast sequences were extracted from Swiss-Prot release 40 by searching for "SUBCELLULAR LOCATION: CHLOROPLAST" in the CC field and "CHLOROPLAST" in the FT field. Proteins encoded in the chloroplast genome were excluded, as were those of algal origin, because cTPs from the green algae *Chlamydomonas reinhardtii* have been shown to be more similar to mTPs than to cTPs from higher plants in terms of length and amino acid composition (Franzén et al. 1990). The chloroplast sequences were truncated to the 130 most N-terminal amino acids before redundancy reduction.

All other sequences in the negative test set were picked at random from the redundancy reduced TargetP training set (available at http://www.cbs.dtu.dk/services/TargetP/datasets/datasets.html) from which the sequences redundancy-reduced on the 112 N-terminal amino acids were used. The mixed negative set contained 50 sequences, 10 destined for each compartment.

Redundancy reduction

Redundancy reduction, that is, removal of homologous sequences, of the positive and negative sets was done in three steps. First, all sequences in a set were pairwise aligned all against all using the full Smith-Waterman algorithm (Smith and Waterman 1981) with the PAM250 scoring matrix as implemented in the *ssearch* program of the FASTA package (Pearson 1990). Based on the distribution of alignment scores, the threshold score above which sequences were considered as too similar for network training was chosen as the value where the actual distribution of scores deviated from the extreme-value distribution expected for a local alignment of random sequences (Pedersen and Nielsen 1997). A pair of proteins whose similarity score is above the chosen cutoff are called "neighbors." The Hobohm algorithm 2 (Hobohm et al. 1992) was applied until no proteins were left that had any neighbors within the cutoff score. This algorithm creates a list of all proteins and their neighbors and then removes the protein that has the largest number of neighbors. Then, the neighbor list is recalculated, and again the protein with the largest number of neighbors is removed, and so on until the list only contains proteins that have no neighbors.

After redundancy reduction, a total of 93 sequences were left in the positive set (50 in the TAT group and 43 in the Sec group).

Cross-validation

Fivefold cross-validation was used during training of the neural networks. Each of the five subsets contained about equal numbers of positive and negative examples, as well as equal numbers of the different types of negative examples, that is, chloroplast, mitochondrial, cytoplasmic, secretory pathway, and nuclear sequences.

Neural network architecture and training

The Billnet (Perantonis and Virvilis 2000) neural network simulator platform (issued under GPL at http://www.iit.demokritos.gr/~vasvir/billnet/) was used for the development of LumenP.

For the recognition of both TAT and non-TAT lumenal proteins, two separate neural networks on top of each other were used. In the first layer network, the input data were as described above, and presented using sparse encoding and a sliding window of size 35 residues. The output of the first layer network is one score per residue, and the outputs corresponding to residues 21 to 110 for each protein (counting from the N-terminus) are forwarded to the second layer network, which outputs one score per protein based on the 90 input values it receives from the first layer network. Networks were separately trained on the TAT and Sec datasets. In the final LumenP predictor, the query protein is processed through both the TAT and Sec networks in parallel, giving two final scores of which the highest is chosen for the prediction.

A standard feed-forward network with a sigmoid transfer function with logistic neurons, one hidden layer, and a sigmoid steepness of 4 was chosen for both the first and second layer network. The number of neurons in the hidden layer were 8 in both the first and second level networks for both the TAT and Sec versions. The back-propagation error method was used as training algorithm and the initial weights were chosen at random. The learning rate was set to 0.001 for all networks, and the number of training cycles to 350 for first layer networks, and 100 or 150 for the TAT and Sec second layer networks, respectively. By choosing a constant number of training cycles for all networks in the cross-validation, we avoid optimizing on the individual test sets. Furthermore, the performance fluctuations were very small in a large region around the chosen training cycle numbers, and test set performance was thus not sensitive to the exact choice of stopping point.

Scoring matrix for cleavage site prediction

A cleavage site scoring matrix was constructed from an alignment of the region around the annotated cleavage sites. The set of 93 redundancy reduced lumenal sequences were used for constructing the alignment. The elements (scores) $s_{i,j}$ of the scoring matrix, where i is the sequence motif position and j the amino acid, were then calculated from the multiple alignment in a standard fashion:

$$s_{i,j} = \log_2 \frac{f_{i,j}}{p_j},$$

where $f_{i,j}$ is the frequency of amino acid j at position i, and p_j is the background frequency of that amino acid in a background set. A simple form of pseudoconts was used: one was added to each count (Laplace's rule). The total amino acid distribution of the 259 lumenal full-length proteins was used as the background.

Performance measures

Prediction performance was measured by determining the sensitivity (number of true positives/[number of true positives + number of false negatives]), specificity (number of true positives/[number of true positives + number of false positives]), and the Matthews' correlation coefficient (Matthews 1975), which is one for a perfect prediction and zero for a completely random assignment.

To further assess the number of predicted false positives, LumenP was also tested on all *Plantae* sequences from Swiss-Prot release 40 annotated as containing a secretory signal peptide (SP) or annotated as chloroplast (but not thylakoid lumen). All entries of plant origin were extracted by searching for "Eukaryota; Viridiplantae" in the OC line, resulting in 5694 entries. From this set, sequences annotated as containing an SP were extracted by searching for the keyword "SIGNAL" in the FT field, resulting in 745 entries. All sequences with "SUBCELLULAR LOCATION: CHLOROPLAST" in the CC field and "CHLOROPLAST" in the FT field were collected and those annotated as thylakoid proteins were excluded, resulting in 878 sequences. Also, all *Arabidopsis* (1034 sequences) and *Oryza* (402 sequences) sequences found in Swiss-Prot release 40 and 40.17, respectively, were analyzed (with proteins with clear annotation of thylakoid localization removed).

Genome-wide datasets

The fully sequenced genomes of *A. thaliana* (The Arabidopsis Genome Initiative 2000; 25,826 ORFs, downloaded from ftp://ftpmips.gsf.de/cress/arabiprot/, version 2002-04-03) and *O. sativa* (Goff et al. 2002; 41,915 ORFs, downloaded from ftp://ftp.tigr.org/pub/data/o_sativa/irgsp/PUBLICATION_RELEASE/GENOME/, version 2002-04-19) were analyzed.

Availability

LumenP prediction is available from the authors by request via e-mail (gunnar@dbb.su.se).

Acknowledgments

This work was supported by grants from the Foundation for Strategic Research and the Swedish Research Council to G.v.H.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.

Berks, B.C., Sargent, F., and Palmer, T. 2000. The Tat protein export pathway. *Mol. Microbiol.* **35:** 260–274.

Emanuelsson, O. and von Heijne, G. 2001. Prediction of organellar targeting signals. *Biochem. Biophys. Acta* **1541:** 114–119.

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

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

Franzén, L.G., Rochaix, J.D., and von Heijne, G. 1990. Chloroplast transit peptides from the green alga *Chlamydomonas reinhardtii* share features with both mitochondrial and higher plant chloroplast presequences. *FEBS Lett.* **260:** 165–168.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). *Science* **296:** 92–100.

Halpin, C., Elderfield, P.D., James, H.E., Zimmermann, R., Dunbar, B., and Robinson, C. 1989. The reaction specificities of the thylakoidal processing peptidase and *Escherichia coli* leader peptidase are identical. *EMBO J.* **8:** 3917–3921.

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

Howe, C.J. and Wallace, T.P. 1990. Prediction of leader peptide cleavage sites for polypeptides of the thylakoid lumen. *Nucleic Acid Res.* **18:** 3417.

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

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

Nielsen, H. and Krogh, A. 1998. Prediction of signal peptides and signal anchors by a hidden Markov model. *Intell. Syst. Mol. Biol.* **6:** 122–130.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10:** 1–6.

O'Donovan, C., Martin, M.J., Gattiker, A., Gasteiger, E., Bairoch, A., and Apweiler, R. 2002. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Bioinformatics* **3:** 275–284.

Pearson, W.R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **183:** 63–98.

Pedersen, A.G. and Nielsen, H. 1997. Neural network prediction of translation initiation sites in eukaryotes. Perspectives for EST and genome analysis. *Intell. Syst. Mol. Biol.* **5:** 226–233.

Peltier, J.B., Friso, G., Kalume, D.E., Roepstorff, P., Nilsson, F., Adamska, I., and van Wijk, K.J. 2000. Proteomics of the chloroplast: Systematic identification and targeting analysis of lumenal and peripheral thylakoid proteins. *Plant Cell* **12:** 319–341.

Peltier, J.B., Emanuelsson, O., Kalume, D.E., Ytterberg, J., Friso, G., Rudella, A., Liberles, D.A., Soderberg, L., Roepstorff, P., von Heijne, G., et al. 2002. Central functions of the lumenal and peripheral thylakoid proteome of Arabidopsis determined by experimentation and genome-wide prediction. *Plant Cell* **14:** 211–236.

Perantonis, S. and Virvilis, V. 2000. Efficient perceptron learning using constrained steepest descent. *Neural Netw.* **13:** 351–364.

Robinson, C., Hynds, P.J., Robinson, D., and Mant, A. 1998. Multiple pathways for the targeting of thylakoid proteins in chloroplasts. *Plant Mol. Biol.* **38:** 209–221.

Schubert, M., Petersson, U.A., Haas, B.J., Funk, C., Schröder, W.P., and Kieselbach, T. 2002. Proteome map of the chloroplast lument of *Arabidopsis thaliana*. *J. Biol. Chem.* **277:** 8354–8365.

Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147:** 195–197.

von Heijne, G. 1983. Patterns of amino acids near signal sequence cleavage sites. *Eur. J. Biochem.* **133:** 17–21.