Amino Acids
DOI 10.1007/s00726-010-0825-7

ORIGINAL ARTICLE

Using increment of diversity to predict mitochondrial proteins of malaria parasite: integrating pseudo-amino acid composition and structural alphabet

Ying-Li Chen · Qian-Zhong Li · Li-Qing Zhang

Received: 12 October 2010 / Accepted: 17 December 2010

Abstract Due to the complexity of *Plasmodium falciparum* (PF) genome, predicting mitochondrial proteins of PF is more difficult than other species. In this study, using the *n*-peptide composition of reduced amino acid alphabet (RAAA) obtained from structural alphabet named Protein Blocks as feature parameter, the increment of diversity (ID) is firstly developed to predict mitochondrial proteins. By choosing the 1-peptide compositions on the N-terminal regions with 20 residues as the only input vector, the prediction performance achieves 86.86% accuracy with 0.69 Mathew's correlation coefficient (MCC) by the jackknife test. Moreover, by combining with the hydropathy distribution along protein sequence and several reduced amino acid alphabets, we achieved maximum MCC 0.82 with accuracy 92% in the jackknife test by using the developed ID model. When evaluating on an independent dataset our method performs better than existing methods. The results indicate that the ID is a simple and efficient prediction method for mitochondrial proteins of malaria parasite.

Y.-L. Chen · Q.-Z. Li (✉) · L.-Q. Zhang
Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot, China
e-mail: qzli@imu.edu.cn

Y.-L. Chen
e-mail: yl74chen@vt.edu

L.-Q. Zhang
e-mail: lqzhang@vt.edu

Y.-L. Chen · L.-Q. Zhang
Department of Computer Science, Virginia Tech, Blacksburg, VA, USA

L.-Q. Zhang
Program in Genetics, Bioinformatics, and Computational Biology, Virginia Tech, Blacksburg, VA, USA

Keywords *Plasmodium falciparum* · Mitochondrial proteins · Increment of diversity · Reduced amino acid alphabet · Hydropathy distribution

Introduction

The malaria caused by *Plasmodium falciparum* (PF) remains the world's most devastating infectious disease, which results in 300–660 million clinical cases and 2–3 million deaths annually and its long-term control and eradication is still a long way off (Snow et al. 2005). The rate of human death and morbidity is increasing in many parts of the developing countries. Thus, it is essential to develop effective drugs and vaccines against this parasite. Mitochondria, commonly known as the powerhouse of a cell, are one of the important organelles of a cell. Mitochondria in plasmodium parasites have many characteristics that distinguish them from mammalian mitochondria (Vaidya and Mather 2005). Mitochondrial proteins of PF are different from human mitochondrial proteins, which make them attractive potential drug targets (Vaidya and Mather 2009). Therefore, the identification of mitochondrial proteins of PF will be helpful for novel potential targets and new drugs against malaria.

Many efforts have been made for predicting protein multi-subcellular locations and mitochondrial specific (Bender et al. 2003; Guda et al. 2004; Kumar et al. 2006; Garg et al. 2005; Chou and Shen 2006a; Rashid et al. 2007; Garg and Raghava 2008; Verma et al. 2010). Several programs have been developed for predicting mitochondrial proteins, such as Target P (Emanuelsson et al. 2000), SignalP 3.0 (Bendtsen et al. 2004), TargetLoc (Höglund et al. 2006), MitoProtII (Claros and Vincens 1996), MITOPRED (Guda et al. 2004) and so on. Based on the

Published online: 30 December 2010

relative frequencies of amino acids in different regions of PF mitochondrial proteins, Bender et al. (2003) developed a neural network system (PlasMit) for the prediction of mitochondrial proteins in malaria parasite. Bender et al. demonstrated that their PF specific method PlasMit performs better than the methods developed for general purpose like TargetP (Emanuelsson et al. 2000) and MitoProtII (Claros and Vincens 1996). Recently, Verma et al. (2010) used Support Vector Machine (SVM) model for predicting mitochondrial proteins of PF based on split amino acid composition (SAAC) and PSSM profile, and obtained better prediction accuracy than previous methods.

On the basis of the Shannon entropy definition, Laxton introduced the concept of measure of diversity (Laxton 1978). The measure of diversity is a kind of information description on discrete state space and a measure of uncertainty of a system. In order to compare the distribution of two species, one defines the increment of diversity (ID) by the difference of the total diversity measure of two systems and the diversity measure of the mixed system. The ID method has been developed and employed for classification in biogeography. Recently, Li's group firstly introduced the ID method to protein prediction, the recognition of protein structural class (Li and Lu 2001; Lin and Li 2007a), the protein superfamily classification (Lin and Li 2007b), the subcellular and subnuclear location (Chen and Li 2007a, b; Li and Li 2008a, b), the defensin family and subfamily classification (Zuo and Li 2009), beta-hairpin and gamma-turn prediction (Hu and Li 2008) and good prediction performances are obtained. It has been shown that the ID is a good index for distinguishing two different sources established by proteins. In this paper, the ID method firstly is applied to predict mitochondrial proteins of malaria parasite, based on pseudo-amino acid composition and structural alphabet. Using this hybrid model, the prediction accuracy and Mathew's correlation coefficient (MCC) are 93.14% and 0.84, respectively, for the re-substitution test, and the prediction accuracy and MCC are 92% and 0.82, respectively, for the jackknife test.

Materials and methods

Dataset

A critical issue in developing mitochondrial protein prediction algorithm of malaria parasite is to find suitable training and testing sets. In this study, the dataset was retrieved from Bender et al. (2003), which consists of 40 mitochondrial proteins called positive examples and 135 proteins of other locations (cytoplasm, secretory, apicoplast) called negative examples while details about these proteins are given in Bender et al. (2003).

The definition of increment of diversity

For a discrete state space X with d dimension X: $[n_1,n_2,\ldots,n_i,\ldots,n_d]$, n_i denotes the absolute frequency of ith state, the Shannon information entropy (Shannon 1948), a measure of uncertainty, denoted by $H(X)$, is defined as:

$$H(X) = -\sum_{i=1}^{d} p_i \log_b p_i \tag{1}$$

where $N = \sum_{i=1}^{d} n_i$, $p_i = {}^{n_i}/_{N}$, p_i indicates probability of ith state.

From the idea of information, the quantity of the measured diversity is called the measure of diversity, denoted by $D(X)$, is defined as:

$$\begin{aligned} D(X) &= -\sum_{i=1}^{d} n_i \log_b p_i = -\sum_{i=1}^{d} n_i \log_b \frac{n_i}{N} \\ &= N\log N - \sum_{i=1}^{d} n_i \log_b n_i \end{aligned} \tag{2}$$

According to the definition of information entropy, combining the formula (1), we get

$$H(X) = -\sum_{i=1}^{d} p_i \log_b p_i = -\sum_{i=1}^{d} \frac{n_i}{N} \log_b \frac{n_i}{N} = \frac{1}{N} D(X) \tag{3}$$

So we have

$$D(X) = N \cdot H(X) \tag{4}$$

$H(X)$ is the information entropy, which indicates a measure of uncertainty associated with a random variable. The measure of diversity $D(X)$ in formula (4) means a kind of information description on state space and a measure of whole uncertainty and total information of a system (Laxton 1978).

In general, for two sources of diversity in the same parameter space of d dimensions X: $[n_1,n_2,\ldots,n_i,\ldots,n_d]$ and Y: $[m_1,m_2,\ldots,m_i,\ldots,m_d]$, the increment of diversity (ID), denoted by ID(X,Y), is defined as:

$$\mathrm{ID}(X,Y) = D(X+Y) - D(X) - D(Y) \tag{5}$$

Here, $D(X+Y)$ is the measure of diversity of the sum of two diversity sources called combination diversity source space.

It can be proved that the increment of diversity (ID(X,Y)) satisfies nonnegativity and symmetry. Therefore, the ID is a quantitative measure of the similarity level of two diversity sources. The higher the similarity of two sources, the smaller the ID.

The ID(X,Y) space can be calculated by using Eq. 5. Then the protein X can be predicted as belonging to the category (mitochondrial (M) or non-mitochondrial (N)) for which the corresponding increment of diversity has the minimum value, and can be formulated as follows:

$$\mathrm{ID}(X, Y^{\xi}) = \mathrm{Min}\left\{\mathrm{ID}(X, Y^{M}), \mathrm{ID}(X, Y^{N})\right\} \quad (\xi = M, N) \tag{6}$$

where ξ can be mitochondrial and non-mitochondrial proteins and the Min means taking the minimum value among those in the parentheses, then the ξ in Eq. 6 will give the protein to which the predicted protein sequence Y should belong.

The local n-peptide compositions

The description of a protein sequence can be based on the n-peptide composition coding. In case of $n = 1$, the coding reduces to the usual amino acid composition, which can be considered as the first-order approximation to the complete protein sequence. For $n = 2$, the coding gives the dipeptide composition. As n increases, the coding provides progressively more detailed sequential information. But in the case of $n \geq 3$, the amount of information parameters increase dramatically, and computation becomes not only impractical but also susceptible to the danger of overfitting. So we chose the case of $n \leq 2$, and split a sequence into three regions: N-terminal, C-terminal and the middle region in between the two terminals, and looked at the local dipeptide composition in segmental fragments of protein sequence.

The hydropathy distribution along protein sequence

It has been shown that the patterns of hydrophobic and hydrophilic residues play a significant role in the definition of global protein structure (Goldenberg 1999). As the hydropathy distribution is broadly conserved in proteins, it has been used to detect analogous as well as distantly related proteins (Russell et al. 1997). The hydropathy distribution along the protein sequence had been recognized as a useful feature for characterization of protein structure in the form of hydropathy profiles (Pánek et al. 2005).

The hydropathy features allow for construction of a high-dimensional hydropathy space, where a protein is represented as points. Similarity of proteins is measured as a distance among the points in the space. Thus, searching the hydropathy space for biologically related proteins is accomplished by using common statistical methods. To obtain the hydropathy characteristics, the amino acids are divided into groups using their individual hydropathies according to the ranges of the hydropathy scale. The three hydropathy characteristics are defined in Table 1. The distribution of amino acids into the three groups was derived from usual classifications of amino acids according to their individual hydropathies (Pánek et al. 2005). In addition, proline, glycine and cysteine are classified into three groups because of their unique backbone properties. The six groups of 20 amino acids are shown in Table 1. So a protein sequence with 20 amino acids can be represented by a sequence with six characters (L (strongly hydrophilic or polar), B (strongly hydrophobic), W (weakly hydrophilic or weakly hydrophobic), P (proline), G (glycine) and C (cysteine)) (Chen and Li 2007a, b). The distribution of the six characters along the protein sequence can be selected as the information parameters of a protein.

Table 1 The hydropathy classification of amino acids

Classification	Abbreviation	Amino acids
Strongly hydrophilic or polar	L	R, D, E, N, Q, K, H
Strongly hydrophobic	B	L, I, V, A, M, F
Weakly hydrophilic or weakly hydrophobic	W	S, T, Y, W
Proline	P	P
Glycine	G	G
Cysteine	C	C

The reduced amino acid alphabet obtained from Protein Blocks

A common way of designing a reduced amino acid alphabet is to cluster amino acids into groups according to specific features. These features may use sequence or structure information. Recently, de Brevern proposed a structural alphabet called Protein Blocks, which is composed of 16 average protein fragments of five residues in length (de Brevern et al. 2000; de Brevern 2005; Joseph et al. 2010). The reduced amino acid alphabet (RAAA, Etchebest et al. 2007) has been applied to various areas of protein annotation successfully (Li and Wang 2007; Nanni and Lumini 2008; Ogul and Mumcuogu 2007; Zuo and Li 2009; Zuo and Li 2010). Compared to the general amino acid composition, the reduced amino acid alphabet not only simplifies the complexity of the protein system, but also improves the ability in finding structurally conserved regions and the structural similarity of entire proteins.

To avoid the complete loss of the sequence-order information, the pseudo-amino acid (PseAA) composition or PseAAC was proposed (Chou 2001, 2005). The essence of Chou's pseudo-amino acid composition is to use a discrete model to represent a protein sample yet without completely losing its sequence-order information. Ever since the concept of Chou's pseudo-amino acid composition was introduced, various PseAAC approaches have been introduced to deal with different problems in proteins and protein-related systems (Cai and Chou 2006; Chou and

Table 2 The scheme for reduced amino acid alphabet based on Protein Blocks method

Size	Protein Blocks method
20	G-I-V-F-Y-W-A-L-M-E-Q-R-K-P-N-D-H-S-T-C
13	G-**IV**-**FYW**-A-L-M-E-**QRK**-P-**ND**-**HS**-T-C
11	G-**IV**-**FYW**-A-**LM**-**EQRK**-P-**ND**-**HS**-T-C
9	G-**IV**-**FYW**-**ALM**-**EQRK**-P-**ND**-**HS**-**TC**
8	G-**IV**-**FYW**-**ALM**-**EQRK**-P-**ND**-**HSTC**
5	G-**IVFYW**-**ALMEQRK**-P-**NDHSTC**

The clustered amino acids are shown by bold values

Shen 2006a, b, 2007; Li and Li 2008b; Wang et al. 2008; Zhang and Fang 2008; Zhang et al. 2008; Zhou et al. 2007). Actually, in this study both the hydropathy distribution along the protein sequence and the composition of reduced amino acid alphabet can be regarded as some variations of Chou's pseudo-amino acid composition.

In this study, the *n*-peptide composition of reduced amino acid alphabet clustered by the Protein Blocks method is used to predict mitochondrial and non-mitochondrial proteins. The distribution of amino acids in Protein Blocks is used to create the clusters of equivalent amino acids according to local structure. The scheme of the reduced amino acid alphabet is shown in Table 2. The feature vector dimensions (*d*) of *n*-peptide composition resulted from different reduced amino acid alphabet sizes are listed in Table 3.

Test and performance assessment

How to evaluate a prediction algorithm is an important issue. Usually, a prediction is evaluated by the re-substitution and jackknife tests. The former reflects the degree of self-consistency and the latter reflects the extrapolating effectiveness of the algorithm.

In the re-substitution test, each of the mitochondrial proteins in the given dataset will be predicted by using the rules derived from the same dataset, the so-called development dataset or training dataset. However, during the process of the re-substitution test, the rule parameters derived from the training data set include the information of the query protein later plugged back in the test. This will certainly underestimate the error and enhance the success rate because the same proteins are used to derive the rule parameters and to test themselves. For some algorithm, there is an overtraining problem. Nevertheless, the re-substitution test is absolutely necessary because it reflects the self-consistency of a prediction method. In other words, the re-substitution test is necessary but not sufficient for evaluating a prediction method.

In statistical prediction, the following three cross-validation tests are often used to examine the power of a predictor: independent dataset test, sub-sampling test and jackknife test. Of these three, the jackknife test is thought the most rigorous and objective one (Chou and Zhang 1995), and hence has been used in many studies (e.g., Feng 2001; Wang et al. 2005; Zhang et al. 2006; Zhou 1998; Zhou and Assa-Munt 2001) for examining the power of various prediction methods. 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. Therefore, the re-substitution test, the jackknife test and the independent dataset test are applied to examine our algorithm.

The predictive capability of the algorithm is estimated by the sensitivity (S_n), specificity (S_p),Matthew's correlation coefficient (MCC) and accuracy (Acc),

$$S_n = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}), \tag{7}$$

$$S_p = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP}), \tag{8}$$

$$\mathrm{Acc} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}) \tag{9}$$

$$\mathrm{MCC} = \frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN})}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN}) \times (\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP})}} \tag{10}$$

where TP denotes the numbers of the correctly recognized positives, FN denotes the numbers of the positives recognized as negatives, FP denotes the numbers of the negatives recognized as positives and TN denotes the numbers of correctly recognized negatives.

Results and discussion

In order to predict mitochondrial proteins, it is very important to choose a set of reasonable information parameters from the sequences. By using various information parameters, our prediction results show that the *n*-peptide compositions of N-terminal region, the hydropathy distribution

Table 3 The RAAA sizes and resulting feature vector dimensions (*d*) used for *n*-peptide composition

n-peptide	The dimension (*d*) with different amino acid alphabet sizes (*S*)					
	$S = 20$	$S = 13$	$S = 11$	$S = 9$	$S = 8$	$S = 5$
$n = 1$	20	13	11	9	8	5
$n = 2$	400	169	121	81	64	25

along the protein sequence, and the n-peptide composition of reduced amino acid alphabet clustered by the Protein Blocks method are very useful for the prediction of the mitochondrial proteins in the malaria parasite.

(1) By choosing the n-peptide composition on the N-terminal with m residues as the information parameters of a diversity source, the predictive results indicate that when $n = 1$ and $m = 20$, sensitivity (S_n), specificity (S_p), Matthew's correlation coefficient (MCC) and accuracy (Acc) are higher. The results are shown by "N-ter" line in Table 4 for the re-substitution test and the jackknife test. Shown in Table 4, by choosing the 1-peptide compositions in the N-terminal region with 20 residues as the only input vector, the prediction performance achieves 86.86% accuracy with 0.69 MCC by the jackknife test. We tried different lengths for both the C-terminal and the middle region and obtained predictive performance lower than N-terminal of a protein. It indicates the importance of the N-terminal region of a protein sequence, because most mitochondrial targeting peptides (mTPs) are located in the N-terminal region of a protein sequence (Emanuelsson et al. 2000; Bender et al. 2003).

(2) For calculating the local features, each protein sequence is split into p parts with same length. We define a protein vector as,

$$\{n_{1,1}, n_{1,2}, \ldots, n_{1,6}, n_{2,1}, n_{2,2}, \ldots, n_{2,6}; \ldots, n_{i,j}, \ldots, n_{p,1}, n_{p,2}, \ldots n_{p,6}\},$$

where $i = 1, \ldots, p$ and $j = 1, \ldots, 6$ are the numbers of the segments and hydropathy characteristics, respectively. In our calculation, 6 corresponding to six different types of amino acids: (L (strongly hydrophilic or polar), B (strongly hydrophobic), W (weakly hydrophilic or weakly hydrophobic), P (proline), G (glycine) and C (cysteine)) which has been described in Table 1. The values of elements $n_{i,j}$ quantify the features. A protein sequence is considered a succession of adjoining segments with the same number of residues. The values of the features are computed by using residues that are successive (i.e., at least 2) and that have the same types of amino acids. Such residues are the subsequences of a protein sequence. Every feature has a value that is equal to the sum of lengths of the subsequences with the same types of amino acids in a segment (Chen and Li 2007a, b).

By choosing $n_{i,j}$ as information parameters, the predictive results indicate that when $p = 20$, sensitivity (S_n), specificity (S_p), Matthew's correlation coefficient (MCC) and accuracy (Acc) are higher. The results are shown by "20 parts" line in Table 4 for the re-substitution test and the jackknife test.

(3) In order to investigate how a particular class or the property of amino acids affects prediction accuracy and determine the optimal amount of information, the 20 amino acid alphabet is reduced to several smaller alphabets according to the Protein Block method, i.e., amino acid pairs with high structural similarity scores are grouped together. Table 3 shows the amino acid alphabet sizes and the feature vector dimensions used for the n-peptide composition. Figure 1 shows the accuracy (Acc) for the jackknife test under different parameter settings for alphabet size (S) and n-peptide compositions ($n = 1, 2$). Thus, the best prediction

Table 4 The prediction results of our method based on the different information parameters for the re-substitution test and the jackknife test

Features	Re-substitution test				Jackknife test			
	S_n (%)	S_p (%)	Acc (%)	MCC	S_n (%)	S_p (%)	Acc (%)	MCC
N-ter	92.50	88.15	89.14	0.74	90.00	85.93	86.86	0.69
20 parts	92.50	88.15	89.14	0.74	87.50	81.48	82.86	0.61
RAAA ($S = 11$, $n = 2$)	87.50	70.37	74.29	0.49	67.50	63.70	64.57	0.26
RAAA ($S = 8$, $n = 2$)	85.00	67.41	71.43	0.44	70.00	62.96	64.57	0.28
N-ter + 20 parts	100	91.11	93.14	0.84	92.50	88.15	89.14	0.74
N-ter + RAAA ($S = 9$, $n = 2$)	100	90.37	92.57	0.83	95.00	86.67	88.57	0.74
N-ter + 20parts + RAAA ($S = 20$, $n = 2$)	100	90.37	92.57	0.83	97.50	89.63	91.43	0.80
N-ter + 20parts + RAAA ($S = 13$, $n = 2$)	100	91.11	93.14	0.84	97.50	88.89	90.86	0.79
N-ter + 20parts + RAAA ($S = 11$, $n = 2$)	**100**	**91.11**	**93.14**	**0.84**	**100**	**89.63**	**92.00**	**0.82**
N-ter + 20parts + RAAA ($S = 9$, $n = 2$)	**100**	**91.11**	**93.14**	**0.84**	**100**	**89.63**	**92.00**	**0.82**
N-ter + 20parts + RAAA ($S = 8$, $n = 2$)	100	90.37	92.57	0.83	**100**	**89.63**	**92.00**	**0.82**
N-ter + 20parts + RAAA ($S = 5$, $n = 2$)	100	90.37	92.57	0.83	97.50	84.44	87.43	0.73

The best results are shown by bold values

S_n, sensitivity; S_p, specificity; Acc, accuracy; MCC, Matthew's correlation coefficient

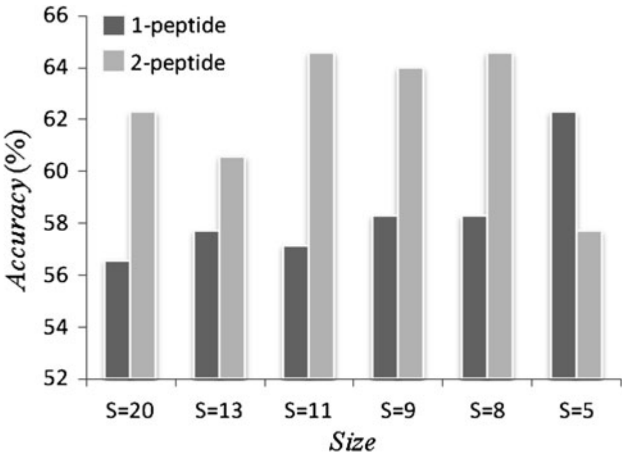

Fig. 1 The accuracy based on n-peptide composition of RAAA with different alphabet size (S) for the jackknife test

accuracies are obtained using 2-peptide composition for the alphabet $S = 11$ and $S = 8$. This result indicates that the dimension of feature vector affects the prediction accuracy. The results are also shown by "RAAA ($S = 11$, $n = 2$)" and "RAAA ($S = 8$, $n = 2$)" lines in Table 4 for both the re-substitution test and the jackknife test.

The hybrid method: a combination of several information parameters from the above three categories

In order to synthesize the effect of the above two or three types of information parameters on predictive results, the hybrid method was developed. Take three types of information parameters for instance, for a protein X, the three increments of diversity $\mathrm{ID}_1(X, Y^{\xi})$, $\mathrm{ID}_2(X, Y^{\xi})$ and $\mathrm{ID}_3(X, Y^{\xi})$between the X and ξ type proteins in the training set can be calculated by using the above three types of information parameters. So we may obtain two increments of diversity for each type of information ($\xi = M, N$),

(1) $\mathrm{ID}_1(X, Y^M)$, $\mathrm{ID}_1(X, Y^N)$
(2) $\mathrm{ID}_2(X, Y^M)$, $\mathrm{ID}_2(X, Y^N)$
(3) $\mathrm{ID}_3(X, Y^M)$, $\mathrm{ID}_3(X, Y^N)$

For the normalization of the IDs obtained by using different information parameters, by normalizing the above three sets of values, respectively, we can obtain the two new IDs by the following method,

$$\mathrm{ID}(X, Y^{\xi}) = \mathrm{ID}_1(X, Y^{\xi}) + \mathrm{ID}_2(X, Y^{\xi}) + \mathrm{ID}_3(X, Y^{\xi}) \quad (\xi = M, N) \tag{11}$$

Protein X is predicted to be the mitochondrial or non-mitochondrial protein for which the corresponding increment of diversity has the minimum value,

$$\mathrm{ID}(X, Y^{\lambda}) = \mathrm{Min}\{\mathrm{ID}(X, Y^M), \mathrm{ID}(X, Y^N)\} \tag{12}$$

where λ can be M or N and the operator Min means taking the minimum value among those in the parentheses, then the λ in Eq. 12 will give the category to which the predicted protein X should belonging.

Based on the hybrid method, the types of 175 PF proteins are predicted by the re-substitution test and the jackknife test in our algorithm. When using a combination of two information parameters as input features, the prediction performance is improved further. As Table 4 shown, the accuracy achieved 89.14% with 0.74 MCC based on "N-ter + 20 parts" by the jackknife test, while the accuracy achieved 88.57% with 0.74 MCC based on "N-ter + RAAA ($S = 9$, $n = 2$)". When using a combination of three information parameters as input features, the results are shown in the "N-ter + 20parts + RAAA" lines in Table 4 with different sizes (S) and n-peptide compositions. The accuracies based on 2-peptide composition of 20 amino acids ($S = 20$, $n = 2$) are 92.57% for the re-substitution test and 91.43% for the jackknife test, respectively. After amino acids reduction, the predictive results show that the Acc values based on 2-peptide composition of 11 and 9 reduced amino acids ($S = 11$, $n = 2$; $S = 9$, $n = 2$) achieve 93.14% for the re-substitution test and 92% for the jackknife test, respectively. Shown in Table 4, the best results when the $S = 11$, $n = 2$, and the $S = 9$, $n = 2$, for the re-substitution test, sensitivity (S_{n}), specificity (S_{p}), accuracy (Acc) and Matthew's correlation coefficient (MCC) are 100%, 91.11%, 93.14% and 0.84. For the jackknife test, the best results are achieved when $n = 2$, and S can be 11, 9, or 8, where sensitivity, specificity, accuracy and Matthew's correlation coefficient are 100%, 89.63%, 92.00% and 0.82, respectively. From the prediction results in Table 4, we can see that the reduced amino acid alphabet of Protein Blocks not only can improve the prediction performance, it also simplifies the amino acid composition of a protein. The reduced n-peptide composition can extract more useful function and structure information than the original n-peptide composition of PF proteins, and can also reduce the dimensions of the feature space.

In order to compute realistic performance of models, it is important to evaluate the performance of models on an independent dataset, not used in training or testing of models. Thus, we evaluated the performance of our model on an independent dataset, which consisted of 24 PF mitochondrial proteins extracted from UniProt, a protein sequence database which strives to provide a high level of manual annotation. Our hybrid model based on "N-ter + 20parts + RAAA ($S = 11$, $n = 2$)" and "N-ter + 20parts + RAAA ($S = 9$, $n = 2$)" correctly predicted 19 mitochondrial proteins. When compared with PlasMit (Bender et al. 2003) and

Table 5 Comparison of prediction performance for different methods on 175 PF proteins (40 mitochondrial and 135 non-mitochondrial proteins)

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
MitoProtII[a]	80.00	73.33	74.86	0.49
Target P[a]	55.00	96.30	86.86	0.60
PlasMit[a]	94.00	89.00	90.00	0.74
PFMpred[b]	97.50	90.37	92.00	0.81
Our method[c]	**100**	**89.63**	**92.00**	**0.82**

[a] As reported by Bender et al. 2003

[b] PSSM + SAAC evaluated using fivefold cross-validation (Verma et al. 2010)

[c] Our method (N-ter + 20parts + RAAA where $n = 2$, $S = 11$, 9, or 8) evaluated using the jackknife test

PFMpred (Verma et al. 2010) methods, which currently both good prediction performance methods, our method correctly predicted more than 5 and 1 mitochondrial proteins, respectively.

In addition to the evaluation on an independent dataset, we also compared our method with existing approaches on 175 PF proteins including 40 mitochondrial and 135 non-mitochondrial proteins (Table 5). Verma et al. (2010) employed the Support Vector Machine (SVM) model for predicting mitochondrial proteins of PF, using split amino acid composition (SAAC) and PSSM profile, the sensitivity (S_n), specificity (S_p), accuracy (Acc) and Matthew's correlation coefficient (MCC) were 97.5%, 90.37%, 92% and 0.81 using fivefold cross-validation. The results in Table 5 show that the sensitivity (S_n), specificity (S_p), accuracy (Acc) and Matthew's correlation coefficient (MCC) which obtained by our ID method based on "N-ter + 20-parts + RAAA ($n = 2$; $S = 11$, 9, or 8)" achieved 100%, 89.63%, 92% and 0.82 using the jackknife test. The specificity in our method is slightly less than the result in Verma et al. (2010), but the sensitivity and MCC values are higher in our method. Moreover, the accuracy in our method using the jackknife test is similar to the SVM model (Verma et al. 2010).

The jackknife test is deemed as the most effective and objective test and does not have the overfitting problem. The high prediction performance also indicates that the algorithm of increment of diversity based on hydropathy distribution and several reduced amino acid alphabets (RAAA) are useful and effective for predicting mitochondrial proteins of *Plasmodium falciparum* (PF).

Acknowledgments The authors would like to thank the reviewers for their comments that help improve the manuscript. This work was supported by the National Natural Science Foundation of China (No. 61063016), the Natural Science Foundation of Inner Mongolia Autonomous Region (No. 200607010101, 20080404MS0105) and the National Science Foundation grant of the United States (No. IIS-0710945).

References

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

Bendtsen JD, Jensen LJ, Blom N, von Heijne G, Brunak S (2004) Feature-based prediction of non-classical and leaderless protein secretion. Protein Eng Des Sel 17:349–356

Cai YD, Chou KC (2006) Predicting membrane protein type by functional domain composition and pseudo amino acid composition. J Theor Biol 238:395–400

Chen YL, Li QZ (2007a) Prediction of the subcellular location of apoptosis proteins. J Theor Biol 245:775–783

Chen YL, Li QZ (2007b) Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. J Theor Biol 248:377–381

Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. Proteins 43:246–255

Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21:10–19

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) Large-scale predictions of Gram-negative bacterial protein subcellular locations. J Proteome Res 5:3420–3428

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

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

de Brevern AG (2005) New assessment of a structural alphabet. In Silico Biol 5:283–289

de Brevern AG, Etchebest C, Hazout S (2000) Bayesian probabilistic approach for prediction backbone structures in terms of protein blocks. Protein Struct Funct Genet 41:271–287

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

Etchebest C, Benros C, Bornot A, Camproux AC, de Brevern AG (2007) A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. Eur Biophys J 36:1059–1069

Feng ZP (2001) Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. Biopolymers 58:491–499

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

Goldenberg DP (1999) Finding the right fold. Nat Struct Biol 6:987–990

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

Höglund A, Doennes P, Blum T, Adolph HW, Kohlbacher O (2006) MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs, and amino acid composition. BMC Bioinform 22:1158–1165

Hu XZ, Li QZ (2008) Using support vector machine to predict β- and γ-turns in proteins. J Comput Chem 29:1867–1875

Joseph AP, Agarwal G, Mahajan S, Gelly JC, Swapna LS, Offmann B, Cadet F, Bornot A, Tyagi M, Valadié H, Schneider B, Etchebest C, Srinivasan N, de Brevern AG (2010) A short survey on Protein Blocks. Biophys Rev 2:137–145

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

Laxton RR (1978) The measure of diversity. J Theor Biol 71:51–67

Li FM, Li QZ (2008a) Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. Amino Acids 34:119–125

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

Li QZ, Lu ZQ (2001) The prediction of the structural class of protein: application of the measure of diversity. J Theor Biol 213: 493–502

Li J, Wang W (2007) Grouping of amino acids and recognition of protein structurally conserved regions by reduced alphabets of amino acids. Sci China C Life Sci 50:392–402

Lin H, Li QZ (2007a) Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. J Comput Chem 28:1463–1466

Lin H, Li QZ (2007b) Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem Biophys Res Commun 354:548–551

Nanni L, Lumini A (2008) A genetic approach for building different alphabets for peptide and protein classification. BMC Bioinform 9:45

Ogul H, Mumcuogu EU (2007) Subcellular localization prediction with new protein encoding schemes. IEEE/ACM Trans Comput Biol Bioinform 24:227–232

Pánek J, Eidhammer I, Aasland R (2005) A new method for identification of protein (sub)families in a set of proteins based on hydropathy distribution in proteins. Proteins Struct Funct Genet 58:923–934

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

Russell RB, Saqi MA, Sayle RA, Bates PA, Sternberg MJ (1997) Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. J Mol Biol 269: 423–439

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423

Snow RW, Guerra CA, Noor AM, Myint HY, Hay SI (2005) The global distribution of clinical episodes of *Plasmodium falciparum* malaria. Nature 434:214–217

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, Varshney GC, Raghava GPS (2010) Prediction of mitochondrial proteins of malaria parasite using split amino acid composition and PSSM profile. Amino Acids 39:101–110

Wang M, Yang J, Xu ZJ, Chou KC (2005) SLLE for predicting membrane protein types. J Theor Biol 232:7–15

Wang T, Yang J, Shen HB, Chou KC (2008) Predicting membrane protein types by the LLDA algorithm. Protein Pept Lett 15: 915–921

Zhang GY, Fang BS (2008) Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition. J Theor Biol 253: 310–315

Zhang SW, Pan Q, Zhang HC, Shao ZC, Shi JY (2006) Prediction of protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and Naive Bayes Feature Fusion. Amino Acids 30:461–468

Zhang TL, Ding YS, Chou KC (2008) Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern. J Theor Biol 250:186–193

Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17:729–738

Zhou GP, Assa-Munt N (2001) Some insights into protein structural class prediction. Proteins Struct Funct Genet 44:57–59

Zhou XB, Chen C, Li ZC, Zou XY (2007) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J Theor Biol 248: 546–551

Zuo YC, Li QZ (2009) Using reduced amino acid composition to predict defensin family and subfamily: Integrating similarity measure and structural alphabet. Peptides 30:1788–1793

Zuo YC, Li QZ (2010) Using K-minimum increment of diversity to predict secretory proteins of malaria parasite based on groupings of amino acids. Amino Acids 38:859–867