BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 14 2006, pages 1717–1722
doi:10.1093/bioinformatics/btl170

Structural bioinformatics

Ensemble classifier for protein fold pattern recognition

Hong-Bin Shen[1,*] and Kuo-Chen Chou[1,2]

[1]Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai 200030, China and [2]Gordon Life Science Institute, San Diego, CA 92130, USA

Received on March 31, 2006; revised on April 26, 2006; accepted on April 27, 2006
Advance Access publication May 3, 2006
Associate Editor: Keith A Crandall

ABSTRACT

Motivation: Prediction of protein folding patterns is one level deeper than that of protein structural classes, and hence is much more complicated and difficult. To deal with such a challenging problem, the ensemble classifier was introduced. It was formed by a set of basic classifiers, with each trained in different parameter systems, such as predicted secondary structure, hydrophobicity, van der Waals volume, polarity, polarizability, as well as different dimensions of pseudo-amino acid composition, which were extracted from a training dataset. The operation engine for the constituent individual classifiers was OET-KNN (optimized evidence-theoretic *k*-nearest neighbors) rule. Their outcomes were combined through a weighted voting to give a final determination for classifying a query protein. The recognition was to find the true fold among the 27 possible patterns.

Results: The overall success rate thus obtained was 62% for a testing dataset where most of the proteins have <25% sequence identity with the proteins used in training the classifier. Such a rate is 6–21% higher than the corresponding rates obtained by various existing NN (neural networks) and SVM (support vector machines) approaches, implying that the ensemble classifier is very promising and might become a useful vehicle in protein science, as well as proteomics and bioinformatics.

Availability: The ensemble classifier, called PFP-Pred, is available as a web-server at http://202.120.37.186/bioinf/fold/PFP-Pred.htm for public usage.

Contact: lifesci-sjtu@san.rr.com

Supplementary information: Supplementary data are available on *Bioinformatics* online.

INTRODUCTION

The avalanche of protein sequences generated in the post-genomic era has challenged us for developing computational methods by which the structural information can be timely extracted from sequence databases. Although the direct prediction of the three-dimensional (3D) structure of a protein from its sequence based on the least free energy principle is scientifically quite sound and some encouraging results already obtained in elucidating the handedness problems and packing arrangements in proteins (see e.g. Chou and Carlacci, 1991; Chou *et al*., 1982, 1984, 1990), it is very difficult to predict its overall fold owing to the notorious local minimum problem. Also, although it is quite successful to predict the 3D structure of a protein according to the homology modeling approach (Chou, 2004; Holm and Sander, 1999), a hurdle exists when the query protein does not have any structure-known homologous protein in the existing databases. Facing this kind of situation, can we find a different approach to predict the fold of a protein? In this paper, we shall resort to the taxonomic approach, whose underpinning is based on the assumption that the number of protein folds is limited (Chou and Zhang, 1995; Dubchak *et al*., 1999; Finkelstein and Ptitsyn, 1987; Murzin *et al*., 1995). Accordingly, predicting the 3D structure of a protein may be first converted to a problem of classification, i.e. identifying which fold pattern it belongs to. The present study was initiated in an attempt to introduce a novel approach, the ensemble classifier, to recognize the fold pattern for a query protein.

*To whom correspondence should be addressed.

MATERIALS AND METHODS

The working (training and testing) datasets studied here were taken from Ding and Dubchak (2001). The original training dataset and testing dataset contain 313 proteins and 385 proteins, respectively. Of these proteins, however, two (i.e. 2SCMC and 2GPS) in the training dataset and two (2YHX_1 and 2YHX_2) in the testing dataset do not have sequence records. These four proteins were excluded for further consideration due to lacking sequence information. Accordingly, we have 311 proteins for training dataset and 383 proteins for testing dataset. The names of the training and testing proteins and their sequences are given in Online Supplementary Materials AI and AII, respectively. None of proteins in the testing dataset has >35% sequence identity to those in the training dataset (Ding and Dubchak, 2001). According to the SCOP database (Andreeva *et al*., 2004; Murzin *et al*., 1995), the proteins in the training and testing datasets (Online Supplementary Materials A) were further classified into the following 27-fold types (Ding and Dubchak, 2001; Dubchak *et al*., 1995, 1999): (1) globin-like, (2) cytochrome c, (3) DNA-binding 3-helical bundle, (4) 4-helical up-and-down bundle, (5) 4-helical cytokines, (6) EF-hand, (7) immunoglobulin-like, (8) cupredoxins, (9) viral coat and capsid proteins, (10) conA-like lectin/glucanases, (11) SH3-like barrel, (12) OB-fold, (13) beta-trefoil, (14) trypsin-like serine proteases, (15) lipocalins, (16) (TIM)-barrel, (17) FAD (also NAD)-binding motif, (18) flavodoxin-like, (19) NAD(P)-binding Rossmann-fold, (20) P-loop, (21) thioredoxin-like, (22) ribonuclease H-like motif, (23) hydrolases, (24) periplasmic binding protein-like, (25) β-grasp, (26) ferredoxin-like and (27) small inhibitors, toxins, lectins. Of the above 27-fold types, types 1–6 belong to all α structural class, types 7–15 to all β class, types 16–24 to α/β class and type 25–27 to $\alpha+\beta$ class. Therefore, the classification of 27 folds is one level deeper than that of 4 structural classes (Cai, 2001; Chou and Zhang, 1995; Zhou, 1998; Zhou and Assa-Munt, 2001). Naturally, it is more challenging and difficult to conduct prediction among the 27-fold types than among the 4 structural classes (Chou, 1995; Chou and Maggiora, 1998).

To deal with the problem, Ding and Dubchak (2001) extracted the following six features from protein sequences: (1) amino acid composition, (2) predicted secondary structure, (3) hydrophobicity, (4) normalized van der Waals volume, (5) polarity and (6) polarizability. Of the above six features, only the amino acid composition contains 20 components, with each representing the occurrence frequency of one of the 20 native amino

acids in a given protein (Chou and Zhang, 1994; Zhou and Doctor, 2003). For the other five features, each contains $3+3+5\times3=21$ components, as detailed in Ding and Dubchak (2001) and Dubchak *et al*. (1999). Based on these multiple parameter sets and majority voting rule trained by the proteins in the training dataset, an overall success rate of 56% was reported (Ding and Dubchak, 2001) in predicting the fold type for the proteins in the testing dataset.

In the present study, in order to avoid completely ignoring the sequence-order effects, the pseudo-amino acid composition (Chou, 2001) was used to replace the conventional amino acid composition (Chou and Zhang, 1993; Nakashima *et al*., 1986) as used in (Ding and Dubchak, 2001). However, rather than using a combined correlation function (Chou, 2001), here the alternate correlation function between hydrophobicity and hydrophilicity (Chou, 2005; Chou and Cai, 2005) is adopted to reflect the sequence-order effects. For reader's convenience, a brief introduction about amphiphilic pseudo-amino acid composition (PseAA) is given below.

Suppose a protein **P** with a sequence of L amino acid residues:

$$\mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7\cdots\mathrm{R}_L, \tag{1}$$

where R1 represents the residue at chain position 1, R2 at position 2 and so forth. The hydrophobicity and hydrophilicity of the constituent amino acids in a protein play a very important role to its folding; e.g. many helices in proteins are amphiphilic that is formed by the hydrophobic and hydrophilic amino acids according to a special order along the helix chain, as illustrated by the 'wenxiang' diagram (Chou *et al*., 1997). Therefore, these two indices may be one of the optimal choices to reflect the sequence-order effects. In view of this, the sequence-order effects can be indirectly and partially, but quite effectively, reflected through the following equations (Fig. 1):

$$\begin{cases} \tau_1 = \frac{1}{L-1}\sum_{i=1}^{L-1} H^1_{i,i+1} \\ \tau_2 = \frac{1}{L-1}\sum_{i=1}^{L-1} H^2_{i,i+1} \\ \tau_3 = \frac{1}{L-2}\sum_{i=1}^{L-2} H^1_{i,i+2} \\ \tau_4 = \frac{1}{L-2}\sum_{i=1}^{L-2} H^2_{i,i+2} \\ \quad\cdots\cdots\cdots\cdots \\ \tau_{2\lambda-1} = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} H^1_{i,i+\lambda} \\ \tau_{2\lambda} = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} H^2_{i,i+\lambda} \end{cases}, \quad (\lambda < L) \tag{2}$$

where $H^1_{i,j}$ and $H^2_{i,j}$ are the hydrophobicity and hydrophilicity correlation functions given by

$$\begin{cases} H^1_{i,j} = h^1(\mathrm{R}_i)\cdot h^1(\mathrm{R}_j) \\ H^2_{i,j} = h^2(\mathrm{R}_i)\cdot h^2(\mathrm{R}_j) \end{cases} \tag{3}$$

where (h^1(R$_i$) and (h^2R$_i$) are, respectively, the hydrophobicity and hydrophilicity values for the ith ($i = 1,2,\ldots,L$) amino acid in Equation (1), and the dot (·) means the multiplication sign. In Equation (2) τ_1 and τ_2 are called the 1st-tier correlation factors that reflect the sequence-order correlation between all the most contiguous residues along a protein chain through hydrophobicity and hydrophilicity, respectively [Figure 1(a1), (a2)], τ_3 and τ_4 are the corresponding 2nd-tier correlation factors that reflect the sequence-order correlation between all the 2nd most contiguous residues [Figure 1(b1),(b2)], and so forth. Note that before substituting the values of hydrophobicity and hydrophilicity into Equation (3), they were all subjected

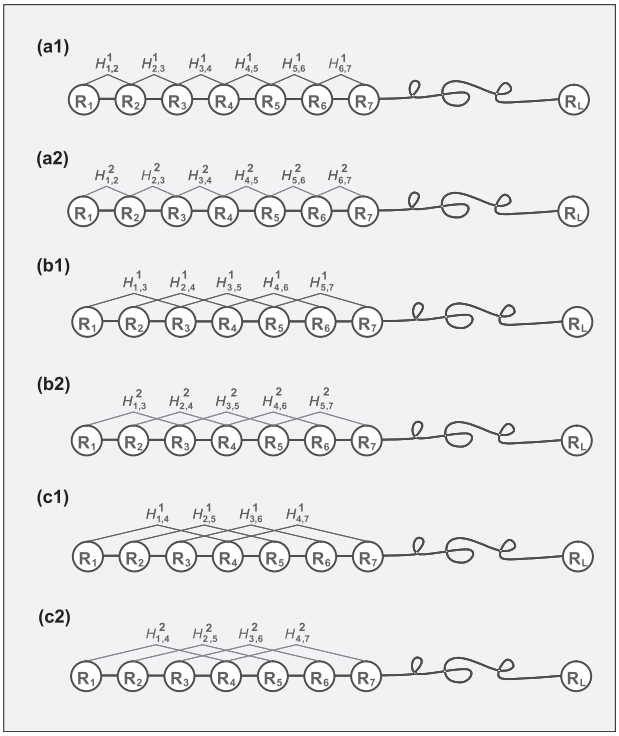

Fig. 1. A schematic drawing to show the amphiphilic correlation along a protein chain, where the values of $H^1_{i,j}$and $H^2_{i,j}$ are given by Equations (3) and (4) and Table 1. The correlation via hydrophobicity is shown in red, while the correlation via hydrophilicity in blue (a colour version of this figure appears in the Supplementary data). Panel (a1/a2) reflects the coupling mode between all the most contiguous residues, panel (b1/b2) that between all the 2nd most contiguous residues, and panel (c1/c2) that between all the 3rd most contiguous residues.

to a standard conversion as described by the following equation:

$$\begin{cases} h_1(\mathrm{R}_i) = \frac{h^0_1(\mathrm{R}_i) - <h^0_1>}{\mathrm{SD}(h^0_1)} \\ h_2(\mathrm{R}_i) = \frac{h^0_2(\mathrm{R}_i) - <h^0_2>}{\mathrm{SD}(h^0_2)} \end{cases} \tag{4}$$

where the symbols h^0_1(R$_i$) and h^0_2(R$_i$) represent the original hydrophobicity value (Tanford, 1962) and hydrophilicity value (Hopp and Woods, 1981) for amino acid R_i, respectively (Table 1); $<h^0_1>$ and $<h^0_2>$ their means over 20 native amino acids; SD(h^0_1) and SD(h^0_2) their standard deviations. The converted hydrophobicity and hydrophilicity values obtained by Equation (4) will have a zero mean value over the 20 native amino acids and will remain unchanged if going through the same conversion procedure again. As we can see from Equations (1–4) as well as Figure 1, a considerable amount of sequence-order information has been incorporated into the 2λ correlation factors through the hydrophobic and hydrophilic values of the amino acid residues along a protein chain. By fusing the 2λ amphiphilic correlation factors into the classical amino acid composition, we have the following augmented discrete form to represent a protein sample **P**:

$$\mathbf{P} = \begin{bmatrix} p_1 \\ \vdots \\ p_{20} \\ p_{20+1} \\ \vdots \\ p_{20+\lambda} \\ p_{20+\lambda+1} \\ \vdots \\ p_{20+2\lambda} \end{bmatrix}, \tag{5}$$

Table 1. The amino acid parameters used for deriving the amphiphilic pseudo-amino acid components [cf. Equation (4)]

Code	Hydrophobicity[a] h_1^0	Hydrophilicity[b] h_2^0
A	0.62	−0.5
C	0.29	−1.0
D	−0.90	3.0
E	−0.74	3.0
F	1.19	−2.5
G	0.48	0.0
H	−0.40	−0.5
I	1.38	−1.8
K	−1.50	3.0
L	1.06	−1.8
M	0.64	−1.3
N	−0.78	2.0
P	0.12	0.0
Q	−0.85	0.2
R	−2.53	3.0
S	−0.18	0.3
T	−0.05	−0.4
V	1.08	−1.5
W	0.81	−3.4
Y	0.26	−2.3

[a]The hydrophobicity values were taken from Tanford (1962).
[b]The hydrophilicity values were taken from Hopp and Woods (1981).

where

$$p_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{2\lambda}\tau_j}, & (1 \le u \le 20) \\ \dfrac{w\tau_u}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{2\lambda}\tau_j}, & (20+1 \le u \le 20+2\lambda) \end{cases} \quad (6)$$

where f_i ($i = 1,2,\ldots 20$) are the normalized occurrence frequencies of the 20 native amino acids in the protein **P**, τ_j the sequence-correlation factor computed according to Equation (2) and w the weight factor. In the current study, we chose $w = 0.5$ to make the results of Equation (6) within the range easier to be handled (w can be of course assigned with other values, but this would not have a big different impact to the final results). Therefore, the first 20 numbers in Equation (5) represent the classic amino acid composition, and the next 2λ discrete numbers reflect the amphiphilic sequence correlation along a protein chain. Such a protein representation is called 'amphiphilic pseudo-amino acid composition', which has the same form as the conventional amino acid composition, but contains much more information. It is through the 2λ pseudo-amino acid components that the sequence order of a protein chain and the distribution of the hydrophobic and hydrophilic amino acids along the chain are indirectly and partially reflected. It should be pointed out that, according to the definition of the classical amino acid composition, all its components must be ≥ 0; it is not always true, however, for the pseudo-amino acid composition: the components corresponding to the sequence correlation factors may also be < 0.

In this study, the OET-KNN (optimized evidence-theoretic k-nearest neighbors) algorithm is adopted as the operation engine of a classifier (Shen and Chou, 2005). For reader's convenience, a brief introduction about OET-KNN classifier and its key equations are given in Appendix A. However, quite different from the case of (Shen and Chou, 2005), now we have many different input types, such as the (20+2λ)D PseAA, 21D predicted secondary structure, 21D hydrophobicity, 21D normalized van der Waals volume, 21D polarity and 21D polarizability (Ding and Dubchak, 2001). Since a basic classifier is defined by one operation engine and one input type, one way to use the information from the multiple input types is to combine the above 6 input types into one and use a [(21×5)+(20+2λ)]D vector to represent it. However, doing so would introduce too many parameters into the input, thereby reducing the cluster-tolerant capacity (Chou, 1999) and cross-validation success rate. Furthermore, the PseAA with a different value of λ will become a different input type. In the present study, λ was assigned with 1, 4, 14 and 30. Therefore, we are actually facing 5 + 4 = 9 different input types (Table 2), and have 9 basic classifiers. To deal with this situation, we shall introduce an ensemble classifier, by which not only the other five features described in (Ding and Dubchak, 2001) but also the pseudo-amino acid compositions with a set of different λ values can be automatically fused into one prediction system.

The framework of ensemble classifier system was established by combining numerous basic classifiers together in order to reduce the variance caused by the peculiarities of a single training set and hence be able to learn a more expressive concept in classification than a single classifier. Illustrated in Figure 2 is the basic framework for an ensemble classifier that consists of

Table 2. List of nine features extracted from protein sequences for fold recognition

Features	Dimension
Pseudo-amino Acid Composition[a]	22
Pseudo-amino Acid Composition[b]	28
Pseudo-amino Acid Composition[c]	48
Pseudo-amino Acid Composition[d]	80
Predicted secondary structure	21
Hydrophobicity	21
Normalized van der Waals volume	21
Polarity	21
Polarizability	21

[a]The effects of the first rank of sequence-order correlation are incorporated [cf. Equation (5) with $\lambda = 1$].
[b]The effects of the first 4 ranks of sequence-order correlation are incorporated [cf. Equation (5) with $\lambda = 4$].
[c]The effects of the first 14 ranks of sequence-order correlation are incorporated [cf. Equation (5) with $\lambda = 14$].
[d]The effects of the first 30 ranks of sequence-order correlation are incorporated [cf. Equation (5) with $\lambda = 30$].

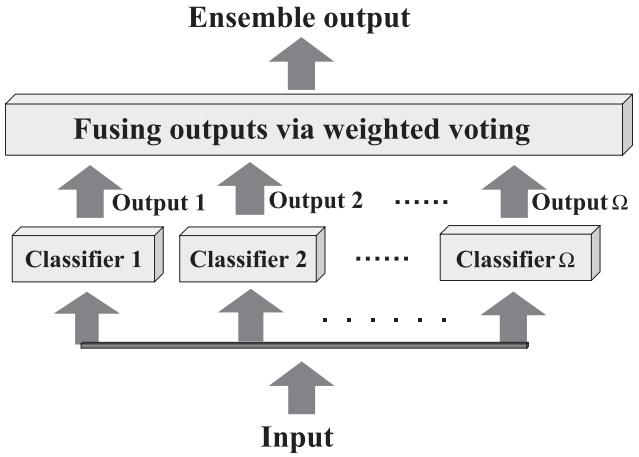

Fig. 2. Flowchart to show how the ensemble classifier $\mathbb{C}$ [Equation (7)] is formed by fusing $\Omega = 9$ basic individual classifiers: $\mathbb{C}_1$, $\mathbb{C}_2, \ldots,$ and $\mathbb{C}_9$. A colour version of this figure appears in the Supplementary data.

$\Omega = 9$ basic classifiers. The final output of the ensemble is the weighted fusion of the outputs produced by the nine individual classifiers, as formulated below.

Suppose the ensemble classifier $\mathbb{C}$ is expressed by

$$\mathbb{C} = \mathbb{C}_1 \oplus \mathbb{C}_2 \oplus \mathbb{C}_3 \oplus \cdots \oplus \mathbb{C}_8 \oplus \mathbb{C}_9 \quad (7)$$

where $\mathbb{C}_1$, $\mathbb{C}_2$, …, $\mathbb{C}_3$ represent the nine basic OET-KNN classifiers (Appendix A) each operating on the input derived from one of the nine features listed in Table 2; i.e. classifier $\mathbb{C}_1$ operates on the 22D PseAA, $\mathbb{C}_2$ on the 28D PseAA, $\mathbb{C}_3$ on the 48D PseAA, $\mathbb{C}_4$ on the 80D PseAA, $\mathbb{C}_5$ on the 21D predicted secondary structure $\mathbb{C}_6$ on the 21D hydrophobicity, $\mathbb{C}_7$ on the 21D normalized van der Waals volume, $\mathbb{C}_8$ on the 21D polarity, and $\mathbb{C}_9$ on the 21D polarizability. In Equation (7) the symbol $\oplus$ denotes the fusing operator. For reader's convenience, the values of the nine input parameter systems (cf. Table 2) for each of the proteins in the training and testing datasets are given in the Online Supplementary Materials BI and BII, respectively.

Thus, the process of how the ensemble classifier $\mathbb{C}$ works by fusing the nine basic classifiers $\mathbb{C}(i)$ ($i = 1,2,\cdots,9$) can be formulated as follows. Suppose

$$Y_j = \sum_{i=1}^{9} w_i \mathbb{R}_i(\mathbf{P}, S_j), (j = 1, 2, \cdots, 27) \quad (8)$$

where S_1 is the set only containing proteins of fold type 1, S_2 the set of fold type 2, and so forth; $\mathbb{R}_i(\mathbf{P},S_j)$ is the belief function or supporting degree for $\mathbf{P}$ belonging to S_j obtained by the ith basic classifier as defined by Equation (A5) in Appendix A; and w_i is the weighted factor, which was assigned in this study with the value of the success rate obtained by the ith single basic classifier $\mathbb{C}_i$, as will be further discussed below.

Thus the query protein $\mathbf{P}$ is predicted belonging to the fold type with which its score of Equation 8 is the highest; i.e. suppose

$$Y_\mu = \mathbf{Max}\{Y_1, Y_2, \cdots, Y_{27}\} \quad (9)$$

where the operator **Max** means taking the maximum one among those in the brackets, and the subscript μ is the very fold type predicted for the query protein $\mathbf{P}$. If there is a tie, the query protein may not be uniquely determined and will be randomly assigned among those with a tie, but cases like that rarely occur.

RESULTS AND DISCUSSION

To demonstrate the power of the ensemble classifier, predictions were conducted based on the same training and testing datasets used by the previous investigators (Chung and Huang, 2003; Ding and Dubchak, 2001). None of proteins in these datasets has >35% sequence identity to any other, and most of proteins in the testing dataset have <25% sequence identity with those in the training dataset (Ding and Dubchak, 2001). The overall success rate in recognizing the fold among the 27 folding types by the ensemble classifier for the 383 proteins in the independent dataset is given in Table 3, where, for facilitating comparison, the success rates by the other approaches are also listed. As can be seen from Table 3, the ensemble classifier, which was formed by fusing nine basic classifiers, obviously outperformed the other approaches.

It is instructive to note that if using each of the nine basic classifiers $\mathbb{C}_1$, $\mathbb{C}_2$, $\mathbb{C}_3$, $\mathbb{C}_4$, $\mathbb{C}_5$, $\mathbb{C}_6$, $\mathbb{C}_7$, $\mathbb{C}_8$, $\mathbb{C}_9$ to do the same prediction, the success rates would be 0.40, 0.44, 0.40, 0.29, 0.42, 0.37, 0.32, 0.29, 0.24, respectively. All of them are significantly lower than the rate of 0.62 = 62% obtained by the ensemble classifier (Table 3), indicating that a strong classifier can be generated by fusing many weak classifiers. Actually, as mentioned above, these single classifier rates were assigned for the weights w_i(i= 1,2,…,9) in Equation (9) to form the ensemble classifier.

Table 3. Overall success rates by different approaches in recognizing the fold types for proteins in the independent testing dataset

Classifier	Success rate(%)
MLP (Multi-Layer Perceptron) (Chung and Huang, 2003)	48.8
GRNN (General Regression Neural Networks) (Chung and Huang, 2003)	44.2
RBFN (Radial Basis Function Networks) (Chung and Huang, 2003)	49.4
NN (Neural Networks)[a] (Ding and Dubchak, 2001)	41.8
SVM (Support Vector Machines)[b] (Ding and Dubchak, 2001)	45.2
SVM[c] (Ding and Dubchak, 2001)	51.1
SVM[d] (Ding and Dubchak, 2001)	56.0
Ensemble Classifier[e]	62.1

[a]The training method for NN is 'one against others'.
[b]The training method for SVM is 'one against others'.
[c]The training method for SVM is 'unique one against others'.
[d]The training method for SVM is 'all against all'.
[e]The ensemble classifier is constructed by nine OET-KNN classifiers [cf. Equation (7)], and the number of neighbors in each OET-KNN classifier is 8.

CONCLUSIONS

An ensemble classifier is formed by a set of basic classifiers, whose individual outcomes are combined in some way, typically through a weighted voting, to give a final determination in classifying a query sample. The current ensemble classifier consists of nine basic individual classifiers. Their operation engine was OET-KNN algorithm, but they were each trained in nine different parameter systems extracted from the training dataset; i.e. 22D PseAA, 28D PseAA, 48D PseAA, 80D PseAA, 21D predicted secondary structure, 21D hydrophobicity, 21D normalized van der Waals volume, 21D polarity and 21D polarizability.

It is instructive to note that although the operation engine adopted here for the basic classifiers is the OET-KNN algorithm, others, such as the covariant discriminant algorithm and SVM algorithm, can also be used to replace the OET-KNN for forming different ensemble classifiers. Moreover, the constituent individual basic classifiers can be driven by completely different operation engines as well, and an ensemble classifier thus formed would become one with a mixture of operation engines. Similarly, we can also design an ensemble classifier by fusing both different input types and different operation engines. It is shown thru the present study that the ensemble classifier formed by fusing different input types, particularly different dimensions of pseudo-amino acid composition [(cf. Equation (5)], is very promising for enhancing the success rate in recognizing the fold type of proteins.

APPENDIX A

The optimized evidence-theoretic *k*-nearest neighbors (OET-KNN) classifier

For reader's convenience, a brief introduction of the OET-KNN classifier is given below. For further explanation, refer to (Shen and Chou, 2005). Let us consider a problem of classifying N entities into 27 classes (fold types), which can be formulated as

$$\mathbb{F} = \{\Phi_1, \Phi_2, \ldots, \Phi_\mu, \ldots, \Phi_{27}\} \quad \text{(A1)}$$

The available information is assumed to consist of a training dataset

$$\mathbb{N} = \{(\mathbf{P}_1, \theta_1), \cdots, (\mathbf{P}_N, \theta_N)\} \quad \text{(A2)}$$

where the N entities $\mathbf{P}_i (i = 1,2,\ldots,N)$ and their corresponding pattern (class) labels $\theta_i (i = 1,2,\ldots,N)$ take values in $\mathbb{F}$ of Equation (A1). According to the KNN (k-nearest neighbors) rule (Cover and Hart, 1967), an unclassified entity $\mathbf{P}$ is assigned to the class represented by a majority of its K-nearest neighbors of $\mathbf{P}$. Owing to its good performance and simple-to-use feature, the KNN rule, also named as 'voting KNN rule', is quite popular in pattern recognition community.

The ET-KNN (evidence theoretic k-nearest neighbors) rule is a pattern classification method based on the Dempster–Shafer theory of belief functions (Denoeux, 1995). In the classification process, each neighbor of a pattern to be classified is considered as an item of evidence supporting certain hypotheses concerning the class membership of that pattern. Based on this evidence, basic belief masses are assigned to each subset concerned. Such masses are obtained for each of the k-nearest neighbors of the pattern under consideration and aggregated using the Dempster's rule of combination (Shafer, 1976). A decision is made by assigning a pattern to the class with the maximum credibility.

Suppose $\mathbf{P}$ is a query protein to be classified, and $S_K^{\mathbf{P}}$ is the set of its k-nearest neighbors in the training dataset $\mathbb{N}$ of Equation (A2). Thus, for any $\mathbf{P}_i \in S_K^{\mathbf{P}}$, the knowledge that $\mathbf{P}_i$ belongs to class $\Phi_\mu \in \mathbb{F}$ can be considered as a piece of evidence that increases our belief that $\mathbf{P}$ also belongs to Φ_μ. According to the basic belief assignment mapping theory (Shafer, 1976), this item of evidence can be formulated by

$$\mathbb{R}(\mathbf{P}_i, \Phi_\mu) = \alpha_0 \exp[-\gamma_\mu^2 D^2(\mathbf{P}_i, \mathbf{P})] \quad \text{(A3)}$$

where α_0 is a fixed parameter, γ_μ is a parameter associated with class Φ_μ and $D^2(\mathbf{P}_i, \mathbf{P})$ is the square Euclidean distance between $\mathbf{P}$ and $\mathbf{P}_i$. In the ET-KNN rule, it was not addressed how to optimally select the parameters. In 1998 an optimization procedure to determine the optimal or near-optimal parameter values was proposed from the data by minimizing an error function (Zouhal and Denoeux, 1998). It was observed that the OET-KNN rule obtained thru such an optimization treatment would lead to a substantial improvement in classification accuracy. The optimal parameter thus obtained for α_0 of Equation A3 is 0.95, and those for γ_μ are given in Table A1.

Table A1. The optimal parameter γ_j ($j = 1, 2, \ldots, 27$) in Equation A3 obtained thru the optimized procedure (Zouhal & Denoeux, 1998) for the 9 basic individual classifiers in Equation 7

	$\mathbb{C}1$	$\mathbb{C}2$	$\mathbb{C}3$	$\mathbb{C}4$	$\mathbb{C}5$	$\mathbb{C}6$	$\mathbb{C}7$	$\mathbb{C}8$	$\mathbb{C}9$
γ_1	0.1028	0.0714	0.0398	0.0434	0.4848	0.1275	0.1977	0.1396	0.0682
γ_2	0.2908	0.0727	0.0523	0.1324	0.0585	0.0784	0.3654	0.2218	0.0387
γ_3	0.0656	0.0490	0.0240	0.0435	0.0469	0.0604	0.2482	0.2988	0.0445
γ_4	0.0888	0.0480	0.0752	0.0597	0.4306	0.1541	0.4092	0.4026	0.0525
γ_5	0.0798	0.0525	0.0274	0.0312	0.4146	0.1510	0.0866	0.1175	0.0543
γ_6	0.0931	0.1176	0.0468	0.0581	0.1298	0.4586	0.0981	0.4580	0.1741
γ_7	0.0954	0.0783	0.0520	0.0543	0.2871	0.0973	0.3723	0.1200	0.0456
γ_8	0.1241	0.1013	0.0475	0.0462	0.1890	0.3587	0.5057	0.1365	0.0385
γ_9	0.1476	0.1076	0.0700	0.0699	0.1386	0.1034	0.3606	0.1101	0.0425
γ_{10}	0.1210	0.0787	0.0436	0.0448	0.4871	0.2142	0.5427	0.1234	0.0795
γ_{11}	0.1002	0.0518	0.0265	0.0840	0.2862	0.2423	0.2530	0.3039	0.0436
γ_{12}	0.1219	0.1014	0.0455	0.0594	0.3810	0.1911	0.3326	0.2352	0.0731
γ_{13}	0.1331	0.0969	0.0449	0.0375	0.2096	0.1164	0.3723	0.1064	0.0645
γ_{14}	0.1033	0.0899	0.0479	0.0484	0.0702	0.2026	0.0665	0.4069	0.1816
γ_{15}	0.1108	0.0875	0.0523	0.0317	0.0961	0.1280	0.4249	0.1442	0.0407
γ_{16}	0.1543	0.1146	0.0687	0.0578	0.1376	0.1298	0.1346	0.1472	0.0505
γ_{17}	0.1665	0.1026	0.0706	0.1815	0.4662	0.1505	0.5627	0.1430	0.0419
γ_{18}	0.6170	0.1548	0.1543	0.0502	0.0977	0.3061	0.1003	0.1334	0.0511
γ_{19}	0.1664	0.1190	0.0616	0.0644	0.5661	0.1416	0.5118	0.1443	0.0419
γ_{20}	0.1478	0.1097	0.0695	0.0682	0.1374	0.1119	0.4803	0.1240	0.0496
γ_{21}	0.1183	0.0812	0.0479	0.1321	0.4207	0.2018	0.2975	0.1517	0.0409
γ_{22}	0.1629	0.1226	0.0759	0.2221	0.1331	0.2484	0.1513	0.5593	0.0998
γ_{23}	0.1553	0.1132	0.0716	0.0663	0.1950	0.1440	0.1581	0.1527	0.0662
γ_{24}	0.1704	0.1144	0.0671	0.0577	0.1430	0.1243	0.1379	0.1454	0.0566
γ_{25}	0.1313	0.1065	0.0519	0.1595	0.2768	0.1490	0.3159	0.1897	0.1788
γ_{26}	0.1517	0.0953	0.0303	0.0831	0.4289	0.3329	0.3937	0.4259	0.0466
γ_{27}	0.0262	0.0153	0.0133	-0.002	0.0093	0.0403	0.2920	0.0425	0.0492

The belief function of **P** belonging to class Φ_μ is a combination of its *k*-nearest neighbors, and can be formulated as

$$\mathbb{R}(\mathbf{P},\Phi_\mu)=(\cdots((\mathbb{R}(\mathbf{P}_1,\Phi_\mu)\oplus\mathbb{R}(\mathbf{P}_2,\Phi_\mu))\oplus\mathbb{R}(\mathbf{P}_3,\Phi_\mu))\oplus\cdots)\oplus\mathbb{R}(\mathbf{P}_\mathrm{K},\Phi_\mu) \quad (A4)$$

where $\oplus$ is called the orthogonal sum, which is commutative and associative. According to Dempster's rule (Shafer, 1976), the belief function of Equation A4 can be expressed as

$$\mathbb{R}(\mathbf{P},\Phi_\mu)=\frac{\sum_{S^{\mathbf{P}}_{\mathrm{K},i}\subseteq S^{\mathbf{P}}_{\mathrm{K}},\,S^{\mathbf{P}}_{\mathrm{K},j}\subseteq S^{\mathbf{P}}_{\mathrm{K}},\,S^{\mathbf{P}}_{\mathrm{K}}\cap S^{\mathbf{P}}_{\mathrm{K},j}=\Phi_\mu}\mathbb{R}(\mathbf{P},S^{\mathbf{P}}_{\mathrm{K},j})\mathbb{R}(\mathbf{P},S^{\mathbf{P}}_{\mathrm{K},j})}{1-\sum_{S^{\mathbf{P}}_{\mathrm{K},i}\subseteq S^{\mathbf{P}}_{\mathrm{K}},\,S^{\mathbf{P}}_{\mathrm{K},j}\subseteq S^{\mathbf{P}}_{\mathrm{K}},\,S^{\mathbf{P}}_{\mathrm{K},i}\cap S^{\mathbf{P}}_{\mathrm{K},j}=\varnothing}\mathbb{R}(\mathbf{P},S^{\mathbf{P}}_{\mathrm{K},j})\mathbb{R}(\mathbf{P},S^{\mathbf{P}}_{\mathrm{K},j})} \quad (A5)$$

where $S^{\mathbf{P}}_{\mathrm{K},i}$ is the *i*-th possible subset of $S^{\mathbf{P}}_{\mathrm{K}}$, and $\subseteq$, $\cap$ and $\varnothing$ are the symbols in set theory, representing 'contained in', 'intersection', and the empty set, respectively.

A decision is made by assigning the query protein **P** to the class with which the belief or credibility function of Equation A5 has the maximum value; i.e. if

$$\mathbb{R}(\mathbf{P},\Phi_\mu)=\mathbf{Max}\{\mathbb{R}(\mathbf{P},\Phi_1),\mathbb{R}(\mathbf{P},\Phi_2),\cdots,\mathbb{R}(\mathbf{P},\Phi_{27})\} \quad (A6)$$

where $\mu=1,2,\ldots,$ or 27 and the operator **Max** means taking the maximum one among those in the brackets, then the class Φ_μ is the class predicted for the query protein.

Conflict of Interest: none declared.

REFERENCES

Andreeva,A. *et al*. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.

Cai,Y.D. (2001) Is it a paradox or misinterpretation. *Proteins*, **43**, 336–338.

Chou,J.J. and Zhang,C.T. (1993) A joint prediction of the folding types of 1490 human proteins from their genetic codons. *J. Theor. Biol.*, **161**, 251–262.

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

Chou,K.C. (1999) A key driving force in determination of protein structural classes. *Biochem. Biophys. Res. Commun.*, **264**, 216–224.

Chou,K.C. (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*, **43**, 246–255 (Erratum: ibid., 2001, Vol.44, 60).

Chou,K.C. (2004) Review: structural bioinformatics and its impact to biomedical science. *Curr. Med. Chem.*, **11**, 2105–2134.

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

Chou,K.C. and Cai,Y.D. (2005) Prediction of membrane protein types by incorporating amphipathic effects. *J. Chem. Inform. Modeling*, **45**, 407–413.

Chou,K.C. and Carlacci,L. (1991) Energetic approach to the folding of alpha/beta barrels. *Proteins*, **9**, 280–295.

Chou,K.C. and Maggiora,G.M. (1998) Domain structural class prediction. *Protein Eng.*, **11**, 523–538.

Chou,K.C. *et al*. (1984) Energetic approach to packing of a-helices: 2. General treatment of nonequivalent and nonregular helices. *J. Am. Chem. Soc.*, **106**, 3161–3170.

Chou,K.C. *et al*. (1990) Review: energetics of interactions of regular structural elements in proteins. *Accounts Chem. Res.*, **23**, 134–141.

Chou,K.C. *et al*. (1982) Structure of beta-sheets: origin of the right-handed twist and of the increased stability of antiparallel over parallel sheets. *J. Mol. Biol.*, **162**, 89–112.

Chou,K.C. and Zhang,C.T. (1994) Predicting protein folding types by distance functions that make allowances for amino acid interactions. *J. Biol. Chem.*, **269**, 22014–22020.

Chou,K.C. and Zhang,C.T. (1995) Review: prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.*, **30**, 275–349.

Chou,K.C. *et al*. (1997) Disposition of amphiphilic helices in heteropolar environments. *Proteins*, **28**, 99–108.

Chung,I.F. and Huang,C.D. (2003) Recognition of structure classification of protein folding by NN and SVM hierarchical learning architecture. In Kaynak,O., Alpaydin,E., Oja,E. and Xu,L. (eds), *Lecture Notes in Computer Sciences*. Springer, Istanbul, Turkey, Vol 2714, pp. 1159–1167.

Cover,T.M. and Hart,P.E. (1967) Nearest neighbour pattern classification. *IEEE Trans. Inform. Theory*, **IT-13**, 21–27.

Denoeux,T. (1995) A *k*-nearest neighbor classification rule based on Dempster–Shafer theory. *IEEE Trans. Syst. Man Cybern.*, **25**, 804–813.

Ding,C.H. and Dubchak,I. (2001) Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, **17**, 349–358.

Dubchak,I. *et al*. (1995) Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl Acad. Sci. USA*, **92**, 8700–8704.

Dubchak,I. *et al*. (1999) Recognition of a protein fold in the context of the structural classification of proteins (SCOP) classification. *Proteins*, **35**, 401–407.

Finkelstein,A.V. and Ptitsyn,O.B. (1987) Why do globular proteins fit the limited set of folding patterns? *Prog. Biophys. Mol. Biol.*, **50**, 171–190.

Holm,L. and Sander,C. (1999) Protein folds and families: sequence and structure alignments. *Nucleic Acids Res.*, **27**, 244–247.

Hopp,T.P. and Woods,K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl Acad. Sci. USA*, **78**, 3824–3828.

Murzin,A.G. *et al*. (1995) SCOP: a structural classification of protein database for the investigation of sequence and structures. *J. Mol. Biol.*, **247**, 536–540.

Nakashima,H. *et al*. (1986) The folding type of a protein is relevant to the amino acid composition. *J. Biochem.*, **99**, 152–162.

Shafer,G. (1976) *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ.

Shen,H.B. and Chou,K.C. (2005) 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.

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

Zhou,G.P. (1998) An intriguing controversy over protein structural class prediction. *Journal of Protein Chemistry*, **17**, 729–738.

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

Zhou,G.P. and Doctor,K. (2003) Subcellular location prediction of apoptosis proteins. *Proteins*, **50**, 44–48.

Zouhal,L.M. and Denoeux,T. (1998) An evidence-theoretic K-NN rule with parameter optimization. *IEEE Trans. Syst. Man Cybern.*, **28**, 263–271.