ELSEVIER

Gene 280 (2001) 19–26

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Deciphering apicoplast targeting signals – feature extraction from nuclear-encoded precursors of *Plasmodium falciparum* apicoplast proteins

Jochen Zuegge[a,b], Stuart Ralph[c], Michael Schmuker[a], Geoffrey I. McFadden[c], Gisbert Schneider[a,b,*]

[a]*F. Hoffmann-La Roche Ltd., Pharmaceuticals Division, CH-4070 Basel, Switzerland*
[b]*Albert-Ludwigs-Universität Freiburg, Institut für Biologie II, Schänzlestrasse 1, D-79104 Freiburg, Germany*
[c]*Plant Cell Biology Research Centre, School of Botany, University of Melbourne, Parkville, Victoria 3052, Australia*

Received 27 July 2001; received in revised form 16 October 2001; accepted 23 October 2001
Received by T. Gojobori

Abstract

The malaria causing protozoan *Plasmodium falciparum* contains a vestigal, non-photosynthetic plastid, the apicoplast. Numerous proteins encoded by nuclear genes are targeted to the apicoplast courtesy of N-terminal extensions. With the impending sequence completion of an entire genome of the malaria parasite, it is important to have software tools in place for prediction of subcellular locations for all proteins. Apicoplast targeting signals are bipartite; containing a signal peptide and a transit peptide. Nuclear-encoded apicoplast protein precursors were analyzed for characteristic features by statistical methods, principal component analysis, self-organizing maps, and supervised neural networks. The transit peptide contains a net positive charge and is rich in asparagine, lysine, and isoleucine residues. A novel prediction system (PATS, *p*redict *a*picoplast-*t*argeted *s*equences) was developed based on various sequence features, yielding a Matthews correlation coefficient of 0.91 (97% correct predictions) in a 40-fold cross-validation study. This system predicted 22% apicoplast proteins of the 205 potential proteins on *P. falciparum* chromosome 2, and 21% of 243 chromosome 3 proteins. A combination of the PATS results with a signal peptide prediction yields 15% potentially nuclear-encoded apicoplast proteins on chromosomes 2 and 3. The prediction tool will advance *P. falciparum* genome analysis, and it might help to identify apicoplast proteins as drug targets for the development of novel anti-malaria agents.

Keywords: Genome; Neural network; Plastid; Principal component analysis; Self-organizing map; Signal peptide

1. Introduction

Malaria is a major world health problem. Approximately 500 million people are infected and 2–3 million of these die annually (WHO, 1997). There is currently no effective vaccine and the parasites are acquiring resistance to the main drugs in use, so it is important that new drugs be developed. A promising new drug target emerged with the identification of a relict chloroplast (apicoplast) in *Plasmodium falciparum*, the causative agent of cerebral malaria. Little is known about the function of this organelle, which likely arose through secondary endosymbiosis. Apicoplasts have been shown to import nuclear-encoded proteins. To date only a handful of such imported apicoplast proteins have been identified, but it seems likely that the apicoplast imports several hundred proteins (Waller et al., 2000). Identification of these proteins would provide insight into apicoplast function and probably help identify new drug targets for the development of novel anti-malaria agents. *Plasmodium falciparum* is the subject of a genome project that is nearing completion. The genome comprises 14 chromosomes with an estimated 18 Mb of DNA, which is thought to encode about 9000 genes (Gardner, 1999). Clearly a proportion of these genes will encode proteins destined for the apicoplast. One approach to identifying targeted gene products is to examine them for leader sequences required for targeting. Targeting of the great majority of proteins into plastids is dependent on N-terminal leader sequences. Within the apicoplast, this leader is removed by a hitherto unknown plastid peptidase (PP) activity (Waller et al., 2000).

Abbreviations: ANN, artificial neural network; cc, Matthews correlation coefficient; Mb, megabase(s); PC, principal component; PCA, principal component analysis; PLS, partial least squares; PP, plastid peptidase; SOM, self-organizing map

* Corresponding author. F. Hoffmann-La Roche Ltd., Pharmaceuticals Division, CH-4070 Basel, Switzerland. Tel.: +41-61-68-70696; fax: +41-61-68-87408.
E-mail address: gisbert.schneider@roche.com (G. Schneider).

0378-1119/01/$ - see front matter
PII: S0378-1119(01)00776-4

Targeting of nuclear-encoded apicoplast proteins apparently commences via the secretory pathway into the endoplasmic reticulum (ER) – courtesy of a classic signal peptide. Subsequent targeting across the inner pair of apicoplast membranes involves a downstream transit peptide. Thus, the N-terminal leader is bipartite, comprising a signal peptide followed by a transit peptide (Fig. 1) (Waller et al., 2000). Deletion experiments combining green fluorescent reporter protein and parts of the leader in *P. falciparum* and the related parasite *Toxoplasma gondii* showed that both components are necessary for successful targeting. Lacking a signal peptide, the proteins accumulated in the cytoplasm of *P. falciparum*, apparently unable to enter the endomembrane system. Lacking a transit peptide, the proteins did enter the endomembrane systems, but failed to be diverted into the apicoplast and were secreted (Waller et al., 2000).

The signal peptide components of the apicoplast-targeted *Plasmodium* proteins resemble classic signal peptides, containing a hydrophobic domain followed by a peptidase cleavage site. These domains can usually be identified using prediction tools such as SignalP or PSORT (Nakai and Kanehisa, 1992; Nielsen et al., 1997). Immediately downstream of the predicted signal peptides, apicoplast-targeted proteins exhibit the general features of chloroplast transit peptides, having a net positive charge. However, unlike plant transit peptides, which are enriched for the hydroxylated residues serine and threonine (Cline and Henry, 1996), *Plasmodium* transit peptides appear enriched in lysine and asparagine. This difference in amino acid composition seems to prevent the existing prediction systems trained to recognize plant transit peptides from identifying *Plasmodium* transit peptides on apicoplast-targeted proteins (Nakai and Kanehisa, 1992; Emanuelsson et al., 2000). We therefore decided to develop a prediction model trained specifically for *Plasmodium* transit peptides. Here we describe the development of this system (PATS, *p*redict *a*picoplast-*t*argeted *s*equences) and report on its predicting performance.

2. Methods

2.1. Sequence retrieval and data sets

Preliminary sequence data for *P. falciparum* were obtained from the Institute for Genomic Research website (http://www.tigr.org), the Sequencing Group at the Sanger Centre website (www.sanger.ac.uk) and the Stanford DNA Sequencing and Technology Center website (http://www-sequence.stanford.edu/group/malaria), which are part of the International Malaria Genome Sequencing Project supported by awards from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, the Burroughs Wellcome Fund and the US Department of Defense. All data sets used in the present work can be obtained via the GECCO!(tm) prediction server on the www at the URL: http://www.modlab.de.

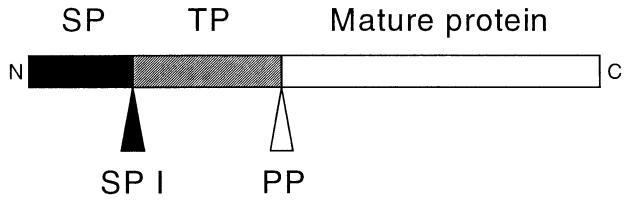

Fig. 1. Schematic of nuclear-encoded *P. falciparum* apicoplast protein precursors containing a bipartite targeting signal. Arrowheads indicate processing peptidase target sites. PP, plastid peptidase; SP, signal peptide; SP I, signal peptidase I; TP, transit peptide.

2.1.1. Positive examples

Location within the apicoplast has only been confirmed experimentally for very few proteins (ACP, FabH and DOXP reductase) (Gleeson, 2000; Waller et al., 2000). Thirty-five likely nuclear-encoded *P. falciparum* apicoplast protein precursors were inferred from similarity to known plastid proteins by pair-wise sequence alignment using TBLASTN 2.1.3 (BLOSUM62; gap existence cost 11; per gap cost 1; Lambda ratio 0.85) (Altschul et al., 1997) against the NCBI *P. falciparum* Blast Database (http://www.ncbi.nlm.nih.gov/Malaria/blastindex.html). Proteins were chosen that possessed N-terminal extensions containing a likely signal peptide followed by a peptide stretch corresponding to a transit peptide. Potential PP cleavage sites (C-terminus of the transit peptide) were deduced from alignments of the mature part with proteins without N-terminal extensions together with Western blot analysis and molecular weight calculation for some of the proteins (Waller et al., 2000). It must be stressed that the exact PP processing sites are still uncertain. Applying the same technique as described above, an additional set of 49 sequences with completely unknown PP cleavage sites – yet very likely apicoplast location – was added to obtain a larger collection of 'positive examples' for feature extraction. The final set of positive examples contained 84 sequences.

2.1.2. Negative examples

Non-apicoplast sequences ('negative examples') were collected from the SWISSPROT database (release of 20 June 2000) using the SRS software (Etzold et al., 1996) (version 5.1.0) for retrieval of all annotated *P. falciparum* sequences. This resulted in 147 entries. We expected that the majority of these sequences were true negatives (non-apicoplast proteins), because no database entries were found containing the words 'apicoplast' or 'plastid'. Two of these sequences, CH60_PLAFG and YB20_PLAFA, were excluded from the list of negative examples. CH60_PLAFG, a chaperonine-60, is annotated as a potential mitochondrial protein in SWISSPROT. Based on our analysis this annotation seems to be incorrect. This assumption is supported by prediction results obtained from the SignalP and TargetP software tools (i.e. it contains a secretory signal; Nielsen et al., 1997; Emanuelsson et al., 2000) and the LocateProtein system (i.e. it is non-mitochondrial; Schneider,

1999). YB20_PLAFA from chromosome 2 belongs to an uncharacterized protein family that spans over the Eubacteria and Eukaryota (UPF0112 family; Gardner et al., 1998). In more recent versions of the SWISSPROT database it is annotated as ISPF_PLAFD, an enzyme containing a secretory signal and being involved in the isoprenoid biosynthesis. Based on this, its apicoplast location can be confirmed. Multiple sequence alignments were produced for the remaining set of 145 sequences using CLUSTAL-W with the BLOSOM62 matrix and standard gap penalties as in the original publication (Thompson et al., 1994). By visual inspection of the alignments, we excluded 57 sequences to limit bias in the sequence set. Fourteen known mitochondrial protein precursors were added to facilitate the extraction of apicoplast-specific targeting signal features. The final set of negative examples ('non-apicoplast') contained 102 sequences.

2.1.3. Chromosome data

Two hundred and five sequences of chromosome 2 of *P. falciparum* were retrieved from the Institute for Genomic Research (TIGR) (Gardner et al., 1998), and 243 sequences of chromosome 3 were retrieved from the Sanger Centre (Bowman et al., 1999).

2.2. Sequence encoding

Each of the 84 likely apicoplast-targeted sequences was dissected into three parts, the anticipated signal and transit peptides, and the mature protein sequence. Subsequent analysis was restricted to the signal and transit peptide portion. The SignalP software (Nielsen et al., 1997) was used to predict the length of the signal peptide in sequences with an unknown signal peptide cleavage site. If SignalP did not find a cleavage site, the first 23 N-terminal amino acids were taken instead. This number of residues represents the average length of a signal peptide, which was calculated on the basis of Nielsen's collection of eukaryotic signal peptides (Nielsen et al., 1997). It is also mirrored in our collection of positive examples containing putative cleavage sites (Fig. 2). The length of transit peptide of the 54 sequences with an unknown PP cleavage site was defined to be 78 residues, because this is the median of the length distribution of the 35 transit peptides with a known PP cleavage site (Fig. 2). The 102 non-apicoplast sequences from the SWISSPROT database were similarly split into two targeting sequence parts, comprising residue positions 1–23 and 24–101, respectively.

In this work we use the terms 'signal peptide part' or 'S-part', and 'transit peptide part' or 'T-part' for the two sequence portions of positive and negative sequences. This was done even though the 102 sequences from the SWISSPROT database most likely will not contain a real signal or transit peptide.

Two different numerical encoding schemes were applied: the amino acid composition and physico-chemical amino acid properties.

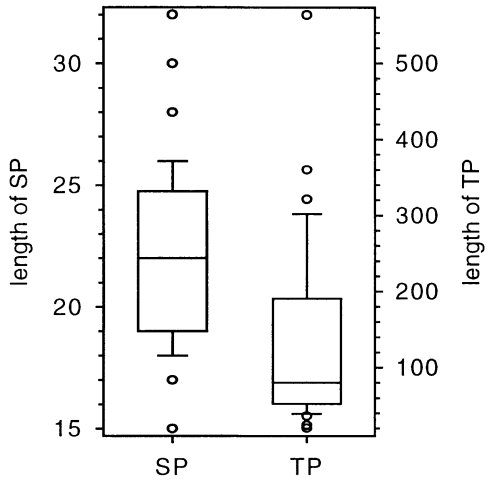

Fig. 2. Lengths of the signal (SP) and transit (TP) peptides of the 35 likely apicoplast-targeted proteins with known PP cleavage site. The box-whisker plots show the median, the 25 and 75% quartiles, the standard deviation and the six most extreme values.

2.2.1. Amino acid composition

For each sequence, two 20-dimensional composition vectors were computed containing the relative residue frequencies of the S- and T-parts. This sequence descriptor was $20 + 20 = 40$-dimensional.

2.2.2. Physico-chemical properties

A set of 19 physico-chemical properties was used to encode the S-part and the T-part, resulting in a 38-dimensional descriptor providing the input for principal component analysis (PCA). The scales were derived from PCA (Jackson, 1991) of 434 amino acid properties (Tomii and Kanehisa, 1996). Analysis of the PCA loadings matrix indicates a correlation of PC1 with hydrophobicity scales, PC2 with secondary structure propensities, and PC3 with the genetic code and residue abundance. The additional principal components did not clearly correlate with single amino acid properties. This result substantiates earlier findings (Schneider and Wrede, 1998). The resulting 38 variables describing a sequence were ranked according to the partial least square (PLS) variable influence on projection (VIP) score (Wold, 1994) as computed by the SIMCA-P 8.0 software package (Umetrics AB, Umeå, Sweden).

2.3. PCA

PCA was used to calculate orthogonal variables from raw data matrices. Our own C-code was written implementing the NIPALS (nonlinear iterative PLS) algorithm for latent variable extraction.

2.4. Projection to latent structures using PLS analysis

PLS analysis was performed in order to rank the amino acid properties according to their ability to discriminate between apicoplast-targeted and non-apicoplast protein

sequences. The ranking was performed utilizing the VIP score provided by the SIMCA-P 8.0 software package (Wold, 1994).

2.5. Self-organizing map (SOM)

Kohonen's SOM algorithm (Kohonen, 1982) was employed to generate two-dimensional projections of the sequence distributions in the spaces spanned by the 40-dimensional composition descriptor and the 38-dimensional property descriptor. We used the NEUROMAP software toolbox (Roche intranet application; Schneider and Wrede, 1998; Schneider, 1999). Toroidal maps containing 64 neurons arranged in a 8×8 rectangular grid were generated for both amino acid composition data and physicochemical property data. Training was aborted after 10^5 optimization cycles (forced stop). The initial learning rate was set to $\tau = 1$, and the initial neighborhood-update radius was $r = 4$. As a distance measure the Manhattan (city block) metric was chosen. For details on the training method, see elsewhere (Schneider and Wrede, 1998).

2.6. Artificial neural network (ANN)

Fully-connected, three-layered, feed-forward networks were used for feature extraction by supervised learning (for details about neural networks see, for example, Hertz et al., 1991). Such systems can be used as nonlinear classifiers. Sigmoidal activation was employed for hidden layer neurons and the single output neuron. Target values were set to one for the 84 nuclear-encoded apicoplast precursors and zero for the 102 other sequences, i.e. the neural network output varied between zero and one, where a value close to one indicates potential apicoplast precursors. All networks were trained using a $(1, \lambda)$ evolution strategy, as implemented in the PROFI software (Schneider and Wrede, 1998; Schneider, 1999). The number of offspring per generation was $\lambda = 500$. Training was stopped after 100 generations. Forty-fold cross-validation was performed with random $8 + 2$ splits of training and test sequences, i.e. 20% cancellation data. Classification and reclassification accuracy was measured by the mean-square-error (mse) and the correlation coefficient according to Matthews (1975).

For the amino acid composition approach, the input layer contained 40 fan-out neurons. Twenty input units received the amino acid composition vector derived from the S-parts, and 20 input units were fed with the amino acid composition vector calculated from the T-parts. Several ANNs with varying numbers of hidden neurons were trained to systematically find the preferred network architecture.

The ANNs that were trained with the property data contained 1–38 fan-out neurons in the input layer, and two to five neurons in the hidden layer. The number of input neurons was systematically increased from 1 to 38 to investigate the influence of the variables on the prediction accuracy.

2.7. Accessibility of the PATS prediction system

The PATS system is based on a 3-4-1 ANN. The www interface accepts FastA sequence format and is accessible through the GECCO!(tm) prediction server on the www at URL: http://www.modlab.de. For further details about the prediction software, see this URL.

3. Results and discussion

A set of 84 apicoplast targeting sequences was compiled and compared to N-terminal parts of 102 non-apicoplast (cytoplasmic, secretory, mitochondrial) sequences. The aim was to extract characteristic targeting signal features and to build a predictive model for *P. falciparum* genome analysis. First, we performed feature extraction by PCA to get an idea of dominant features. PCA and SOM projections were then used to visualize the distribution of apicoplast and non-apicoplast sequences in descriptor space. The identical sets of sequences were used in all experiments to obtain comparable results. Finally, two types of neural networks were trained based either on the amino acid composition or a ranked list of properties providing the input.

3.1. Feature extraction by PCA

The PCA based on amino acid frequencies revealed the following features. After varimax rotation, the first principal component (PC1, eigenvalue = 3.1, 7.7% explained variance) correlates with a high asparagine content in the T-part (loading = 0.7) and a low content of aspartic acid in the S-part (loading = −0.66). The second principal component (PC2, eigenvalue = 2.7, 6.7% explained variance) correlates with a low histidine content in the S-part (loading = −0.73). These findings are in perfect agreement with the observed prevalence of amino acids in apicoplast targeting peptides (Fig. 3), and the amino acid composition observed in secretory signal peptides (von Heijne, 1985).

The second PCA was based on the property descriptor. After varimax rotation, the first principal component (PC1, eigenvalue = 5.4, 14.2% explained variance) correlates with amino acid property component 3 (loading = 0.77) of the transit peptide part, which correlates with general genetic code and residue abundance. The second principal component (PC2, eigenvalue = 4.1, 10.9% explained variance) correlates with amino acid property component 6 (loading = −0.73) of the signal peptide part, which correlates with amino acid frequencies within known secondary structure motifs.

Although this analysis offers some hints as to what the important features of the apicoplast targeting signal might be, one must be careful with the interpretation of loadings here, since the fraction of variance explained by the individual principal components is low.

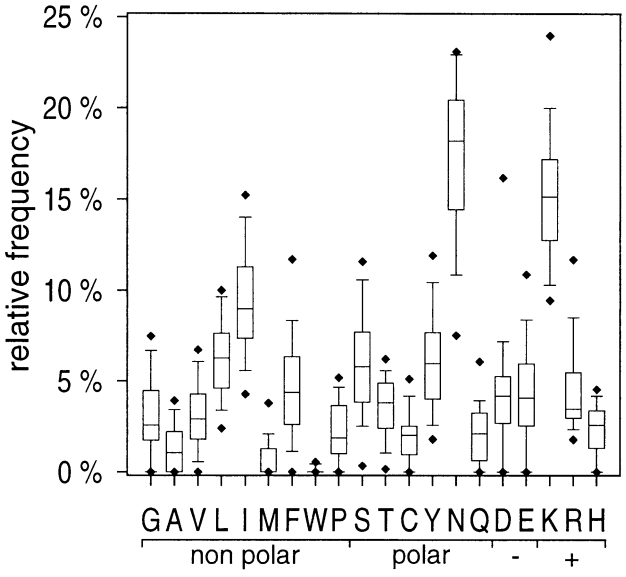

Fig. 3. Average amino acid composition of *P. falciparum* transit peptides ($N = 35$). This box-whisker plot represents the median, 25 and 75% quartiles, and extreme values.

3.2. Mapping of sequences by PCA and SOM and identification of outliers

PCA leads to a linear projection of a high-dimensional space, whereas the SOM projections are inherently nonlinear. The scatter plots resulting from the two first principal components (PC1, PC2) clearly show a separation of apicoplast and non-apicoplast data for both encoding schemes (Fig. 4a,b). This observation is substantiated by the SOM projections, where the apicoplast and non-apicoplast data occupy separated areas (Fig. 4c,d).

Both mappings led to the identification of three potential non-apicoplast proteins, which fall in the 'apicoplast area' of all calculated maps: ASP_PLAFS, EBA1_PLAFC and S230_PLAFO. ASP_PLAFS is annotated as hypothetical aspartic acid-rich protein precursor and is coded on the reverse strand of a histidine-rich protein. A BLAST2 search in the SWISSPROT database with this query did not result in any other significant hits, thereby favoring its hypothetical nature. It remains unclear if this hypothetical protein would actually be targeted to the apicoplast. Both EBA1_PLAFC and S230_PLAFO are known to be non-apicoplast-targeted. EBA1_PLAFC is the gene for erythro-

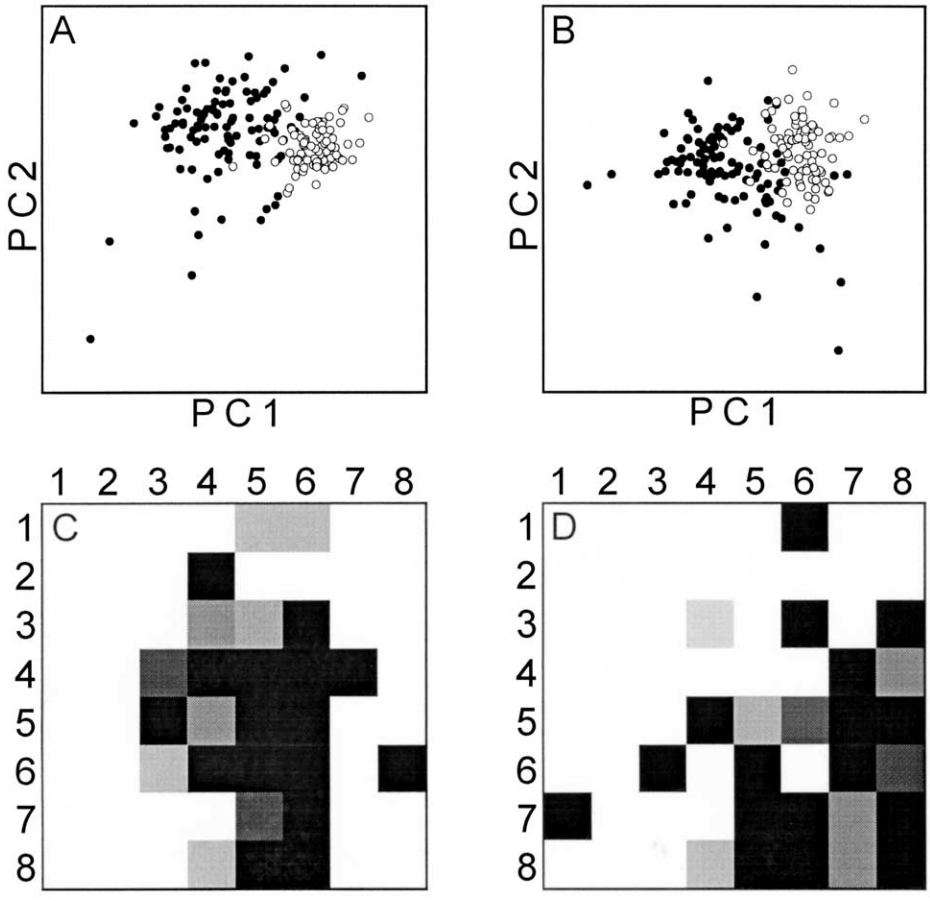

Fig. 4. Distributions of nuclear-encoded apicoplast-targeted and putative non-apicoplast sequences. Sequences were described by amino acid composition (a,c) and physico-chemical properties (b,d). PCA projections are shown in (a,b) (white, apicoplast; black, non-apicoplast). SOM projections are shown in (c,d), where gray shading indicates the location of apicoplast proteins (dark, many; white, none). Note that the SOMs form a torus.

cyte binding antigen EBA-175, which is located in the microneme organelle. The existence of a secretion signal in the two precursors is evident, and there is experimental proof of the cell surface location of S230_PLAFO.

Several additional sequences cluster together with the known apicoplast proteins in some of the maps: ABRA_PLAFC, GARP_PLAFF, ARP2_PLAFA, PF2L_PLAFP, PF12_PLAFA, RPOB_PLAFA and ERD2_PLAFA. All but RPOB_PLAFA and ERD2_PLAFA have SWISSPROT annotations related to 'antigen' or 'surface protein', or are involved in cell–cell recognition. ABRA_PLAFC has also been found on the surface of the parasitophorous vacuole. GARP_PLAFF, a glutamic acid-rich sequence, might also be an apicoplast-targeted protein. It contains a signal peptide, and the amino acid composition of the putative transit peptide part classifies it as apicoplast-targeted. Both ARP2_PLAFA and PF2L_PLAFP are incomplete sequences annotated as a 'malaria antigen'. PF12_PLAFA is annotated as a 'membrane antigen'. RPOB_PLAFA is annotated as 'DNA directed RNA polymerase'. ERD2_PLAFA is a receptor required for protein retention in the ER, and immunolocalization studies indicated that this protein is concentrated in the *cis* Golgi. RPOB_PLAFA and ERD2_PLAFA were the only obviously false-predicted proteins that fell in the 'apicoplast area' in some of the maps. Both sequences are considered to lack a signal peptide according to the SignalP software. We observed that the SOM clustering results slightly depend on the training conditions chosen, e.g. the learning update radius or the total number of update cycles. Thus, one might also attribute the mis-classification of RPOB_PLAFA, for example, as an artifact resulting from SOM training errors. Such problems are known for conventional Kohonen-type SOM (Kohonen, 1982).

Three of our presumably positive examples, DNAJ (EMBL ID: AB016024), '50s rpl7/12' (PlasmoDB entry no. 89141541), and 'Leucine aminopeptidase' (PlasmoDB entry no. 89143622) (The Plasmodium Genome Consortium, 2001), do not cluster together with the other apicoplast-targeted sequences. DNAJ is a heat shock protein (Watanabe, 1997). The '50s rpl7/12' (neuron 8/6 in Fig. 4c) and 'Leucine aminopeptidase' were expected to be apicoplast-targeted, because of their predicted signal peptides, their N-terminal extensions and their homology to known plastid proteins. However, mitochondrially targeted copies of these proteins may also exist. Therefore, it might be possible that these two proteins are indeed non-apicoplast-targeted.

In summary, the two-dimensional projections of high-dimensional descriptor space showed that our choice of descriptors provides a useful starting point for subsequent neural network training. Both encoding schemes seem to be equally suited. Furthermore, we were able to identify some examples of proteins, which were probably mis-classified in our data sets. We did not remove them from our database, because these singletons could also be attributed to mapping errors.

3.3. Neural network training

An ANN-based prediction system was developed to classify unknown protein sequences from *P. falciparum*. First, a system was developed using the amino acid composition patterns. A three-layered ANN containing two hidden units (40-2-1 architecture) was trained with the complete *P. falciparum* data set (84 positive, 102 negative examples; 40-dimensional input vector). The Matthews correlation coefficient on the training data was $cc_{train} = 0.98$, with one false-positive (ASP_PLAFS) and one false-negative prediction (DNAJ). The test data prediction yielded $cc_{test} = 0.75$. A system with three hidden neurons yielded $cc_{train} = 0.99$ and $cc_{test} = 0.76$. ANNs with additional hidden neurons showed no mis-predictions of the training patterns ($cc_{train} = 1.0$), and reach test data accuracy of $cc_{test} = 0.77$. From the cross-validation study results we assume that this behavior is due to over-fitting (Table 1).

A common way to minimize the effect of over-fitting is to diminish the dimensionality of the ANN input vector. We achieved this by changing the encoding scheme of the S- and T-parts from amino acid abundance to 19 physicochemical descriptors. The resulting 38 input dimensions of each sequence were sorted by their VIP score in a PLS model. ANNs with the full set of descriptors showed very similar behavior to the ANNs that were trained with the amino acid composition vectors (Table 1). Systematically changing the dimension of the input vectors and hidden neurons resulted in a 3-4-1 net topology yielding $cc_{train} = 0.91$ (97% correct prediction) and $cc_{test} = 0.87$ (93% correct prediction) (Fig. 5). The three input dimensions corresponded to the first amino acid property of the S-part (hydrophobicity), and the third and fifth properties of the T-part. The ANN showed the highest Matthews correlation coefficient and the lowest mse for cross-validation test data of all tested networks (Table 1).

Table 1
Neural network prediction (average values obtained from 40-times cross-validation)

Network architecture	Training data		Test data	
	mse $\times 10^{-3}$	cc	mse $\times 10^{-3}$	cc
Networks based on the amino acid composition				
40-2-1	8.4	0.98	112.5	0.75
40-3-1	6.9	0.99	107.3	0.76
40-4-1	5.2	0.99	102.5	0.77
40-5-1	5.5	0.99	102.0	0.76
Networks based on physico-chemical properties				
38-2-1	4.2	0.99	113.2	0.76
38-3-1	4.7	0.99	101.4	0.77
38-4-1	1.8	1.00	103.2	0.76
38-5-1	2.4	1.00	98.7	0.77
3-2-1	45.2	0.90	54.4	0.86
3-3-1	42.1	0.91	55.7	0.87
3-4-1	41.0	0.91	54.7	0.87
4-5-1	39.4	0.91	59.2	0.86

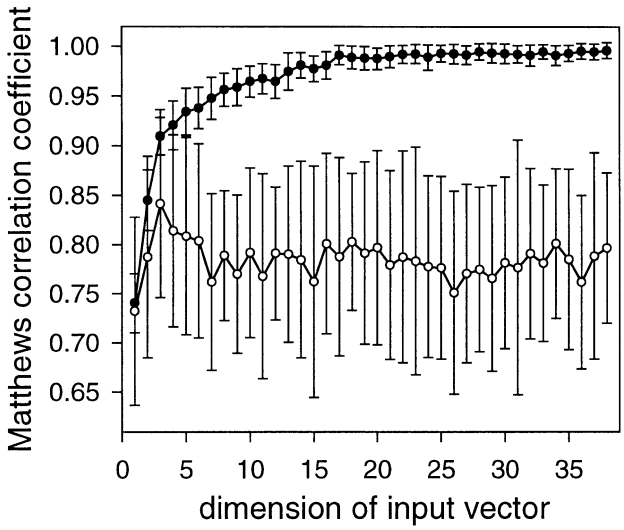

Fig. 5. Mean values and standard deviations of the Matthews correlation coefficient (cc) for ANNs with different input vector dimensions. The property values providing the network input were sorted by VIP rank. Dark circles show values for the training data, while white circles indicate values from the cross-correlation testing data.

A network with this 3-4-1 topology was trained with a complete *P. falciparum* data set to form the PATS prediction system. For reclassification of the training data this ANN reached a correlation coefficient of cc = 0.91. Three false-negatives (Leucine aminopeptidase, '50s rpl7/12' and DNAJ) and five false-positives (EBA1_PLAFC, ERD2_PLAFA, PF12_PLAFA, RPOB_PLAFA and S230_PLAFO) were observed. These proteins were already identified as outliers in one or more SOM or PCA projections. Using the SignalP software to reduce the number of false-positives, the reclassification power of the whole prediction system increases to cc = 0.94, because ERD2_PLAFA and RPOB_PLAFA do not contain a clear signal peptide.

3.4. Analysis of P. falciparum chromosome data

How many apicoplast-targeted proteins are to be expected on the *P. falciparum* genome? It was initially estimated that plant chloroplasts contain between 1000 and 5000 proteins, the vast majority of which are nuclear-encoded (Martin and Herrmann, 1998). Analysis of the finished *Arabidopsis* genome predicts 3574 plastid-targeted proteins (The Arabidopsis Genome Initiative, 2000). In *Plasmodium* the apicoplast will lack the photosynthetic enzymes (Wilson et al., 1996) and the enzymes of the shikimate pathway for aromatic amino acids (Keeling et al., 1999), but may have some others that chloroplasts lack. At a very rough guess, we expect that the nuclear genome will encode for between 500 and 1500 apicoplast proteins (Waller et al., 2000). The entire *Plasmodium* genome contains an estimated 9000 proteins, so between 6 and 17% of all proteins could be apicoplast-targeted. The *P. falciparum* project is ongoing and the complete genome sequence will be available in the near future. So far, only the data for chromosomes 2 and 3 are available for sequence analysis (Gardner et al., 1998; Bowman et al., 1999). Chromosome 2 encodes 205 recognized proteins. Our PATS system recognized 45 (22%) of these sequences as potentially apicoplast-targeted. On chromosome 3 there are 243 proteins encoded, of which 51 (21%) were classified as potentially apicoplast-targeted by PATS. Subsequent analysis with the SignalP software excluded 11 proteins from chromosome 2 and 16 proteins from chromosome 3, because of the lack of a signal peptide. Manual inspection of the resulting 69 (15%) protein sequences from both chromosomes led to the assumption that 38 of these are indeed very likely to be apicoplast-targeted. This means that at least 8.5% of the proteins encoded on chromosome 2 and 3 likely are apicoplast-targeted. This number is well within the rough estimate discussed above.

In addition to those proteins targeted to the apicoplast via the canonical bipartite presequence, another population of apicoplast-targeted proteins may exist, employing hitherto unrecognized targeting motifs. As is the case for plant chloroplasts, this may apply particularly to membrane proteins (Schleiff and Soll, 2000). It has been estimated that as many as several hundred proteins are targeted to the *Arabidopsis* chloroplast in processes independent of transit peptides (Abdallah et al., 2000). The detection of such proteins in *Plasmodium* awaits a proteomic analysis of isolated apicoplasts or a systematic tagging of predicted proteins from the genome project.

The PATS prediction system complements existing prediction tools available for plant chloroplast transit peptides, which can be predicted at a reasonable level of accuracy, e.g. by neural network models like ChloroP (Emanuelsson et al., 2000) or knowledge-based systems like PSORT (Nakai and Kanehisa, 1992). Nevertheless, there is room for further improvement. In the current version of the software, no attempts are made to predict the actual lengths of the S- and T-parts, although the software can handle differing lengths for each predicted sequence. While highly sophisticated systems for the recognition of signal peptides and their cleavage sites are available (Nakai and Kanehisa, 1992; Nielsen et al., 1997), no such system is available for the cleavage site of transit peptides. Such a system could employ hidden Markov models, which were shown to be particularly suited for this kind of problem (Hughey and Krogh, 1996). Another approach could make use of the Shannon information content of apicoplast-targeted sequences, either encoded through amino acid abundance or physico-chemical properties (Zuegge et al., 2001). Unfortunately, all of these methods require a substantial number of known examples and there is currently little information on apicoplast leader cleavage sites. We are confident that the current version of the PATS software will help to identify new unknown apicoplast-targeted sequences, and thereby enable the development of improved transit peptide prediction systems.

Acknowledgements

We thank Karin Zuegge for the helpful comments on the manuscript and careful editing.

References

Abdallah, F., Salamini, F., Leister, D., 2000. A prediction of the size and evolutionary origin of the proteome of chloroplasts of Arabidopsis. Trends Plant Sci. 5, 141–142.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Bowman, S., et al., 1999. The complete nucleotide sequence of chromosome 3 of Plasmodium falciparum. Nature 400, 532–538.

Cline, K., Henry, R., 1996. Import and routing of nucleus-encoded chloroplast proteins. Annu. Rev. Cell Dev. Biol. 12, 1–26.

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

Etzold, T., Ulyanov, A., Argos, P., 1996. SRS: information retrieval system for molecular biology data banks. Methods Enzymol. 266, 114–128.

Gardner, M.J., 1999. The genome of the malaria parasite. Curr. Opin. Genet. Dev. 9, 704–708.

Gardner, M.J., et al., 1998. Chromosome 2 sequence of the human malaria parasite Plasmodium falciparum. Science 282, 1126–1132.

Gleeson, M.T., 2000. The plastid in Apicomplexa: what use is it? Int. J. Parasitol. 30, 1053–1070.

Hertz, J., Palmer, R.G., Krogh, A.S., 1991. Introduction to the Theory of Neural Computation, Addison-Wesley, Redwood City, CA.

Hughey, R., Krogh, A., 1996. Hidden Markov models for sequence analysis: extension and analysis of the basic method. Comput. Appl. Biosci. 12, 95–107.

Jackson, J.E., 1991. A User's Guide to Principal Components, Wiley, New York.

Keeling, P.J., Palmer, J.D., Donald, R.G., Roos, D.S., Waller, R.F., McFadden, G.I., 1999. Shikimate pathway in apicomplexan parasites. Nature 397, 219–220.

Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biol. Cybern. 43, 59–69.

Martin, W., Herrmann, R.G., 1998. Gene transfer from organelles to the nucleus: how much, what happens, and why? Plant Physiol. 118, 9–17.

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

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.

Schleiff, E., Soll, J., 2000. Travelling of proteins through membranes: translocation into chloroplasts. Planta 211, 449–456.

Schneider, G., 1999. How many potentially secreted proteins are contained in a bacterial genome? Gene 237, 113–121.

Schneider, G., Wrede, P., 1998. Artificial neural networks for computer-based molecular design. Prog. Biophys. Mol. Biol. 70, 175–222.

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

The Plasmodium Genome Consortium, 2001. PlasmoDB: an integrative database of the Plasmodium falciparum genome. Tools for accessing and analyzing finished and unfinished sequence data. The Plasmodium Genome Database Collaborative. Nucleic Acids Res. 29 (1), 66–69.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Tomii, K., Kanehisa, M., 1996. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng. 9, 27–36.

von Heijne, G., 1985. Signal sequences. The limits of variation. J. Mol. Biol. 184, 99–105.

Waller, R.F., Reed, M.B., Cowman, A.F., McFadden, G.I., 2000. Protein trafficking to the plastid of Plasmodium falciparum is via the secretory pathway. EMBO J. 19, 1794–1802.

Watanabe, J., 1997. Cloning and characterization of heat shock protein DnaJ homologues from Plasmodium falciparum and comparison with ring infected erythrocyte surface antigen. Mol. Biochem. Parasitol. 88, 253–258.

WHO, 1997. World malaria situation in 1994. Part I. Population at risk. Wkly. Epidemiol. Rec. 72, 269–274.

Wilson, R.J., Denny, P.W., Preiser, P.R., Rangachari, K., Roberts, K., Roy, A., Whyte, A., Strath, M., Moore, D.J., Moore, P.W., Williamson, D.H., 1996. Complete gene map of the plastid-like DNA of the malaria parasite Plasmodium falciparum. J. Mol. Biol. 261, 155–172.

Wold, S., 1994. Exponentially weighted moving principal component analysis and projections to latent structures. Chemometr. Intell. Lab. Syst. 23, 149–161.

Zuegge, J., Ebeling, M., Schneider, G., 2001. H-BloX: visualizing alignment block entropies. J. Mol. Graph. Model. 19, 303–305.