Chloroplast transit peptide prediction: a peek inside the black box

Andrew I. Schein*, Jessica C. Kissinger[1] and Lyle H. Ungar

University of Pennsylvania Department of Computer and Information Science, 556 Moore Building,
200 S. 33rd Street, Philadelphia, PA 19104-6389, USA and [1]Department of Biology, Goddard Laboratories,
University of Pennsylvania, Philadelphia, PA 19104-6018, USA

Received March 16, 2001; Revised and Accepted July 1, 2001

ABSTRACT

Previous work in predicting protein localization to the chloroplast organelle in plants led to the development of an artificial neural network-based approach capable of remarkable accuracy in its prediction (ChloroP). A common criticism against such neural network models is that it is difficult to interpret the criteria that are used in making predictions. We address this concern with several new prediction methods that base predictions explicitly on the abundance of different amino acid types in the N-terminal region of the protein. Our successful prediction accuracy suggests that ChloroP uses little positional information in its decision-making; an unexpected result given the elaborate ChloroP input scheme. By removing positional information, our simpler methods allow us to identify those amino acids that are useful for successful prediction. The identification of important sequence features, such as amino acid content, is advantageous if one of the goals of localization predictors is to gain an understanding of the biological process of chloroplast localization. Our most accurate predictor combines principal component analysis and logistic regression. Web-based prediction using this method is available online at http://apicoplast.cis.upenn.edu/pclr/.

INTRODUCTION

The chloroplast organelle in plant cells contains its own genome, but many of its genes have been transferred to the plant cell nuclear genome during the course of evolution. The protein products of these nuclear-encoded plastid genes are targeted back to the chloroplast organelle. Transit peptide sequences located in the N-terminus of these proteins facilitate the transfer from the cytoplasm where they are synthesized, back to the chloroplast organelle (1; reviewed in 2). Identification of proteins that contain an N-terminal chloroplast targeting sequence has been difficult. First, many proteins contain N-terminal extensions that are used by the cell to target the protein to any of several destinations, e.g. the mitochondria. Secondly, there is no known motif, or obvious sequence pattern that identifies proteins as chloroplast targeted. The exact nature of the 'biological' signal that the cell interprets is not yet known.

We are concerned with the problem of predicting the presence of a chloroplast transit peptide (cTP) based on N-terminal amino acid sequence. What we generally desire from computational predictors are accuracy and intuition for discriminating between sequence types in the data (i.e. biological insight). Previous work by Emanuelsson, Nielsen and von Heijne (ENH) resulted in a neural network-based model called ChloroP (3). ChloroP examines amino acid content at specific positions in order to predict localization to the chloroplast. ChloroP's accuracy and availability via an Internet web site have made it the gold standard for chloroplast transit peptide prediction.

However, it is very difficult to determine exactly how the neural network 'used' what it 'saw' to make its prediction. Our goal is to provide as an alternative to ChloroP a set of methods for which we can say with certainty what intuition lies beneath the decision mechanism. Pilot simulations of ChloroP led us to posit the hypothesis that amino acid distribution alone provides useful information for localization prediction. The ChloroP first-layer network looks at a 'sliding window' of sequence, and we observed a smooth trend in prediction as the window slid from the N- and C-termini. This smoothness indicates that ChloroP does not use motifs (relative positions of certain amino acids) in its decision-making. In a neural network model with only two internal nodes, as is the case here, one would expect to see the effects of a motif coming in and out of input frame reflected in the predictions. This observation led us to look at amino acid distribution as an indicator of chloroplast targeting sequences.

We present a new model using a principal component logistic regression (PCLR). Unlike ChloroP, our PCLR model uses the distribution of amino acids in the N-terminus of the protein with no positional information as input to its prediction function. In addition to the 20 amino acids there is one other input corresponding to the sequence's amino acid sample variance. We use the same method of encoding sequence data as we experiment with two additional machine learning algorithms: stepwise logistic regression and a neural network. Where possible, we identify amino acids (features) from our models that are used in the prediction algorithm. Our models are benchmarked against ChloroPv1.1 on an independent test set, and receiver operating characteristic (ROC) curves (4)

*To whom correspondence should be addressed. Tel: +1 215 735 9167; Fax: +1 215 898 0587; Email: ais@gradient.cis.upenn.edu comparing PCLR and ChloroPv1.1 prediction accuracy are presented.

MATERIALS AND METHODS

Training and test sets

Olof Emanuelsson (Stockholm Bioinformatics Center) supplied the 150 sequence ChloroP data set (3), which we randomly divided into 20 pairs of training and validation sets for the purpose of setting the parameters of our method. Training sets consisted of 124 sequences, and validation sets consisted of the remaining 26 sequences with each set containing equal numbers of in-class (e.g. cTP) and out-of-class examples. Note that when we use the phrase 'validation set testing' we refer to testing done on a partition of the training set. For final testing, we downloaded the TargetP training set (5), and used SWISS-PROT accession numbers to remove those sequences already contained in the ChloroP training set. The TargetP training set consisted of 371 mitochondrial transit peptides (mTP), 269 secretory pathway/signal peptide (SP), 48 'nuclear' (Nuc), and 87 'cystolic' (Cyt) sequences from which we removed 17, 14, 9 and 10 sequences respectively. The SP, Nuc and Cyt sequences were all from the TargetP 'plant set'. From the 141 cTP sequences we removed 28 redundant sequences. These were the only sequences removed, and the remaining test set contains 113 in-class and 725 out-of-class sequences.

Encoding a protein

For our PCLR, logistic regression, and neural network models, the input size is 21. The first 20 inputs consist of percentages of amino acid composition in the first 55 positions of the protein sequence. The 21st input is a measure of variance of the particular protein's amino acid distribution in the first 55 positions. Our methods performed similarly on the validation sets with sequence lengths between 45 and 60, but ultimately a length of 55 was chosen for our study, based on sum of squared errors (SSE) measurements.

Principal component logistic regression

Principal components analysis is a method of factoring co-linearity out of data and reducing dimensionality for a machine learning algorithm (6). We performed principal component analysis and subsequent stepwise logistic regression on the first 12 components (ordered by decreasing eigenvalue magnitude) on the principal component matrix using the R statistics package (7). We transformed testing data into the training data principal component space before generating prediction results.

The logistic regression always makes predictions between (0,1), but we require a threshold to use for classification. Based on 'total number correct' counts during validation set testing we chose a decision threshold of 0.42 for classification (e.g. a prediction of 0.41 means our method predicts 'non-chloroplast targeting'). After deciding on a number of principal components to consider and the classification threshold, we trained PCLR on the entire ChloroP training set. The resulting predictor, principal components, and regression coefficients are available online at http://apicoplast.cis.upenn.edu/pclr/.

Logistic regression

We attempted a standard stepwise logistic regression in addition to the principal component stepwise logistic regression to see if a simpler model would provide equal performance. In the R package we used the same input to the logistic regression as in the PCLR case. A decision threshold of 0.40 was selected during validation set testing and then used on the TargetP test set.

Neural network

We used NevProp4r1, a standard feed-forward neural network with sigmoidal hidden units and one sigmoidal output unit (http://www.scs.unr.edu/nevprop). We used the same inputs as in the PCLR case described above. The number of hidden units was varied from 1 to 12, with peak performance occurring with 4 hidden units and decreasing performance soon after. A weight decay of 0.005 was chosen based on validation set performance. For training, we picked a maximum iteration of 700, and used NevProp's auto-train switch to pick a good stopping point. Based on validation set performance (total number correct), we chose a classification threshold of 0.59.

The ChloroP neural network architecture

The ChloroP architecture is described in Emanuelsson *et al.* (3); however, for clarification and comparative purposes a brief description is included. ChloroP consists of two neural networks where the output of the first network against a set of different inputs feeds into the second neural network for a final prediction. The input to the first network consists of a sliding window of 51 amino acids from the first 100 positions of a protein. There are 100 ordered windows per protein, and they start so that the first window consists of the first 51 amino acids of the protein sequence. Shifting the previous window to the right one place forms each subsequent window. As windows overlap an area past position 100, 'blank' amino acids feed into the predictor. 100 of these windows feed into the first layer, and so 100 predictions are made.

The first network consists of 1020 input units, 2 hidden units and 1 output unit. The rather large number of input units is the result of using categorical data in a neural network. There are 20 possible attributes (amino acids) in a position, and so each position has 20 input units. Only one of these units is turned on (denoted by '1.0'): the other 19 are left at '0.0'. Hence, a window of 51 positions requires $51 \times 20 = 1020$ input units. Compounding this explosion in input size are the 100 windows per protein sequence that feed into the first layer network. All together, it takes 102 000 total inputs to the first-layer network to make a prediction on a single protein. The second layer network has 100 input units, 10 hidden units and 1 output unit. For both networks, sigmoidal units are used in hidden and output layers.

We benchmarked the ChloroP model using the web-accessible ChloroPv1.1 release located at http://www.cbs.dtu.dk/services/ChloroP/. We used the classification threshold 0.50 as suggested by Emanuelsson *et al.* (3).

RESULTS

Benchmark results

In order to give an accurate assessment of our new models against each other and ChloroPv1.1, we trained them on the original ChloroP training data (3). It is important to test such models on new data not present in the training set in order to give an unbiased estimate of prediction ability. Our testing data came from the TargetP set (5), from which we removed any sequences already present in our training set. [Though the TargetP details (5) were not available until after the completion of our work, the datasets were available online months ahead of time.]

In order to describe the benchmark results, a mixture of terms and abbreviations from both molecular biology and machine learning is employed in the data analysis. We adopt abbreviations for categories of protein in training and test sets: sequences containing a cTP, sequences containing a mTP, sequences containing a SP but no additional signaling information, nuclear-localized proteins (Nuc) and cytosolic sequences (Cyt). The inputs to the prediction methods are often referred to as features; each sequence is encoded as a 21 component 'feature vector' for predication. All four machine learning methods benchmarked here make predictions within a (0,1) interval, where predictions closer to 1 indicate the presence of a cTP. A decision threshold in the (0,1) range is used for 'classification' of sequences as containing a cTP (cTP class) or not containing a cTP (non-cTP class). The decision threshold for our tests is determined by training set performance (see Materials and Methods).

We benchmarked the following methods: web-based ChloroPv1.1, PCLR, our own neural network predictor, and logistic regression. Each of the 838 sequences in the test set were classified by each method as either cTP-containing or not (Table 1). Also shown is a SSE providing an additional standard for comparing different prediction methods. SSE measures how close a predictor is to getting the right answer (0 or 1):

$$\mathrm{SSE} = \sum_{i} (y_i - \hat{y}_i)^2 \cdot$$

The summation is over the protein sequences (i) in the test set, while y and $\hat{y}$ refer to the correct answer and prediction respectively. A smaller SSE indicates a closer correct prediction (to 0 or 1) in general. ChloroP's predictions fall in the [0.40,0.59] interval, whereas our methods make predictions in the wider range: (0,1). Hence, there is often a lower SSE for our methods even where ChloroP classifies more accurately. The wider (0,1) range of prediction is useful when an estimated probability of chloroplast localization is desired from the predictor in addition to a classification.

In the field of machine learning sensitivity and specificity are often used to compare two different prediction methods. Sensitivity is defined here as the percentage of cTP-containing sequences correctly classified, and specificity is the percentage of cTP-predicted sequences that are correctly classified. ChloroP's predictions generate a sensitivity of 0.87 and a specificity of 0.37 while our PCLR model achieved a sensitivity of 0.82 and a specificity of 0.30. Figure 1 shows an ROC curve comparing PCLR against ChloroPv1.1 on test set data. ROC curves illustrate the trade-off between sensitivity and specificity for different classification thresholds. PCLR was chosen for this analysis over logistic regression because of its

Table 1. Classification and SSE results on the test set by ChloroP v1.1 and our methods: PCLR, logistic regression and the neural network

ChloroP v1.1		cTP	mTP	SP	Cyt	Nuc
Predicted Class	cTP	99	131	33	4	4
	non-cTP	24	223	222	73	35
	SSE	23.7	86.1	57.1	15.5	7.9
PCLR						
Predicted Class	cTP	93	174	30	5	5
	non-cTP	20	180	225	72	34
	SSE	16.4	99.6	15.2	2.5	3.1
Logistic Regression						
Predicted Class	cTP	94	208	26	1	3
	non-cTP	19	146	226	76	36
	SSE	19.2	146.2	16.8	1.8	0.9
Neural Network						
Predicted Class	cTP	71	233	102	7	5
	non-cTP	42	121	153	70	34
	SSE	26.6	170.7	74.4	6.1	4.1

The rows display predictions (cTP versus non-cTP) for the different categories of sequences in the data. The five columns of data are filled in order with results for sequences containing a chloroplast transit peptide, a mitochondrial transit peptide, and a signal (secretory) peptide (with no additional signal), followed by cystolic and nuclear-localized proteins.

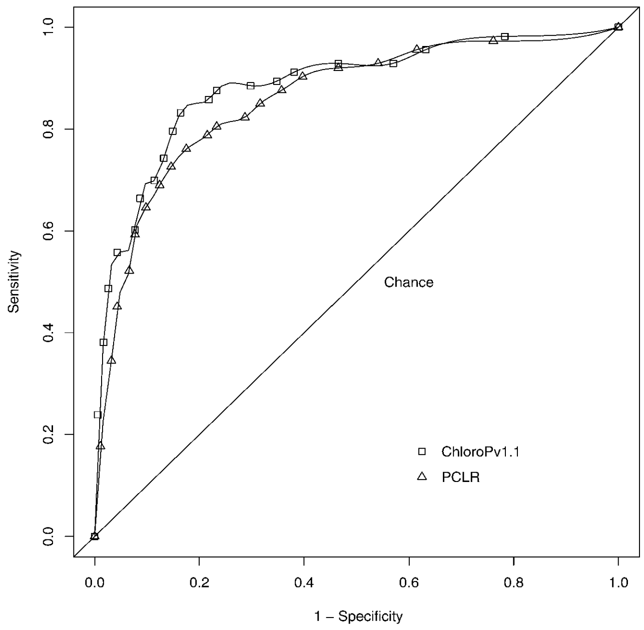

Figure 1. ROC curves for ChloroPv1.1 and PCLR show how these methods trade sensitivity for specificity with different classification thresholds. A greater area under a ROC curve indicates superior prediction ability. Classification thresholds are available in Table 2.

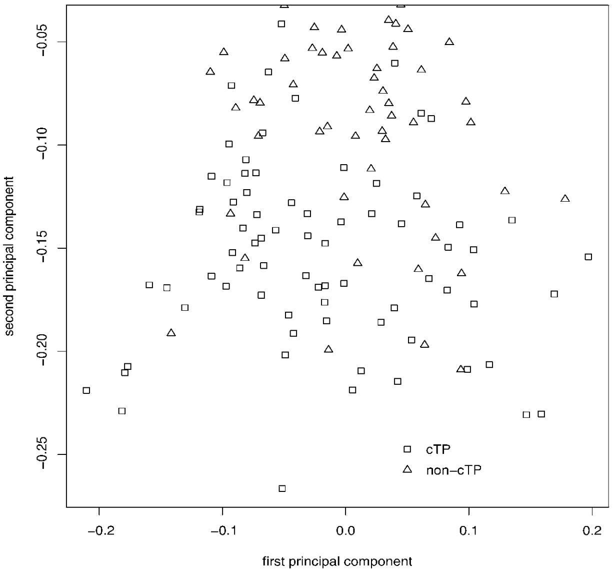

Figure 2. Plot of the training set with respect to its first two principal components. These two components were selected for classification during the stepwise regression.

Table 2. Sensitivity levels at various classification thresholds for our PCLR method

Threshold	0.85	0.82	0.78	0.72	0.68	0.62	0.55	0.51	0.43	0.39	0.33	0.27
Sensitivity	0.51	0.55	0.60	0.65	0.69	0.73	0.77	0.79	0.83	0.85	0.88	0.90

The corresponding specificities for the sensitivities below are available in Figure 1.

superior SSE and classification performance (summed over columns of Table 1). Due to the number of variables present in testing, it is difficult to decide which of our methods performs ‘best’: PCLR or logistic regression. For instance, PCLR does better at predicting sequences containing mTPs, but logistic regression is better at predicting nuclear-localized proteins. PCLR has a better total SSE (summed over columns of Table 1) even when mTPs are not included in the total, and so we have focused our attention on this method.

Figure 2 shows a plot of the training data with respect to the two largest principal components used during the regression. ‘Large’ here refers to eigenvalue magnitude. We see that these vectors visibly separate the chloroplast targeting and non-targeting sequences, showing that the 21 inputs (amino acid frequency and sample variance) discriminate between cTP and non-cTP sequences. If it were possible to represent on paper, a plot consisting of all dimensions selected during the regression would show much more separation.

Selecting features from PCLR

It is valuable to understand in intuitive terms what criteria are used by these prediction methods for the classification of proteins to see if we can gain any biological insight into chloroplast targeting. Our approach is to identify amino acids that contribute heavily to the classification function (i.e. are useful for discriminating among the various N-terminal types). In the PCLR model, by examining the principal components (see Materials and Methods) for our training data, we are able to observe the weight given to each amino acid for each vector *V*. In order to extract interesting amino acids from the eigenvectors selected during the stepwise regression process, we multiplied the eigenvectors by their regression weights and summed the results into a single vector *V*. At this point we could see relative magnitudes of the different features and their sign of contribution: positive indicating cTP, and negative indicating non-cTP. We then squared and added up the 21 *V* components to get a total sum of squared contribution. Then, comparing the squared value of each component of *V* to the total sum of squared contribution, we picked the components/features that stand out as contributing the most. Aspartic acid, glutamic acid, proline and serine together contribute more than 46% of the total sum of squared weight. Aspartic acid and glutamic acid influence the decision negatively, while proline and serine are indicative of cTPs.

Selecting features from logistic regression

Stepwise logistic regression selected eight amino acids as informative for classification. Serine and arginine are given positive cTP weight, while aspartic acid, cysteine, glutamic acid, histidine, isoleucine and tryptophan are all given negative weight. The variance input was also selected and given negative weight in the decision process. Positive contribution indicates cTP, while negative contribution indicates non-cTP.

Selecting features from our neural network

In order to select key amino acids for prediction we used a neural network input relevance score:

$$\text{Relevance}_{\text{i}} = \frac{(\text{sum of square weights of the } i\text{th input group})}{(\text{sum of square weights of all input groups}).}$$

This definition of relevance captures the degree to which each of the 21 inputs is weighted in the decision. It is not the complete picture, as it does not take into account second layer weights. Also, this definition of relevance says nothing about whether the contribution is ultimately positive or negative. Four amino acids were selected as important to classification: they have input relevance above 11% each and account for more than 58% of total relevance. The four amino acids are asparagine, glutamine, isoleucine and serine.

DISCUSSION

The peculiarities of the amino acid distribution of chloroplast and mitochondrial targeting sequences have been known for over 10 years (8), but until now this knowledge has not been leveraged so successfully in its ability to actually predict chloroplast localization. Serine has been identified in the past as abundant in cTPs in comparison to the entire chloroplast targeting protein (8). In the same study, aspartic and glutamic acid were reported as under-represented in cTPs. In contrast to this previous approach of identifying ‘interesting’ amino acids, our methods select amino acids that discriminate between the N-termini of different classes of proteins. The three amino acid overlap in the results of our logistic regression, PCLR methods and earlier work by von Heijne and co-workers (9) establish our two methods as capable of producing and confirming insightful observations about a data set. The inferior accuracy of our neural network model implies that we should discount from consideration the amino acids extracted from this model. In the case of cTPs, our results partly confirm what the *Arabidopsis thaliana* community has already discovered. In future pattern recognition applications, we hope that classification accuracy is recognized from the start as an intuitive and useful method for selecting 'interesting' amino acids.

Signal prediction and extraction have been active research areas in recent years. The SignalP tool, which is built from several neural networks, provides excellent prediction for targeting of the endoplasmic reticulum and location of signal peptide cleavage sites (9,10). Recent work using adaptive encoding neural networks has identified features useful for prediction of the signal peptide cleavage site (11). An alternative method uses amino acid frequency at relative positions in a probabilistic model (12). The problem of predicting mitochondrial transit peptides has also been tackled using a combination of principal component analysis and linear discriminant analysis (13,14).

There are previous studies that have used amino acid distribution for chloroplast organelle prediction as well. A covariant discriminant (CD) algorithm was developed to predict among 12 different subcellular localizations including the chloroplast (15). After adjusting for sequence composition in the CD test set (e.g. percentage of mTPs), we estimate that our PCLR method represents a substantial improvement of accuracy in predicting chloroplast localization. (CD was unavailable for independent testing.) Based on published results (15), we can say that our test set sensitivity is at least 10% greater than the CD 'self-consistency' (testing on training data) performance. No attempt has yet been made to interpret the decision mechanisms of the CD method for cTP classification, and this prevents us from comparing amino acid weightings. Some studies of hybrid methods combining amino acid composition with limited positional data include Chou's work with pseudo-amino acid composition (16,17). Other researchers have combined an 'expert system' with machine learning clustering methods in using amino acid frequencies among other features to predict localization (18,19). The published accuracy is not yet competitive with our method.

In performing our experiment, we have learned that amino acid frequency can account substantially for the prediction accuracy of a ChloroP-like neural network architecture for predicting localization. When a system such as ChloroP with over 2000 parameters is trained on a challenging data set it is dangerous to conjecture on an intuitive explanation without experimentation. Our frequency-based models are a first step in such experimentation. Consider another feature that one might hypothesize as important for localization prediction: the transit peptide cleavage site. By combining our method with a transit peptide cleavage-site predictor such as the scoring matrix method of ENH or alternatively a hidden markov model, we will gain a sense of how much each set of features can contribute to correct prediction ability. Separating prediction features in this matter can give greater insight into the prediction mechanism.

The recognition of amino acid frequencies as important factors in protein localization prediction should have an important effect on the way the transit peptide prediction problem is addressed in the future. The original ChloroP training set was homology-reduced by the Hobohm motif-finding algorithm (20), which is based on sequence similarity, but our findings suggest that similarity of amino acid counts could be a more important tool for making training sets sufficiently representative of the population. In addition, discrimination between mTPs and cTPs, a problem that cripples prediction accuracy for all methods benchmarked here, is given a specific definition: cTPs and mTPs have similar amino acid frequencies. In the short term we have focused on developing purely distribution-based techniques because they generalize more easily beyond the chloroplast transit peptide problem. For instance, the organism *Toxoplasma gondii* uses transit peptides to target proteins to the apicoplast organelle (21). Transit peptide cleavage sites (another potential prediction feature) for these sequences have not yet been identified, but a prediction tool is still needed for screening.

PCLR and logistic regression methods may be applied to other pattern recognition problems on a modern desktop computer using built-in routines of various commercial data-manipulation packages (e.g. Splus®, Matlab®). Once training sequences have been encoded as amino acid frequency data, training and testing can occur within seconds. For these reasons, we recommend PCLR and logistic regression to small laboratories that wish to experiment with pattern recognition algorithms.

ACKNOWLEDGEMENTS

We would like to thank David Roos for discussions on the design of this experiment. Our thanks go also to Olof Emanuelsson for help with data set preparation and explanation of ChloroP details. A.I.S. was supported by NIH Training Grant T32-HG-00046.

REFERENCES

1. Robinson,C., Hynds,P., Robinson,D. and Mant,A. (1997) Multiple pathways for the targeting of thylakoid proteins in chloroplasts. *Plant Mol. Biol.*, **38**, 209–221.
2. Jean-Benoît,P., Friso,G., Kalume,D., Roepstorff,P., Nilsson,F., Adamska,I. and van Wijk,K. (2000) Proteomics of the chloroplast: systematic identification and targeting analysis of lumenal and peripheral thylakoid proteins. *Plant Cell*, **12**, 319–341.
3. 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.
4. Lang,T. and Secic,M. (1997) *How to Report Statistics in Medicine*. American College of Physicians, Philadelphia, PA.
5. 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.
6. Rawlings,J., Pantula,S. and Dickey,D. (1998) *Applied Regression Analysis, a Research Tool.* Springer-Verlag, New York, NY.
7. Ihaka,R. and Gentleman,R. (1996) R: A Language for Data Analysis and Graphics. *J. Comput. Graph. Stat.*, **5**, 299–314.
8. von Heijne,G., Steppuhn,J. and Herrman,R. (1989) Domain structure of mitochondrial and chloroplast targeting peptides. *Eur. J. Biochem.*, **180**, 535–545.
9. Nielsen,H., Englebrecht,J., Brunak,S. and von Heijne,G. (1997) A neural network method for the identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural Syst.*, **8**, 581–599.

10. Nielsen,H., Englebrecht,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.
11. Jagla,B. and Schuchhardt,J. (2000) Adaptive encoding neural networks for the recognition of human signal peptide cleavage sites. *Bioinformatics*, **16**, 245–250.
12. Chou,K. (2001) Prediction of protein signal sequences and their cleavage sites. *Proteins*, **42**, 136–139.
13. Claros,M. (1995) MitoProt: a Macintosh application for studying mitochondrial proteins. *Comput. Appl. Biosci.*, **11**, 441–447.
14. Claros,M. and Vincens,P. (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.*, **241**, 779–786.
15. Chou,K. and Elrod,D. (1999) Protein subcelluar localization prediction. *Protein Eng.*, **12**, 107–118.
16. Chou,K. (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*, **43**, 246–255.
17. Chou,K. (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem. Biophys. Res. Commun.*, **278**, 477–483.
18. Nakai,K. and Kanehisa,M. (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, **14**, 897–911.
19. Nakai,K and Horton,P. (1997) Better prediction of protein cellular localization sites with the *k* nearest neighbors classifier. *Ismb*, **5**, 147–152.
20. Hobohm,U., Scharf,M., Schneider,R and Sander,C. (1992) Selection of representative protein data sets. *Protein Sci.*, **1**, 409–417.
21. Waller,R., Keeling,P., Donald,R., Striepen,B., Handman,E., Lang-Uunasch,N., Cowman,A., Besra,G., Roos,D. and McFadden,G. (1998) Nuclear-encoded proteins target to the plastid in *Toxoplasma gondii* and *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA*, **95**, 12352–12357.