Amino Acids
DOI 10.1007/s00726-009-0381-1

ORIGINAL ARTICLE

Prediction of mitochondrial proteins of malaria parasite using split amino acid composition and PSSM profile

Ruchi Verma · Grish C. Varshney · G. P. S. Raghava

Received: 16 March 2009 / Accepted: 20 October 2009

Abstract The rate of human death due to malaria is increasing day-by-day. Thus the malaria causing parasite *Plasmodium falciparum* (PF) remains the cause of concern. With the wealth of data now available, it is imperative to understand protein localization in order to gain deeper insight into their functional roles. In this manuscript, an attempt has been made to develop prediction method for the localization of mitochondrial proteins. In this study, we describe a method for predicting mitochondrial proteins of malaria parasite using machine-learning technique. All models were trained and tested on 175 proteins (40 mitochondrial and 135 non-mitochondrial proteins) and evaluated using five-fold cross validation. We developed a Support Vector Machine (SVM) model for predicting mitochondrial proteins of *P. falciparum*, using amino acids and dipeptides composition and achieved maximum MCC 0.38 and 0.51, respectively. In this study, split amino acid composition (SAAC) is used where composition of N-termini, C-termini, and rest of protein is computed separately. The performance of SVM model improved significantly from MCC 0.38 to 0.73 when SAAC instead of simple amino acid composition was used as input. In addition, SVM model has been developed using composition of PSSM profile with MCC 0.75 and accuracy 91.38%. We achieved maximum MCC 0.81 with accuracy 92% using a hybrid model, which combines PSSM profile and SAAC. When evaluated on an independent dataset our method performs better than existing methods. A web server PFMpred has been developed for predicting mitochondrial proteins of malaria parasites (http://www.imtech.res.in/raghava/pfmpred/).

Keywords *Plasmodium falciparum* · Mitochondria · Support vector machine · Position specific scoring matrix · Online server

R. Verma · G. P. S. Raghava (✉)
Bioinformatics Centre, Institute of Microbial Technology, Sector 39-A, Chandigarh, India
e-mail: raghava@imtech.res.in

R. Verma
e-mail: ruchi@imtech.res.in

G. C. Varshney
Cell biology and Immunology, Institute of Microbial Technology, Sector 39-A, Chandigarh, India
e-mail: girish@imtech.res.in

Introduction

The human malaria causing parasite *Plasmodium falciparum* (PF) has been playing much havoc to the mankind, in spite of the several efforts being made to curb this deadly disease (Gardner et al. 2002). The rate of human death and morbidity is increasing in many parts of the developing countries. Thus, there is a need to understand the critical pathways in malaria parasite in order to develop better drugs and vaccines. Fortunately, now whole genomic data of PF are available due to advancement in sequence technology (Gardner et al. 2002). This poses a major challenge for researchers to annotate genome of PF particularly functional proteins. The functional annotation of PF proteins using experimental techniques is not practically feasible as the process is costly and time-consuming. There is a need to develop bioinformatics tools to elucidate the functions of PF proteins. Mitochondria, commonly known as powerhouse of the cell, are one of the important organelle of the cell. The importance of mitochondrial proteins is reflected by the fact that children keep dying from mysterious illness that has been traced to tiny structures

Published online: 12 November 2009

called mitochondria. Thus it is important to identify or annotate mitochondrial proteins. Mitochondria in plasmodium parasites have many characteristics that distinguish them from mammalian mitochondria. Mitochondrial proteins of PF are different than human mitochondrial proteins; this makes PF mitochondrial protein as attractive potential drug targets (Vaidya and Mather 2009; Mather and Vaidya 2008; Vaidya and Mather 2005). Thus prediction of mitochondrial proteins is very important to identify novel potential target and new drugs against malaria. In past, number of methods, including multi-subcellular localization and mitochondria specific, has been developed (Guda et al. 2004; Kumar et al. 2006). It is likely that these methods developed for eukaryotic proteins will also be valid for malaria parasite. Recently, it has been shown in number of studies that organism-specific methods perform better than generalized methods (Garg et al. 2005; Chou and Shen 2006a; Rashid et al. 2007; Bender et al. 2003). The organism and the organelle-specific methods are more accurate and precise in predicting the particular and specific query sequence. Recently, Bender et al. (2003) analyzed the amino acid composition of mitochondrial proteins of PF and observed unique composition pattern, which is significantly different than composition pattern of eukaryotic proteins. Based on these observations, they developed a method PlasMit (Bender et al. 2003) for predicting mitochondrial proteins in malaria parasite. Bender et al. demonstrated that their PF specific method PlasMit performs better than methods developed for general purpose like TargetP (Emanuelsson et al. 2000) and MitoProtII (Claros and Vincens 1996). However, there is need to develop PF-specific method for predicting PF mitochondrial proteins with high accuracy which can be achieved by taking some more features. In this study, a systematic attempt has been made to predict PF mitochondrial proteins with high accuracy. In this study we used Support Vector Machine (SVM) technique, which is well-known technique. This technique has been used successfully in past for various type of classification/prediction methods including subcellular localization, structural classification, signal sequences (Cai et al. 2005, 2002; Garg and Raghava 2008).

Methods

Dataset

The dataset used in this study was retrieved from Bender et al. 2003, which consists of 40 mitochondrial proteins called positive examples and 135 proteins of other locations (cytoplasm, extracellular, apicoplast) called negative examples while details about these proteins are given in Bender et al. We had also checked the homologies between the 40 mitochondrial proteins using CD-HIT (Li and Godzik 2006) and observed high homology. To remove the homologous sequences from the benchmark dataset, a cut-off threshold of 25% was imposed to exclude those proteins from the benchmark datasets that have equal to or greater than 25% sequence identity to any other in a same subset (Chou and Shen 2006b; Chou and Shen 2007a, b; Chou and Shen 2008a). However, in this study we did not use such a stringent criterion because the currently available data for mitochondrial proteins do allow us to do so. Otherwise, the numbers of proteins for some subsets would be too few to have statistical significance.

We also created an independent or blind dataset, which consists of 24 PF mitochondrial proteins. These 24 PF proteins were obtained from UniProt database of ExPASy available at URL http://www.expasy.org/sprot/. We extracted these proteins from SRS version (http://www.expasy.org/srs5bin/cgi-bin/wgetz) of UniProt for organism *Plasmodium falciparum* using following query "[libs = {swiss_prot trembl}-Organism: *Plasmodium falciparum**] & [libs-Comment: mitochondr*]". These sequences are available from website of PFMpred. This independent dataset was used to evaluate performance of existing method and modules developed in this study.

Evaluation

Among the independent dataset test, sub-sampling (e.g., 5- or 10-fold cross-validation) test, and jackknife test, which are often used for examining the accuracy of a statistical prediction method (Chou and Zhang 1995), the jackknife test was deemed the most objective that can always yield a unique result for a given benchmark dataset, as elucidated in past (Cai et al. 2002). Therefore, the jackknife test has been increasingly and widely adopted by investigators to test the power of various prediction methods (Chen et al. 2008; Chou and Shen 2008b, 2009; Ding et al. 2009; Li and Li 2008; Shen et al. 2009; Xiao et al. 2009a, b; Kaur and Raghava 2003; Ding and Zhang 2008; Ding et al. 2009). However, to reduce the computational time, we adopted the 5-fold cross validation in this study. The dataset was randomly divided into five equal sets, out of which four sets were used for training and the remaining one for testing. This procedure was repeated five times by changing the test dataset, so that each set was used for training as well as testing. The final performance was calculated by averaging over all five sets.

Performance measures:

Finally we computed the performance of our method using following performance measures (Kumar et al. 2006; Garg et al. 2005).

(a) *Sensitivity or coverage of positive examples* It is percent of mitochondrial proteins correctly predicted mitochondrial.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100$$

(b) *Specificity or coverage of negative examples* It is percent of non-mitochondrial proteins correctly predicted non-mitochondrial.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100$$

(c) *Accuracy* It is percentage of correctly predicted proteins (mitochondrial and non-mitochondrial proteins).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100$$

(d) *Mathew's correlation coefficient* (*MCC*) It is considered to be the most robust parameter of any class prediction method. MCC equal to 1 is regarded as perfect prediction while 0 for completely random prediction.

$$\text{MCC} = \frac{(\text{TP} * \text{TN}) - (\text{FP} * \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

where TP and TN are truly or correctly predicted mitochondrial and non-mitochondrial proteins, respectively. FP and FN are wrongly predicted mitochondrial and non- mitochondrial proteins, respectively.

Amino acid and dipeptide composition

The aim of calculating composition of proteins is to transform the variable length of protein sequence to fixed length feature vectors. This is necessary for the kind of SVM classification scheme we adopt. The information of proteins can be encapsulated to a vector of 20 dimensions using amino acid composition of the protein. In addition to amino acid composition, dipeptide composition was also used for classification that gave a fixed pattern length of 400. The advantage of dipeptide composition over amino acid composition is that it encapsulates information about the fraction of amino acids as well as their local order.

Split amino acid composition (SAAC)

In Split Amino Acid Composition method (Chou and Shen 2006a, b) protein sequence is divided in parts and composition of each part is calculated separately. In our SAAC model each protein is divided into three parts; (i) 20 amino acids of N-termini, (ii) 20 amino acids of C-termini, and (iii) region between these two terminuses.

PSSM profile

In order to utilize multiple sequence alignment, we computed PSSM profile for each protein using PSI-BLAST. Sequence was searched against nr database with three iterations. Intermediate PSI-BLAST generated PSSMs were used for developing SVM. The matrix had $20 \times M$ elements (excluding dummy residue X), where M is the length of protein. Each element in the matrix represents the frequency of occurrence of each of the amino acid at a given position in the alignment. We generated a vector of dimension 400 from PSSM matrix where composition of occurrences of each type of amino acid corresponding to each type of amino acids in protein sequence; it means for each column we will have 20 values thus making a matrix of dimension 20×20 for PSSM (Rashid et al. 2007; Chou and Shen 2007a; Chou and Shen 2008a, b; Shen and Chou 2007a, b, 2009).

Hybrid models

In addition, number of hybrid models were developed using combination of two or more than two types of information. These hybrid modules include combination of (i)amino acid (20) and dipeptide composition (400), where a vector of dimension 420 presents information; (ii) PSSM and amino acid composition; (iii) PSSM and dipeptide composition.

Support vector machine

In this study, we implemented SVM using SVM_light package (Joachims 1999; Chou and Shen 2007b, c) which allows choosing number of parameters and kernels (e.g., linear, polynomial, radial basis function, sigmoid) or any user-defined kernel. Assuming that we have number of patterns xi ε R^{d} ($I = 1, 2, \ldots\ldots N$) with corresponding target values yi ε {⟨target value⟩} that is, mitochondrial proteins were used a positive training examples and non-mitochondrial for negative examples. Here the target value is $+1$ for mitochondrial and -1 for non-mitochondrial proteins. The value of epsilon determines the level of accuracy of the approximated function. It relies entirely on the target values in the training set. If epsilon is larger than the range of the target values then we cannot expect a good result. If epsilon is zero, we can expect over fitting. Epsilon must therefore be chosen to reflect the data in some way.

Results

We analyzed the amino acid composition of both mitochondrial and non-mitochondrial proteins. As shown in

Fig. 1, it has been observed that amino acid glutamic acid, isoleucine, and tyrosine are more abundant in mitochondrial proteins than non-mitochondrial proteins. In contrast aspartic acid, asparagines, serine, and valine are more abundant in non-mitochondrial proteins than mitochondrial proteins. This means that mitochondrial and non-mitochondrial proteins can be predicted based on their amino acid composition. We developed a SVM model using amino acid composition and got maximum MCC 0.38 with accuracy of 74.14%. Similarly, SVM models were developed using dipeptide composition which achieved maximum MCC 0.50 with accuracy of 82.76%.

In past it has been shown that composition of parts of a protein provides more information than composition of whole protein (Kumar et al. 2006). Based on this observation, methods have been developed using split amino acid composition (Verma et al. 2008). In SAAC, first sequence is divided in parts then composition of each part is computed separately and finally SVM model is developed using composition of all parts of proteins. In this study, SAAC is used to develop SVM models for predicting PF mitochondrial proteins. A protein is divided into three parts: (i) 20 amino acids of the N terminus, (ii) 20 amino acids of the C terminus, and (iii) the region between these two regions. We achieved MCC 0.73 using SAAC-based SVM model (Table 1) with rbf kernel and g 0.0001, c 6, j4. The SAAC-based model performs better than composition models because it uses the composition of three parts of protein independently. In SAAC model, we tried different lengths from N- and C-terminal and found best performance using 20 residues from each terminus of protein.

The evolutionary information has been used successfully in developing prediction methods for example in prediction of protein secondary structure (Kaur and Raghava 2004a, b), nucleotide-binding proteins (Kumar et al. 2008; Kumar et al. 2007), subcellular localization of proteins. Thus, we also used evolutionary information for predicting PF mitochondrial proteins. In this approach evolutionary information was extracted in the form of PSSM profile from multiple alignments of proteins similar to query sequence. We performed PSI-BLAST search of query protein against non-redundant database. Finally a SVM model was developed using composition of PSSM profile which could achieve maximum MCC 0.75.

In this study, a number of hybrid models have been developed. Firstly, a hybrid model was developed using amino acid and dipeptide composition and achieved MCC up to 0.48. In this model, a vector of dimension 420 presents composition (400 dipeptides + 20 amino acid). Secondly, a SVM model has been developed using dipeptide and PSSM profile composition and achieved MCC 0.78 with accuracy 92.57%. Here, dimension of input vector is 800, 400 for dipeptide, and 400 for PSSM composition. Finally, a hybrid model has been developed using PSSM and SAAC and achieved maximum MCC 0.81 with accuracy 92.00%. In this model, input vector is of dimension 460, 60 for SAAC, and 400 for PSSM composition. This model was found to be more accurate compared to other models developed in this study.

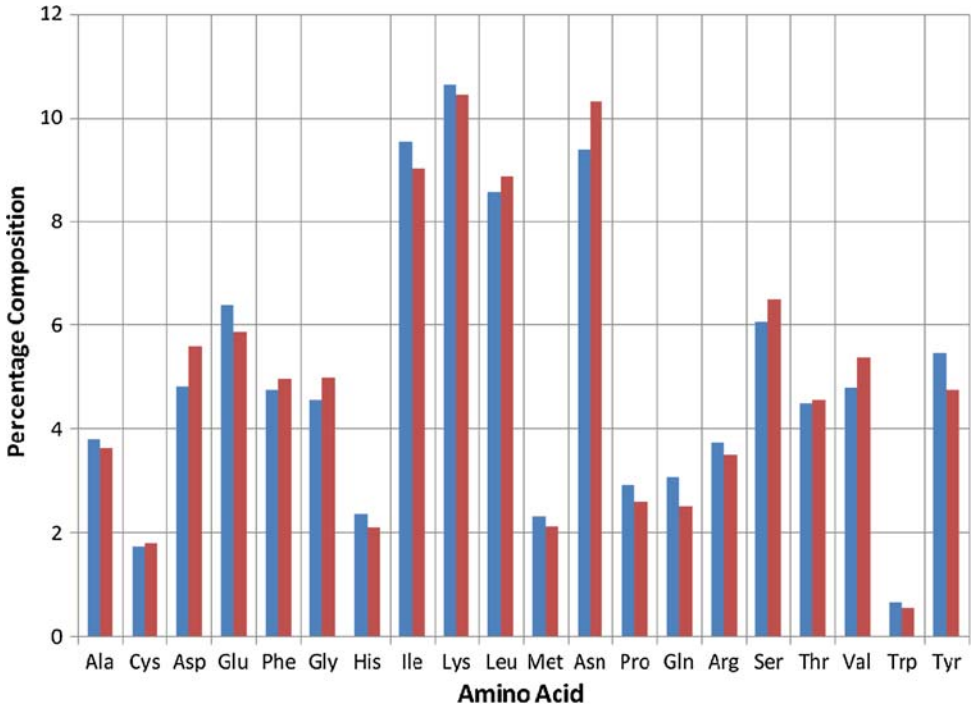

Fig. 1 Comparison of amino acid composition of 40 mitochondrial and 135 non-mitochondrial proteins. Blue and red bars represent mitochondrial and non-mitochondrial proteins, respectively

Table 1 Performance of various SVM modules developed using various types of compositions; amino acids, dipeptides, SAAC, and PSSM profile composition

Thrs	Amino acids		Dipeptides		Split amino acids		PSSM	
	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
−1.0	55.75	0.29	52.30	0.35	60.92	0.43	66.09	0.48
−0.9	59.77	0.33	56.32	0.37	64.94	0.46	72.41	0.53
−0.8	60.34	0.32	59.20	0.39	68.39	0.50	73.56	0.54
−0.7	60.92	0.31	62.64	0.41	72.41	0.54	75.29	0.56
−0.6	61.49	0.32	66.09	0.44	74.71	0.57	78.16	0.59
−0.5	63.79	0.34	71.26	0.50	76.44	0.57	78.16	0.58
−0.4	66.67	0.37	72.41	0.48	81.03	0.63	80.46	0.60
−0.3	68.39	0.39	77.01	0.50	83.33	0.65	83.33	0.63
−0.2	68.97	0.35	78.16	0.51	83.91	0.65	82.18	0.57
−0.1	70.69	0.35	79.89	0.50	85.63	0.68	82.76	0.56
0.0	71.84	0.36	81.03	0.49	87.36	0.71	85.06	0.60
0.1	**74.14**	**0.38**	**82.76**	**0.50**	87.93	0.70	89.08	0.68
0.2	75.29	0.36	81.61	0.44	89.66	0.73	89.66	0.70
0.3	75.86	0.32	84.48	0.51	**90.23**	**0.73**	**91.38**	**0.75**
0.4	77.59	0.33	83.33	0.47	89.08	0.68	89.66	0.70
0.5	78.16	0.32	82.76	0.44	89.08	0.68	86.21	0.58
0.6	77.59	0.28	80.46	0.33	88.51	0.66	85.06	0.54
0.7	79.89	0.34	81.03	0.37	86.78	0.59	83.91	0.50
0.8	78.74	0.28	79.31	0.27	85.06	0.53	82.18	0.43
0.9	79.31	0.29	78.16	0.19	83.33	0.47	80.46	0.35
1.0	78.74	0.24	77.01	0.07	82.76	0.44	78.16	0.20

Bold font shows performance at default threshold

Thrs Threshold, *ACC* Accuracy, *MCC* Matthews correlation coefficient

Benchmarking of models

In order to compute realistic performance of models, it is important to evaluate performance of models on an independent dataset, not used in training or testing of models. Thus, we computed the performance of our model on an independent dataset, which consisted of 24 PF mitochondrial proteins extracted from UniProt, a curated protein sequence database which strives to provide a high level of annotation. Our hybrid model correctly predicted 18 mitochondrial proteins at default threshold. We also evaluated the performance of the popular existing methods like TargetP (Emanuelsson et al. 2000), MitoPred (Guda et al. 2004), MitPred (Kumar et al. 2006), ESLpred (Bhasin and Raghava 2004), PlasMit (Bender et al. 2003) on our independent dataset. As shown in Table 2, performance of most of general purpose methods was very poor on 24 PF proteins. It is also possible that some of proteins in independent dataset have similarity with proteins in training dataset. We performed BLAST search and detected that 14 proteins have no-detectable homology (e-value 0.0001) with proteins in training proteins. We also computed performance of methods on these 14 proteins. Organism-specific method PlasMit performed better than most of existing method because it is developed specially for PF. The performance of Mitpred is higher than any other method on non-similar proteins including PFMpred because Mitpred also uses PFAM database for searching profiles. The proteins we called non-similar in this study have no similarity with proteins in our main dataset but there are chances that these proteins or similar proteins are present in PFAM database. As shown in Table 2, the performance of Mitpred is lower than PFMpred in absence of PFAM search. In addition to evaluation on an independent dataset, we also evaluated existing methods on 175 PF proteins used in this study to develop our SVM models (Table 3). In order to make fair comparison, PFMpred and PlasMit were evaluated using cross-validation technique as both methods were developed on this dataset. Both organism-specific methods PFMpred and PlasMit performed better than any other existing methods. PFMpred even performed better than PlasMit. This demonstrates the

Table 2 Performance of existing methods on an independent dataset of 24 PF mitochondrial proteins at a default threshold

Method	Correctly predicted proteins	Correctly predicted non-similar proteins
ESLpred	8	7
Mitopred	10	5
Mitpred	15 (10)	11 (5)[a]
MitoprotII	12	5
PlasMit	14	6
TargetP	6	3
PFMpred	18	8

In addition performance also checked on 14 non-similar proteins

[a] values presented in parenthesis represented the performance of Mitpred without PFAM search

Table 3 Benchmarking of mitochondrial prediction methods on 175 PF mitochondrial proteins (40 mitochondrial and 135 non-mitochondrial proteins) used in this study for developing SVM models

Method	Sensitivity	Specificity	Accuracy	MCC
Eslpred	57.50	73.33	69.71	0.27
Mitopred	55.00	87.41	80.00	0.43
Mitpred	62.50	88.89	82.86	0.51
[a]MitoProtII	80.00	74.07	75.43	0.46
[a]Target P	55.00	96.30	86.86	0.60
[a]PlasMit	94.00	89.00	90.00	0.74
[b]PFMpred	97.50	91.04	92.00	0.81

[a] As reported by Bender et al. 2003

[b] Present method (PSSM + SAAC) evaluated using five-fold cross validation

Table 4 All proteins of *P. falciparum* predicted using various methods; proteins predicted mitochondrial proteins are shown in table

Method	Total proteins used for prediction	Number of mitochondrial proteins	Percentage of mitochondrial proteins (%)
Mitopred[a]	445	41	9.21
Mitpred	5460	560	10.26
MitoProtII	5460	882	16.15
TargetP	5460	170	3.11
PlasMit	5460	1,191	21.81
PFMpred	5460	1,432	25.39

[a] All proteins could not be predicted by this method due to limitation of time

importance of developing organism-specific methods. In order to estimate number of mitochondrial proteins in *P. falciparum*, we used various methods. As shown in Table 4, the percentage of PF proteins predicted as mitochondrial protein by different methods varies from 3 (TargetP) to 25% (PFMpred).

Web-server

In order to assist the scientific community, we developed a web server PFMpred for predicting mitochondrial proteins from protein sequences. This server will assist the researcher in annotating genome of malaria parasite. PFMpred is available for public from http://www.imtech.res.in/raghava/pfmpred/ for academic use. This server allows user to submit multiple sequences in FASTA format for predicting PF mitochondrial proteins using SAAC. PFMpred allows user to predict one sequence at a time in case user wishes to predict using our best model (Hybrid model), which combines SAAC- and PSSM- based models.

Discussion

Functional annotation of genomes is one of the major challenges in post genomic era. In last one decade, a large number of bioinformatics methods have been developed to predict function of proteins. Most of these methods are indirect methods where they predict important properties of a protein that include prediction of protein structure, subcellular localization, class of protein, and antigenic properties (Garg et al. 2005; Chou and Shen 2006a, b; Rashid et al. 2007; Bender et al. 2003; Emanuelsson et al. 2000; Claros and Vincens 1996). One of the important classes of protein resides in mitochondria called powerhouse of a cell. In this study, we made systematic attempt to predict mitochondrial proteins of *P. falciparum* significant composition biasness was observed in mitochondrial proteins (Fig. 1), which encouraged us to develop SVM models using amino acid and

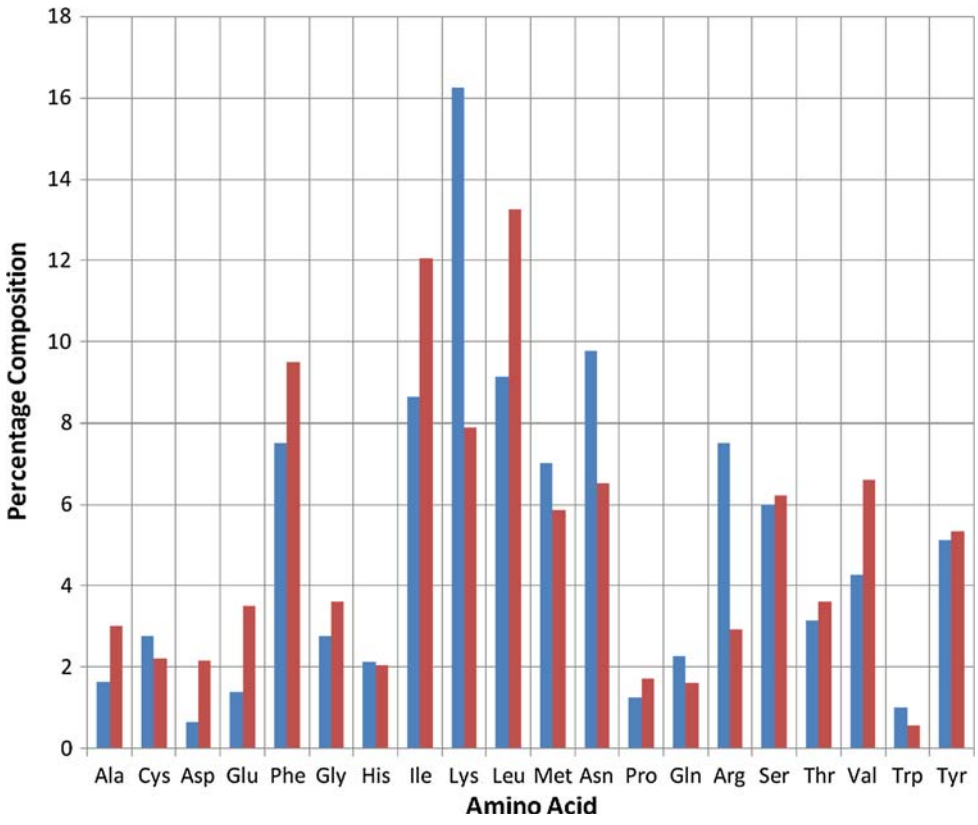

Fig. 2 Amino acid compositon of N-terminal (first 20 amino acids) in 40 mitochondrial (blue) and 135 non-mitochondrial (red) proteins of *P. falciparum* parasite

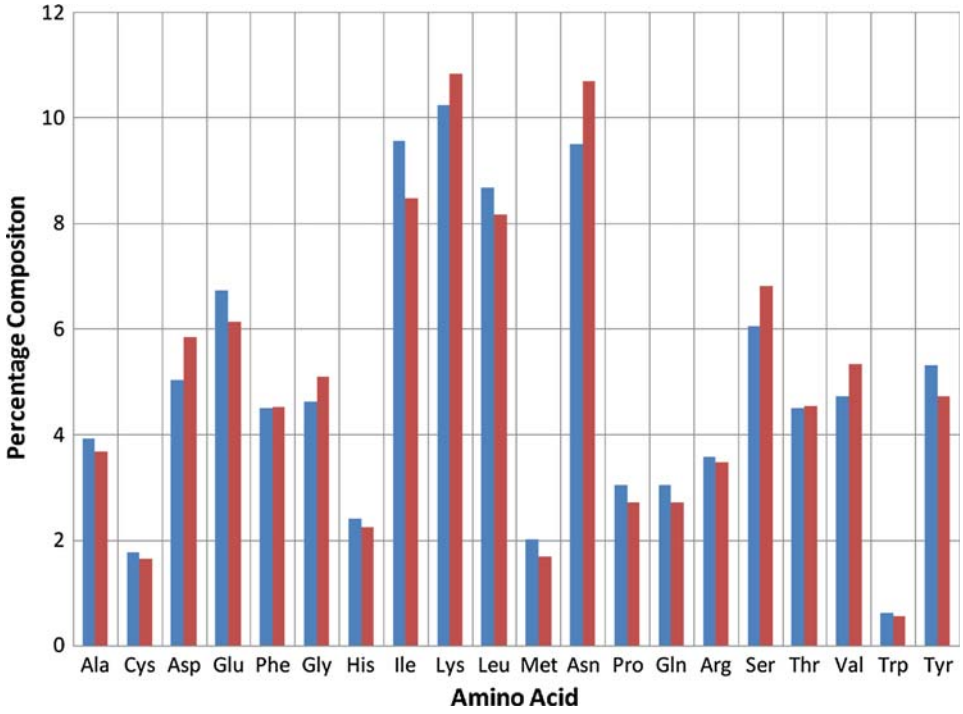

Fig. 3 Amino acid compositon of middle region amino acids in 40 mitochondrial (blue) and 135 non-mitochondrial (red) proteins of *P. falciparum* parasite

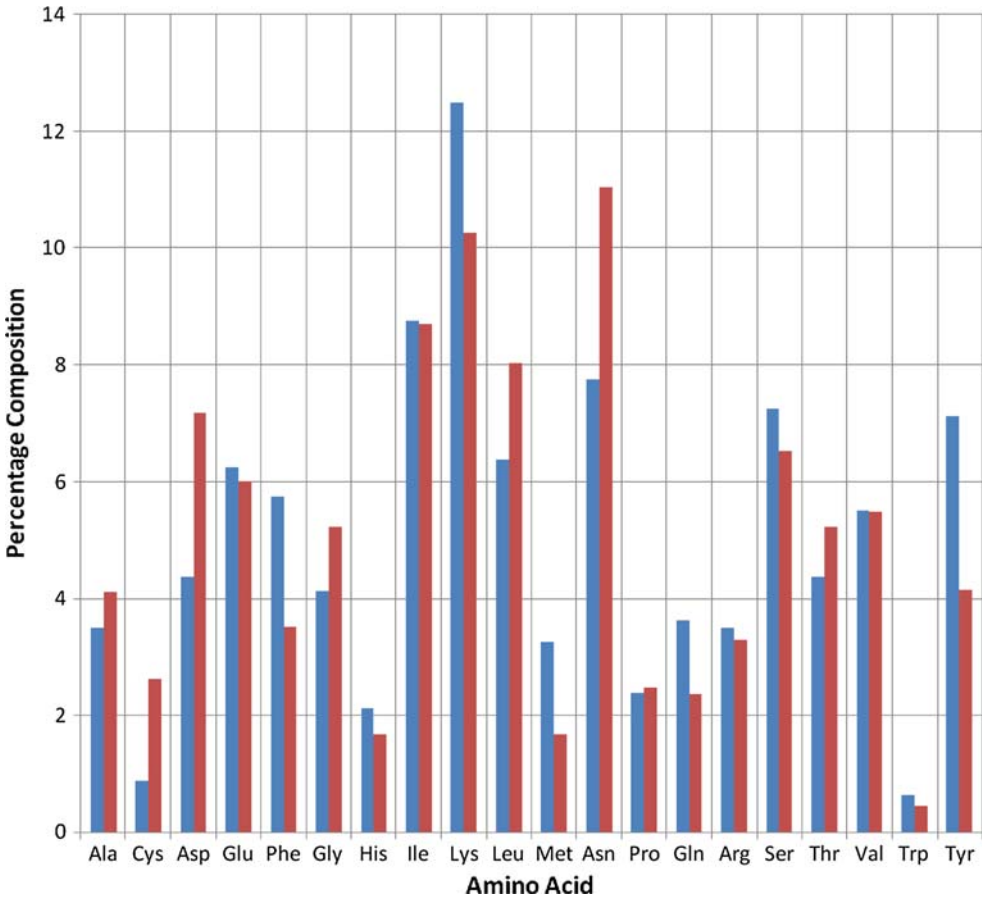

Fig. 4 Amino acid compositon of C-terminal (last 20 amino acids) in 40 mitochondrial (blue) and 135 non-mitochondrial (red) proteins of *P. falciparum* parasite

Table 5 Performance of SVM-based hybrid models developed using (i) PSSM and dipeptide compositions and (ii) SAAC and PSSM composition

Thrs	Dipeptide and PSSM composition		SAAC and PSSM composition	
	ACC	MCC	ACC	MCC
−1.0	42.86	0.25	69.71	0.51
−0.9	49.71	0.31	74.29	0.55
−0.8	54.29	0.33	77.14	0.58
−0.7	60.57	0.37	80.57	0.62
−0.6	69.14	0.46	85.14	0.69
−0.5	80.57	0.59	86.86	0.72
−0.4	88.00	0.70	88.57	0.74
−0.3	91.43	0.76	89.71	0.76
−0.2	**92.57**	**0.78**	**92.00**	**0.81**
−0.1	89.71	0.69	90.86	0.77
0.0	87.43	0.61	89.71	0.73
0.1	87.43	0.62	89.14	0.70
0.2	86.29	0.58	89.71	0.71
0.3	84.57	0.51	86.86	0.62
0.4	82.29	0.42	87.43	0.62
0.5	82.29	0.43	87.43	0.62
0.6	81.14	0.38	86.86	0.60
0.7	80.57	0.35	86.86	0.60
0.8	79.43	0.28	84.00	0.49
0.9	79.43	0.28	84.00	0.49
1.0	79.43	0.28	82.29	0.42

Bold font shows performance at default threshold

Thrs Threshold, *ACC* Accuracy, *MCC* Matthews correlation coefficient

dipeptide composition. The composition-based SVM performs reasonably well. It has been reported in previous studies that mitochondrial proteins have signal mostly in terminal region (Kumar et al. 2006). In this study, we compared N/C-terminal composition of mitochondrial and non-mitochondrial proteins and observed significant differences. Based on these observations, SVM models using SAAC have been developed. It is interesting to note that performance of these SAAC-based SVM models was much higher when compared to composition-based models. We analyzed the composition of N-terminal, C-terminal, and middle region of mitochondrial and non-mitochondrial proteins, in order to understand why SAAC-based method performs better than composition based model. Figures 2, 3, and 4 show the composition of N-terminal, middle, and C-terminal regions of both positive and negative types of proteins. As shown in Fig. 2, composition of 20 residues of N-terminal of mitochondrial protein is significantly different than non-mitochondrial proteins. These figures clearly indicate the importance of terminal regions in prediction particularly of N-terminal.

It has been shown in past that evolutionary information of a protein is important for predicting its function and has been extensively used for predicting RNA interacting residues (Kumar et al. 2008), irregular secondary structure (Kaur and Raghava 2004a), and subcellular localization of proteins (Kaur and Raghava 2004b). Thus, we also made an attempt to utilize the evolutionary information for predicting PF mitochondrial proteins. In order to utilize evolutionary information we generated PSSM profile and developed SVM model using composition of PSSM profile. As shown in Table 5, PSSM-based model performs better than any other model including SAAC (Table 5). Finally, we developed hybrid models and achieved maximum performance using SAAC and PSSM profile. Our hybrid model performs better than existing methods when evaluated on an independent dataset (Table 2). Also, similarity search was done using blast between the 24 independent dataset and 40 mitochondrial proteins (Table 4).

The major limitation of our models is that they have been benchmarked on small dataset. We used same dataset as used by Bender et al. 2003, this dataset have high similarity. Unfortunately dataset size is too small (40 proteins), so it is not possible to develop models on non-redundant proteins. In order to overcome limitation we tested our model on an independent dataset of 24 proteins, here we remove redundant proteins from independent dataset and again evaluate methods. In summary we have tried our best to evaluate our models in given condition. We hope our method will complement the existing methods in functional annotation of malaria proteomes. It is possible to improve the performance of PFMpred further by integrating it with Gene Ontology (GO) (Ashburner et al. 2000) annotation, which describes the function of genes and gene products across species. Recently, GO has been integrated successfully with subcellular localization and sub nuclearization methods (Guo et al. 2006; Huang et al. 2008; Chou and Shen 2007d). Thus, in future, attempt will be made to integrate GO annotation with PFMpred in order to make it more accurate.

Acknowledgments The authors gratefully acknowledged the financial support provided by the Council of Science and Industrial Research (CSIR) and Department of Biotechnology (DBT), Government of India. This paper has IMTECH communication number 048/2007.

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29

Bender A, van Dooren GG, Ralph SA, McFadden GI, Schneider G (2003) Properties and prediction of mitochondrial transit peptides from *Plasmodium falciparum*. Mol Biochem Parasitol 132:59–66

Bhasin M, Raghava GPS (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res 32:W414–W419

Cai YD, Liu XJ, Xu XB, Chou KC (2002) Prediction of protein structural classes by support vector machines. Comput Chem 26:293–296

Cai YD, Lin S, Chou KC (2005) Support vector machines for prediction of protein signal sequences and their cleavage sites. Peptides 24:159–161

Chen C, Chen LX, Zou XY, Cai PX (2008) Predicting protein structural class based on multi-features fusion. J Theor Biol 253:388–392

Chou KC, Shen HB (2006a) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J Proteome Res 5:1888–1897

Chou KC, Shen HB (2006b) Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem Biophys Res Commun 347:150–157

Chou KC, Shen HB (2007a) MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Commun 360:339–345

Chou KC, Shen HB (2007b) Large-scale plant protein subcellular location prediction. J Cell Biochem 100:665–678

Chou KC, Shen HB (2007c) Review: recent progresses in protein subcellular location prediction. Anal Biochem 370:1–16

Chou KC, Shen HB (2007d) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J Proteome Res 6:1728–1734

Chou KC, Shen HB (2008a) ProtIdent: a web server for identifying proteases and their types by fusing functional domain and sequential evolution information. Biochem Biophys Res Commun 376:321–325

Chou KC, Shen HB (2008b) Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms. Nat Protoc 3:153–162

Chou KC, Shen HB (2009) FoldRate: a web-server for predicting protein folding rates from primary sequence. Open Bioinform J 3:31–50. Accessible at http://www.bentham.org/open/tobioij/)

Chou KC, Zhang CT (1995) Review: prediction of protein structural classes. Crit Rev Biochem Mol Biol 30:275–349

Claros MG, Vincens P (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur J Biochem 241:770–786

Ding YS, Zhang TL (2008) Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier. Pattern Recognit Lett 29:1887–1892

Ding YS, Zhang TL, Gu Q, Zhao PY, Chou KC (2009) Using maximum entropy model to predict protein secondary structure with single sequence. Protein Pept Lett 16:552–560

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

Gardner MJ et al (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419:498–511

Garg A, Raghava GPS (2008) ESLpred2: improved method for predicting subcellular localization of eukaryotic proteins. BMC Bioinform 9:503

Garg A, Bhasin M, Raghava GPS (2005) Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. J Biol Chem 280:14427–14432

Guda C, Fahy E, Subramaniam S (2004) MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. Bioinformatics 20:1785–1794

Guo J, Lin Y, Liu X (2006) GNBSL: a new integrative system to predict the subcellular location for Gram-negative bacteria proteins. Proteomics 6:5099–5105

Huang WL, Tung CW, Ho SW, Hwang SF, Ho SY (2008) ProLoc-GO: utilizing informative gene ontology terms for sequence-based prediction of protein subcellular localization. BMC Bioinform 9:80

Joachims T (1999) Making large-scale SVM learning practical. In: Scholkopf B, Burges C, Smola A (eds) Advances in Kernel methods—support vector learning. MIIT Press, Cambridge, MA; London, England

Kaur H, Raghava GPS (2003) Prediction of beta-turns in proteins from multiple alignment using neural network. Protein Sci 12:627–634

Kaur H, Raghava GPS (2004a) A neural network method for prediction of beta-turn types in proteins using evolutionary information. Bioinformatics 16:2751–2758

Kaur H, Raghava GPS (2004b) Role of evolutionary information in prediction of aromatic-backbone NH interactions in proteins. FEBS Lett 564:47–57

Kumar M, Verma R, Raghava GPS (2006) Prediction of mitochondrial proteins using support vector machine and hidden markov model. J Biol Chem 281:5357–5363

Kumar M, Gromiha MM, Raghava GPS (2007) Identification of DNA-binding proteins using support vector machines and evolutionary profiles. BMC Bioinform 8:463

Kumar M, Gromiha MM, Raghava GPS (2008) Prediction of RNA binding sites in a protein using SVM and PSSM profile. Proteins 71:189–194

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659

Li FM, Li QZ (2008) Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. Protein Pept Lett 15:612–616

Mather MW, Vaidya AB (2008) Mitochondria in malaria and related parasites: ancient, diverse and streamlined. J Bioenerg Biomembr 40:425–433

Rashid M, Saha S, Raghava GPS (2007) Support vector machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. BMC Bioinform 8:337

Shen HB, Chou KC (2007a) EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Commun 364:53–59

Shen HB, Chou KC (2007b) Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Eng Des Sel 20:561–567

Shen HB, Chou KC (2009) QuatIdent: a web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information. J Proteome Res 8:1577–1584

Shen HB, Song JN, Chou KC (2009) Prediction of protein folding rates from primary sequence by fusing multiple sequential features. J Biomed Sci Eng 2:136–143. Accessible at http://www.srpublishing.org/journal/jbise/)

Vaidya AB, Mather MW (2005) A post-genomic view of the mitochondrion in malaria parasites. Curr Top Microbiol Immunol 295:233–250

Vaidya AB, Mather MW (2009) Mitochondrial evolution and functions in malaria parasites. Annu Rev Microbiol 63:249–267

Verma R, Tiwari A, Kaur S, Varshney GC, Raghava GPS (2008) Identification of proteins secreted by malaria parasite into erythrocyte using SVM and PSSM profiles. BMC Bioinform 9:201

Xiao X, Wang P, Chou KC (2009a) GPCR-CA: a cellular automaton image approach for predicting G-protein-coupled receptor functional classes. J Comput Chem 30:1414–1423

Xiao X, Wang P, Chou KC (2009b) Predicting protein quaternary structural attribute by hybridizing functional domain composition and pseudo amino acid composition. J Appl Crystallogr 42:169–173