Available online at www.sciencedirect.com
SCIENCE DIRECT®
Molecular & Biochemical Parasitology 132 (2003) 59–66
MOLECULAR & BIOCHEMICAL PARASITOLOGY

Properties and prediction of mitochondrial transit peptides from *Plasmodium falciparum*

Andreas Bender[a], Giel G. van Dooren[b], Stuart A. Ralph[b], Geoffrey I. McFadden[b], Gisbert Schneider[a,*]

[a] *Johann Wolfgang Goethe-Universität Frankfurt, Institut für Organische Chemie und Chemische Biologie, Marie-Curie-Strasse 11, D-60439 Frankfurt, Germany*
[b] *Plant Cell Biology Research Centre, School of Botany, University of Melbourne, Parkville, Vic. 3010, Australia*

Received 8 May 2003; received in revised form 30 July 2003; accepted 30 July 2003

Abstract

A neural network approach for the prediction of mitochondrial transit peptides (mTPs) from the malaria-causing parasite *Plasmodium falciparum* is presented. Nuclear-encoded mitochondrial protein precursors of *P. falciparum* were analyzed by statistical methods, principal component analysis and supervised neural networks, and were compared to those of other eukaryotes. A distinct amino acid usage pattern has been found in protein encoding regions of *P. falciparum*: glycine, alanine, tryptophan and arginine are under-represented, whereas isoleucine, tyrosine, asparagine and lysine are over-represented compared to the SwissProt average. Similar patterns were observed in mTPs of *P. falciparum*. Using principal component analysis (PCA), mTPs from *P. falciparum* were shown to differ considerably from those of other organisms. A neural network system (*Plas*Mit) for prediction of mTPs in *P. falciparum* sequences was developed, based on the relative amino acid frequency in the first 24 N-terminal amino acids, yielding a Matthews correlation coefficient of 0.74 (90% correct prediction) in a 20-fold cross-validation study. This system predicted 1177 (22%) mitochondrial genes, based on 5334 annotated genes in the *P. falciparum* genome. A second network with the same topology was trained to give more conservative estimate. This more stringent network yielded a Matthews correlation coefficient of 0.51 (84% correct prediction) in a 10-fold cross-validation study. It predicted 381 (7.1%) mitochondrial genes, based on 5334 annotated genes in the *P. falciparum* genome.

Keywords: Neural network; Principal component analysis; Protein targeting; Sequence analysis; Transit peptide

1. Introduction

The malaria-causing parasite *Plasmodium falciparum* is a major cause of human death and morbidity in many regions of the world. The recent publication of the complete nuclear genome of *P. falciparum* is a significant advance in studying the biology of the parasite [1]. With a wealth of genome data available, it is now important to develop tools to make sense of these data.

Several methods for the prediction of subcellular locations of nuclear-encoded proteins in eukaryotic organisms were developed over the past decade. Different approaches to this issue have been conceived, such as artifical neural networks in case of TargetP [2] or linear discriminant techniques applied to physicochemical parameters in case of MitoProtII [3]. In cases where neural networks were employed, relative amino acid fractions were most commonly used for describing the input data [4,5]. TargetP and MitoProtII yielded Matthews correlation coefficients [6] of 0.46 and 0.66, respectively, when applied to human protein test sequences [7]. The scenario in *P. falciparum* is quite different. Using a set of 40 putative mitochondrially targeted peptides (positive data set) and 135 cytosolic, extracellular and apicoplast peptides (negative data set), the programs performed as follows. MitoProtII achieved a Mathews coefficient of cc = 0.49 with a sensitivity (positive predictions divided by total number of positive sequences) of 0.8 (32/40) and a selectivity (true positive divided by total number of positive predictions) of 0.47 (32/68). TargetP (plant-network) achieved a Mathews coefficient of cc = 0.60 with a sensitivity of 0.55 (22/40) and a selectivity of 0.81 (22/27). Low selectivity in case of MitoProtII and low sensitivity in case of TargetP made the development of a new method for the prediction of mitochondrial transit peptides (mTPs) in *P. falciparum*

* Corresponding author. Tel.: +49-69-798-29821; fax: +49-69-798-29826.
E-mail address: G.Schneider@chemie.uni-frankfurt.de (G. Schneider).

0166-6851/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.molbiopara.2003.07.001

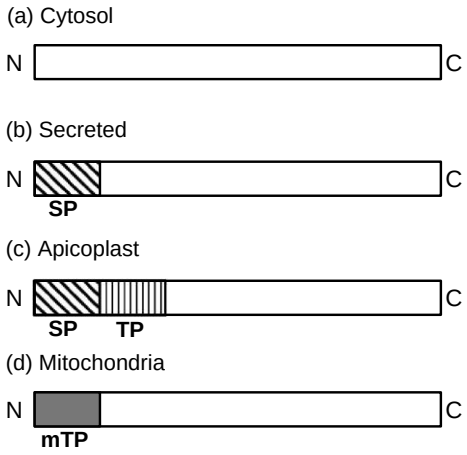

Fig. 1. Schematic of protein precursors targeted to four different cellular locations. Targeting signals are indicated by grey shading. SP: signal peptide, TP: transit peptide, mTP: mitochondrial transit peptide.

necessary. This new tool complements earlier work on the prediction of apicoplast-targeted sequences in *P. falciparum* [8].

The mitochondrial genome of *P. falciparum* is only 6 kb in size—the smallest yet discovered—and holds only three protein-coding genes [9]. The great majority of its proteins are nuclear encoded and have to be imported. Proteins are generally targeted to their destination via appropriate targeting signals (Fig. 1). Cytosolic proteins do not contain targeting signals and therefore remain in the cytoplasm after translation. Secreted proteins contain an N-terminal targeting signal—the signal peptide—whereas apicoplastical proteins contain an N-terminal signal peptide in conjunction with a transit peptide. Most mitochondrial targeting signals are located at the N-terminus of the nascent amino acid chain, although examples of internal and C-terminal targeting motifs are also known [7]. N-terminal mTPs of other organisms generally have a positive net charge due to a significantly higher content of Arg; they also feature an enrichment in Ala and Ser and a lower content of negatively charged residues [7]. They are thought to form positively charged, amphiphilic α-helices as a dominant characteristic [10], mTPs interact with multiprotein translocase complexes at the outer and inner mitochondrial membranes, and are necessary and sufficient to faciliate protein translocation across the mitochondrial membranes into the mitochondrial matrix [11]. Once in the matrix, mTPs are removed by a mitochondrial processing peptidase (MPP) [12].

Here we describe how mitochondrial transit petides in *P. falciparum* differ from those found in other eukaryotic organisms. We describe the development and performance of a software tool capable of predicting mTPs of *P. falciparum*, based on their N-terminal amino acid composition. This tool is called *Plas*Mit (for *Plasmodium* mitochondrial transit peptide prediction). In cross-validation studies it was shown to outperform established tools such as TargetP and MitoProtII when applied to *P. falciparum* sequences.

2. Materials and methods

2.1. Sequence retrieval and data sets

Both positive data, i.e. mitochondrial transit peptides taken from the protein precursors, and negative data, i.e. cytosolic proteins, secreted and apicoplast protein precursors, were compiled based on sequence similarity studies, and, in several cases, from direct experimental observation. Sequence data for *P. falciparum* were obtained from the Institute for Genomic Research website (http://www.tigr.org). All data sets used in this study can be downloaded from URL http://www.modlab.de.

2.1.1. Positive examples

Sequences were included in the set of positive examples when they fulfilled at least one of the following criteria.

(a) They are homologous to proteins exclusively or usually found in mitochondria of other organisms (e.g. proteins of the citric acid cycle, the electron transport chain, ubiquinone biosynthesis).
(b) In constructing a phylogenetic tree, they branch with proteins known to be mitochondrial. The protein of interest was included in a multiple sequence alignment using ClustalW [13] and Pima 1.4 [14] of the BCM SearchLauncher (http://searchlauncher.bcm.tmc.edu). The aligned sequences were imported into PAUP 4.0b and adjusted manually. For the phylogenetic tree analysis, regions with poor sequence conservation were removed. Neighbour-joining trees were produced and boostrap values using 500–1000 replicated were obtained. Values above 70% bootstrap support for a mitochondrial grouping and were used as an indication for mitochondrial proteins.
(c) Proteins that have been experimentally shown to localize in the mitochondrion (only dihydroorotate dehydrogenase and δ-aminolaevulinic acid synthase (ALAS) fall into this category [15,16]).

A total of 40 positive examples was collected this way.

2.1.2. Negative examples

Negative examples, analogously, included proteins that were experimentally assigned to “non-mitochondrial” locations in *P. falciparum*, and proteins based on sequence similarity studies using proteins usually found in “non-mitochondrial” compartments of other eukaryotic organisms. One hundred and thirty-five negative sequences were incorporated into the negative data set.

Table 1
Amino acid usage in *P. falciparum* compared to other organisms

One-letter amino acid code	Overall residue frequencies (%)			Residue frequencies in mTPs (%)		
	P. falciparum	SwissProt database 40.28	*P. falciparum*: SwissProt ratio	*P. falciparum*	Sample of 282 eukaryotic mTPs	*P. falciparum*: other eukaryote ratio
A	2.00	7.70	0.26	1.70	13.20	0.13
C	1.80	1.60	1.13	2.70	1.80	1.50
D	6.40	5.28	1.21	0.60	0.50	1.20
E	7.00	6.50	1.08	1.80	0.90	2.00
F	4.40	4.08	1.08	7.20	3.60	2.00
G	3.00	6.90	0.43	2.60	5.70	0.46
H	2.50	2.30	1.09	2.30	1.50	1.53
I	9.10	5.90	1.54	8.80	3.00	2.93
K	11.70	6.00	1.95	15.80	4.10	3.85
L	7.60	9.60	0.79	8.80	13.30	0.66
M	2.30	2.37	0.97	6.00	5.60	1.07
N	13.90	4.30	3.23	9.70	2.10	4.62
P	2.00	4.91	0.41	1.70	4.40	0.39
Q	2.70	3.90	0.69	2.10	3.10	0.68
R	2.90	5.20	0.56	7.50	12.20	0.61
S	6.30	7.00	0.90	6.80	11.30	0.60
T	4.20	5.50	0.76	3.10	5.50	0.56
V	4.00	6.70	0.60	4.10	5.70	0.72
W	0.40	1.24	0.32	0.90	1.20	0.75
Y	5.80	3.10	1.87	5.90	1.10	5.36

In columns 2–4, bulk coding sequence from *P. falciparum* is compared to SwissProt release 40.28. In columns 5–7, *P. falciparum* mTPs are compared to 282 mTPs from other eukaryotes.

2.1.3. *P. falciparum* Chromosome data

Five thousand three hundred and thirty-four annotated genes from the PlasmoDB database were used [1,17,18].

2.1.4. Codon usage in *P. falciparum*

Amino acid usage in the peptide encoding regions of *P. falciparum* was calculated from the "PlasmoDB—*Plasmodium falciparum* Codon Usage Table" given at the PlasmoDB website (http://www.plasmodb.org). The sample contained annotated peptide encoding regions from Chromosome 2 and Chromosome 3. The calculated residue frequencies are given in Table 1.

2.2. Sequence encoding

Our data collection suggests that mTPs of *P. falciparum* vary greatly in length (varying from 23 to 169 amino acids, on average 64 amino acids with a standard deviation of 48 amino acids). Therefore, in all cases a fixed number of N-terminal amino acids was used in our analysis. Their length in other eukaryotes, taken from a sample of 422 mitochondrial transit peptides from SwissProt, was determined to have a median length of 31 amino acids, with the first quartile at 24 and the third quartile at 42 amino acids. All data sets were cut to these three different lengths. We found that all sequences, apart from one cytosolic peptide, did not have more than 50% pair-wise sequence identity (with respect to different amino acids among the first 24, 31 or 42 residues, respectively, of all sequences from the same localization group. Comparisons were performed using JalView [19]). Therefore, no further redundancy reduction seemed to be necessary and all sequences were used for further analysis.

For each sequence, a 20-dimensional composition vector was computed containing the relative residue frequencies among the first 24 N-terminal amino acids. These vectors were used for both principal component analysis (PCA) and neural network training.

2.3. PCA

The principal component analysis tool of Statistica was used to calculate linear independent variables from the raw data matrices [20].

2.4. Artificial neural network (ANN)

Fully-connected, three-layered, feed-forward networks were used for feature extraction by supervised learning. The neural network code was generated by Statistica [20]. In the hidden layer, a hyperbolic activation function, and in the output layer, a logistic activation function, were used. All networks were trained using the standard Back-Propagation (BP) algorithm [21]. Twenty-fold cross validation was performed with random splits of 89 sequences in the training set and 43 sequences in the select and test sets, respectively. The training set was used for ANN parameter optimization, whereas the select set was used for evaluation of learning progress and stopping conditions. Prediction performance was evaluated using the test set. The condition for continuation of training was an improvement of the fraction of correctly predicted sequences within 10^4 iterations. Ran-

dom initialization of network weights and a fixed learning rate of 0.01 were used. The input layer contained 20 fan-out neurons. Several ANNs with varying numbers (1–50) of hidden neurons were trained to systematically find the preferred network architecture. Genetic variable selection was performed using the best performing architecture [22], leading to a final network architecture with 13 neurons in the input layer.

2.5. Accessibility of the PlasMit prediction system

The final *Plas*Mit system is based on a 13-3-1 ANN. The www interface accepts FastA sequence format and is accessible at http://gecco.org.chemie.uni-frankfurt.de. For technical details about the prediction software, see this URL.

3. Results and discussion

A set of 40 mTPs from *P. falciparum* was compiled and compared to N-terminal parts of 135 non-mitochondrial (61 cytoplasmic, 21 secretory and 53 apicoplast) sequences. The aim was to extract characteristic mTP features and to build a predictive model for *P. falciparum* genome analysis. First we performed principal component analysis to get an idea of dominant features and the data distribution. Then residue frequencies of both the overall (genome-derived) and of the first 24 N-terminal amino acids of mitochondrial transit peptides of *P. falciparum* were compared with those of other eukaryotes. In addition, relative frequencies of amino acid groups with respect to important physicochemical properties—small, hydrophobic, negatively charged, positively charged, polar—were compared among proteins of *P. falciparum*. Neural networks with variable numbers of neurons in the hidden layer were trained using the first 24, 31 and 42 amino acids for calculation of relative amino acid frequencies. Finally, the best performing neural network was used to predict the number of mitochondrially targeted proteins based on the annotated genome of *P. falciparum*. An additional network using the same topology was trained to give a lower number of false-positive results.

3.1. Feature extraction by principal component analysis

The PCA based on relative amino acid frequencies (Fig. 2) suggests that there is a difference between mitochondrial transit peptides of *P. falciparum* and those of other eukaryotic organisms. After varimax rotation, the first principal component (PC1, 17.4% explained variance) correlates with histidine (loading = 0.79) and alanine content (loading = −0.64). Both correlations are in accordance with the observed prevalence of amino acids in mTPs (Table 1). The second principal component (PC2, 8.3% explained variance) correlates with a serine (loading = 0.66) and glutamic acid content (loading = −0.43). This is again in accordance with the observed prevalence of amino acids of mTPs (Table 1).

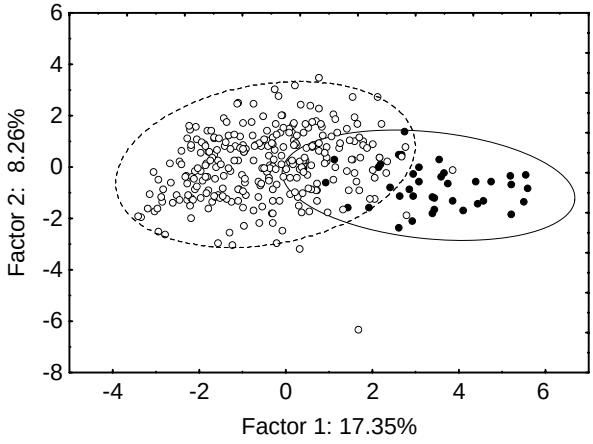

Fig. 2. Principal component analysis of the relative amino acid frequencies of the first 24 N-terminal amino acids reveals differences between mTPs from *P. falciparum* and mTPs from other eukaryotes. In light of these differences it is unsurprising that generic tools (TargetP and MitoProtII) recognise *P. falciparum* mTPs only poorly. Solid dots: mTPs from *P. falciparum*, circles: mTPs from other eukaryotes. The 95% confidence ellipsoids are shown.

It should be stressed that the direction a PC takes, with respect to the original variables, is arbitrary. When PCs are calculated for the same data set using two different software packages, it is not unusual to find that the signs of the loadings of the variables on corresponding PCs (e.g. the first PC from the two programs) are reversed. The component just defines that the “serine-glutamic acid direction” bears high information content with respect to the given data, but the signs are arbitrary. In conclusion, Fig. 2 suggests that there are sufficient differences between mTPs from *P. falciparum* and mTPs from other organisms to render the application of established tools like TargetP and MitoProtII unviable.

PCA was primarily used to visualize the underlying data distribution revealing that a distinction between positive and negative examples seemed possible. A linear classifier was tested based on the first two principal components yielding 12 erroneous classifications (7%). For the final classification system artificial neural networks were used because of their superior performance in this case and with our datasets (2% error; *vide infra*).

In addition to classification by using relative amino acid frequencies, we also attempted classification by using a PCA of several hundred amino acid properties (not shown). Classification by using physicochemical properties did not outperform classification by relative amino acid frequencies. We therefore remained with the relative amino acid frequency encoding.

3.2. Comparison of residue usage in P. falciparum mTPs and its whole genome to the residue usage in other organisms

Our data indicate considerable differences in amino acid residue usage between *P. falciparum* and other eukaryotic

organisms (Table 1). Glycine, alanine, tryptophan and proline occur less than half as often in the sample of peptide encoding regions from Chromosomes 2 and 3 of *P. falciparum*, compared to residue frequencies in the SwissProt database, version 40.28. This is also true for glycine, alanine and proline, comparing mTPs of *P. falciparum* to mTPs of other organisms. Asparagine, lysine, tyrosine and isoleucine occur more than 1.5 times as often in the annotated regions of *P. falciparum* than in the reference database mentioned above. This observation holds when mTPs of *P. falciparum* and mTPs of other eukaryotes are compared. In addition, phenylalanine occurs twice as often in mTPs of *P. falciparum*. Much of the difference in amino acid usage between *P. falciparum* and other eukaryotes in both mTPs and in the overall genome can be explained by relating amino acid usage to the low G + C nucleotide bias in *P. falciparum*. The G + C content of *P. falciparum* protein-coding regions is estimated to be 24% [1]. This makes *P. falciparum* the eukaryote with the lowest G + C content so far sequenced.

Lobry has shown a connection between G + C content of DNA and relative amino acid frequency for 59 bacterial genomes [23]. It was found that G + C-depleted genomes contain low amino acid frequencies for alanine, arginine, glycine and proline. This is likely due to the codons that encode these residues usually requiring two or more G or C nucleotide bases. In *P. falciparum* these four residues are similarly depleted. Glycine (ratio of 0.44), alanine (0.26), arginine (0.55) proline (0.41), occur less than half or about half as frequently as in the SwissProt database, version 40.28 (Table 1). Lobry also found that G + C-depleted genomes contained relatively higher amino acid frequencies for asparagine, isoleucine, leucine, lysine and tyrosine. This is likely due to the codons that encode these residues usually requiring one or fewer G or C nucleotide bases. In *P. falciparum* asparagine (ratio of 3.2), isoleucine (1.5), lysine (2.0) and tyrosine (1.9) occur more frequently compared to the reference SwissProt database (Table 1). As the only exception in this case, Leucine (0.67) occurs less frequently in the *P. falciparum* genome than in the reference database. The residue usage pattern in *P. falciparum* can thus be partly explained by its genomic base composition, and it is reasonable to assume that this nucleotide bias also contributes to the unsatisfactory performance of TargetP and MitoProtII when applied to *P. falciparum* mTPs.

3.3. Relative frequencies of amino acids of P. falciparum proteins with different locations

Cytosolic proteins generally lack N-terminal targeting motifs (Fig. 1). Secretory proteins, which in *P. falciparum* are targeted to numerous destinations both inside and outside the parasite cell, usually contain a hydrophobic N-terminal signal peptide (Fig. 1). This signal peptide encodes entry to the endomembrane system at the endoplasmic reticulum. Proteins targeted to the apicoplast represent a subset of secretory proteins, with an asparagine- and lysine-rich transit peptide following the signal peptide (Fig. 1) [24]. This transit peptide directs apicoplast proteins from the secretory pathway into the apicoplast [24,25].

Comparison of the physicochemical properties of amino acids that comprise the N-terminus of proteins targeted to these various destinations reveal distinct features (Fig. 3). mTPs differ from the other two classes of N-terminal targeting sequences (secretory and apicoplast) in their lower content of hydrophobic residues. Proteins targeted to mitochondria have the highest amount of positively charged amino acids as well as a scarcity of negatively charged residues, resulting in the highest average positive net charge. Secretory, apicoplast and cytosolic sequences each contain some positively charged residues, but cytosolic proteins also contain a considerable number of negatively charged residues, resulting in a net charge that is relatively neutral compared to apicoplast or secretory sequences.

3.4. Artificial neural network (ANN) training

An ANN-based prediction system was developed to classify protein sequences from *P. falciparum*, based on relative amino acid frequencies of the first 24, 31 and 42 N-terminal residues. Three-layered ANN containing 1–50 hidden units were trained in a 20-fold cross validation with the *P. falciparum* data set (40 positive, 135 negative examples; 89 sequences in training and 43 in each select and test set). A network using 3 hidden neurons was chosen as the best network, because it achieved a high Matthews coefficient (cross validated cc = 0.74), together with low variance ($\sigma^2 = 0.10$) and a low number of hidden neurons.

Genetic selection of variables was performed to reduce the set of input descriptors. The parameters chosen in Statistica were 100 iterations, 100 children per iteration, mutation rate = 1 and a crossover probability of 0.1. Seven amino acids—Cys, Gln, His, Ser, Thr, Trp and Tyr—were found not to improve classification results and were consequently omitted, resulting in a 13-dimensional input vector of relative amino acid frequencies being used (Ala, Arg, Asp, Asn, Glu, Gly, Ile, Leu, Lys, Met, Phe, Pro, Val).

A 20-fold cross validation employing the 13-dimensional input vectors was performed, using an improvement in the classification of the select set as a criterion to end learning. On average, training was ended after 436 epochs, yielding a Matthews coefficient of cc = 0.74, with on average 90% correct prediction. Sensitivity was 0.94 with a selectivity of 0.68. This means that 94% of positive sequences were detected and slightly more than two out of three (68%) positive predictions were true positives.

Using all 175 sequences for training, a Matthews coefficient of cc = 0.92 was achieved, with a sensitivity of 0.98 and a selectivity of 0.91. Of the 175 sequences, only 5 were not correctly classified. The proteins M1 family aminopeptidase, clathrin coat assembly protein, vacuolar proton-pumping pyrophosphatase-2 and the knob-associated histidine-rich protein (KAHRP) were “false positives”,

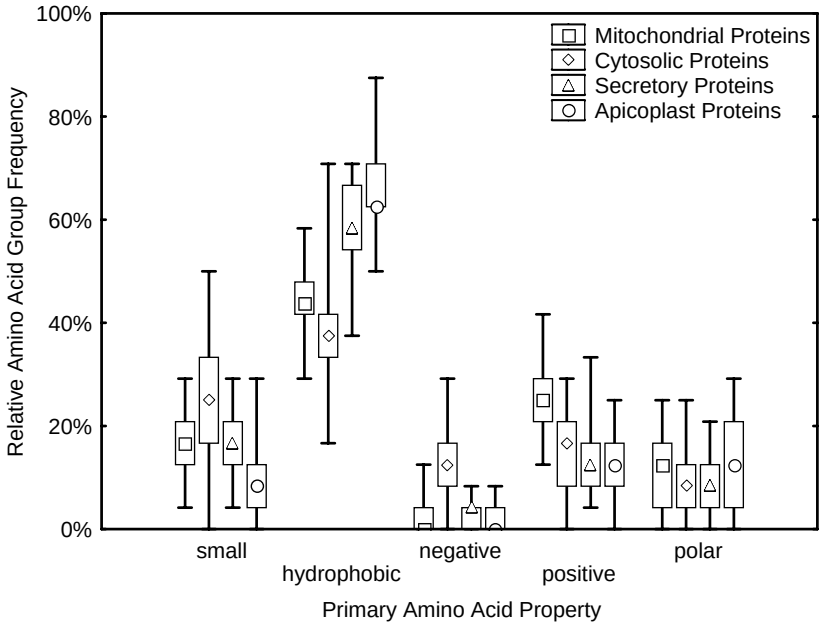

Fig. 3. Relative amino acid distribution of proteins of different locations in *P. falciparum*, grouped by physicochemical properties. The plots show the medians (symbol), quartiles (box), and ranges (whiskers). As in other organisms, all N-terminal targeting signals have a remarkably low content of negatively charged residues. mTPs from *P. falciparum* possess the highest positive net charge, due to the highest content of positively charged residues. In contrast to other organisms, *P. falciparum* generally uses lysine rather than arginine to achieve the positive charge, both overall and within its mTPs.

whereas fumarase class 1 was the only "false negative". Of the overpredicted sequences, both the M1 family aminopeptidase and the vacuolar proton-pumping pyrophosphatase-2 contain large basic N-terminal extensions compared to homologues and may be incorrectly annotated. Indeed the annotating authors in both cases observed some intracellular punctate immunolocalization for their respective proteins [26,27], which with hindsight might indicate mitochondrial localization. KAHRP is known to target to the plasma membrane of the host red blood cell inside which *P. falciparum* lives, and contains an unusual internal signal peptide [28]. Our focus on the N-terminal part of the amino acid sequence does not account for internal signal peptides, so sequences like KAHRP will be classified incorrectly. The reason for the misclassification of the one remaining false-positive and the one false-negative sequence is unknown.

To reduce the number of false-positive results, the relative penalty for false positives to false negatives was set to 3, and the net was retrained in a 10-fold cross-validation study, using again the network with 3 neurons in the hidden layer. A Matthews correlation coefficient of cc = 0.51 was obtained, with only one sequence being overpredicted (vacuolar proton-pumping pyrophosphatase 2), while 26 sequences were underpredicted. To further improve prediction results, in may be possible to incorporate additional information, such as secondary structure elements [21,29]. The role of amphiphilic α-helices in mTPs of other eukaryotes is reasonably well understood [30,31], and it will be fascinating to determine whether mTPs of *P. falciparum* also exhibit α-helix characteristics.

3.5. Analysis of *P. falciparum* chromosome data

We used the network trained with all 175 sequences to analyze the predicted protein-coding sequences from the entire *P. falciparum* genome. Of the 5334 annotated genes, 1177 genes encoding proteins with potential mitochondrial transit peptides (22%) were predicted. This number seems high compared to other organisms. *Saccharomyces cerevisiae* contains about 6000 genes within its nuclear genome [32], with about 500 containing mitochondrial transit peptides (slightly above 8%). In the case of *Arabidopsis thaliana*, automated tools predicted as few as 349 and as many as 2897 mitochondrially targeted genes from a genome size of over 25,000 genes [33]. Therefore, the estimation given here should be regarded as an upper limit to the number of nuclear-encoded mitochondrial proteins in *P. falciparum*. Using the more stringent net with a high penalty for false-positive sequences, 381 potential mitochondrial precursor sequences were predicted, corresponding to 7.1% of the *P. falciparum* genome. This is in a more realistic range, compared to the numbers from other organisms. This number of mTPs is considerably higher than the 246 (4.7%) sequences predicted by TargetP and MitoProtII, as reported in earlier work [1].

The mitochondrion of *P. falciparum* is a little-studied organelle. Most researchers have focused on the role of the *P. falciparum* mitochondrion in electron transport and pyrimidine biosynthesis [15,34–36]. The recent publication of the *P. falciparum* genome and the development of *Plas*Mit as a tool for identifying mitochondrial proteins now provide

us with the basis for future research into the *P. falciparum* mitochondrion. For example, homology searching identified numerous components of the citric acid cycle, which were included in the positive data set. However, the enzymes catalyzing some of the reactions in the cycle are not clear, e.g. there are several candidates for the enzyme catalyzing the dehydrogenation of malate to oxaloacetate. These include a malate dehydrogenase and malate quinone oxidoreductase [1]. *Plas*Mit predicts malate dehydrogenase to be cytosolic rather than mitochondrial, while malate quinone oxidoreductase is predicted to be mitochondrial with a high confidence level, potentially complementing the loss of malate dehydrogenase from the mitochondria. *Plas*Mit also predicts ferrochelatase, the ultimate enzyme in de novo haem biosynthesis, to be mitochondrial. Thus, the first (aminolaevulinate synthase, ALAs) and last enzymes of haem biosynthesis are predicted to be mitochondrial, while it appears that one or more of the remaining enzymes in the pathway may be plastidic [37]. Using *Plas*Mit it will be possible to begin assembling metabolic pathways that putatively occur in the organelle.

Acknowledgements

This research was supported by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften, Frankfurt. The authors are grateful to Jochen Zuegge for helpful discussions, and Akhil Vaidya for advice in compiling the list of putative mitochondrial proteins.

References

[1] Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 2002;419:498–511.
[2] Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 2000;300:1005–16.
[3] Claros MG, Vincens P. Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur J Biochem 1996;241:779–86.
[4] Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res 1998;26:2230–6.
[5] Schneider G. How many potentially secreted proteins are contained in a bacterial genome? Gene 1999;237:113–21 [ibid. 1999;240:245].
[6] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405:442–51.
[7] Emanuelsson O, von Heijne G, Schneider G. Analysis and prediction of mitochondrial targeting peptides. Methods Cell Biol 2001;65:175–87.
[8] Zuegge J, Ralph S, Schmuker M, McFadden GI, Schneider G. Deciphering apicoplast targeting signals—feature extraction from nuclear-encoded precursors of *Plasmodium falciparum* apicoplast proteins. Gene 2001;280:19–26.
[9] Feagin JE. Mitochondrial genome diversity in parasites. Int J Parasitol 2000;30:371–90.
[10] Hammen PK, Weiner H. Mitochondrial leader sequences: structural similarities and sequence differences. J Exp Zool 1998;282:280–3.
[11] Neupert W, Brunner M. The protein import motor of mitochondria. Nat Rev Mol Cell Biol 2002;3:555–65.
[12] Ito A. Mitochondrial processing peptidase: multiple-site recognition of precursor proteins. Biochem Biophys Res Commun 1999;30:611–6.
[13] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL-W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–80.
[14] Smith RF, Smith TF. Automatic generation of primary sequence patterns from sets of related protein sequences. Proc Natl Acad Sci USA 1990;87:118–22.
[15] Krungkrai J. Purification, characterization and localization of mitochondrial dihydroorotate dehydrogenase in *Plasmodium falciparum*, human malaria parasite. Biochim Biophys Acta 1995;1243:351–60.
[16] Varadharajan S, Dhanasekaran S, Bonday ZQ, Rangarajan PN, Padmanaban G. Involvement of delta-aminolaevulinate synthase encoded by the parasite gene in de novo haem synthesis by *Plasmodium falciparum*. Biochem J 2002;367:321–7.
[17] The *Plasmodium* Genome Database Collaborative. PlasmoDB: an integrative database of the *Plasmodium falciparum* genome. Nucleic Acids Res 2001;29: 66–9.
[18] Kissinger JC, Brunk BP, Crabtree J, Fraunholz MJ, Gajria B, Milgram AJ, et al. The *Plasmodium* genome database. Nature 2002;419:490–2.
[19] Clamp M. Jalview—A Java Multiple Alignment Editor. http://www.ebi.ac.uk/~michele/jalview/, as implemented in the ClustalW service at the European Bioinformatics Institute, http://www.ebi.ac.uk/clustalw; 1998.
[20] StatSoft, Inc. 2001. STATISTICA (data analysis software system), version 6. www.statsoft.com. Distributor: StatSoft Inc., Tulsa, OK; www.statsoft.com.
[21] Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. In: Rumelhart DE, McClelland JL, editors. Parallel distributed processing: explorations in the microstructure of cognition, vol. 1. Cambridge: MIT Press; 1986. p. 318–62.
[22] Goldberg DE. Genetic algorithms. Reading: Addison Wesley; 1989.
[23] Lobry JR. Influence of genomic G + C content on average amino-acid composition of proteins from 59 bacterial species. Gene 1997;205:309–16.
[24] Foth BJ, Ralph SA, Tonkin CJ, Struck NS, Fraunholz MJ, Roos DS, et al. Dissecting apicoplast targeting in the malaria parasite *Plasmodium falciparum*. Science 2003;299:705–8.
[25] Waller RF, Reed MB, Cowman AF, McFadden GI. Protein trafficking to the plastid of *Plasmodium falciparum* is via the secretory pathway. EMBO J 2000;19:1794–802.
[26] McIntosh MT, Drozdowicz YM, Laroiya K, Rea PA, Vaidya AB. Two classes of plant-like vacuolar-type H(+)-pyrophosphatases in malaria parasites. Mol Biochem Parasitol 2001;114:183–95.
[27] Allary M, Schrevel J, Florent I. Properties. Parasitology, stage-dependent expression and localization of *Plasmodium falciparum* M1 family zinc-aminopeptidase 2002;125:1–10.
[28] Wickham ME, Rug M, Ralph SA, Klonis N, McFadden GI, Tilley L, et al. Trafficking and assembly of the cytoadherence complex in *Plasmodium falciparum*-infected human erythrocytes. EMBO J 2001;20:5636–49.
[29] Nair R, Rost B. Sequence conserved for subcellular localization. Protein Sci 2002;11:2836–47.
[30] von Heijne G. Mitochondrial targeting sequences may form amphiphilic helices. EMBO J 1986;5:1335–42.
[31] Roise D. Recognition and binding of mitochondrial presequences during the import of proteins into mitochondria. J Bioenerg Biomembr 1997;29:19–27.
[32] Kurland CG, Andersson SGE. Origin and evolution of the mitochondrial proteome. Microbiol Mol Biol Rev 2000;64:786–820.

[33] The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 2000;408:796–815.
[34] Srivastava IK, Morrisey JM, Darrouzet E, Daldal F, Vaidya AB. Resistance mutations reveal the atovaquone-binding domain of cytochrome b in malaria parasites. Mol Microbiol 1999;33:704–11.
[35] Takeo S, Kokaze A, Ng CS, Mizuchi D, Watanabe J, Tanabe K, et al. Succinate dehydrogenase in *Plasmodium falciparum* mitochondria: molecular characterization of the SDHA and SDHB genes for the catalytic subunits, the flavoprotein (Fp) and iron-sulfur (Ip) subunits. Mol Biochem Parasitol 2000;107:191–205.
[36] Uyemura SA, Luo S, Moreno SN, Docampo R. Oxidative phosphorylation, Ca(2+) transport, and fatty acid-induced uncoupling in malaria parasites mitochondria. J Biol Chem 2000;275: 9709–15.
[37] Sato S, Wilson RJ. The genome of *Plasmodium falciparum* encodes an active delta-aminolevulinic acid dehydratase. Curr Genet 2002;40:391–8.