

Using *K*-minimum increment of diversity to predict secretory proteins of malaria parasite based on groupings of amino acids

Yong-Chun Zuo · Qian-Zhong Li

Received: 27 December 2008 / Accepted: 1 April 2009 / Published online: 22 April 2009
© Springer-Verlag 2009

Abstract Due to the complexity of *Plasmodium falciparum* genome, predicting secretory proteins of *P. falciparum* is more difficult than other species. In this study, based on the measure of diversity definition, a new *K*-nearest neighbor method, *K*-minimum increment of diversity (*K*-MID), is introduced to predict secretory proteins. The prediction performance of the *K*-MID by using amino acids composition as the only input vector achieves 88.89% accuracy with 0.78 Mathew's correlation coefficient (MCC). Further, the several reduced amino acids alphabets are applied to predict secretory proteins and the results show that the prediction results are improved to 90.67% accuracy with 0.83 MCC by using the 169 dipeptide compositions of the reduced amino acids alphabets obtained from Protein Blocks method.

Keywords Secretory proteins · Increment of diversity · Reduced amino acids alphabets · Amino acid and dipeptide composition · Prediction performance

Introduction

While several species of *Plasmodium* cause disease in humans (including *P. vivax*, *P. malariae*, *P. ovale* and *P. knowlesi*), *Plasmodium falciparum* (*P. falciparum*) is by far the deadliest. The malaria caused by *P. falciparum* remains the world's most devastating tropical infectious disease, which results in 300–500 million clinical cases and

1–2 million deaths annually and its long-term control and eradication is still a long way off (Snow et al. 2005; Winzeler 2008). The potential for developing effective drugs and vaccines against this parasite is thus considerable. Parasite secretes an array of proteins within the host erythrocyte and beyond to facilitate its own survival within the host cell and for immunomodulation (Verma et al. 2008). These proteins secreted by parasite can serve as potential drug or vaccine targets. Therefore, the identification of secretory proteins of *P. falciparum* will be helpful for drug design and combination (Birkholtz et al. 2008).

Recently, An increasing amount of studies have indicated that mathematical/computational approaches, such as molecular docking (Chou et al. 2003; Wang et al. 2008a), pharmacophore modeling (Chou et al. 2006; Sirois et al. 2004), protein subcellular location prediction (Chou and Shen 2007d, 2008a; Shen and Chou 2007c), protein structural class prediction (Chou 1995, 2000; Chou and Cai 2004; Chou and Zhang 1995; Xiao et al. 2008), identification of proteases and their types (Chou and Shen 2008b; Shen and Chou 2009), protein cleavage site prediction (Chou 1993, 1996; Shen and Chou 2008a), and signal peptide prediction (Chou and Shen 2007e; Shen and Chou 2007b), can be timely provide very useful information and insights for both basic research and drug design and hence are widely welcome by science community. The present study was devoted to develop a novel computational approach for predicting secretory proteins that would be particularly effective in characterizing the properties of selected proteins. Several programs have been developed for predicting secretory proteins, such as SecretomeP-2.0 (Bendtsen et al. 2004), TargetP (Emanuelsson et al. 2000), SRTpred (Garg and Raghava 2008) and the work of Liu et al. (2007). When *P. falciparum* genome sequence was published in 2002, it was revealed that the nucleotide composition was unusually

Y.-C. Zuo · Q.-Z. Li (✉)
Laboratory of Theoretical Biophysics,
School of Physical Science and Technology,
Inner Mongolia University, 010021 Hohhot, China
e-mail: qzli@imu.edu.cn

AT-rich ($\sim 80\%$ AT on average) and the proteins of *P. falciparum* are more complex than other species. Thus, the prediction of secretory proteins for *P. falciparum* is more difficulty than other species (Gardner et al. 2002). Recently, Verma et al. first developed the SVM models for predicting secretory proteins of malaria parasite, and achieved good prediction accuracy (Verma et al. 2008).

On the basis of the Shannon entropy definition, Laxton introduced the concept of measure of diversity (Laxton 1978), which is a parallel definition with Shannon entropy. The measure of diversity is a kind of information description on discrete state space and a measure of whole 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 are successfully developed and employed for classification in biogeography. Recently, the Li's group firstly introduced the ID 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 2007; Li and Li 2008a, b), beta-hairpin and gamma-turn prediction (Hu and Li 2008) and good prediction performances are obtained. It can be proved that the ID is a good index for distinguishing two different sources established by proteins. In this paper, based on the ID and K -nearest neighbor method, the K -minimum increment of diversity (K -MID) is developed to predict secretory protein of malaria parasite. Using amino acid composition, the prediction accuracy and Mathew's correlation coefficient (MCC) are 88.69% and 0.78 when $K = 5$, higher than the SVM models. In order to investigate how a particular class or property of amino acids affects prediction accuracy and examine the effect of special amino acid with different biochemical properties, several different reduced amino acids alphabets are introduced in this study. The results indicate that the 20 amino acids can be clustered into about ten reduced amino acid groups. And by using reduced amino acids obtained from Protein Blocks method, the best prediction performance is obtained.

Materials and methods

Datasets

A critical issue in developing secretory protein prediction algorithm of malaria parasite is lack of suitable training and testing sets. In this study, the 252 secretory proteins and 252 non-secretory proteins were constructed by Verma et al. (2008). From the literature Verma et al. collected total 267 secretory proteins consisting of 208

secretory proteins (119 Rifins, 22 Stevors, 67 PfEMP1); 6 experimentally proven proteins (PF10_0159, PFE0040c, PFB0100c, PFB0095c, AAD31511, AAC47454). Another set of 3 experimentally proved secretory proteins (PFD1175w, PFD1170c, PFB0100c); more 7 proteins (PFI1755c, PFE0055c, PFI1780w, PFE0360c, PF10_0321, PF14_0607, PFE0355c); 4 REX proteins (PFI1740c, PFI1755c, PFI1760w, PFI1735c); 2 PIESTPs (PFC0435c, PFE0060w); clag9 (PFI1730w); Sbp1 (PFE0065w) and 35 maurer's cleft associated proteins. These all sum up to 267 secretory proteins. They got 252 non redundant secretory proteins after removing redundant proteins using program PROSET. The 252 non-secretory proteins are extracted from two sources, 197 non-secretory proteins are extracted from Swiss-Prot using SRS with query "*Plasmodium falciparum* (organism) but not secreted (comment)" and the remaining 55 non-secretory proteins are extracted nuclear proteins from PlasmoDB and randomly picked up 55 proteins from ~ 300 nuclear proteins.

The definition of increment of diversity

For a discrete state space X with d dimension $X: \{n_1, n_2, \dots, n_d, \dots, n_d\}$, n_i denotes the times of i th state, the Shannon information entropy (Shannon 1948), a measure of uncertainty and denoted by $H(X)$, is defined as:

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

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

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

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

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

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

So we have

$$D(X) = N \cdot H(X) \quad (4)$$

$H(X)$ is the information entropy, which indicates a measure of the 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 $X: \{n_1, n_2, \dots, n_i, \dots, n_d\}$ and $Y: \{m_1, m_2, \dots, m_i, \dots, m_d\}$, the increment of diversity (ID), denoted by $ID(X, Y)$, is defined as:

$$ID(X, Y) = D(X + Y) - D(X) - D(Y) \quad (5)$$

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

It is easily proved that the increment of diversity ($ID(X, Y)$) satisfies nonnegative 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 K -minimum increment of diversity (K -MID) classifier

The K -nearest neighbor (K -NN) technique has become extremely popular for a variety of forest inventory mapping and estimation applications, such as protein subcellular localization (Chou and Shen 2006a, b, 2007a, b; Shen and Chou 2007b, c; Shen et al. 2007), subnuclear protein localization (Shen and Chou 2005a), protein structural classification (Shen et al. 2005; Zhang et al. 2008a, b), protein fold pattern (Shen and Chou 2006), membrane protein type (Shen and Chou 2005b; Shen et al. 2006; Chou and Shen 2007c), enzyme main and sub functional classification (Shen and Chou 2007a) as well as signal peptide (Chou and Shen 2007e). Much of this popularity may be attributed to the non-parametric, multivariate features of the technique, its intuitiveness, and its ease of use. The query protein should be classified by a majority vote of its neighbors, with the protein being assigned to the class most common amongst its K nearest neighbors. K is a positive integer, typically small. If $K = 1$, then the protein is simply assigned to the class of its nearest neighbor. Although different distance measures can be used for this, such as Euclidean distance, Hamming distance (Mardia et al. 1979) and Mahalanobis distance (Chou 1995), the Euclidean distance is mostly used. In this paper, the similarity measure of ID is used for predicting secretory protein.

For an arbitrary protein sequence X to be predicted, the increment of diversity (ID) between the sequence and to all stored sequences of the diversity sources established by secretory proteins (S) or non-secretory proteins (N) are computed. K -minimum IDs are selected and the average ID, denoted by K -MID(X, Y), is calculated as follows:

$$K\text{-MID}(X, Y) = \frac{1}{K} \sum_{i=1}^K ID(X, Y) \quad (6)$$

The $ID(X, Y)$ can be calculated by using Eq. 5. Then the protein X can be predicted as belonging to the category (secretory (S) or non-secretory (N)) for which the corresponding K -MID has the minimum value, and can be formulated as follows:

$$K\text{-MID}(X, Y^\xi) = \text{Min}\{K\text{-MID}(X, Y^S), K\text{-MID}(X, Y^N)\} \\ (\xi = S, N) \quad (7)$$

where ξ can be secretory and non-secretory proteins and the Min means taking the minimum value among those in the parentheses, then the ξ in Eq. 7 will give the protein to which the predicted protein sequence Y should belong.

Reduced amino acids alphabets

It has been found that some residues are similar in their physicochemical features, and can be clustered into groups because they play similarly structural or functional roles in proteins (Regan and Degradó 1988; Kamtekar 1993; Henikoff and Henikoff 1992). The reduced amino acids not only simplify the complexity of the protein system, but also improve the ability in finding structurally conserved regions and the structural similarity of entire proteins. In recent years, several alphabet reduction techniques have been applied to protein prediction, such as intrinsically disordered proteins prediction (Weathers et al. 2004), recognition of protein structurally conserved regions (Li and Wang 2007), subcellular localization prediction (Oğul and Mumcuoğlu 2007) and peptide and protein classification (Nanni and Lumini 2008). To investigate how a particular class or property of amino acids affects prediction accuracy, several reduced amino acid alphabets with different clustering approaches were discussed in this study. These cluster methods include Miyazawa–Jernigan matrix (MJM) based on inter-residue contact energies (Rakshit and Ananthasuresh 2008), Markov models of evolution (Susko and Roger 2007), BLOSUM62 matrix (Li et al. 2003), Protein Blocks based on local protein structures (Etchebest et al. 2007) and BLOSUM50 similarity matrix (Henikoff and Henikoff 1992).

Protein sequence representation

The amino acid composition (AAC) representation of a given sequence is composed by 20 different amino acids with a variety of shapes, size and chemical properties. The AAC representation has recently been widely utilized in predicting protein function annotation. To avoid completely lose the sequence-order information, the pseudo-amino acid (PseAA) composition or PseAAC was proposed (Chou

2001; Chou 2005). The essence of Chou's pseudo-amino acid composition is to use a discrete model to represent a protein sample yet without complete losing its sequence-order information. Ever since the concept of Chou's pseudo-amino acid composition was introduced, various PseAAC approaches have been stimulated to deal with different problems in proteins and protein-related systems (Chen et al. 2009; Ding and Zhang 2008; Jiang et al. 2008; Li and Li 2008b; Lin 2008; Lin et al. 2008; Wang et al. 2008b; Zhang and Fang 2008; Zhang et al. 2008a, b; Zhou et al. 2007). Owing to its wide usage, recently a very flexible PseAA composition generator, called "PseAAC" (Shen and Chou 2008b), was established at the website <http://chou.med.harvard.edu/bioinf/PseAAC/>, by which users can generate 63 different kinds of PseAA composition. In this study, the amino acid compositions (AAC) and dipeptide compositions (DPC) of the reduced amino acid alphabets are selected to test the K -MID algorithm.

Test and assessment

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling test, and jackknife test (Chou and Zhang 1995). However, as elucidated by Chou and Shen (2008a, b) and demonstrated in Chou and Shen (2007d), among the three cross-validation methods, the jackknife test is deemed the most objective that can always yield a unique result for a given benchmark dataset, and hence has been increasingly used by investigators to examine the accuracy of various predictors (Jiang et al. 2008; Li and Li 2008b; Lin 2008; Lin et al. 2008; Yang and Chou 2008; Zhang and Fang 2008; Zhang et al. 2008a, b; Zhou 1998; Zhou and Assa-Munt 2001; Zhou and Doctor 2003; Zhou et al. 2007). During the process of jackknife test, each protein is singled out in turn as a test sample, the remaining proteins are used as training set to calculate test sample's membership and predict the class. The prediction performance was evaluated by the sensitivity (Sn), specificity (Sp), positive predictive value (PPV), accuracy (Acc) and Mathew's correlation coefficient (MCC), which defined as follows:

$$Sn = TP / (TP + FN) \quad (8)$$

$$Sp = TN / (TN + FP) \quad (9)$$

$$PPV = TP / (TP + FP) \quad (10)$$

$$Acc = (TP + TN) / (TP + FN + TN + FP) \quad (11)$$

MCC

$$= \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TN + FN) \times (TP + FP) \times (TN + FP)}} \quad (12)$$

where TP denotes the number of the correctly predicted secretory proteins, FN denotes the number of the secretory proteins predicted as non-secretory proteins, FP denotes the number of the non-secretory proteins predicted as secretory proteins, and TN denotes the number of correctly predicted non-secretory proteins.

Result and discussion

To investigate the best K value for predicting secretory proteins, test has been done by using of the various values of minimum increment of diversity (MID) K (from 1 to 20). The prediction results compared with SVM Models based on the 20 amino acid composition (AAC) are shown in the Table. 1. For different values of K , it is shown that the prediction ability is improved along with the K increase, up to the peak when K equals to 5, and decrease when the $K > 11$. The prediction accuracy (Acc) and MCC are not changed significantly at 6–11. The performance of prediction achieves 88.89% Acc with 0.78 MCC when $K = 5$, better than the best results achieved by the SVM models with 85.66% Acc and 0.72 MCC when Thr = 0.4. Therefore, in the following calculations, the $K = 5$ is used as the operation parameters.

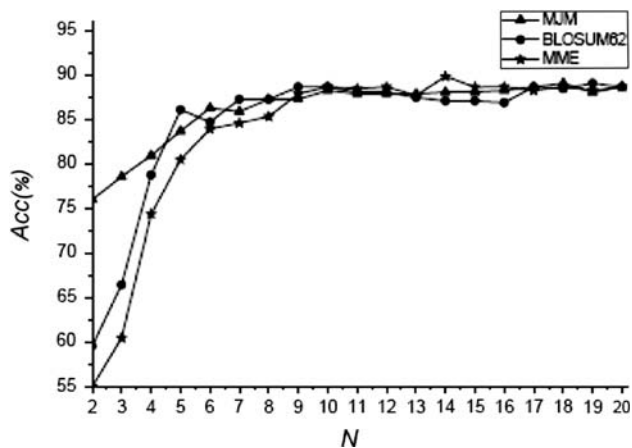
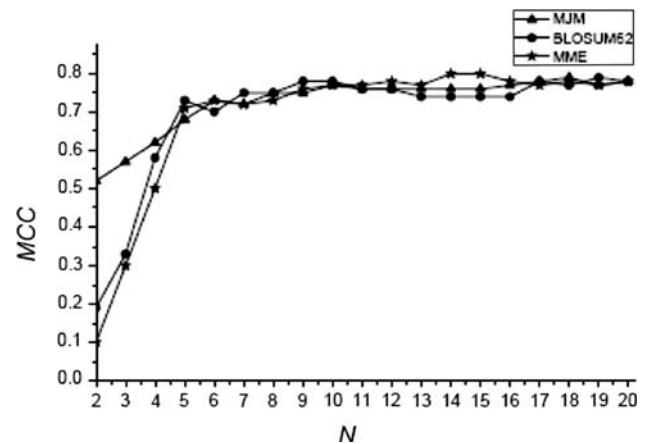
In order to investigate how a particular class or property of amino acids affects prediction accuracy and to determine the minimal amount of information needed for prediction, three latest reduced amino acids methods, Miyazawa–Jernigan matrix (MJM) (Rakshit and Ananthasuresh 2008), Markov models of evolution (MME) (Susko and Roger 2007) and BLOSUM62 matrix (Li et al. 2003), are applied to predict the secretory proteins of *P. falciparum*. The prediction accuracy (Acc) and MCC with the different number of alphabets N are shown in Fig. 1 and 2, respectively. The results show that the Acc and MCC do not present significant change when the N at 10–20. The Acc or MCC with the number of alphabets around 10 performed similarly with the N from 12 to 20. This regular is similar to other studies, such as disorder protein prediction ($N = 10$), and structure conservative regions prediction ($N = 9$). The conclusion indicates that the amino acids content with similar features of protein sequence can be clustered properly.

Based on the above discussion, two latest reduced amino acids alphabets based on Protein Blocks (Etchebest et al. 2007) and BLOSUM50 substitution matrix (Weathers et al. 2004) methods are used to further determine the optimization of alphabets for secretory protein prediction (Table. 2). The reduced amino acid alphabet obtained from Protein Blocks method is a kind of structural alphabet which is composed of 16 average protein fragments of five residues in length. Because the reduced amino acid

Table 1 Prediction result of *K*-MID compared with the SVM models based on amino acid composition

<i>K</i>	<i>K</i> -MID(20 AAC)					Thr	SVM Models(20 AAC)				
	Sn(%)	Sp(%)	PPV(%)	Acc(%)	MCC		Sn(%)	Sp(%)	PPV(%)	Acc(%)	MCC
1	83.33	90.08	89.36	86.71	0.74	−1.0	94.84	27.60	–	61.35	0.30
2	84.13	91.27	90.60	87.70	0.76	−0.9	93.25	33.20	–	63.35	0.33
3	83.52	93.25	92.51	88.53	0.77	−0.8	92.86	37.60	–	65.34	0.37
4	83.73	93.65	92.95	88.69	0.78	−0.7	90.48	43.20	–	66.93	0.38
5	84.52	93.25	92.61	88.89	0.78	−0.6	89.29	48.00	–	68.73	0.41
6	83.73	93.25	92.54	88.49	0.77	−0.5	88.89	58.40	–	73.71	0.50
7	83.73	93.25	92.54	88.49	0.77	−0.4	87.70	65.60	–	76.69	0.55
8	82.94	93.25	92.48	88.10	0.77	−0.3	86.90	72.00	–	79.48	0.60
9	82.54	93.25	92.44	87.90	0.76	−0.2	85.71	78.00	–	81.87	0.64
10	82.54	93.25	92.44	87.90	0.76	−0.1	85.32	80.80	–	83.07	0.66
11	82.54	93.65	92.86	88.10	0.77	0.0	83.33	84.00	–	83.67	0.67
12	82.14	92.86	92.00	87.50	0.75	0.1	81.75	85.20	–	83.47	0.67
13	82.14	93.25	92.41	87.70	0.76	0.2	80.56	86.00	–	83.27	0.67
14	82.14	92.86	92.00	87.50	0.75	0.3	79.76	89.60	–	84.66	0.70
15	81.75	92.86	91.96	87.30	0.75	0.4	78.97	92.40	–	85.66	0.72
16	81.75	93.25	92.38	87.50	0.76	0.5	77.78	93.20	–	85.46	0.72
17	81.35	93.25	92.34	87.30	0.75	0.6	76.98	94.40	–	85.66	0.72
18	80.95	92.86	91.89	86.90	0.74	0.7	76.19	95.60	–	85.86	0.72
19	80.95	92.06	91.07	86.51	0.73	0.8	73.02	96.00	–	84.46	0.71
20	80.95	92.06	91.07	86.51	0.73	0.9	70.24	96.40	–	83.27	0.69
21	80.16	92.06	90.99	86.11	0.73	1.0	65.48	98.80	–	82.07	0.68

The bold values show the best results

**Fig. 1** Prediction accuracy of *K*-MID method by using different reduced amino acids methods**Fig. 2** Prediction Mathew's correlation coefficient of *K*-MID method by using different reduced amino acids methods

alphabet obtained from Protein Blocks method can extract more useful information in secretory protein sequences, eliminate some useless information and reduces the dimension of the feature space, the Protein Blocks method has been successful used to analyze longer protein fragments and to predict functional regions. And the results have proven their efficiency both in description and prediction of longer fragments (Etchebest et al. 2007;

de Brevern 2005), such as local protein structures prediction (Benros et al. 2006), outer membrane proteins analysis (Martin et al. 2008) and backbone structure prediction of proteins (de Brevern 2005). The prediction results of *K*-MID based on the amino acids composition of the reduced alphabets with $K = 5$ are shown in Table. 3.

As Table 3 shown, the sizes of 13, 11 and 9 vectors achieve 88.89, 87.70 and 87.50% accuracy (Acc) for

Table 2 Two schemes for reducing amino acid alphabet used in our study

Size	Protein Blocks	Size	BLOSUM50 substitution matrix
20	G-I-V-F-Y-W-A-L-M-E-Q-R-K-P-N-D-H-S-T-C	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	15	G- FY -W-A-ILMV-E-Q- RK -P-N-D-H-S-T-C
11	G- IV -FYW-A-LM- EQRK -P-ND-HS-T-C	10	G- FYW -A-ILMV- RK -P- EQND -H-ST-C
9	G- IV -FYW-ALM- EQRK -P-ND-HS-TC	8	FWY -CILMV-GA-ST-P- EQND -H-KR
8	G- IV -FYW-ALM- EQRK -P-ND-HSTC	4	FWY -CILMV-AGSTP- EQNDH KR
5	G- IVFYW -ALME QRK -P-NDHSTC	3	FWYCILMVAGSTPHNQ -DE-KR

The clustered amino acids are shown by bold values

Table 3 The prediction performance of *K*-MID method for different vector sizes of reduced amino acids alphabets with *K* = 5

<i>K</i> = 5	Protein Blocks					BLOSUM50 substitution matrix				
Size	13	11	9	8	5	15	10	8	4	3
Sn(%)	85.32	85.32	84.92	85.71	82.14	84.92	86.11	82.14	75.00	76.98
Sp(%)	92.46	90.08	90.08	88.49	77.38	92.06	89.29	87.70	85.71	80.56
PPV(%)	91.88	89.58	89.54	88.16	78.41	91.45	88.93	86.97	84.00	79.84
Acc(%)	88.89	87.70	87.50	87.10	79.76	88.49	87.70	84.92	80.36	78.77
MCC	0.78	0.75	0.75	0.74	0.60	0.77	0.75	0.70	0.61	0.58

The bold values show the best results

Protein Block method, and the sizes of 15, 10 and 8 vectors achieved 88.49, 87.70 and 84.92% accuracy for BLOSUM50 substitution matrix method. The best results are 88.89% Acc and 0.78 MCC with the 13 vector size, the same to the prediction performance of 20 amino acid compositions. When using dipeptide composition as the input features, the prediction performance is improved further. The accuracy achieved 89.88% with 0.81 MCC based on 100 dipeptide compositions (DPC) of BLOSUM50 substitution matrix reduced alphabets. The best prediction accuracy is up to 90.67% by using the 169 DPC of Protein Blocks reduced alphabets with 0.83 MCC. In summary, the suitable reduced amino acids alphabets can improve the predict accuracy by clustering the similar amino acids and the reduce alphabets also can reduce the dimensions of the feature space.

In order to examine the performance of our method, some comparisons with the SVM models program are made, and the prediction results of two methods based on different features are listed in Table 4. The results in Table 4 show that prediction accuracy(Acc) obtained by our *K*-MID method based on 20 AAC achieves 88.89% with 0.78 MCC, about 3.2% higher than the SVM models with 85.66% Acc and 0.72 MCC. For the prediction based on 400 dipeptide compositions, the *K*-MID method achieves 34.92% sensitivity (Sn) with 100% specificity (Sp) and the SVM models only achieve 24.21% Sn with the 99.60% Sp, about 10.71% higher than the SVM models. The sensitivity (Sn) achieves 79.76% with 100% specificity

Table 4 Comparisons of *K*-MID method with the SVM models for secretory protein prediction

Features	Sn(%)	Sp(%)	PPV(%)	Acc(%)	MCC
20 AAC (SVM)	78.97	92.40	–	85.66	0.72
20 AAC (<i>K</i> -MID, <i>K</i> = 5)	84.52	93.25	92.61	88.89	0.78
400 DPC (SVM)	24.21	99.60	–	61.75	0.36
400 DPC (<i>K</i> -MID, <i>K</i> = 5)	34.92	100.00	100.00	67.46	0.46
PSSM profiles (SVM)	82.94	99.21	–	91.07	0.83
Split amino acids (SVM)	58.73	99.60	–	79.00	0.64
Hydrophobic (SVM)	82.94	95.22	–	89.07	0.79
pI at 25°C (SVM)	82.54	94.44	–	88.49	0.78
13 AAC(<i>K</i> -MID, <i>K</i> = 5)	82.94	96.03	95.43	89.48	0.80
169 DPC (<i>K</i> -MID, <i>K</i> = 5)	81.75	99.60	99.52	90.67	0.83
10 AAC (<i>K</i> -MID, <i>K</i> = 5)	86.11	89.29	88.93	87.70	0.75
100 DPC (<i>K</i> -MID, <i>K</i> = 5)	79.76	100.00	100.00	89.88	0.81

The best results are shown by bold values

(Sp) by using 100 dipeptide compositions of reduced amino acids alphabet. The best prediction performance of SVM models achieves 91.07% Acc with 82.94% Sn, 99.21% Sp and 0.83 MCC by using the PSSM profiles obtained by PSI-BLAST. Based on the 169 dipeptide compositions of reduced amino acids alphabet as the only input vectors, the *K*-MID method achieved 90.67% Acc with 81.75% Sn, 99.60% Sp, 99.52% PPV, and 0.83 MCC, which are similar to the SVM models. The surprising good prediction performance indicates that the *K*-MID method is indeed a good predictor for secretory proteins annotation.

Conclusion

For protein prediction and classification, most of the existing methods are based on a group of features that possess kinds of discriminative information from the protein sequence. In this study, the *K*-MID method is firstly developed to predict secretory protein of malaria parasite. The successful prediction performance indicates that amino acid composition and ID combined with *K*-nearest neighbor method are quite suitable to predict secretory protein. The reduced amino acids alphabets can reduce the dimension of inputting vector and improve the prediction accuracy. The results obtained in our study have also demonstrated that amino acid alphabet obtained from Protein Blocks method has the ability of abstracting useful functional and conservative information and it is suitable for secretory protein prediction. When compared with the work of Verma et al. (2008), the results show that the sensitivity in our method is less than result in Verma et al. (2008), but the specificity is higher in our results. Moreover, the overall accuracy is higher in our method than results in the work of Verma et al. (2008). We hope this algorithm will assist annotation of protein function and help for drug and vaccine design against malaria caused by *P. falciparum*.

Acknowledgments The author would like to thank R. Verma for sharing the datasets. This work was supported by the National Natural Science Foundation of China (No. 30560039), the Natural Science Foundation of Inner Mongolia Autonomous Region (No. 200607010101) and project for excellent subject-directors of Inner Mongolia Autonomous Region (No. 20060702).

References

- 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. doi:[10.1093/protein/gzh037](https://doi.org/10.1093/protein/gzh037)
- Benros C, de Brevern AG, Etchebest C, Hazout S (2006) Assessing a novel approach for predicting local 3D protein structures from sequence. *Proteins* 62:680–685. doi:[10.1002/prot.20815](https://doi.org/10.1002/prot.20815)
- Birkholtz LM, Blatch G, Coetzer TL, Hoppe HC, Human E, Morris EJ, Ngcete Z, Oldfield L, Roth R, Shonhai A, Stephens L, Louw AI (2008) Heterologous expression of plasmodial proteins for structural studies and functional annotation. *Malar J* 7:197. doi:[10.1186/1475-2875-7-197](https://doi.org/10.1186/1475-2875-7-197)
- Chen YL, Li QZ (2007) Prediction of the subcellular location of apoptosis proteins. *J Theor Biol* 245:775–783. doi:[10.1016/j.jtbi.2006.11.010](https://doi.org/10.1016/j.jtbi.2006.11.010)
- Chen C, Chen L, Zou X, Cai P (2009) Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. *Protein Pept Lett* 16:27–31. doi:[10.2174/092986609787049420](https://doi.org/10.2174/092986609787049420)
- Chou KC (1993) A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. *J Biol Chem* 268:16938–16948
- Chou KC (1995) A novel approach to predicting protein structural classes in a (20–1)-D amino acid composition space. *Proteins* 21:319–344. doi:[10.1002/prot.340210406](https://doi.org/10.1002/prot.340210406)
- Chou KC (1996) Review: prediction of HIV protease cleavage sites in proteins. *Anal Biochem* 233:1–14. doi:[10.1006/abio.1996.0001](https://doi.org/10.1006/abio.1996.0001)
- Chou KC (2000) Review: prediction of protein structural classes and subcellular locations. *Curr Protein Pept Sci* 1:171–208. doi:[10.2174/1389203003381379](https://doi.org/10.2174/1389203003381379)
- Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins* 43:246–255. doi:[10.1002/prot.1035](https://doi.org/10.1002/prot.1035)
- Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* 21:10–19. doi:[10.1093/bioinformatics/bth466](https://doi.org/10.1093/bioinformatics/bth466)
- Chou KC, Cai YD (2004) Predicting protein structural class by functional domain composition. *Biochem Biophys Res Commun* 321:1007–1009. doi:[10.1016/j.bbrc.2004.07.059](https://doi.org/10.1016/j.bbrc.2004.07.059)
- Chou KC, Shen HB (2006a) Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. *Biochem Biophys Res Commun* 347:150–157. doi:[10.1016/j.bbrc.2006.06.059](https://doi.org/10.1016/j.bbrc.2006.06.059)
- Chou KC, Shen HB (2006b) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic *K*-nearest neighbor classifiers. *J Proteome Res* 5:1888–1897. doi:[10.1021/pr060167c](https://doi.org/10.1021/pr060167c)
- Chou KC, Shen HB (2007a) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J Proteome Res* 6:1728–1734. doi:[10.1021/pr060635i](https://doi.org/10.1021/pr060635i)
- Chou KC, Shen HB (2007b) Large-scale plant protein subcellular location prediction. *J Cell Biochem* 100:665–678. doi:[10.1002/jcb.21096](https://doi.org/10.1002/jcb.21096)
- Chou KC, Shen HB (2007c) 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. doi:[10.1016/j.bbrc.2007.06.027](https://doi.org/10.1016/j.bbrc.2007.06.027)
- Chou KC, Shen HB (2007d) Review: recent progresses in protein subcellular location prediction. *Anal Biochem* 370:1–16. doi:[10.1016/j.ab.2007.07.006](https://doi.org/10.1016/j.ab.2007.07.006)
- Chou KC, Shen HB (2007e) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem Biophys Res Commun* 357:633–640. doi:[10.1016/j.bbrc.2007.03.162](https://doi.org/10.1016/j.bbrc.2007.03.162)
- Chou KC, Shen HB (2008a) Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms. *Nat Protocols* 3:153–162. doi:[10.1038/nprot.2007.494](https://doi.org/10.1038/nprot.2007.494)
- Chou KC, Shen HB (2008b) 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. doi:[10.1016/j.bbrc.2008.08.125](https://doi.org/10.1016/j.bbrc.2008.08.125)
- Chou KC, Zhang CT (1995) Review: prediction of protein structural classes. *Crit Rev Biochem Mol Biol* 30:275–349. doi:[10.3109/10409239509083488](https://doi.org/10.3109/10409239509083488)
- Chou KC, Wei DQ, Zhong WZ (2003) Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. *Biochem Biophys Res Commun* 308:148–151. doi:[10.1016/S0006-291X\(03\)01342-1](https://doi.org/10.1016/S0006-291X(03)01342-1)
- Chou KC, Wei DQ, Du QS, Sirois S, Zhong WZ (2006) Review: progress in computational approach to drug development against SARS. *Curr Med Chem* 13:3263–3270. doi:[10.2174/09298670677873077](https://doi.org/10.2174/09298670677873077)
- de Brevern AG (2005) New assessment of a structural alphabet. *Silico Biol* 5:283–289
- 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. doi:[10.1016/j.patrec.2008.06.007](https://doi.org/10.1016/j.patrec.2008.06.007)
- Emanuelsson O, Nielsen H, Brunak S, Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* 300:1005–1016. doi:[10.1006/jmbi.2000.3903](https://doi.org/10.1006/jmbi.2000.3903)
- 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. doi:[10.1007/s00249-007-0188-5](https://doi.org/10.1007/s00249-007-0188-5)
- Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shalom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419:498–511. doi:[10.1038/nature01097](https://doi.org/10.1038/nature01097)
- Garg A, Raghava GPS (2008) A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search. *In Silico Biol* 8:0012
- Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 89:10915–10919. doi:[10.1073/pnas.89.22.10915](https://doi.org/10.1073/pnas.89.22.10915)
- Hu XZ, Li QZ (2008) Using support vector machine to predict β - and γ -turns in proteins. *J Comput Chem* 29:1867–1875. doi:[10.1002/jcc.20929](https://doi.org/10.1002/jcc.20929)
- Jiang X, Wei R, Zhang TL, Gu Q (2008) Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. *Protein Pept Lett* 15:392–396. doi:[10.2174/092986608784246443](https://doi.org/10.2174/092986608784246443)
- Kamtekar S (1993) Protein design by binary patterning of polar and nonpolar amino acids. *Science* 262:1680–1685. doi:[10.1126/science.8259512](https://doi.org/10.1126/science.8259512)
- Laxton RR (1978) The measure of diversity. *J Theor Biol* 71:51–67. doi:[10.1016/0022-5193\(78\)90302-8](https://doi.org/10.1016/0022-5193(78)90302-8)
- Li FM, Li QZ (2008a) Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. *Amino Acids* 34:119–125. doi:[10.1007/s00726-007-0545-9](https://doi.org/10.1007/s00726-007-0545-9)
- 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. doi:[10.2174/092986608784966930](https://doi.org/10.2174/092986608784966930)
- 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. doi:[10.1006/jtbi.2001.2441](https://doi.org/10.1006/jtbi.2001.2441)
- 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. doi:[10.1007/s11427-007-0023-3](https://doi.org/10.1007/s11427-007-0023-3)
- Li T, Fan K, Wang J, Wang W (2003) Reduction of protein sequence complexity by residue grouping. *Protein Eng* 16:323–330. doi:[10.1093/protein/gzg044](https://doi.org/10.1093/protein/gzg044)
- Lin H (2008) The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. *J Theor Biol* 252:350–356. doi:[10.1016/j.jtbi.2008.02.004](https://doi.org/10.1016/j.jtbi.2008.02.004)
- 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. doi:[10.1002/jcc.20554](https://doi.org/10.1002/jcc.20554)
- 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. doi:[10.1016/j.bbrc.2007.01.011](https://doi.org/10.1016/j.bbrc.2007.01.011)
- Lin H, Ding H, Guo FB, Zhang AY, Huang J (2008) Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. *Protein Pept Lett* 15:739–744. doi:[10.2174/092986608785133681](https://doi.org/10.2174/092986608785133681)
- Liu H, Yang J, Liu DQ, Shen HB, Chou KC (2007) Using a new alignment kernel function to identify secretory proteins. *Protein Pept Lett* 14:203–208. doi:[10.2174/092986607779816087](https://doi.org/10.2174/092986607779816087)
- Mardia KV, Kent JT, Bibby JM (1979) Multivariate analysis chapter 11: discriminant analysis; chapter 12: multivariate analysis of variance; chapter 13: cluster analysis. Academic Press, London pp 322–381
- Martin J, de Brevern AG, Camproux AC (2008) In silico local structure approach: a case study on outer membrane proteins. *Proteins* 71:92–109. doi:[10.1002/prot.21659](https://doi.org/10.1002/prot.21659)
- Nanni L, Lumini A (2008) A genetic approach for building different alphabets for peptide and protein classification. *BMC Bioinform* 9:45. doi:[10.1186/1471-2105-9-45](https://doi.org/10.1186/1471-2105-9-45)
- Oğul H, Mumcuoğlu EU (2007) Subcellular localization prediction with new protein encoding schemes. *IEEE/ACM Trans Comput Biol Bioinform* 24:227–232
- Rakshit S, Ananthasuresh GK (2008) An amino acid map of inter-residue contact energies using metric multi-dimensional scaling. *J Theor Biol* 250:291–297. doi:[10.1016/j.jtbi.2007.09.032](https://doi.org/10.1016/j.jtbi.2007.09.032)
- Regan L, Degradó WF (1988) Characterization of a helical protein designed from first principles. *Science* 241:976–978. doi:[10.1126/science.3043666](https://doi.org/10.1126/science.3043666)
- Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:379–423
- Shen HB, Chou KC (2005a) Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. *Biochem Biophys Res Commun* 337:752–756
- Shen HB, Chou KC (2005b) Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. *Biochem Biophys Res Commun* 334:288–292. doi:[10.1016/j.bbrc.2005.06.087](https://doi.org/10.1016/j.bbrc.2005.06.087)
- Shen HB, Chou KC (2006) Ensemble classifier for protein fold pattern recognition. *Bioinformatics* 22:1717–1722. doi:[10.1093/bioinformatics/btl170](https://doi.org/10.1093/bioinformatics/btl170)
- Shen HB, Chou KC (2007a) EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. *Biochem Biophys Res Commun* 364:53–59. doi:[10.1016/j.bbrc.2007.09.098](https://doi.org/10.1016/j.bbrc.2007.09.098)
- Shen HB, Chou KC (2007b) Signal-3L: a 3-layer approach for predicting signal peptide. *Biochem Biophys Res Commun* 363:297–303. doi:[10.1016/j.bbrc.2007.08.140](https://doi.org/10.1016/j.bbrc.2007.08.140)
- Shen HB, Chou KC (2007c) Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. *Protein Eng Des Sel* 20:39–46. doi:[10.1093/protein/gzl053](https://doi.org/10.1093/protein/gzl053)
- Shen HB, Chou KC (2007d) Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. *Biochem Biophys Res Commun* 355:1006–1011. doi:[10.1016/j.bbrc.2007.02.071](https://doi.org/10.1016/j.bbrc.2007.02.071)
- Shen HB, Chou KC (2008a) HIVcleave: a web-server for predicting HIV protease cleavage sites in proteins. *Anal Biochem* 375:388–390. doi:[10.1016/j.ab.2008.01.012](https://doi.org/10.1016/j.ab.2008.01.012)
- Shen HB, Chou KC (2008b) PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. *Anal Biochem* 373:386–388. doi:[10.1016/j.ab.2007.10.012](https://doi.org/10.1016/j.ab.2007.10.012)
- Shen HB, Chou KC (2009) Identification of proteases and their types. *Anal Biochem* 385:153–160. doi:[10.1016/j.ab.2008.10.020](https://doi.org/10.1016/j.ab.2008.10.020)

- Shen HB, Yang J, Liu XJ, Chou KC (2005) Using supervised fuzzy clustering to predict protein structural classes. *Biochem Biophys Res Commun* 334:577–581. doi:[10.1016/j.bbrc.2005.06.128](https://doi.org/10.1016/j.bbrc.2005.06.128)
- Shen HB, Yang J, Chou KC (2006) Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. *J Theor Biol* 240:9–13. doi:[10.1016/j.jtbi.2005.08.016](https://doi.org/10.1016/j.jtbi.2005.08.016)
- Shen HB, Yang J, Chou KC (2007) Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. *Amino Acids* 33:57–67. doi:[10.1007/s00726-006-0478-8](https://doi.org/10.1007/s00726-006-0478-8)
- Sirois S, Wei DQ, Du QS, Chou KC (2004) Virtual screening for SARS-CoV protease based on KZ7088 pharmacophore points. *J Chem Inf Comput Sci* 44:1111–1122. doi:[10.1021/ci034270n](https://doi.org/10.1021/ci034270n)
- 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. doi:[10.1038/nature03342](https://doi.org/10.1038/nature03342)
- Susko E, Roger AJ (2007) On reduced amino acid alphabets for phylogenetic inference. *Mol Biol Evol* 24:2139–2150. doi:[10.1093/molbev/msm144](https://doi.org/10.1093/molbev/msm144)
- 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. doi:[10.1186/1471-2105-9-201](https://doi.org/10.1186/1471-2105-9-201)
- Wang JF, Wei DQ, Chen C, Li Y, Chou KC (2008a) Molecular modeling of two CYP2C19 SNPs and its implications for personalized drug design. *Protein Pept Lett* 15:27–32. doi:[10.2174/092986608783330305](https://doi.org/10.2174/092986608783330305)
- Wang T, Yang J, Shen HB, Chou KC (2008b) Predicting membrane protein types by the LLDA algorithm. *Protein Pept Lett* 15:915–921. doi:[10.2174/092986608785849308](https://doi.org/10.2174/092986608785849308)
- Weathers EA, Paulaitis ME, Woolf TB, Hoh JH (2004) Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. *FEBS Lett* 576:348–352. doi:[10.1016/j.febslet.2004.09.036](https://doi.org/10.1016/j.febslet.2004.09.036)
- Winzeler EA (2008) Malaria research in the post-genomic era. *Nature* 455:751–756. doi:[10.1038/nature07361](https://doi.org/10.1038/nature07361)
- Xiao X, Wang P, Chou KC (2008) Predicting protein structural classes with pseudo amino acid composition: an approach using geometric moments of cellular automaton image. *J Theor Biol* 254:691–696. doi:[10.1016/j.jtbi.2008.06.016](https://doi.org/10.1016/j.jtbi.2008.06.016)
- Yang ZR, Chou KC (2008) Correlation of metabolic pathways with the primary structure in acetylated proteins. *Open Bioinform J* 2:90–96. doi:[10.2174/1875036200802010090](https://doi.org/10.2174/1875036200802010090)
- 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. doi:[10.1016/j.jtbi.2008.03.015](https://doi.org/10.1016/j.jtbi.2008.03.015)
- Zhang GY, Li HC, Fang BS (2008a) Predicting lipase types by improved Chou's pseudo-amino acid composition. *Protein Pept Lett* 15:1132–1137. doi:[10.2174/092986608786071184](https://doi.org/10.2174/092986608786071184)
- Zhang TL, Ding YS, Chou KC (2008b) Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern. *J Theor Biol* 250:186–193. doi:[10.1016/j.jtbi.2007.09.014](https://doi.org/10.1016/j.jtbi.2007.09.014)
- Zhou GP (1998) An intriguing controversy over protein structural class prediction. *J Protein Chem* 17:729–738. doi:[10.1023/A:1020713915365](https://doi.org/10.1023/A:1020713915365)
- Zhou GP, Assa-Munt N (2001) Some insights into protein structural class prediction. *Proteins* 44:57–59. doi:[10.1002/prot.1071](https://doi.org/10.1002/prot.1071)
- Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. *Proteins* 50:44–48. doi:[10.1002/prot.10251](https://doi.org/10.1002/prot.10251)
- 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. doi:[10.1016/j.jtbi.2007.06.001](https://doi.org/10.1016/j.jtbi.2007.06.001)