# Prediction of Enzyme Subfamily Class *via* Pseudo Amino Acid Composition by Incorporating the Conjoint Triad Feature

Yong-Cui Wang[1,2], Xiao-Bo Wang[1], Zhi-Xia Yang[3,*] and Nai-Yang Deng[1,*]

[1]*College of Science, China Agricultural University, Beijing, China, 100083;* [2]*Key Laboratory of Adaptation and Evolution of Plateau Biota, Northwest Institute of Plateau Biology, Chinese Academy of Science, Xining, China, 810001;* [3]*College of Mathematics and System Science, Xinjiang University, Urumuchi, China, 830046*

**Abstract:** Predicting enzyme subfamily class is an imbalance multi-class classification problem due to the fact that the number of proteins in each subfamily makes a great difference. In this paper, we focus on developing the computational methods specially designed for the imbalance multi-class classification problem to predict enzyme subfamily class. We compare two support vector machine (SVM)-based methods for the imbalance problem, AdaBoost algorithm with RBFSVM (SVM with RBF kernel) and SVM with arithmetic mean (AM) offset (AM-SVM) in enzyme subfamily classification. As input features for our predictive model, we use the conjoint triad feature (CTF). We validate two methods on an enzyme benchmark dataset, which contains six enzyme main families with a total of thirty-four subfamily classes, and those proteins have less than 40% sequence identity to any other in a same functional class. In predicting oxidoreductases subfamilies, AM-SVM obtains the over 0.92 Matthew's correlation coefficient (MCC) and over 93% accuracy, and in predicting lyases, isomerases and ligases subfamilies, it obtains over 0.73 MCC and over 82% accuracy. The improvement in the predictive performance suggests the AM-SVM might play a complementary role to the existing function annotation methods.

**Keywords:** Enzyme subfamily class prediction, conjoint triad feature, imbalance problem, support vector machine.

## 1. INTRODUCTION

Enzymes, as one of the largest and most important group of all proteins, participate in maintaining and regulation of the metabolic states of the cells. According to the definition of Enzyme Commission (EC) number [1], all enzymes can be classified into six main families: oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases. And each family can be further classified into a number of subfamily classes. For a newly sequenced protein, which enzyme family and subfamily should it belong to? This is the most important problem due to the fact that it is related to the function protein as well as its specificity and molecular mechanism. However, experimentally determining the enzyme family and subfamily is still time-consuming and costly, the computational methods then have become a viable alternative to experimental approaches.

Introduction of novel mathematical approaches and physical concepts into molecular biology, such as Mahalanobis distance [2,3], pseudo amino acid composition [4], complexity measure factor [5,6], graph and diagram analysis [7-12], cellular automaton [13-16], grey theory [17], geometric moments [18], surface diffusion-controlled reaction [19], and ensemble classifier [20], as well as a series of user-friendly web-servers summarized in Table **3** in [21], can significantly stimulate the development of biological and medical science. Here, we would like to introduce a novel mathematical approach for predicting the enzyme subfamily class.

The most common computational method is transferring an enzymatic annotation between two globally aligned protein sequences, but it has been reported to significantly drop under 40% sequence identity [22]. To remedy this, many machine learning-based methods were successfully used. Especially, as an excellent machine learning method, support vector machines (SVMs) motivated by statistical learning theory [23,24], have been provided state-of-the-art performance in particular in computational biology [25]. Furthermore, SVM-based machine learning algorithm was used in predicting protein subcellular location [26], membrane protein type [27,28], protein structural class [29], specificity of GalNAc-transferase [30], HIV protease cleavage sites in protein [31], beta-turn types [32], protein signal sequences and their cleavage sites [33], alpha-turn types [34], catalytic triads of serine hydrolases [35], B-cell epitope prediction [36], *et al*. Here, we would like to use SVM-based methods to predict enzyme subfamilies for proteins with low homologies to known enzymes.

One key problem for using SVM-based methods is to construct a number of features to represent a given protein sequence. In previous studies, the amino acid composition (AAC) representation has been widely utilized in many predicting problems [37-39], including enzyme family and subfamily class [40]. Owing to lack of the sequence order information, some modified versions of AAC, such as pseudo amino acid composition (Pse-AAC) [41] and amphiphilic pseudo-amino acid composition (Am-Pse-AAC) [42] have

*Address correspondence to these authors at the College of Science, China Agricultural University, Beijing, China, 100083; Tel: 86-13718691409; Fax 86-0991- 8588010; E-mail: dengnaiyang@cau.edu.cn and College of Mathematics and System Science, Xinjiang University, Urumuchi, China, 830046; Tel 86-13681517962; Fax 86-10- 62561963;
E-mail: xjyangzhx@sina.com

**0929-8665/10 $55.00+.00** 

been developed. However, both Pse-AAC and Am-Pse-AAC have some parameters to be determined, and need the properties of physio-chemistry of amino acids. Recently, a much simple feature describing method for protein-protein interaction (PPI) prediction has been proposed [43]. The authors have shown that SVM with the conjoint triad feature (CTF) outperformed other sequence-based PPI prediction methods. The CTF considers not only properties of one amino acid but also its vicinal amino acids and treats any three continuous amino acids as an unit. That is, it contains not only the composition of amino acids but also sequence-order effect. It has also successfully been used in prediction of DNA- and RNA-binding proteins [44]. Inspired by these, in this paper, we introduce the CTF into our predictive model.

Only used sequence-based feature may not enough to generate the encouraging predictive results. One way to improve the performance is to develop automated classifiers specially designed for the problems to be solved. And it should be point out that the prediction of enzyme subfamily class is an imbalance multi-class classification problem due to the fact that the number of proteins in each subfamily makes a great difference. Unfortunately, when faced with imbalanced problem, the performance of SVM drops significantly [45]. So it is essential to use modified versions of SVM specially designed for imbalance classification problem to predict enzyme subfamily class. In this paper, we compare two modified versions of SVM, AdaBoost algorithm with RBFSVM (SVM with RBF kernel) and SVM with arithmetic mean (AM) offset (AM-SVM) in enzyme subfamily classification. AdaBoost algorithm, which can produce an accurate predictive rule by combining rough and moderately inaccurate rules-of-thumb [46], has lots of advantages such as high efficiency, high detection rate and low false positive rate [47]. Furthermore, it has been proofed that AdaBoost with RBFSVM component classifier, called as AdaBoostSVM, has better performance than standard SVM on imbalance classification problem [48]. While AM-SVM is SVM with arithmetic mean (AM) offset. It has been proofed that by introducing an offset parameter, the decision boundary can be modified on imbalance classification problem [49,50]. And a simple way to obtain the offset is to calculate the AM of support vectors (SVs)' decision values [49].

AdaBoostSVM and AM-SVM are validated on an enzyme benchmark dataset covering six enzyme main families with a total of thirty-four subfamily classes and any two proteins in a same functional class have less than 40% sequence identity. The improvement in predictive performance suggests that AdaBoostSVM and AM-SVM are all better than the Standard SVM and moreover, AM-SVM has much well performance. These results show that our SVM-based method based only on the sequence information performs well in predicting members of the enzyme subfamily, which is an extremely imbalance multi-class classification problem. We therefore hope our SVM-based method will be a useful tool for some other imbalance computational biological problem, including determination of subcellular location, prediction of membrane protein types and so on.

The paper is structured as follows. We begin by encoding the proteins by the CTF. Then we introduce the benchmark dataset and give the description of AdaboostSVM and AM-SVM in Materials and Methods section. In Results section, we compare our method with standard SVM on the benchmark dataset. Lastly, the discussions and conclusions are presented.

## 2. MATERIALS AND METHODS

### 2.1. Input Feature:The Conjoint Triad Feature (CTF)

Construction of feature vectors for each data dominates the learning capability of the SVM-based methods. Since the CTF considers not only the composition of amino acids but also sequence-order effect, so this efficient and simple encoding scheme is under our consideration here.

Now let us address to construct the CTF. Based on the dipoles and volumes of the side chains, the 20 amino acids can be classified into seven classes: {A, G, V }, {I, L, F, P}, {Y, M, T, S}, {H, N, Q, W}, {R, K}, {D, E}, {C}. Thus, a $(7\times7\times7=)$343-dimension vector is used to represent a given protein, where each element of this vector is the frequency of the corresponding conjoint triad appearing in the protein sequence. The detail description of the CTF can be see in [32].

### 2.2. SVM for Binary Classification Problem

At first, we briefly introduce the SVM for binary classification problem.

Given training examples $(x_i, y_i)$ for $i = 1, \ldots, l$, where $x_i$ is a vector in the input space $R^n$ and $y_i$ denotes the corresponding class label taking a value of +1 or -1. Let $\phi : R^n \to \mathrm{H}$ be a mapping from the input space to a Hilbert space $\mathrm{H}$. The SVM is to find a hyperplane $(w \cdot \phi(x)) + b = 0$ which can separate the two classes with the maximal margin and minimal training errors in the Hilbert space. By applying kernel function to replace the inner product in $\mathrm{H}$, the corresponding decision function is

$$\mathrm{sgn}(g(x)) = \mathrm{sgn}(\sum_{i=1}^{l} \alpha_i^* y_i K(x_i, x) + b^*),$$

where $\alpha^*$ is the solution of the following optimization problem

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{j=1}^{l} \alpha_j,$$

$$\text{s.t.} \quad \sum_{i=1}^{l} y_i \alpha_i = 0, 0 \leqslant \alpha_i \leqslant C,\ i = 1, \cdots, l,$$

and $b^*$ can be obtained as follows:

If there exists $\alpha_j^* \in (0, C), j = 1, \ldots, l$, then

$$b^* = y_j - \sum_{i=1}^{l} y_i \alpha_i^* K(x_i, x_j).$$

We define the decision value of $x$ as the value of $g(x)$.

Notice that, enzyme subfamily classification problem is a multi-class classification problem. For convenience, the

multi-class classification problem is represented as follows: Given the training set

$$T = \{(x_1, y_1), \cdots, (x_l, y_l)\} \in (R^n \times \{1, \cdots, m\})^l, \quad (1)$$

where m is the number of classes. According to the idea of one versus one SVM, we can construct $(m\text{-}1)m/2$ two-class training subsets based on training set (1), then $(m\text{-}1)m/2$ binary classifiers are constructed by implementing the binary SVM. Specially, for every $(i, j) \in \{(i, j) | i < j, i, j = 1, \cdots, m\}$, a training subset $T^{i\text{-}j}$, which contains the class $i$ and class $j$ examples, is constructed. By implementing the algorithm of binary SVM on $T^{i\text{-}j}$, the sub-classifier $f^{i-j}(x)$ is calculated as follows:

$$f^{i-j}(x) = \begin{cases} i, & g^{i-j}(x) > 0; \\ j, & else, \end{cases}$$

where $g^{i-j}(x)$ is the decision value of $x$. At last, we can get the decision function $f(x) = \{f^{i-j} | i < j, i, j = 1, \cdots, m\}$, for the training set (1).

By using the major vote rule, the predictive label of $x$ can be obtained.

### 2.3. AdaBoostSVM

For dealing with imbalance problem, we introduce a modified version of AdaBoost.M1 (a kind of AdaBoost algorithm for multi-class classification problem) [51], called as AdaBoostSVM, to predict enzyme subfamily class. The AdaBoostSVM generates a set of component classifiers, and combines the component classifiers into a single prediction rule. The pseudo code of AdaBoostSVM is described as follows:

Algorithm 2.1(AdaBoostSVM)

Input:

1. Training set (1);

2. Integer $K$ specifying number of iteration.

Initialize the weight: $u_i^0 = \frac{1}{l}$ for all $i$, set $k = 0$.

While $k \leq K$ :

- Ranking the weight vector of $u^k$ in descending order.

- Choosing the top $N$ samples according to the order of weight vector.

  If the selected samples are the same as the last ones, then break, return $k$, else,

  set $k = k + 1$;

- Implement one versus one RBFSVM on the chosen samples and output the decision function $f^k$;

- Calculate the training error rate $e^k$, set $v^k = \frac{1-e^k}{e^k}$;

- Assign the $k$th decision function to a weight of $p^k = \frac{1}{2} \ln v^k$;

- Update the weight vector as follows:

$$u_i^{k+1} = u_i^k + v^k I(f^k(x_i) \neq y_i),$$

  where $I(\cdot)$ is the indicate function.

end While.

Output the final classification rule: set $q = \arg_k \max(p^k)$, $H_{final}(x) = f^q(x)$;

### 2.4. SVM Classifier with AM Offset

In this subsection, we introduce another modified version of SVM to modify the decision boundary by introducing an offset parameter $\delta$. Specially, the offset $\delta$ is calculated by the AM of SVs' decision value [38]. The mathematical expression of $\delta$ is as follows: $\delta = \frac{\sum_{i=i}^{s} S_i}{s}$, where $S_1, \cdots S_s$ are decision values of SVs, $s$ is the number of SVs.

After introducing the offset parameter $\delta$, the decision value of a input $x$ is reformulated as follows: $g'(x) = \sum_{i=1}^{s} \alpha_i^* y_i K(x_i, x) + b^* - \delta.$

The decision function then becomes:

$$\mathrm{sgn}(g'(x)) = \mathrm{sgn}(\sum_{i=1}^{l} \alpha_i^* y_i K(x_i, x) + b^* - \delta).$$

By using one versus one idea, the predictive label of the input protein can be obtained. We denote this method as AM-SVM.

### 2.5. Dataset

The benchmark dataset used to validate the performance of our method is collected from the literature [52]. The sequences in this dataset have less than 40% sequence identity to any other in a same functional class. The detail information of this dataset can be found in [52]. In addition, for avoiding the extreme subfamily bias, those subfamilies which contain less than 40 proteins are excluded in our validation. Finally there are six main functional classes and thirty-four subfamily classes (i.e., twelve for oxidoreductases, seven for transferases, five for hydrolases, four for lyases, four for isomerases and two for ligases) in the benchmark dataset.

### 2.6. Parameters Selection and Experimental Protocol

The performance of SVM heavily depends on the combination of several parameters. Specially, both AdaboostSVM and AM-SVM involve two classes parameters: the trade-off parameter C and RBF kernel function parameter γ should be appropriately chosen, as C controls the trade-off between maximizing the margin and minimizing the training error, and γ dominates the generalization ability of SVM by regulating the amplitude of the RBF kernel function. We optimize them by using a grid search. To minimize the overfitting of the prediction models, 3-fold cross-validation is performed on the training dataset. The cross accuracy is used to select the parameters. And for Standard SVM, the optimal C and γ are 100 and 0.25 respectively, for AdaBoostSVM, the optimal C and γ are 10 and 0.125 respectively and for AM-SVM, the optimal C and γ are 10 and 0.25 respectively. With respect to the Adaboost algorithm parameters K and N, we fix them as ten and 0.8 times of current training samples respectively.

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 [53], the jackknife test was deemed the most objective that can always yield a unique result for a given benchmark dataset, as demonstrated by Eq.50 of [54]. Therefore, the jackknife test has been increasingly and widely adopted by investigators to test the power of various prediction methods (see, e.g. [55-72]). However, to reduce the computational time, we adopted the 10-fold cross-validation in this study as done by many investigators with SVM as the prediction engine. That is, each main functional class dataset is split into 10 subsets of roughly equal size, each subset is then taken in turn as a test set, and we train Standard SVM, AdaBoostSVM and AM-SVM on the remaining nine sets. Note that, we treat each main functional class independently and construct the predictor only on the one of them.

## 3. RESULTS

In this section, we evaluate the performance of our methods for prediction of enzyme subfamily classes. The experiments are implemented by Libsvm(version 2.88) [73].

### 3.1. Comparison with Alternative Methods

Here, Standard SVM (one versus one binary SVM) is introduced for comparison. Standard SVM has been illustrated in method section.

We plot the distributions of accuracy with respect to the subfamily classes for each method on the six main functional classes respectively in Fig. (**1**). Except for the Ec2.8 and Ec3.1 sub-classes, AM-SVM obtains the best accuracies, AdaBoostSVM comes close second, and Standard SVM becomes the last in identifying all sub-classes of all six main families. That is because that, both AM-SVM and AdaBoostSVM take the imbalance property into account, perform better than Standard SVM. These result suggests that the more properties of dataset itself are incorporated into the predictive model, the more better results can be expected.

From Fig. (**1**) we can see that, except for the Ec2.8 subclass, AM-SVM outperforms AdaBoostSVM in identifying all sub-classes of all six main families. That is, although both AdaboostSVM and AM-SVM are specially designed for the imbalance problem, AM-SVM seems more reliable. In addition, comparing with standard SVM, the accuracy from AdaBoostSVM is not increased dramatically as that from AM-SVM. The reason for the different results will be discussed here. In [37], the authors have shown that by designing parameter adjusting strategies, AdaBoost algorithm with RBFSVM component demonstrates better generalization performance than Standard SVM on imbalanced classification problems. That is, the parameter $\gamma$ has been adjusted in each AdaBoost algorithm iteration. However, to reduce the computational time, we fix the parameter $\gamma$ in each iteration, so the accuracy from AdaBoostSVM is not increased dramatically as the expected.

The CTF was initially developed for predicting protein interactions, although it achieved better outcome than AAC for PPI prediction, its effect in protein class prediction will be established by comparison in predicting enzyme subfamily classes. In Table **1**, we list the accuracies from AM-SVM with CTF and AAC respectively in predicting subfamily classes of oxidoreductases. We denote AM-SVM with CTF and AAC as $AM-SVM_{CTF}$ and $AM-SVM_{AAC}$ respectively. From Table **1**, we can see that, except for Ec1.3 and Ec1.11, the CTF achieves better results than AAC in predicting subfamily classes of oxidoreductases. So we believe that the CTF is also efficient in enzyme subfamily classs prediction.

### 3.2. Other Evaluation Criterion

Moreover, the Matthew's correlation coefficient (MCC) (Matthews, 1975) is used to evaluation the performance of predictive methods. MCC allows us to overcome the shortcoming of accuracy on imbalanced data [74]. For example, if the number of the positive samples are much larger than that of the negative samples, a classifier is easy to predict all samples as positive. Significantly it is not a good classifier because it predicts all negative samples incorrectly. In this case, the accuracy and MCC of the positive class are 100% and 0, respectively. Therefore, MCC is a better measure for imbalanced data classification.

Set $M \in R^m \times R^m$ as the confusion matrix of the prediction result, where $M_{ij}$ $(1 \le i, j \le m)$ represents the number of proteins that actually belong to class *i* but are predicted as class *j*. We further set

$$p_k = M_{kk}, q_k = \sum_{i=1,i\neq k}^{m} \sum_{j=1,j\neq k}^{m} M_{ij},$$

$$r_k = \sum_{i=1,i\neq k}^{m} M_{ik}, s_k = \sum_{j=1,j\neq k}^{m} M_{kj},$$

where $k(k = 1, \cdots m)$ is the index of a particular class, $m$ is the number of classes ($m$ = 12, 7, 5, 4, 4, 2 for oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases respectively). For class $k$, if the samples which belong to the class $k$ are treated as the positive samples, while the other samples are treated as the negative ones, then $p_k$ represents the number of true positive samples, $q_k$ represents the number of true negative samples, $r_k$ represents the number of false positive samples, while $s_k$ represents the number of false negative samples. Based on the equations above, the MCC of class $k(MCC_k)$ is

$$MCC_k = \frac{p_k q_k - r_k s_k}{\sqrt{(p_k + s_k)(p_k + r_k)(q_k + s_k)(q_k + r_k)}}$$

The distributions of MCCs with respect to the subfamily classes for each method on the six main functional classes are drawn respectively in Fig. (**2**). From Fig. (**2**) we can see that, expect for Ec3.4 and Ec4.6 sub-classes, AM-SVM obtains the best MCCs, AdaBoostSVM comes close second,

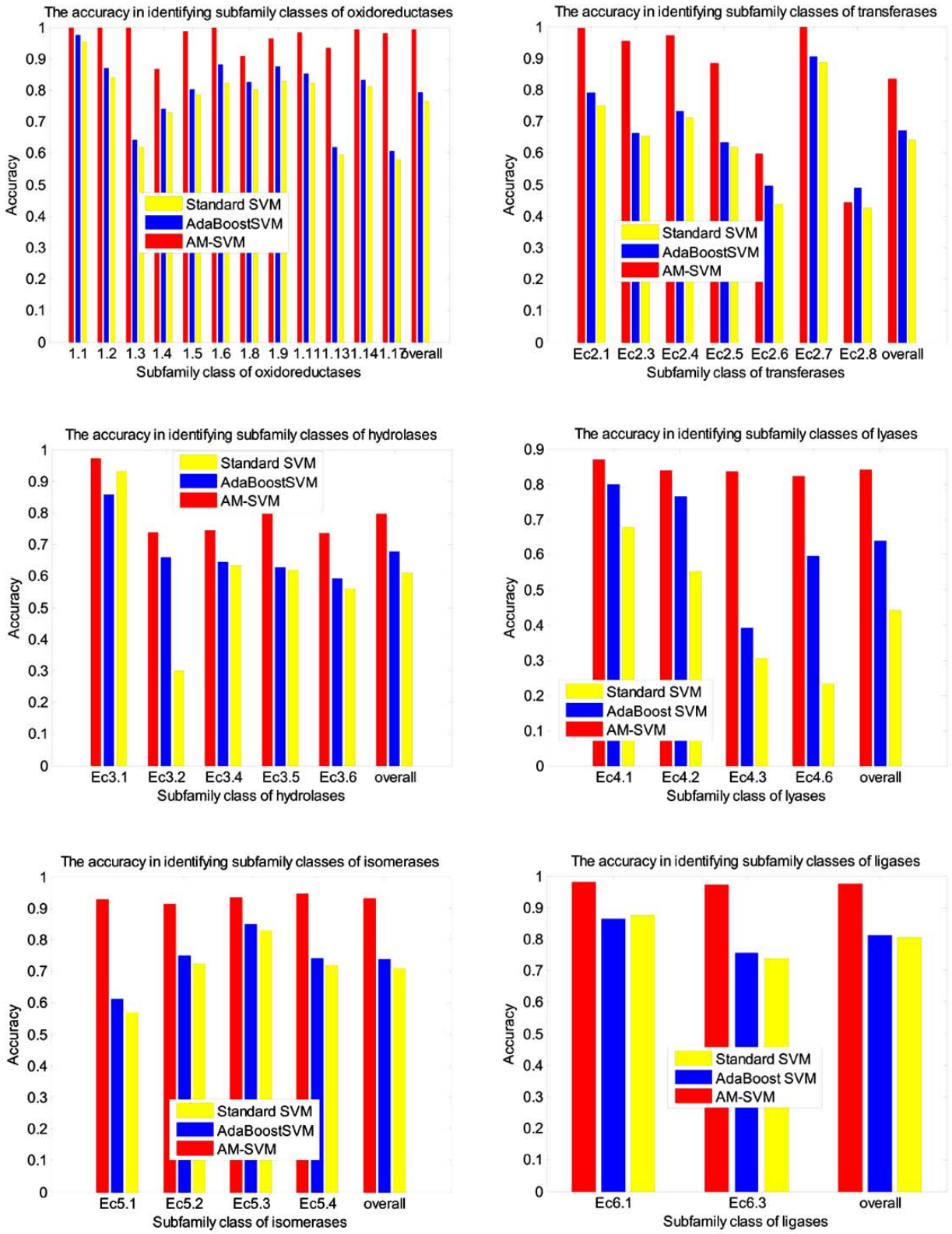


**Figure 1.** The distributions of accuracy with respect to the subfamily classes for each method on oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases respectively.

and Standard SVM becomes the third in identifying all sub-classes of oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases. Fig. (**2**) also shows that although both AdaboostSVM and AM-SVM are specially designed for the imbalance problem, AM-SVM seems more reliable. AM-SVM obtains over 92% MCC, and the best MCC can be

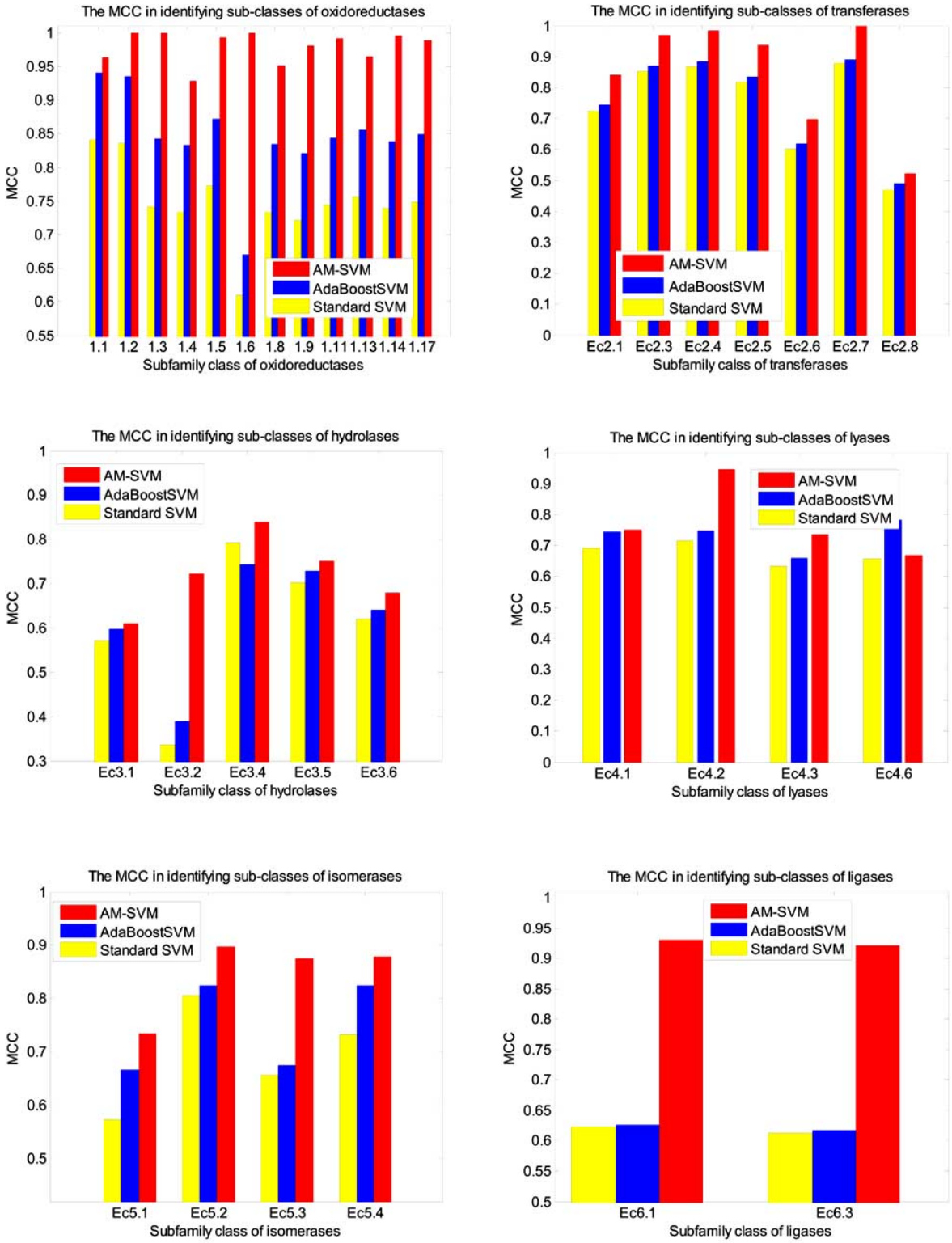


**Figure 2.** The distributions of MCC with respect to the subfamily classes for each method on oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases respectively.

close to 1 in predicting sub-classes of oxidoreductases , while AdaboostSVM obtain over 83% MCC.

In addition, we list the MCCs from AM-SVM with CTF and AAC respectively in predicting subfamily classes of oxidoreductases in Table **2**. As Table **1** shown, except for Ec1.3 and Ec1.11, the CTF achieves better results than AAC in predicting subfamily classes of oxidoreductases. So we have the reason to believe that the CTF is also efficient in enzyme subfamily classs prediction.

**Table 1. The Accuracy of Various Features for Oxidoreductases Subfamily Classes**

| Subfamily | $AM-SVM_{AAC}$ (%) | $AM-SVM_{CTF}$ (%) |
|---|---|---|
| Ec1.1 | 100 | 100 |
| Ec1.2 | 100 | 100 |
| Ec1.3 | 100 | 99.2 |
| Ec1.4 | 86.6 | 97.2 |
| Ec1.5 | 98.8 | 100 |
| Ec1.6 | 100 | 100 |
| Ec1.8 | 90.9 | 100 |
| Ec1.9 | 96.3 | 96.3 |
| Ec1.11 | 98.3 | 96.6 |
| Ec1.13 | 93.3 | 100 |
| Ec1.14 | 99.3 | 100 |
| Ec1.17 | 97.6 | 100 |
| Overall | 96.8 | 99.1 |

**Table 2. The MCC of the Various Features for Oxidoreductases Subfamily Classes**

| Subfamily | $AM-SVM_{AAC}$ | $AM-SVM_{CTF}$ |
|---|---|---|
| Ec1.1 | 0.96 | 0.98 |
| Ec1.2 | 1 | 1 |
| Ec1.3 | 1 | 0.99 |
| Ec1.4 | 0.93 | 0.98 |
| Ec1.5 | 0.94 | 1 |
| Ec1.6 | 1 | 1 |
| Ec1.8 | 0.96 | 1 |
| Ec1.9 | 0.98 | 0.98 |
| Ec1.11 | 0.99 | 0.98 |
| Ec1.13 | 0.97 | 1 |
| Ec1.14 | 0.99 | 1 |
| Ec1.17 | 0.99 | 1 |

## 4. DISCUSSION AND CONCLUSION

In this paper, for predicting enzyme subfamilies for proteins with low homologies to known enzymes, the sequence-base feature and the SVM-based model specially designed for the imbalance classification problem are introduced. Two modified version of SVM: AdaBoostSVM and AM-SVM are validated on a benchmark dataset proposed in [52]. This benchmark dataset covers six enzyme main family classes with a total of thirty-four subfamily classes and any two proteins in a same subfamily class have less than 40% identity. As a result, comparing with the Standard SVM, the accuracy from AdaBoostSVM without parameter adjusting strategies, is not increased dramatically as that from AM-SVM. These results imply that, although AM-SVM combines only sequence-based feature: the CTF, the promising results can be expected due to it considering the properties of problem to be solved. Furthermore, AM-SVM may be suitable to be a tool for some other imbalance classification biological problem, including determination of subcellular location, prediction of membrane protein types and so on.

Efficient feature construction is important in determining the performance of a predictive method. The results in this paper suggest that, comparing with the AAC, the CTF display its promising prospects (Table **1**, **2**). According to a recent comprehensive review [75], the CTF belongs to a different mode of pseudo amino acid composition (PseAAC). That is, it contains not only the composition of amino acids but also sequence-order effect. Thus future work can focus on introducing some encoding features which consider the sequence-order information, including K-spaced amino acid pairs [76] and so on. Another way to improve the feature construction methods is to integrate more data sources into encoding features (such as functional domain and evolution information) and use efficient kernel methods to fuse different information [77]. In addition, we can define a different similarity measure for each data source and thereby incorporate more prior information into the design of the classifier [78].

In this paper, we just compare two automated classifiers on imbalance classification problem: AdaBoostSVM and AM-SVM. There are many modified versions of SVM suitable on imbalance problem, such as Multisurface proximal support vector machine classification [79], and its extended versions: Twin SVM [80] and Nonparallel plane proximal classifier (NPPC) [81]. In the future, we can apply these models to facilitate the predictive task, and compare the performance of them on predicting enzyme family and subfamily class.

Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful predictors [32], we shall make efforts in our future work to provide a web-server for the method presented in this paper.

## ACKNOWLEDGEMENTS

This work is supported by the Key Project of the National Natural Science Foundation of China (No.10631070), the National Natural Science Foundation of China (No. 10801131, No. 10801112, No.10971223) and the Ph.D Graduate Start Research Foundation of Xinjiang University Funded Project (No.BS080101).

## REFERENCES

[1] Webb, E.C. *Enzyme Nomenclature*. Academic Press: San Diego, CA, 1992.

[2] Chou, K. C. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. *Proteins Struct. Func. Genet.*, **1995**, *21*, 319-344.

[3] Chou, K. C.; Zhang, C. T. Predicting protein folding types by distance functions that make allowances for amino acid interactions. *J. Biol. Chem.*, **1994**, *269*, 22014-22020.

[4] Chou, K. C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curr. Proteomics*, **2009**, *6*, 262-274.

[5] Xiao, X.; Shao, S.; Ding, Y.; Huang, Z.; Huang, Y.; Chou, K. C. Using complexity measure factor to predict protein subcellular location. *Amino Acids*, **2005**, *28*, 57-61.

[6] Xiao, X.; Shao, S. H.; Huang, Z. D.; Chou, K. C. Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. *J. Comput. Chem.*, **2006**, *27*, 478-482.

[7] Zhou, G. P.; Deng, M. H. An extension of Chou's graphical rules for deriving enzyme kinetic equations to system involving parallel reaction pathways. *Biochem. J.*, **1984**, *222*, 169-176.

[8] Myers, D.; Palmer, G. Microcomputer tools for steady-state enzyme kinetics. *Bioinformatics (original: Computer Applied Bioscience)*, **1985**, *1*, 105-110.

[9] Chou, K. C. Graphical rules in steady and non-steady enzyme kinetics. *J. Biol. Chem.*, **1989**, *264*, 12074-12079.

[10] Chou, K. C. Review: Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. *Biophys. Chem.*, **1990**, *35*, 1-24.

[11] Chou, K. C.; Kezdy, F. J.; Reusser, F. Review: Steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. *Anal. Biochem.*, **1994**, *221*, 217-230.

[12] Andraos, J. Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: new methods based on directed graphs. *Can. J. Chem.*, **2008**, *86*, 342-357.

[13] Xiao, X.; Shao, S.; Ding, Y.; Huang, Z.; Chen, X.; Chou, K. C. An application of gene comparative image for predicting the effect on replication ratio by HBV virus gene missense mutation. *J. Theor. Biol.*, **2005**, *235*, 555-565.

[14] Xiao, X.; Shao, S. H.; Chou, K. C. A probability cellular automaton model for hepatitis B viral infections. *Biochem. Biophys. Res. Commun.*, **2006**, *342*, 605-610.

[15] Xiao, X.; Chou, K. C. Digital coding of amino acids based on hydrophobic index. *Protein Pept. Lett.*, **2007**, *14*, 871-875.

[16] Xiao, X.; Wang, P.; Chou, K. C. GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes. *J. Comput. Chem.*, **2009**, *30*, 1414-1423.

[17] Xiao, X.; Lin, W. Z.; Chou, K. C. Using grey dynamic modeling and pseudo amino acid composition to predict protein structural classes. *J. Comput. Chem.*, **2008**, *29*, 2018-2024.

[18] Xiao, X.; Wang, P.; Chou, K. C. Predicting protein structural classes with pseudo amino acid composition: an approach using geometric moments of cellular automaton image. *J. Theor. Biol.*, **2008**, *254*, 691-696.

[19] Chou, K. C.; Zhou, G. P. Role of the protein outside active site on the diffusion-controlled reaction of enzyme. *J. Am. Chem. Soc.*, **1982**, *104*, 1409-1413.

[20] Chou, K. C.; Shen, H. B. Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J. Proteome Res.*, **2007**, *6*, 1728-1734.

[21] Chou, K. C.; Shen, H. B. Review: recent advances in developing web-servers for predicting protein attributes. *Natural Science*, **2009**, *2*, 63-92 (openly accessible at http://www.scirp.org/journal/NS/).

[22] Tian, W.; Skolnick, J. How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.*, **2003**, 333, 863-882.

[23] Vapnik, V. The Nature of Statistical Learning Theory. Springer: USA, 1995.

[24] Vapnik, V. *Statistical Learning Theory*. Wiley: USA, 1998.

[25] Sch¨okopf, B.; Tsuda, K.; Vert, J.P. *Kernel Methods in Computational Biology*. MIT Press: Cambridge, MA, **2004**, 71-92.

[26] Chou, K. C.; Cai, Y. D. Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.*, **2002**, *277*, 45765-45769.

[27] Cai, Y. D.; Zhou, G. P.; Chou, K. C. Support vector machines for predicting membrane protein types by using functional domain composition. *Biophys. J.*, **2003**, *84*, 3257-3263.

[28] Cai, Y. D.; Pong-Wong, R.; Feng, K.; Jen, J. C. H.; Chou, K. C. Application of SVM to predict membrane protein types. *J. Theor. Biol.*, **2004**, *226*, 373-376.

[29] Cai, Y. D.; Liu, X. J.; Xu, X. B.; Chou, K. C. Prediction of protein structural classes by support vector machines. *Comput. Chem.*, **2002**, *26*, 293-296.

[30] Cai, Y. D.; Liu, X. J.; Xu, X. B.; Chou, K. C. Support vector machines for predicting the specificity of GalNAc-transferase. *Peptides*, **2002**, *23*, 205-208.

[31] Cai, Y. D.; Liu, X. J.; Xu, X. B.; Chou, K. C. Support Vector Machines for predicting HIV protease cleavage sites in protein. *J. Comput. Chem.*, **2002**, *23*, 267-274.

[32] Cai, Y. D.; Liu, X. J.; Xu, X. B.; Chou, K. C. Support vector machines for the classification and prediction of beta-turn types. *J. Pept. Sci.*, **2002**, *8*, 297-301.

[33] Cai, Y. D.; Lin, S.; Chou, K. C. Support vector machines for prediction of protein signal sequences and their cleavage sites. *Peptides*, **2003**, *24*, 159-161.

[34] Cai, Y. D.; Feng, K. Y.; Li, Y. X.; Chou, K. C. Support vector machine for predicting alpha-turn types. *Peptides*, **2003**, *24*, 629-630.

[35] Cai, Y. D.; Zhou, G. P.; Jen, C. H.; Lin, S. L.; Chou, K. C. Identify catalytic triads of serine hydrolases by support vector machines. *J. Theor. Biol.*, **2004**, *228*, 551-557.

[36] Chen, J.; Liu, H.; Yang, J.; Chou, K. C. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids*, **2007**, *33*, 423-428.

[37] Hua, S.; Sun, Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **2001**, *17*, 721-728.

[38] Lei, Z.; Dai, Y. An SVM-based system for predicting protein subnuclear localizations. *BMC Bioinformatics*, **2005**, 6,291-298.

[39] Zhou, G.P.; Assa-Munt, N. Some insights into protein structural class prediction. *Proteins*, **2001**, *44*, 57-59.

[40] Chou, K.C.; Elrod, D.W. Prediction of enzyme family classes. *J. Proteome Res.*, **2003**, *2*, 183-190.

[41] Chou, K.C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins Struct. Funct. Genet.*, **2001**, *43*, 246-255.

[42] Chou, K.C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, **2005**, *21*, 10-19.

[43] Shen, J.W.; Zhang, J.; Luo, X. M.; Zhu, W.L.; Yu, K.Q.; Chen, K.X.; Li, Y.X.; Jiang, H.L. Predicting protein-protein interactions based only on sequences information. *Proc. Nat. Acad. Sci.*, **2007**, *104*, 4337-4341.

[44] Shao, X.J.; Tian, Y.J.; Wu, L.Y.; Wang, Y.; Deng, N.Y. Predicting DNA- and RNA-binding proteins from sequences with kernel methods. *J. Theor. Biol.*, **2009**, *258*, 289-293.

[45] Wu, G.; Chang, E. *Class-Boundary Alignment for Imbalanced Dataset Learning*. In ICML 2003 Workshop on Learning from Imbalanced Data Sets II, Washington, DC.

[46] Schapire, R.E. A brief int roduction to boosting. In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, **1999**, *14*(14)*1401*-1406.

[47] Kearns, M.; Valiant, L.G. Cryptographic limitations on learning boolean formulae and finite automata. *J. ACM*, **1994**, *41*, 67-95.

[48] Li, X.C.; Wang, L.; Sung, E. AdaBoost with SVM-based component classifiers. *App. Artif. Intell.*, **2008**, *21*, 785-795.

[49] Li, B.; Hu, J.; Hirasawa, K.; Sun, P.; Marko, K. Support vector machine with fuzzy decision-making for real-world data classification. In IEEE World Congress on Computational Intelligence, Int. Joint Conf. on Neural Networks, Canada, 2006.

[50] Li, B.; Hu, J.L.; Hirasawa, K. Support vector machine classifier with WHM offset for unbalanced data. *J. Adv. Comput. Intell. Intell. Inform.*, **2008**, *12*, 94-101.

[51] Freund, Y.; Schapire, R.E. Experiments with a new boosting algorithm. In Proceedings of the Thirteenth International Conference on Machine Learning, **1996**, 148-156, Morgan Kaufmann.

[52] Shen, H.B.; Chou, K.C. EzyPred: A top-down approach for predicting enzyme functional classes and subclasses. *Biochem. Biophys. Res. Commun.*, **2007**, *364* , 53-59.

[53] Chou, K. C.; Zhang, C. T. Review: Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.*, **1995**, *30*, 275-349.

[54] Chou, K. C.; Shen, H. B. Review: Recent progresses in protein subcellular location prediction. *Anal. Biochem.*, **2007**, *370*, 1-16.

[55] Zhou, G. P. An intriguing controversy over protein structural class prediction. *J. Protein Chem.*, **1998**, *17*, 729-738.

[56] Zhou, X. B.; Chen, C.; Li, Z. C.; Zou, X. Y. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J. Theor. Biol*., **2007**, *248*, 546-551.

[57] Lin, H. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. *J. Theor. Biol*., **2008**, *252*, 350-356.

[58] Jiang, X.; Wei, R.; Zhang, T. L.; Gu, Q. Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. *Protein Pept. Lett.*, **2008**, *15*, 392-396.

[59] Li, F. M.; Li, Q. Z. Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. *Protein Pept. Lett.*, **2008**, *15*, 612-616.

[60] Lin, H.; Ding, H.; Feng-Biao Guo, F. B.; Zhang, A. Y.; Huang, J. Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. *Protein Pept. Lett.*, **2008**, *15*, 739-744.

[61] Ding, Y. S.; Zhang, T. L. 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*., **2008**, *29*, 1887-1892.

[62] Chen, C.; Chen, L.; Zou, X.; Cai, P. Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. *Protein Pept. Lett.*, **2009**, *16*, 27-31.

[63] Ding, H.; Luo, L.; Lin, H. Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. *Protein Pept. Lett.*, **2009**, *16*, 351-355.

[64] Zeng, Y. H.; Guo, Y. Z.; Xiao, R. Q.; Yang, L.; Yu, L. Z.; Li, M. L. Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. *J. Theor. Biol.*, **2009**, *259*, 366-372.

[65] Qiu, J. D.; Huang, J. H.; Liang, R. P.; Lu, X. Q. Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform. *Anal. Biochem.*, **2009**, *390*, 68-73.

[66] Shen, H. B.; Song, J. N.; Chou, K. C. Prediction of protein folding rates from primary sequence by fusing multiple sequential features. *J.Biomed. Sci. Eng. (JBiSE)*, **2009**, *2*, 136-143 (openly accessible at http://www.srpublishing.org/journal/jbise/).

[67] Chou, K. C.; Shen, H. B. FoldRate: A web-server for predicting protein folding rates from primary sequence. *Open Bioinform. J*., **2009**, *3*, 31-50 (openly accessible at http://www.bentham.org/open/tobioij/).

[68] Chou, K. C.; Shen, H. B. Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J. Proteome Res.*, **2007**, *6*, 1728-1734.

[69] Chou, K. C.; Shen, H. B. ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information. *Biochem. Biophys. Res. Commun.*, **2008**, *376*, 321-325.

[70] Esmaeili, M.; Mohabatkar, H.; Mohsenzadeh, S. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *J. Theor. Biol*., **2010**, *263*, 203-209.

[71] He, Z. S.; Zhang, J.; Shi, X. H.; Hu, L. L.; Kong, X. G.; Cai, Y. D.; Chou, K. C. Predicting drug-target interaction networks based on functional groups and biological features. *PLoS ONE*, **2010**, *5*, e9603.

[72] Chou, K. C.; Shen, H. B. A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0 *PLoS ONE*, 2010, *5*, e9931; openly accessible at http://www.plosone.org/article/info%9933Adoi%9932F9910.1371%9932 Fjournal.pone.0009931.

[73] Hsu, Chih-wei.; Hsu, C.W., Chang, C.C.; Lin, C.J. A practical guide to support vector classfication, 2007. Available from: *http://www.csie.ntu.edu.tw/ cjlin*.

[74] Pu, X.; Guo, J.; Leunga, H.; Lin, Y.L. Prediction of membrane protein types from sequences and position-specific scoring matrices. *J. Theor.l Biol*., **2007**, *247*, 259-265.

[75] Chou, K. C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curr. Proteomics*, **2009**, *6*, 262-274.

[76] Wang, X.B.; Wu, L.Y.; Wang, Y.C.; Deng, N.Y. Prediction of palmitoylation sites using the composition of k-spaced amino acid pairs. *Protein Eng. Des. Select.*, **2009**, *22*(11), 707-712.

[77] Ong, C.S.; Smola, A.J.; Williamson, R.C. *Hyperkernels*. Becker, S., Thrun, S., Obermayer, K. Advances in neural information processing systems 15, MIT Press: Cambridge, MA, **2003**, 495-502.

[78] Guan, Y.; Myers, C.; Hess, D.; Barutcuoglu, Z.; Caudy, A.; Troyanskaya, O. Predicting gene function in a hierarchical context with an ensemble of classifiers, *Genome Biol*., **2008**, *9*(S3).

[79] Mangasarian, O.L.; Wild, E.W. Multisurface proximal support vector machine classification via generalized eigenvalues. *IEEE Trans. Pattern Anal.Machine Intell.*, **2006**, *28*, 69-74.

[80] Jayadeva Khemchandani, R.; Chandra, S. Twin support vector-machines for Pattern classification. *IEEE Trans.Pattern Anal. Machine Intell*., **2007**, *29*, 905-910.

[81] Ghorai, S.; Mukherjee, A.; Dutta, P. K. Nonparallel plane proximal classifier. *Signal Process.*, **2008**, *89*, 510-522.

**Received: March 28, 2010** **Revised: June 17, 2010** **Accepted: July 08, 2010**