MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **66** (2011) 445-462

ISSN 0340 - 6253

Predicting Protein Functional Class with the Weighted Segmented Pseudo-Amino Acid Composition Moment Vector

Xinyuan Zhou[1*],Xi Li [2], Man Li[3], Xinguo Lu[2]

[1] *Department of Computer Science and Technology, Changsha University Changsha Hunan, 410003, China*

[2] *School of information science and technology, Hunan University, Changsha Hunan, 410082, China*

[3] *Department of Computer Science, Hunan University of Chinese Medicine, Changsha Hunan, 410208, China*

(Received September 27, 2010)

Abstract

Predicting protein function at the proteomic-scale is one of the fundamental goals in cell biology and proteomics. In this paper, we proposed a new method for characterizing protein sequences—the Weighted Segmented Pseudo-amino acid composition Moment Vector (W-SPsAA-MV). From protein sequences, the encoding method of W-SPsAA-MV is applied to protein functional class prediction associated with the nearest neighbor algorithm (NNA) and covariant discriminant (CD) classifier. The experiment results show that our new method is efficient to predict functional class of query proteins when protein-protein interaction information is limited.

[*] Corresponding author. E-mail address: xyzhou@yeah.net (X.Y. Zhou)

1. Introduction

The number of sequenced nucleotide sequences encoding proteins is growing at an extraordinarily fast rate, but the rate of sequence acquisition far surpasses one of protein function determination. Protein function prediction becomes an important issue in the post-genome era as the gap between the amount of sequence information and functional identification widens.

With the accumulation of sequence information, attention has been paid to the development of automatic and reliable methods for the prediction of protein function from sequence. Generally speaking, protein function depends on its' structure, which depends on protein sequence. At present, this fact has been widely recognized and has become the theoretical basis of protein functional prediction based on protein sequence[1].

Various computational approaches for the prediction of protein functional classes from protein sequences and analysis of protein sequences have been developed. Among them, the sequence similarity-based approach is the most famous. It is generally believed that proteins of the high sequences similarity are homologous proteins which have the same or similar structures and function[2].This method is mainly achieved through sequence alignment, such as FASTA [3]and BLAST[4]. However, not all homologous proteins share analogous function[5]. Therefore, the sequence similarity-based method is not suitable for the prediction of protein functional classification in the case of poor similarity, especially for orphan sequences.

Instead of using the sequence similarity-based approach, methods that rely on an alternative representation of proteins were proposed in some papers[6-12]. These methods are mainly extract some characteristics contained in protein sequence and use data mining or machine learning methods to predict protein function class. Ross[6] used data mining prediction(DMP) method to predict protein functional classification from sequences. DMP also can work when protein function is unknown. Cai was used of some properties of amino acids[7], including amino acid composition,

hydrophobicity, plarity, charge and other properties to represent the protein sequence as specific feature vector, which contains three parts, the composition (C), transition(T) and distribution(D).The SVM is used to predict protein functional classes in this method. In 2007, Wang proposed EBGW approach and combined with nearest neighbor algorithm to predict protein function[8]. And this method has been successfully applied to the study of protein structure prediction[9]. The concept of pseudo amino acid composition was proposed by Chou to predict protein subcellular location and structural classes and so on[10,11].

In this paper, we adopt the Weighted Segmented Pseudo-amino acid composition Moment Vector (W-SPsAA-MV) to represent the sample of a protein via a discrete model. We obtain good results using W-SPsAA-MV associate with nearest neighbor algorithm (NNA) and covariant discriminant (CD) classifier respectively in our study.

2. Dataset

We load down the 1818 proteins from ftp://ftpmips.gsf.de/yeast/ which was used in GOM[13], EBGW[8] and GE[12] method. There are 1377 proteins as known proteins among them. The eighteen functional categories of all proteins were shown in Table 1, including the unclassified proteins. The W-SPsAA-MV coding vector dimension must be less than the number of samples of each functional categories subset when we adopt CD classifier. Because the W-SPsAA-MV is 58-dimensional vector, we can't assign to the functional category whose sample size smaller than 58 for adopting CD classifier. Therefore, for CD classifier, the Num. 3 and Num. 17 functional categories in Table 1 will not be considered, because they are only 26 and 5 samples respectively. In this way, there are 1373 proteins with known function for CD classifier.

Table 1 The numbers of each functional class in dataset

Num	Functional class	total
1	Metabolism	408
2	Energy	95
3	Development (Systemic)	26
4	Cell Type Differentiation	204
5	Protein Systhesis	98
6	Interaction with the Environment	172
7	Cell Fate	143
8	Biogenesis of Cellular Components	324
9	Transcription	427
10	Protein Fate(folding, modification, destination)	452
11	Cell Cyle and DNA Processing	441
12	Protein with Binding function of Cofactor requirement	458
13	Cellular Transport,Transport Facilities and Transport routes	331
14	Regulation of Metabolism and Protein function	115
15	Cellular Communication/Signal Transduction Mechanism	110
16	Cell Rescue, Defense and Virulence	201
17	Transposable elements, Viral and Plasmid proteins	5
18	Unclassified protein	441

3. Methods

3.1. The Weighted Segmented Pseudo-amino acid composition Moment Vector (W-SPsAA-MV)

The feature extraction of protein sequences is that every protein sequence was represented as a specific feature vector. In this paper, the Weighted Segmented Pseudo-amino acid composition Moment Vector (W-SpsAA-MV) is proposed as the characterization of protein sequences. W-SpsAA-MV includes not only the amino acid composition information, but also take into account the physiochemical properties of amino acids, position and local information. This encoding vector

mainly contains three parts, first part is the conventional amino acid composition, second part is the segmented amino acid coding, these two parts are called Segmented Pseudo-amino acid composition (SpsAA), and the third part is the moment vector of amino acid (MV). Given a protein P, its W-SpsAA-MV can be generally formulated as:

$$P = [p_1, p_2, \quad p_{20}, p_{20+1}, \quad p_{20+\lambda}, p_{20+\lambda+1}, \quad p_{20+\lambda+20}]^T \tag{1}$$

where the first 20 elements $p_1, p_2, \quad p_{20}$ are associated with 20 components in the conventional amino acid composition. The middle λ elements $p_{20+1}, \quad p_{20+\lambda}, p_{20+\lambda+1}$ are the segmented amino acid coding. The last 20 elements are the moment vector of amino acid. Given a protein P with L amino acid residues,

$$R_1 R_2 R_3 R_4 R_5 R_6 R_7 \quad R_L \tag{2}$$

where R_1 represents the first residue in the sequence, R_2 represents second, and so forth. The conventional amino acid composition of protein P can be formulated as:

$$P_{AA} = \begin{bmatrix} f_1 \\ f_2 \\ \\ f_{20} \end{bmatrix} \tag{3}$$

where f_1 is the occurrence frequency of amino acid A in the protein P, f_2 that of amino acid C, and so forth. Here, without loss of generality, the single codes of 20 native amino acids are used according to their alphabetical order.

The second part of W-SpsAA-MV is the segmented amino acid coding. At first, based on physiochemical characteristics such as residue's hydrophobic property and charged property and so on, we can classify residues into four different classes according to Dandekra[14].

Neutral or non-polar amino acid: A_1={G,A,V,L,I,M,P,F,W};

Neutral or polar amino acid: A_2={Q,N,S,T,Y,C};

Acidic amino acid: A_3={D,E};

Alkalescent amino acid: A_4={H,K,R}.

And then we grouped these physiochemical properties with every two classes.

Thus, three groups can be obtained as follows: { A_1, A_2} vs {A_3, A_4}, {A_1, A_3} vs {A_2, A_4} and {A_1, A_4} vs{A_2, A_3}.We set three rules as described below:

Rule 1: Suppose protein P shown in formula (2), and we transform protein sequence P into numerical sequence in this way(Zhang, et al,2005,Wang,et al.,2007):

$$H_1(R_i)=\begin{cases}1 & R_i \in \{A_1, A_2\} \\ 0 & R_i \in \{A_3, A_4\}\end{cases} \tag{4}$$

$$H_2(R_i)=\begin{cases}1 & R_i \in \{A_1, A_3\} \\ 0 & R_i \in \{A_2, A_4\}\end{cases} \tag{5}$$

$$H_3(R_i)=\begin{cases}1 & R_i \in \{A_1, A_4\} \\ 0 & R_i \in \{A_2, A_3\}\end{cases} \tag{6}$$

Hence, protein sequence P can be represented as three binary sequences S^i=H_i(P)=H_i(R_1) H_i(R_2) … H_i(R_L), (i=1,2,3). We call S^1, S^2, S^3 the 1-eigen sequence, 2-eigen sequence and 3-eigen sequences, respectively.

Rule 2: For eigen sequence S with length L, let M a positive integer, we can divided S into M subsequence fragments, and was written as S_1, S_2, …, S_M(1≤M≤L), where M is the number of fragments. An example of the segmentation process can be seen in Figure 1.

```
S(n): 10101100101101011101001110010011010101 00
S(6): 101011
S(7): 0010110
S(7): 1011101
S(6): 001110
S(7): 0100110
S(7): 1010100
```

Figure 1. The segmentation of characteristic sequence

In the Figure 1, the eigen sequence S with length 40 was segmented into 6 eigen subsequence fragments. The length of k-th eigen subsequence was defined as:

$$n_k=\left\lfloor \frac{k\times L}{M} \right\rfloor-\left\lfloor \frac{(k-1)\times L}{M} \right\rfloor \tag{7}$$

where L is the length of eigen sequence S, and M is the number of segment.

Rule 3: For every eigen sequence S, we defined the segmented subsequence coding as the frequency of ‘1’ in every eigen subsequence fragments. In this way, we

can obtain a M-dimensional vector from a eigen sequence. We can obtain three eigen sequence from a protein sequence according to Rule 1, therefore, a 3M-dimentional vector can be gained, which was called the segmented amino acid composition. This vector was defined as:

$$P_{ss} = [\theta_1^{(1)}, \quad , \theta_M^{(1)}, \theta_1^{(2)}, \quad , \theta_M^{(2)}, \theta_1^{(3)}, \quad , \theta_M^{(3)}]^T \tag{8}$$

where the first M elements is the frequency of '1' in M segments of 1-eigen sequence, the middle M elements that of 2-eigen sequence, and the last M elements that of 3-eigen sequence.

The third part of W-SpsAA-MV is the moment vector of amino acid (MV)[15-17], which was defined as:

$$P_{MV} = \begin{bmatrix} a_1^{(1)} \\ a_2^{(1)} \\ \\ a_{20}^{(1)} \end{bmatrix} \tag{9}$$

where 20 elements in P_{MV} was formulated as:

$$a_i^{(1)} = \frac{w}{L \times (L-1)} \sum_{j=1}^{l} p_{ij} \qquad i = (1, 2, \quad 20) \tag{10}$$

where p_{ij} means the i-th type amino acid in the protein sequence P located in j-th. And a_i is formulated according with the alphabetical order. L is the length of protein sequence P, and w is the weight factor.

Supposing w=1, for protein P_1: AAACCC and protein P_2: ACACAC, the MV of them according to formula (10) can be represented as:

$$P_{P1} = (0.5, 0.5, 0, \quad , 0, 0.5, 1.25, 0, \quad , 0)^T \tag{11}$$

$$P_{P2} = (0.5, 0.5, 0, \quad , 0, 0.75, 1.0, 0, \quad , 0)^T \tag{12}$$

In summary, the W-SpsAA-MV in Eq (1) can be uniquely derived by normalizing the 40+λ elements in Eqs (3), (8) and (9) according to the following equations:

$$p_u = \begin{cases} w_1 f_u, & (1 \le u \le 20) \\ w_2 \theta_k, & (1 \le k \le 3M) \\ \dfrac{1}{L(L-1)}(w_3 a_t), & (1 \le t \le 20) \end{cases} \tag{13}$$

where w_1, w_2 and w_3 are the weight factors which were used to regulate the degree of influence of each feature on the classification system. For different datasets and conditions, we choose different w_1, w_2 and w_3 value to achieve the best result. M is the number of segment of eigen sequence, we chose M=6 to achieve the best result for our experimental dataset. Therefore, we can get 58-dimentional vector for W-SPsAA-MV.

3.2. Nearest neighbor algorithm (NNA)

The NNA[18] is well known in pattern recognition community due mainly to its good result and its simple-to-use feature. The NNA assign to an unclassified sample point the classification of the nearest of a set of classified points. In this section we introduced how the NNA was used to predict protein functional class in terms of W-SPsAA-MV.

The classification process of the NNA is as follows. Suppose a set of proteins $\{P_1, P_2, \ldots, P_n\}$ that have been classified into categories $\{C_1, C_2, \ldots, C_m\}$, from which an unknown protein P can be classified into those categories using the NNA. First, the nearest neighbor of protein P is given by the following equation:

$$nn(P) = P_k \tag{14}$$

where P is the W-SPsAA-MV of protein, And P_i is determined by the following equation:

$$D(P, P_k) = \min_{i=1}^{n} D(P, P_i) \tag{15}$$

Where n is the number of known proteins. And $D(P, P_i)$ is defined as

$$D(P, P_i) = 1 - (P \cdot P_i) / (\|P\| \|P_i\|) \tag{16}$$

where $P \cdot P_i$ is the dot product of P and P_i. The $\|P\|$ and $\|P_i\|$ mean the module of P and P_i respectively. Obviously, when $P \equiv P_i$, then $D(P, P_i) = 0$.

According to the rule of NNA, if and only if

$$\min[D(P,P_i)\,|\,(P_i \in C_1)] \le \min[D(P,P_i)\,|\,(P_i \in C_2)] \le \quad \le \min[D(P,P_i)\,|\,(P_i \in C_m)] \qquad (17)$$

Here, we use '≤' because some proteins belong to more than one category.

3.3. Covariant Discriminant (CD) classifier

The CD classifier has been applied to protein structural class and subcellular location prediction[19-21], and achieved satisfactory results. In this paper, the CD classifier is used to predict protein functional classification based on W-SPsAA-MV discrete model.

Suppose there are N proteins (P_1,P_2,…,P_N) which have been classified into δ functional categories subset, such as:

$$S = S_1 \cup S_2 \cup S_3 \cup \quad \cup S_\delta \qquad (18)$$

where each subset S_m(m=1,2,…,δ) is composed of proteins with the same functional class. There are N_m proteins in the functional subset S_m. Since every protein has more than one function, different functional subsets maybe contain the same proteins. Obviously,

$$N \ge N_1 + N_2 + N_3 + \quad + N_\delta \qquad (19)$$

where N is the total number of proteins. N_1 means the number of proteins in functional subset S_1 and N_2 means the number of proteins in functional subset S_2, and so forth.

According to Eq (1), we can suppose that the u-th protein in the subset S_m is formulated by

$$P_m^u = [p_{m,1}^u, p_{m,2}^u, \quad , p_{m,20}^u, p_{m,21}^u, \quad , p_{m,20+\lambda}^u, p_{m,21+\lambda}^u, \quad , p_{m,40+\lambda}^u]^T \qquad (20)$$

where $p_{m,j}^u$ (j=1,2,…,40+λ) means the j-th component of the u-th protein in subset S_m. The standard vector for the subset S_m is represented by

$$\overline{P}_m = [\overline{p}_{m,1}, \overline{p}_{m,2}, \quad , \overline{p}_{m,20}, \overline{p}_{m,21}, \quad , \overline{p}_{m,20+\lambda}, \overline{p}_{m,21+\lambda}, \quad , \overline{p}_{m,40+\lambda}]^T \qquad (21)$$

where

$$\overline{p}_{m,i} = \frac{1}{N_m}\sum_{u=1}^{N_m} p_{m,i}^{u} \qquad (i=1,2,\ldots,40+\lambda) \tag{22}$$

where N_m is the number of proteins in the m-th functional subset S_m. $\overline{P}_m$ can be considered as the standard protein of the subset S_m. According to the definition of W-SPsAA-MV, the sample of a query protein P should be given by

$$P = [p_1, p_2, \quad p_{20}, p_{21}, \quad p_{20+\lambda}, p_{21+\lambda}, \quad p_{40+\lambda}]^T \tag{23}$$

The element for the query protein P can be derived by Eq (13). In this way, the similarity between a unknown protein P and $\overline{P}_m$ is formulated by Eq (24), which is CD function.

$$S(P,\overline{P}_m) = D_{Mah}^2(P,\overline{P}_m) + |C_m| \qquad (m=1,2,\ldots,M) \tag{24}$$

In Eq (24), $D_{Mah}^2(P,\overline{P}_m)$ is the squared Mahalanobis distance between the query protein P and $\overline{P}_m$,

$$D_{Mah}^2(P,\overline{P}_m) = (P-\overline{P}_m)^T C_m^{-1}(P-\overline{P}_m) \tag{25}$$

where T is the transpose operator, and C_m^{-1} is the inverse matrix of C_m, which is a $(40+\lambda)\times(40+\lambda)$-dimensional matrix.

$$C_m = \begin{pmatrix} c_{1,1}^m & & c_{1,40+\lambda}^m \\ & & \\ c_{40+\lambda,1}^m & & c_{40+\lambda,40+\lambda}^m \end{pmatrix} \tag{26}$$

where C_m is the covariance matrix for the subset S_m and the $c_{i,j}^m$ is given by

$$c_{i,j}^m = \frac{1}{N_m - 1}\sum_{u=1}^{N_m}(p_{m,i}^u - \overline{p}_{m,i})(p_{m,j}^u - \overline{p}_{m,j}),\ (i,j=1,2,\ldots,40+\lambda) \tag{27}$$

In Eq (24), $|C_m|$ is the determinant of matrix C_m. The smaller the value $D(P,\overline{P}_m)$, the greater the similarity between P and $\overline{P}_m$. A query protein P may have one or more functional classes, but we generally can only predict a functional category by calculating the minimum of Eq (24). Therefore, the query protein will be assigned to ξ ($\xi \geq 1$) functional categories through the following function:

$$u = Min_{\xi}\{F(P,\overline{P}_1), F(P,\overline{P}_2), \quad , F(P,\overline{P}_{\delta})\} \tag{28}$$

where u is the functional set which was assigned to the query protein P. Min_{ξ} means the nearest ξ point between the query protein P and $\overline{P}_m$ (m=1,2,…, δ). For our data set, we take ξ=5 and also lists the predict results when ξ=1,2,3,4,5 respectively.

3.4. Evaluation method

In this paper, to prove the effectiveness of W-SPsAA-MV associated with NNA method, we compare the predict results of our approach and GOM, EBGW and GE method. We examined the prediction quality by the test which was used in the other four methods. Here, the yeast proteome is divided into eight groups according to the number of partner protein, and about 40% proteins of each group are separated from the original dataset as testing test, the remaining proteins as training set. The goal of the method is to predict functional classes of the proteins in the test set using that of the remaining proteins. After random sampling for many times for every group, we consider the average success rate as the last success rate.

For W-SPsAA-MV associated with CD classifier, we adopt self-consistency examination, Jackknife examination and 5-fold cross validation examination to verify its effectiveness.

4. Results and discussion

4.1. The result and discussion for W-SPsAA-MV and NNA

All datasets are calculated using our W-SPsAA-MV method and the nearest neighbor algrithm (NNA), and the results were tested by the evaluation method as mentioned above. In W-SPsAA-MV, different values w_1, w_2, w_3 and M have a significant impact on the final results. After repeated test, we find that M=6 is the best choice for the dataset. And the value of w_1, w_2 and w_3 were shown in Table 2 for the different conditions. The comparison result between our method and GOM, EBGW and GE method is shown in Table 2. The overall accuracy of protein classification is

improved from GOM, EBGW and GE method in some cases. The “K” in Table 2 denotes the number of interacting partners, and “N_k” is the number of proteins, which has k interacting partners function is known

Table 2 The comparison results between our method and different algorithms

K	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K>7
n_K	670	247	159	99	63	34	26	79
GOM[a](%)	__	61	76	77	86	89	94	94
EBGW[b](%)	__	66	72	79	76	79	80	87
GE[c](%)	60.3	66.5	66	78	76.7	74.7	90	77.5
w_1	4	9	2	3	5	3	7	1
w_2	0.5	0.4	0.8	0.2	0.8	0.2	0.5	0.4
w_3	5	2	10	5	3	6	8	5
our method(%)	62.5	68	68	78	78.5	80	85	82.5

[a]Comes from[13].

[b]Comes from[8].

[c]Comes from[12].

From Table 2, we can find that the accuracy of our method is better than that of GOM method when K=2 and 4 with 7% and 1% improvement, respectively. Especially when K=1, the accuracy of our method achieve to about 62.5%, while the GOM approach can’t make prediction. Therefore, the results show that W-SpsAA-MV and NNA can be used to predict protein functional class as a complement method, when the protein-protein interaction information is limited.

Besides, from Table 2, we find that the success rate of our method is 62.5%, 68%, 78.5%, 80% and 85% when K=1, 2, 5, 6 and 7, respectively, which is better than EBGW method. We can see from Table 2, our method’s prediction overall accuracy is better than the GE method in most cases. Therefore, the results demonstrate that the performance of W-SpsAA-MV is superior to EBGW and GE as a whole.

During our study, we find that the choice of M, w_1, w_2 and w_3 is very important,

different value of M, w_1, w_2 and w_3 will have a major impact on the last result. So we need test many times to selet optimal values.

4.2. The result and discussion for W-SPsAA-MV and CD classifier

The Table 3, 4 and 5 give the prediction results for W-SpsAA-MV and CD classifier. We take w_1=5, w_2=0.1 and w_3=1 respectively for CD classifier with self-consistency examination, Jackknife examination and 5-fold cross validation examination.

Table 3 The self-consistency results using W-SpsAA-MV and CD classifier

N	N ≥ 1	N ≥ 2	N ≥ 3	N ≥ 4	N ≥ 5
number	1373	1090	677	408	215
k=1(%)	70.8	—	—	—	—
k=2(%)	84.6	50.8	—	—	—
k=3(%)	91.1	70.7	41.2	—	—
k=4(%)	95.4	81.6	63.1	35.3	—
k=5(%)	97.1	88.5	76.5	55.4	28.4

Table 4 The jackknife test results using W-SpsAA-MV and CD classifier

N	N ≥ 1	N ≥ 2	N ≥ 3	N ≥ 4	N ≥ 5
number	1373	1090	677	408	215
k=1(%)	40.1	—	—	—	—
k=2(%)	59.8	17.7	—	—	—
k=3(%)	72.7	25.3	8.0	—	—
k=4(%)	81.6	51.0	21.4	4.4	—
k=5(%)	87.7	63.4	38.3	13.2	3.3

Table 5 The 5-fold cross-validation results using W-SpsAA-MV and CD classifier

N	N≥1	N≥2	N≥3	N≥4	N≥5
k=1(%)	38.3	—	—	—	—
k=2(%)	60.7	17.1	—	—	—
k=3(%)	73.5	35.8	7.6	—	—
k=4(%)	81.4	52.5	21.0	3.8	—
k=5(%)	88.1	64.9	37.8	12.2	4.3

In Table 3, the “number” denotes the number of proteins which have more than N functional categories. According to different N value, we can find the number of protein functional categories is uncertain, some only have a functional class, and some have more than five functional categories. The “k” denotes the probability of accurately predict k functional classes for query proteins. When k=1 and N≥1, we assign a nearest functional category to proteins of test set and reach 70.8% prediction accuracy. When k=2 and N≥1, two functional categories was assigned to proteins of test set, the probability of accurately predict one of which is 84.6%. When k=2 and N≥2, the probability which this two functional categories were accurately predict is 50.8% for CD classifier.

From Figure 2, we find every curve gradually rise with the increase of k value, which means that when N value is unchanged, as the k value increase, the prediction accuracy increase significantly. Especially, when N≥1, the overall accuracy has increased from 70.8% to 97.1%. This is because when the predicted functional categories increase, the probability will undoubtedly increase as long as one of these functional categories is accurate. From the vertical axis of Fig.1, we can find when k

is constant, the prediction accuracy will less and less with the N value increase, which demonstrate that the more the number of functional categories which was predict correctly with CD classifier, the less the success rate. For example, when k=5, the success rate has reduced from 97.1% to 28.4% with the N value increase. Therefore, it is very difficult to predict accurately all functional categories for a query protein.

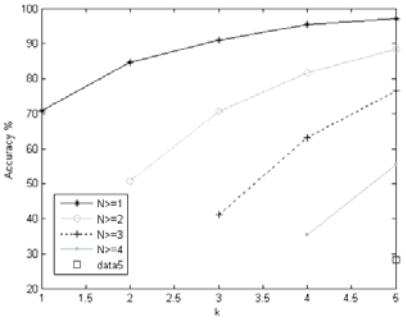

Figure 2. Curve for self-consistency using CD classifier.

Jackknife test also called leave-one-out test, which is considered the most objective and most rigorous evaluation method. The jackknife test result using CD classifier with W-SPsAA-MV is shown in Table 4.

In addition, we adopt 5-fold cross-validation to evaluate our method. The Table 5 shows the accuracy using 5-fold cross-validation.

From Table 4 and Table 5, we can find that the predicted result is similar by comparing the jackknife test and 5-fold cross-validation. In order to directly compare the predicted result of three evaluation methods, Figure 3 shows the predicted result curves when $N \geq 1$.

From Figure 3, we can find the accuracy of self-consistency is best and that of jackknife test and 5-fold cross-validation are fairly. Self-consistency contains the test samples themselves so that extracted features are more consistent with the test samples, so the accuracy is better than that of other two assessment methods. We also find from Figure 3 the accuracy of jackknife test and 5-fold cross-validation grow faster than that of self-consistency examination.

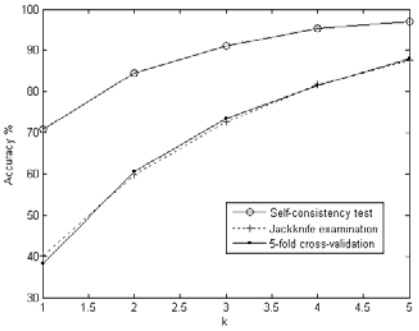

Figure 3. Accuracy of self-consistency using CD classifier.

5. Conclusion

In this paper, a new method of characterizing protein sequence named W-SPsAA-MV is proposed. We adopt the nearest neighbor algorithm (NNA) and covariant discriminant (CD) classifier respectively based on W-SPsAA-MV to predict protein functional classes.The experimental result show that the accuracy of W-SPsAA-MV with NNA is better than GOM, SWN-BA, EBGW and GE methods in some cases. In addition, by analyzing the experimental results with CD classifier, we can find the effect of predict is not very satisfactory when we assign a small number of functional classes to query proteins. While the predicted effect is significantly improved obviously with the increase of the number of predicted functional categories increase. So, the results demonstrate that W-SPsAA-MV is convenient to calculate and effective on the problem of protein functional class prediction. Furthermore, our approach is applicable in the case of litte and no information of protein-protein interaction.

Acknowledgement

This work is supported by the National Nature Science Foundation of China (Grant 60973082, 60873184) the National Nature Science Foundation of Hunan province (Grant 07JJ5080) and the Planned Science and Technology Project of Hunan Province (Grant 2009FJ3195).

References

[1] H. Shen, K. Chou, Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifer and pseudo amino acid composition, *Biochem. Biophys. Res. Commun.* **337** (2005) 752–756.

[2] J. Whisstock, A. M. Lesk, Prediction of protein function from protein sequence and structure*, Rev. Biophys.* **36** (2003) 307–340.

[3] W. Pearson, D. Lipman, Improved tools for biological sequence comparison, *PANS* **85** (1998) 2444–2448.

[4] S. Altachul, T. Madden, A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucleic Acids Res.* **25** (1997) 3389–3402.

[5] S. Benner, S. Chamberlin, D. Liberles, S. Covindarajan, L. Kencht, Functional inferences from reconstructed evolutionary biology involving recitified databases-an evolutionarily grounded approach to functional genomics, *Res. Microbiol.* **151** (2000) 97–106.

[6] D. Ross, K. Andreas, C. Anmanda, L. Dehaspe, Genome scale prediction of protein functional class from sequence using data mining, *ACM* **233** (2000) 384–389.

[7] Z. Cai, L. Han, Z. Ji, X. Chen, Y. Chen, SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence, *Nucleic Acids Res.* **31** (2003) 3692–3697.

[8] X. Wang, Z. Wang, W. Wang, Z. Zhang, A method of encoding based on grouped weight for protein function prediction*, China J. Bioinformatics* **5** (2007) 25–27.

[9] Z. Zhang, Z. Wang, Y. Wang, A new encoding scheme to improve the performance of protein structural class prediction, *Lecture Notes Comput. Sci.* **36** (2005) 1164–1173.

[10] T. Zhang, Y. Ding, K. Chou, Prediction protein structural classes with pseudo-amino acid composition: Approximate entropy and hydrophobicity pattern, *J. Theor. Biol.* **250** (2008) 186–193.

[11] K. Chou, H. Shen, Recent progress in protein subcellular location prediction. *Anal. Biochem.* **370** (2007) 1–16.

[12] X. Li, B. Liao, Y. Shu, Q. Zeng, Protein functional class prediction using global encoding of amino acid sequence, *J. Theor. Biol.* **261** (2009) 290–293.

[13] A. Vazquez, A. Flammimi, A. Maritan, A. Vespignani, Global function prediction from protein-protein interaction networks, *Nature Biotech.* **21** (2003) 697–700.

[14] T. Dandekra, B. Snel, M. Huynen, P. Bork, Conservation of gene order: A fingerprint of proteins that physically interact, *J. Trends Biochem.* **23** (1998) 324–328.

[15] L. Kurgan, L. Homaeian, Prediction of structural classes for protein sequences and domains-Impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy, *Pattern Recogn.* **39** (2006) 2323–2343.

[16] K. Kedarisetti, L. Kurgan, S. Dick, Classifier ensembles for protein structural class prediction with varying homology, *Biochem. Biophys. Res. Commun.* **348** (2006) 981–983.

[17] L. Kurgan, K. Chen, Prediction of protein structural class for the twilight zone sequences, *Biochem. Biophys. Res. Commun.* **357** (2007) 453–460.

[18] R. Duda, P. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

[19] Y. Cai, A. Diog, Prediction of Saccharomyces cerevisiae protein functional class from functional domain composition, *Bioinformatics* **20** (2004) 1292–1300.

[20] K. Chou, Y. Cai, Predicting protein-protein interactions from sequences in a hybridization space, *J. Proteome Res.* **5** (2006) 316–322.

[21] K. Chou, H. Shen, Predicting protein subcellular location by fusing multiple classifiers, *J. Cell. Biochem.* **99** (2006) 517–527.