Available online at www.sciencedirect.com
ScienceDirect
BioSystems 90 (2007) 405–413
BioSystems
www.elsevier.com/locate/biosystems

Accurate prediction of enzyme subfamily class using an adaptive fuzzy *k*-nearest neighbor method

Wen-Lin Huang [a], Hung-Ming Chen [a], Shiow-Fen Hwang [a], Shinn-Ying Ho [b,c,*]

[a] *Institute of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan*
[b] *Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, Taiwan*
[c] *Institute of Bioinformatics, National Chiao Tung University, Hsinchu, Taiwan*

Received 15 June 2006; received in revised form 15 October 2006; accepted 22 October 2006

Abstract

Amphiphilic pseudo-amino acid composition (Am-Pse-AAC) with extra sequence-order information is a useful feature for representing enzymes. This study first utilizes the *k*-nearest neighbor (*k*-NN) rule to analyze the distribution of enzymes in the Am-Pse-AAC feature space. This analysis indicates the distributions of multiple classes of enzymes are highly overlapped. To cope with the overlap problem, this study proposes an efficient non-parametric classifier for predicting enzyme subfamily class using an adaptive fuzzy *r*-nearest neighbor (AFK-NN) method, where *k* and a fuzzy strength parameter *m* are adaptively specified. The fuzzy membership values of a query sample *Q* are dynamically determined according to the position of *Q* and its weighted distances to the *k* nearest neighbors. Using the same enzymes of the oxidoreductases family for comparisons, the prediction accuracy of AFK-NN is 76.6%, which is better than those of Support Vector Machine (73.6%), the decision tree method C5.0 (75.4%) and the existing covariant-discriminate algorithm (70.6%) using a jackknife test. To evaluate the generalization ability of AFK-NN, the datasets for all six families of entirely sequenced enzymes are established from the newly updated SWISS-PROT and ENZYME database. The accuracy of AFK-NN on the new large-scale dataset of oxidoreductases family is 83.3%, and the mean accuracy of the six families is 92.1%.

Keywords: Amino acid composition; Enzyme subfamily class prediction; Fuzzy theory; *k*-Nearest neighbor; Support vector machine

1. Introduction

Enzymes can be classified into six families according to specific molecular functions and acting objects (Webb, 1992): oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases. Each family can be further classified into a number of subfamilies. Table 1 displays the numbers of subfamilies and functions of each family, as given in the ENZYME database (Release 38).

For a novel enzyme sequence, determining its family or subfamily class is an important task, because it gives direct evidence of its specific molecular functions and the objects on which they act (Webb, 1992). Such a task is usually performed with using biochemical analysis of either eukaryotic and prokaryotic genomes, or microarray chips (Chou and Elrod, 2003). These experimental methods are both time-consuming and expensive. With the explosion of protein entries in databanks, understanding the functions of many enzymes from large-scale sequencing projects is of priority concern. Thus,

* Corresponding author at: 75 Bo-Ai Street, Hsinchu, Taiwan, R.O.C. Tel.: +886 3 5712121x56905; fax: +886 3 5729288.
E-mail address: syho@mail.nctu.edu.tw (S.-Y. Ho).

0303-2647/$ – see front matter © 2006 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2006.10.004

Table 1
The number of subfamilies and its function for each family obtained from Release 38.0 (September 2005) of the ENZYME database and SWISS-PROT databank (Version 48.5, November 2005, http://tw.expasy.org/enzyme)

Enzyme family	No. of Subfamily	No. of samples	Function
Oxidoreductases	20	10,184	Roughly take responsible for catalyzing oxidoreduction reactions
Transferases	9	30,947	Transferring a group from one compound to another
Hydrolases	9	45,271	Responsible for catalyzing the hydrolysis of various bonds
Lyases	6	51,054	Cleaving C–C, C–O, C–N and other bonds by means other than hydrolysis or oxidation
Isomerases	5	54,487	Catalyzing geometrical or structural changes within one molecule
Ligases	6	60,682	Catalyzing the joining together of two molecules coupled with the hydrolysis of a pyrophosphate bond in ATP or a similar triphosphate

accurately predicting the enzyme family or subfamily class from its amino acid sequence is highly desired.

Most previous research about sequence analysis has focused on extracting a number of effective features and developing accurate classifiers from these effective features to distinguish the sequences from different class instances (Nakai and Kanehisa, 1992; Hua and Sun, 2001; Cai et al., 2002; Huang and Li, 2004; Chou, 2005). Previous studies generally used two major categories of feature representations, sequence sorting signals (Nielsen et al., 1999; Nakai, 2000) and amino acids composition (AAC). The AAC representation of a given enzyme sequence is denoted by a 20-dimensional vector which consists of occurrence frequencies of the amino acids.

The AAC representation has recently been widely utilized in predicting protein structural classes (Bahar et al., 1997; Zhou and Assa-Munt, 2001), subcellular localizations (Cedano et al., 1997; Nakai, 2000; Hua and Sun, 2001), subnuclear localizations (Lei and Dai, 2005), and enzyme family or subfamily class (Chou and Elrod, 2003; Chou, 2005). Owing to the lack of sequence order information in the conventional AAC feature, some improved versions of AAC such as pseudo-amino acid composition (Pse-AAC, Chou and Elrod, 2003) and amphiphilic pseudo-amino acid composition (Am-Pse-AAC, Chou, 2005) have been developed.

With regard to establishing efficient classifiers for prediction problems of biological and medical data, some professional classifiers were proposed such as Support Vector Machine (SVM, Cortes and Vapnik, 1995; Hua and Sun, 2001; Cai et al., 2002; Lei and Dai, 2005), fuzzy *k*-nearest neighbor (*k*-NN, Keller et al., 1985; Bezdek et al., 1993; Leszczynski et al., 1999; Huang and Li, 2004), neural network (Chandonia and Karplus, 1995; Cai et al., 2000), C5.0 decision tree (Quinlan, 2003), and covariant-discriminate algorithm (CDA, Chou, 2005).

Chou (2005) used CDA with the Am-Pse-AAC feature to predict the enzyme subfamily class, and achieved a prediction accuracy of 70.6% using enzymes of the oxidoreductases family. To design a more accurate classifier, the distribution of enzymes in the Am-Pse-AAC feature space using the *k*-NN rule were analyzed in this study (Cover and Hart, 1967). Analysis results show that the *k* nearest neighbors of a sample often belong to several enzyme subfamily classes, revealing that the distributions of multiple classes of enzymes are highly overlapped.

After investigating the abilities of three state-of-the-art classifiers based on *k*-nearest neighbor, SVM, and C5.0, in coping with the overlap problem, this study propose an adaptive fuzzy *k*-nearest neighbor (AFK-NN) classifier using the Am-Pse-AAC feature for predicting the enzyme subfamily class, where *k* and a fuzzy strength parameter *m* are adaptively specified. The fuzzy membership value of a query sample Q is dynamically determined according to the position of Q and its weighted distances to the *k*-nearest neighbors.

Using the same dataset of oxidoreductases family with 16 subfamilies and 2640 enzymes for comparisons, the prediction accuracy of AFK-NN was found to be 76.6% using a jackknife test, which is better than 66.5% for a standard *k*-NN classifier, 75.4% for C5.0 and 73.6% for SVM. The three proposed classifiers AFK-NN, C5.0, and SVM are all better than CDA (Chou, 2005) with 70.6%. This result indicates that AFK-NN performs well in predicting members of the oxidoreductases family, which has a large number 16 of subfamilies with a high overlap distribution.

Table 2
The 16 subfamily classes of oxidoreductases family with 2640 enzymes, obtained from Chou and Elrod (2003)

Subfamily class	Groups acted by the enzyme	No. of samples (Chou and Elrod, 2003)
1	CH–OH group	314
2	Aldehyd/oxo group	216
3	CH–CH group	194
4	CH–NH_2 group	130
5	CH–NH group	112
6	NADH/NADPH	305
7	Other nitrogenous compounds	64
8	Sulfur group	59
9	Heme group	254
10	Diphenols and related substances	94
11	Peroxide	154
12	Single donors	94
13	Paired donors	257
14	Superoxide radicals	155
15	–CH_2 group	84
16	Reduced ferredoxin	154
Total		2640

All six families of entirely sequenced enzymes derived from the ENZYME database (Version 38, Bairoch, 2000) and SWISS-PROT (Version 48.5, Bairoch and Apweiler, 2000) were further tested to evaluate the generalization ability of AFK-NN. The accuracy of AFK-NN on the new large-scale dataset of oxidoreductases family increased to 83.3%, and the mean accuracy of the six families was as high as 92.1%, which is also slightly better than 91.2% for C5.0 and 91.7% for SVM.

2. Materials and methods

2.1. Datasets

For comparison the dataset used in previous investigations (Chou and Elrod, 2003; Chou, 2005) were used in this study. The dataset of the oxidoreductases family has 2640 enzymes belonging to 16 subfamilies, where each subfamily acts on a different target, as shown in Table 2. The sequence lengths of all enzymes are larger than 20.

To evaluate the generalization ability of the proposed AFK-NN, six datasets were established from all six enzyme families from Release 38.0 of the ENZYME database (Bairoch, 2000), where an enzyme commission (EC) number and a primary accession number were assigned to each enzyme. According to the accession numbers, the protein sequences of enzymes were obtained from SWISS-PROT (Version 48.5, Bairoch and Apweiler, 2000). The subfamilies were selected through the following screening procedure (Chou and Elrod, 2003): (1) remove the enzymes having more than one EC number, (2) delete the sequences identical to any of the others in the dataset, and (3) remove the enzymes with the sequence length not larger than 20. For each enzyme family, the numbers of subfamilies and samples of each family are listed in Table 1. Notably, the new dataset of the oxidoreductases family has 20 subfamilies and 10184 enzymes.

2.2. Amphiphilic pseudo-amino acid composition

The amino acid composition (AAC) of a protein sequence is a 20-dimensional vector, reflecting the normalized occurrence frequencies p_i of the 20 native amino acids. A protein is given with a sequence of L amino acids $R_1R_2R_3\cdots R_L$, where R_i represents the amino acid at chain position i, $1 \le i \le L$. The AAC feature of a protein can be expressed as a vector $\mathbf{P}_{\mathrm{AAC}}$ in a 20-dimensional space:

$$\mathbf{P}_{\mathrm{AAC}} = [p_1, \ldots, p_{20}]^{\mathrm{t}}. \tag{1}$$

The hydrophobic and hydrophilic values of proteins play a crucial role in protein folding and interaction with its environment, which are involved in the Am-Pse-AAC feature (Chou, 2005). The Am-Pse-AAC feature of a protein is expressed as a vector P, which consists of $20+2\lambda$ components. The first 20 components are the AAC features and the next 2λ ones are a set of correlation factors that reveal the physicochemical properties hydrophobicity and hydrophilicity along a protein sequence. The vector $\mathbf{P}$ of the Am-Pse-AAC feature is represented as follows:

$$\mathbf{P} = [p_1, \ldots, p_{20}, p_{20+1}, \ldots, p_{20+\lambda}, p_{20+\lambda+1}, \ldots, p_{20+2\lambda}]^{\mathrm{t}}, \tag{2}$$

$$p_u = \frac{\omega\pi_{u-20}}{\sum_{i=1}^{20} p_i + \omega\sum_{j=1}^{2\lambda}\pi_j}, \quad 21 \le u \le 20+2\lambda, \tag{3}$$

where ω is a weight factor, and π_j is the jth-tier sequence-correlation factor calculated based on the following equation:

$$\pi_1 = \frac{1}{L-1}\sum_{i=1}^{L-1} H_{i,i+1}^1, \qquad \pi_2 = \frac{1}{L-1}\sum_{i=1}^{L-1} H_{i,i+1}^2,$$
$$\pi_3 = \frac{1}{L-2}\sum_{i=1}^{L-2} H_{i,i+2}^1, \qquad \pi_4 = \frac{1}{L-2}\sum_{i=1}^{L-2} H_{i,i+2}^2, \ldots,$$
$$\pi_{2\lambda-1} = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} H_{i,i+\lambda}^1, \quad \pi_{2\lambda} = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} H_{i,i+\lambda}^2, \quad \lambda < L, \tag{4}$$

where $H_{i,j}^1 = h^1(R_i)h^1(R_j)$ and $H_{i,j}^2 = h^2(R_i)h^2(R_j)$. The terms π_1 and π_2 are called the first-tier sequence-correlation factors between all the most contiguous amino acids along a protein chain with hydrophobic and hydrophilic attributes, respectively, and $\pi_{2\lambda-1}$ and $\pi_{2\lambda}$ are the corresponding λth-tier sequence-correlation factors between all the λ contiguous amino acids. $h^1(R_i)$ and $h^2(R_i)$ are the corresponding hydrophobic and hydrophilic values for the ith amino acid in the protein,

subject to a standard conversion computed by

$$h^1(R_i) = \frac{h_0^1(R_i) - \Gamma_1}{\sqrt{\sum_{u=1}^{20}[h_0^1(\Re_u) - \Gamma_1]^2/20}}, \quad \Gamma_1 = \sum_{u=1}^{20} h_0^1(\Re_u)/20,$$

$$h^2(R_i) = \frac{h_0^2(R_i) - \Gamma_2}{\sqrt{\sum_{u=1}^{20}[h_0^2(\Re_t) - \Gamma_2]^2/20}}, \quad \Gamma_2 = \sum_{t=1}^{20} h_0^1(\Re_t)/20, \tag{5}$$

where $\Re_t$ ($t = 1, \ldots, 20$) are the 20 native amino acids based on the alphabetical order of their single-letter codes, A, C, D–I, K–N, P–T, V, W and Y. And $h_0^1(\Re)$ and $h_0^2(\Re)$ are the original hydrophobic and hydrophilic values of the amino acid $\Re$ (Tanford, 1962; Hopp and Woods, 1981).

2.3. Analysis of sample distribution

The distribution of samples should be analyzed before designing an accurate classifier. The settings $\lambda = 9$ and $\omega = 0.5$ were used to obtain the best prediction accuracy of CDA for multiple-class prediction problems, as recommended by Chou (2005). Therefore, the sample distribution of all $N = 2640$ enzymes of the oxidoreductases family having $C = 16$ subfamilies in the 38-dimensional feature space was analyzed using a k-NN rule.

Each of the $N = 2640$ enzymes is considered as a query sample in turn. The number δ of categories to which the k nearest neighbors of a query sample belong are first counted, where $1 \le \delta \le \min(C, k)$. Let N_δ be the number of query samples, where their k nearest neighbors belong to δ categories. Let the sample ratio $\varphi_\delta = N_\delta N$ where $\sum \varphi_\delta = 1$. Fig. 1(a) indicates the statistical result of the sample ratio with $k = 10$. The figure shows that there are only 26.90% ($=\varphi_1$) of samples are surrounded by $k = 10$ nearest neighbors belonging to one category. However, the $k = 10$ nearest neighbors belong to more than three classes in 53.64% ($=\sum \varphi_i$, $4 \le i \le 10$) of the samples. This result reveals that the distributions of $C = 16$ classes in the Am-Pse-AAC space are highly overlapped. The overlap problem must be concerned when designing a classifier for prediction.

The number n of enzymes belonging to the same subfamily of the query sample from the $k = 10$ nearest neighbors are also counted where $0 \le n \le k$. Let C_n be the number of samples which has n of the k nearest neighbors belonging to the same subfamily of the query sample where $C_0 + C_1 + \cdots + C_k = N$. Let the sample ratio $\wp_n = C_n/N$. Fig. 1(b) illustrates the statistic result of $\wp_n$. The case $\wp_0 = \wp_1 = 0$ indicates that each query sample can always find at least two of $k = 10$ nearest neighbors belonging to the same subfamily. Therefore, an adaptive fuzzy k-nearest neighbor (AFK-NN) method is proposed while considering the distribution property.

Fig. 2 shows a typical query sample O83491 belonging to the second subfamily (class 2), which is adopted as an example to illustrate the high-overlap distribution. For visualization, the two features p_1 and p_2 having the best significant discrimination

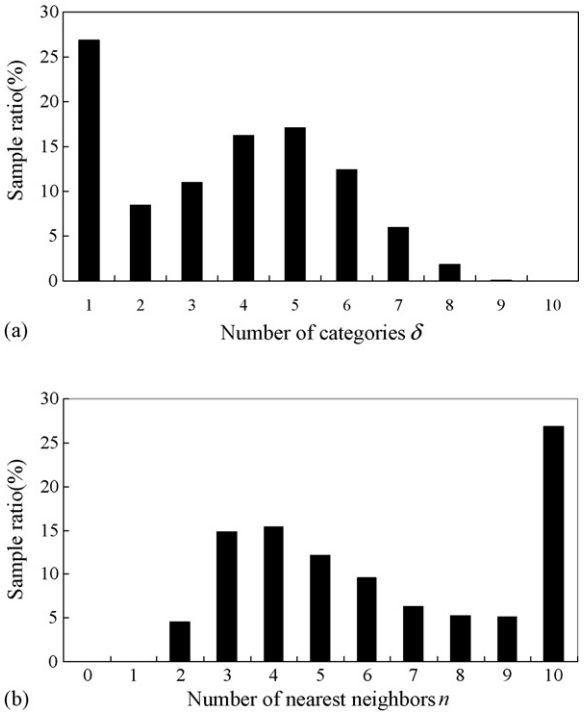

Fig. 1. Statistic results of oxidoreductases family in the Am-Pse-AAC space with $\lambda = 9$ using a k-NN rule with $k = 10$ (a) the sample ratio φ_δ and (b) the sample ratio $\wp_n$.

capability are selected from the top two ranks of rank sum test (Snedecor and Cochran, 1989). Fig. 2(a) shows that the nearest neighbors of O83491 belong to many classes. Fig. 2(b) indicates that the query sample is surrounded by 10 nearest neighbors belonging to five subfamily classes, namely P17445 of class 1, P33327 of class 2, P0807 of class 3, P40875, O79677, O84970, O85274, O99826, and P00390 of class 4, and P43083 of class 13.

2.4. Proposed AFK-NN

The proposed AFK-NN classifier assigns fuzzy membership values $r_c(\mathbf{P})$ of a query sample $\mathbf{P}$ to each class c as follows

$$r_c(\mathbf{P}) = \frac{\sum_{j=1}^{k} r_c(\mathbf{P}^j)(||\mathbf{P} - \mathbf{P}^j||^{-2/(m-1)})}{\sum_{j=1}^{k} ||\mathbf{P} - \mathbf{P}^j||^{-2/(m-1)}}, \quad c = 1, 2, \ldots, C. \tag{6}$$

In the above equation, a fuzzy strength parameter m is used to determine the weighting of the distance when calculating the contribution of each of the k nearest neighbors to the membership value, and $||\mathbf{P} - \mathbf{P}^j||$ is an Euclidean distance between $\mathbf{P}$ and one of its nearest neighbors $\mathbf{P}^j$. Various definitions of $r_c(\mathbf{P}^j)$ can be chosen depending on the applications. In this study, let $r_c(\mathbf{P}^j) = 1$ if $\mathbf{P}^j$ belongs to class c; otherwise, $r_c(\mathbf{P}^j) = 0$. After calculating the membership values of the query sample, $\mathbf{P}$ is categorized into the class having the highest membership

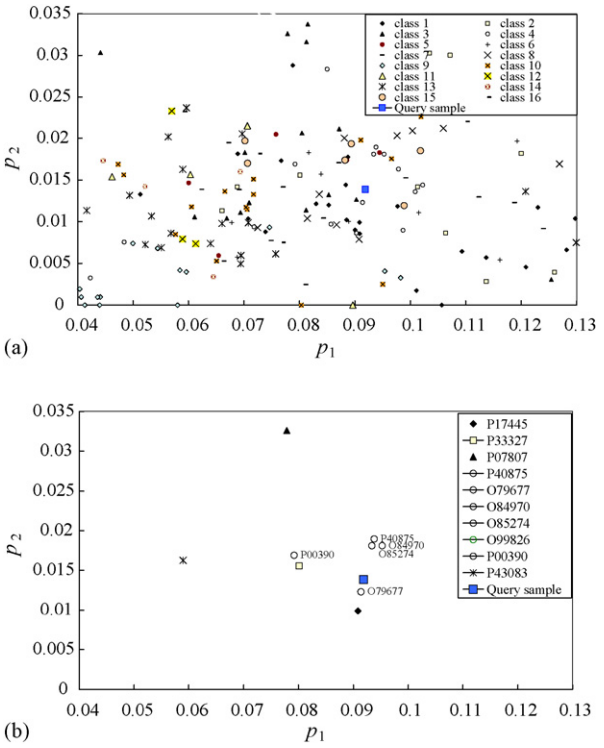

Fig. 2. A typical example for illustrating high-overlap distribution of enzymes in the Am-Pse-AAC feature space where p_1 and p_2 are two most informative features. (a) The nearest neighbors of the query sample O83491 belong to a large number of categories. (b) The $k = 10$ nearest neighbors of O83491 belong to five categories where the two square symbols belong to the second subfamily.

value. The best values of both parameters m and k for AFK-NN are determined based on the training dataset of an individual enzyme family and the used feature set.

Consider Fig. 2(b) as an example to illustrate merit of AFK-NN. If a standard k-NN classifier is applied, then the query sample O83491 of class 2 would be classified to class 4 because six of the $k = 10$ nearest neighbors belong to class 4. However, AFK-NN would correctly classify the query sample **P** into class 2 because P33327 of class 2 has the smallest distance to P such that the largest membership value is $r_c(\mathbf{P})$ with $c = 2$.

3. Results

3.1. Comparison with CDA

For comparison with the CDA classifier (Chou, 2005), this study used the dataset of the oxidoreductases family (Table 2) and the Am-Pse-AAC feature with $\lambda = 9$ and $\omega = 0.5$, as used in Chou's work, since it has the best prediction accuracy of CDA. The general settings $k = 10$ and $m = 1.05$ of the fuzzy k-NN classifier used by (Huang and Li, 2004) were adopted to evaluate AFK-NN. The prediction accuracy of AFK-NN was 74.88%, which is better than 70.61% for CDA using a jackknife test (Chou, 2005). Fig. 3 gives the prediction accuracies of AFK-NN using various values of the parameter λ of the Am-Pse-AAC feature, where $\lambda < L$ and $L > 20$ for this used dataset. The highest accuracy of AFK-NN was 76.29% with $\lambda = 20$, which is higher than 74.88% for $\lambda = 9$. AFK-NN can effectively apply the extra information by using a larger value of λ.

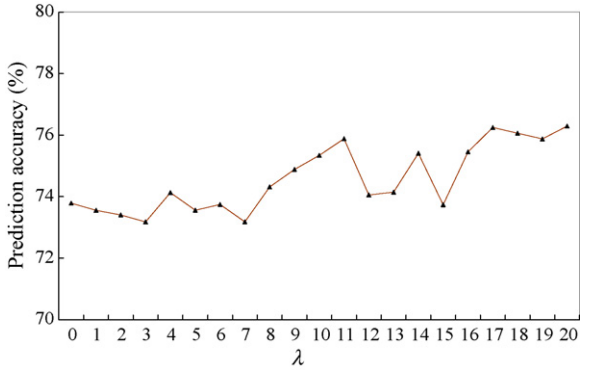

Fig. 3. Prediction accuracies of AFK-NN using various values of the parameter λ of the Am-Pse-AAC feature on the 2640 enzymes of oxidoreductases family.

To investigate the best values of the combination of k and λ, the features with three typical values of λ were evaluated: AAC (the case of $\lambda = 0$), and Am-Pse-AAC with $\lambda = 9$ and 20. Fig. 4 depicts the prediction accuracies for the three features, with $1 \le k \le 30$ and $m = 1.05$. The prediction accuracies are not changed significantly at $k \ge 15$. The performance of Am-Pse-AAC with $\lambda = 20$ was significantly batter than that of the other two features. The best prediction accuracy was 76.41% with $m = 1.05$, $k = 19$ and $\lambda = 20$. Further examining the accuracy by greedily tuning the $m = 1.1, 1.15, \ldots, 1.65,$

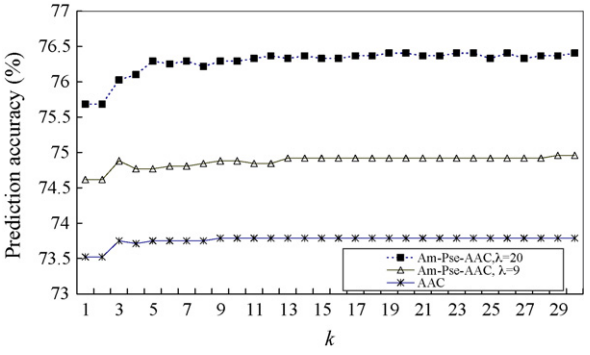

Fig. 4. Prediction accuracies of AFK-NN using k-NN rule and three typical features of amino acid composition.

Table 3
Prediction accuracies of enzyme subfamily class on the dataset of oxidoreductases family shown in Table 2 using a jackknife test

Classifier	AAC (%)	Am-Pse-AAC (λ = 9) (%)	Am-Pse-AAC (λ = 20) (%)
CDA (Chou, 2005)	63.64	70.61	NA
AFK-NN[a]	74.40	74.88	76.63
AFK-NN ($m = 1.05$, $k = 10$)	73.80	74.88	76.29
k-NN ($k = 10$)	64.15	66.46	64.42
C5.0 with boosting	73.90	75.40	75.40
SVM	71.78	73.64	73.22

NA: not available, CDA obtained the best accuracy using λ = 9 (Chou, 2005).
[a] (m, k) = (1.15, 9), (1.05, 10), and (1.1, 19) for AAC, Am-Pse-AAC with λ = 9 and 20, respectively.

Table 4
Prediction rates distribution of C5.0 decision tree with boosting algorithm for two compositions: performed with classical AAC and Am-Pse-AAC

BT	AAC (CF = 50) (%)	Am-Pse-AAC (λ = 9, CF = 30) (%)	Am-Pse-AAC (λ = 20, CF = 40) (%)
0	52.80	53.60	51.60
50	71.60	72.60	73.00
100	72.50	74.90	74.40
150	72.50	74.20	74.40
200	73.90	74.60	74.40
250	73.10	74.80	75.40
300	73.00	75.40	74.70
350	73.10	75.30	74.40
400	73.30	75.00	75.20

yielded a slightly improved accuracy of 76.63% with $m = 1.1$.

3.2. Comparison with other classifiers

Two efficient methods involving C5.0 decision tree (Quinlan, 2003) and SVM (Hua and Sun, 2001; Cai et al., 2002; Joachims, 1999) were further investigated by using the dataset of the oxidoreductases family, as listed in Table 2. A standard k-NN classifier is also applied for revealing the effect of fuzziness. Table 3 summarizes the results of these compared methods using the same jackknife test. Accuracies of C5.0 and SVM were obtained from the performance of proper settings of control parameters. Tables 4 and 5, respectively, present detailed results of C5.0 and SVM.

The conventionally adopted decision tree method C5.0 is based on a non-parametric type of regression fitting approach, which is suitable for an unknown data distribution. Another advantage is that it effectively manages large datasets and the issues of high dimensionality. One approach to avoiding overfitting in decision tree learning is tree pruning. The parameter CF of confidence level used to prune the decision tree affects both tree size and accuracy, which can be properly tuned to avoid overfitting. The adaptive boost-

Table 5
Selected prediction accuracies of SVM using a radial basis kernel function with proper settings of a kernel parameter γ

AAC		Am-Pse-AAC (λ = 9)		Am-Pse-AAC (λ = 20)	
γ	Accuracy (%)	γ	Accuracy (%)	γ	Accuracy (%)
240	71.02	50	72.88	20	69.92
250	71.78	60	73.64	30	71.93
260	71.44	70	73.56	40	73.22
270	71.63	80	73.18	50	72.84
280	71.74	90	72.56	60	71.85
290	71.78	100	72.69	70	70.72
300	71.59	110	72.31	80	69.84

Table 6
Prediction accuracies (%) of AFK-NN obtained from the best settings of parameters m and k where $m = 1.05, 1.1, \ldots, 1.65$ and $k = 10, \ldots, 30$

Enzyme family	Accuracy (m, k)		
	AAC	Am-Pse-AAC ($\lambda = 9$)	Am-Pse-AAC ($\lambda = 20$)
Oxidoreductases	80.53 (1.2, 30)	82.13 (1.1, 28)	83.34 (1.1, 23)
Transferases	79.94 (1.2, 17)	81.09 (1.15, 27)	82.64 (1.05, 15)
Hydrolases	90.34 (1.05, 10)	91.90 (1.1, 30)	92.28 (1.1, 15)
Lyases	97.30 (1.2, 28)	97.56 (1.1, 28)	97.87 (1.2, 23)
Isomerases	97.97 (1.2, 16)	98.03 (1.1, 26)	98.30 (1.1, 23)
Ligases	97.22 (1.2, 29)	97.58 (1.15, 28)	97.90 (1.1, 28)
Mean	90.55 (1.2, 22)	91.38 (1.1, 28)	92.09 (1.1, 21)

ing algorithm (Freund and Schapire, 1997) improves the classification process by generating a number of classifiers from training data. Due to exploitation of groups of hypotheses with independent errors, the main advantage of boosting is that it increases the overall accuracy of classification, and to reduce both variance and bias of the classification. The parameter BT of boosting trail controls the total number of classifiers. The proper values of CF and BT are problem-dependent. Through various settings of CF (=10, 20, . . ., 100) and BT (=0, 50, . . ., 400), Fig. 4 shows that the best settings for the features AAC and Am-Pse-AAC with $\lambda = 9$, 20 are CF = 50, 30, and 40, respectively. The worst results of BT = 0 reveal that boosting is effective. The highest accuracy was 75.40% for the Am-Pse-AAC features with $\lambda = 9$ and 20. The prediction accuracy using AAC was 73.9%, which was worse than that using Am-Pse-AAC.

SVM is a powerful machine learning method to handle classification, prediction, and regression problems. SVM maps original feature vectors into one either linearly or non-linearly higher dimensional feature space through a nonlinear transformation by using one of various kernel functions. Within the feature space, SVM seeks an optimal hyper-plane separating samples of two classes, called binary SVM. Multi-class classification can be performed simply by using a series of binary SVMs. The use of a large number of binary SVMs seems not to be effective enough for dealing with classification having a large number of classes (Huang and Li, 2004). This study utilized SVM^{light} (http://svmlight.joachims.org/old/svm_light_v4.00.html) using a radial basis kernel function $\exp(-\gamma||x^i - x^j||^2)$, where x^i and x^j are training samples and γ is a kernel parameter. The cost parameter C of SVM was set to a default value of 1.0. The values of $\gamma = 10, 20, \ldots, 500$ were evaluated to find the proper setting of γ. The best prediction accuracies were 73.64% and 73.22% using Am-Pse-AAC with $\lambda = 9$ and 20, respectively. The accuracy of SVM using AAC was 71.78%, worse than that using Am-Pse-AAC.

The following conclusions can be drawn from the results in Table 3:

(a) AFK-NN has the highest prediction accuracy 76.63% for the enzyme subfamily class using the Am-Pse-AAC feature with $\lambda = 20$. AFK-NN is better than the standard k-NN classifier with $k = 10$ and fuzzy k-NN classifier with fixed settings $m = 1.05$ and $k = 10$.
(b) AFK-NN can effectively use of extra information when the value of λ is increased from 9 to 20, while the other compared methods have performance of $\lambda = 9$ better than or equal to that of $\lambda = 20$.

Table 7
Prediction accuracies of C5.0 using the proper settings of CF and BT according to computer simulation

Enzyme family	AAC (CF = 50, BT = 200) (%)	Am-Pse-AAC ($\lambda = 9$, CF = 30, BT = 300) (%)	Am-Pse-AAC ($\lambda = 20$, CF = 40, BT = 250) (%)
Oxidoreductases	83.40	83.70	83.80
Transferases	81.00	81.90	81.40
Hydrolases	89.60	89.40	89.10
Lyases	96.40	97.31	97.00
Isomerases	96.46	97.43	96.90
Ligases	96.28	97.16	96.97
Mean	90.52	91.15	90.86

Table 8
Prediction accuracies of SVM using the proper settings of γ according to computer simulation

Enzyme family	AAC ($\gamma = 250$) (%)	Am-Pse-AAC ($\lambda = 9$, $\gamma = 60$) (%)	Am-Pse-AAC ($\lambda = 20$, $\gamma = 40$) (%)
Oxidoreductases	80.05	83.48	83.98
Transferases	79.40	81.86	83.74
Hydrolases	88.61	90.89	91.38
Lyases	95.82	97.11	97.70
Isomerases	97.46	97.93	98.28
Ligases	95.78	97.26	94.97
Mean	89.52	91.42	91.68

(c) The three presented methods AFK-NN, C5.0 and SVM are all better than the existing CDA method (Chou, 2005) using the same features.

3.3. Performance of all six families

To evaluate the generalization abilities of the above-mentioned three efficient methods AFK-NN, C5.0 and SVM, the large-scale datasets of all six enzyme families derived from the newly updated SWISS-PROT and ENZYME database (Table 1) were tested and the results using a jackknife test are given in Tables 6–8. From Table 6, the prediction accuracy 83.34% for Am-Pse-AAC with $\lambda = 20$ was obtained using the best settings (m, k) = (1.1, 23) of AFK-NN for the new oxidoreductases family, where the settings $k = 10, \ldots, 30$, and $m = 1.05$, $1.1, \ldots, 1.65$ are evaluated.

Table 6 shows that the mean accuracies performed by using the classical AAC and Am-Pse-AAC with $\lambda = 9$ were 90.55% and 91.38%, respectively. The accuracy of Am-Pse-AAC with $\lambda = 20$ was better than that with $\lambda = 9$ for each of the six families. The mean accuracy 92.09% for AFK-NN using Am-Pse-AAC with $\lambda = 20$ had the best prediction performance of enzyme subfamily class, better than 90.72% for AFK-NN using the fixed settings $m = 1.05$ and $k = 10$ for all subfamilies. The fixed settings of parameters $\lambda = 20$, $m = 1.1$ and $k = 21$ were recommended for fuzzy k-nearest neighbor classifiers using Am-Pse-AAC to predict all enzyme families, based on the simulation results.

Table 7 summaries the prediction accuracies of C5.0 using appropriate settings of CF and BT, obtained from computer simulation of Table 4. The mean accuracies of AAC and Am-Pse-ACC with $\lambda = 9$ and 20 were 90.52%, 91.15% and 90.86%, respectively. Because C5.0 utilized the most informative features from part of all features in **P**, the difference ($0.63\% = 91.15 - 90.52\%$) of mean accuracy between the best feature (Am-Pse-ACC with $\lambda = 9$) and the worst feature (AAC) is smaller than those of AFK-NN and SVM which use all features in **P**.

Table 8 shows the prediction accuracies of SVM using proper setting of $\gamma = 250$, 60, 40 from Table 5 for the three features AAC and Am-Pse-AAC with $\lambda = 9$ and 20, respectively. Five of six families performed better using Am-Pse-AAC with $\lambda = 20$. The best mean accuracy from Am-Pse-AAC was 91.68% obtained with $\lambda = 20$, and was slightly smaller than 92.09% for AFK-NN.ConclusionsFor predicting the enzyme subfamily class, the existing CDA method (Chou, 2005) using the Am-Pse-AAC feature with the best setting $\lambda = 9$ yielded an accuracy of 70.61% for the dataset of the oxidoreductases family. The oxidoreductases family has a distribution property in the Am-Pse-AAC feature space: a large number 16 of subfamilies, a fairly small number of enzymes for each subfamily and high-overlap distribution. This study presents three efficient classifiers the k-NN based classifier AFK-NN, decision tree based C5.0 with boosting and SVM. All of the classifiers AFK-NN, C5.0, and SVM have accuracies 74.88%, 75.40% and 73.64%, respectively, perform better than CDA (Chou, 2005) using the same feature and jackknife test. Due to this distribution property, AFK-NN using the feature Am-Pse-AAC with $\lambda = 20$ yields the best accuracy 76.63%.

Large-scale datasets of all six enzyme families, oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases, obtained from Release 38 of the ENZYME database and SWISS-PROT (Version 48.5), were established to perform a comprehensive experiment for evaluating the three efficient methods AFK-NN, C5.0 and SVM. The numbers of enzymes for six enzyme families are much larger than 2640 in the old dataset of the oxidoreductases family. Therefore, the overfitting problem can be effectively avoided to lower both variance and bias of classification using a jackknife test. Carefully setting the control parameters by effectively searching the feature space means that no method is significantly superior to the others. This study found that the mean performance of AFK-NN is slightly better than C5.0 and SVM, where the parameters k

and m of AFK-NN are adaptively specified for each family.

The proposed AFK-NN supports dynamic computation of membership values to predict the category to which the query samples belong to. Even though the distributions of enzyme families in a feature space are highly overlapped, AFK-NN performs well in predicting the enzyme subfamily class. The major concern of AFK-NN is its long computation time. An evolutionary design of optimal k-NN classifiers by minimizing the sizes of the reference and feature sets while maximizing accuracy can be utilized (Ho et al., 2002). The minimized reference set can significantly reduce the computation time of prediction. The future work for advancing AFK-NN will use an intelligent evolutionary algorithm (Ho et al., 2002; Ho et al., 2004) to simultaneously minimize the reference set and optimize the control parameters k and m of AFK-NN and λ of the Am-Pse-AAC feature.

Acknowledgement

The authors would like to thank the National Science Council of the Republic of China, Taiwan for financially supporting this research under Contract No. NSC 95-2627-B-009-002.

References

Bahar, I., Atilgan, A.R., Jernigan, R.L., Erman, B., 1997. Understanding the recognition of protein structural classes by amino acid composition. Proteins 29, 172–185.

Bairoch, A., 2000. The ENZYME Database in 2000. Nucl. Acids 28, 304–305.

Bairoch, A., Apweiler, R., 2000. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucl. Acids 25, 31–36.

Bezdek, J.C., Hall, L.O., Clarke, L.P., 1993. Review of MR image segmentation techniques using pattern recognition. Med. Phys. 20, 1033–1048.

Cai, Y.D., Li, Y.X., Chou, K.C., 2000. Using neural networks for prediction of domain structural classes. Biochim. Biophys. Acta 1476, 1–2.

Cai, Y.D., Liu, X.J., Xu, X.B., Chou, K.C., 2002. Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. J. Cell. Biochem. 84, 343–348.

Cedano, J., Aloy, P., P'erez-Pons, J.A., Querol, E., 1997. Relation between amino acid composition and cellular location of proteins. J. Mol. Biol. 266, 594–600.

Chandonia, J.M., Karplus, M., 1995. Neural networks for secondary structure and structural class prediction. Protein Sci. 4, 275–285.

Chou, K.C., 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21, 10–19.

Chou, K.C., Elrod, D.W., 2003. Prediction of enzyme family classes. J. Proteome Res. 2, 183–190.

Cortes, C., Vapnik, V., 1995. Support-vector network. Machine Learning 20, 273–297.

Cover, T.M., Hart, P.E., 1967. Nearest neighbor pattern classification. IEEE Trans. Inform. Theory 13, 21–27.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 119–139.

Ho, S.-Y., Liu, C.-C., Liu, S., 2002. Design of an optimal nearest neighbor classifier using an intelligent genetic algorithm. Pattern Recogn. Lett. 23, 1495–1503.

Ho, S.-Y., Shu, L.-S., Chen, J.-H., 2004. Intelligent evolutionary algorithms for large parameter optimization problems. IEEE Trans. Evol. Comput. 8, 522–541.

Hopp, T.P., Woods, K.R., 1981. Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. U.S.A. 78, 3824–3828.

Hua, S., Sun, Z., 2001. Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17, 721–728.

Huang, Y., Li, Y., 2004. Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics 20, 21–28.

Joachims, T., 1999. Making large-scale SVM learning practical. In: Schölkopf, B., Burges, C., Smola, A. (Eds.), Advances in Kernel Methods—Support Vector Learning. MIT-Press.

Keller, J.M., Gray, M.R., Givens, J.A., 1985. A fuzzy k-nearest neighbor algorithm. IEEE Trans. Syst. Man Cybern. 15, 580–585.

Lei, Z., Dai, Y., 2005. An SVM-based system for predicting protein subnuclear localizations. BMC Bioinform. 6, 291–298.

Leszczynski, K., Cosby, S., Bissett, R., Provost, D., Boyko, S., Loose, S., Mvilongo, E., 1999. Application of a fuzzy pattern classifier to decision making in portal verification of radiotherapy. Phys. Med. Biol. 44, 253–269.

Nakai, K., 2000. Protein sorting signals and prediction of subcellular localization. Adv. Protein Chem. 54, 277–344.

Nakai, K., Kanehisa, M., 1992. A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics 14, 897–911.

Nielsen, H., Brunak, S., von Heijne, G., 1999. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. Protein Eng. 12, 3–9.

Quinlan, J.R., 2003. C5.0 Online Tutorial. http://www.rulequest.com.

Snedecor, G.W., Cochran, W.G., 1989. Statistical Methods, 8th ed. Iowa State University Press, pp. 142–144.

Tanford, C., 1962. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. J. Am. Chem. Soc. 84, 4240–4274.

Webb, E.C., 1992. Enzyme Nomenclature. Academic Press, San Diego, CA.

Zhou, G.P., Assa-Munt, N., 2001. Some insights into protein structural class prediction. Proteins 44, 57–59.