Subcellular localization prediction of apoptosis proteins based on the data mining for amino acid index database

Zhuo-xing Shi[1], Qi Dai[1], Ping-an He[2], Yu-hua Yao[*1], Bo Liao[*3]
[1]College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China
[2]College of Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China
[3] College of Information Science and Engineering, Hunan University, Changsha Hunan, 410082, China.
zhuoxing1988@163.com, daiailiu2004@yahoo.com.cn, pinganhe@zstu.edu.cn, yaoyuhua2288@163.com, boliao@yeah.net

***Abstract*—In this work, based on the ACF model and the SVM classifier, succeeded on trials mining information that it's more effective to analyze the subcellular localization prediction of apoptosis proteins when adopting hydrophobicity property. This information is obtained in three benchmark datasets by using the ACF model and SVM to scan the AAindex database, which contains 544 kinds of amino acids. The contribution of this work is that it first did a comprehensive research on the effectiveness of the amino acid index for the subcellular localization of apoptosis proteins.**

Keywords—Database mining; Amino acid index; Subcellular localization prediction; Apoptosis protein.

I. INTRODUCTION

Apoptosis is a fundamental process controlling normal tissue homeostasis by regulating a balance between cell proliferation and death [1]. When apoptosis malfunctions, a variety of formidable diseases can ensue: blocking apoptosis is associated with cancer [2] and autoimmune diseases, while unwanted apoptosis can possibly lead to ischemic damage [3] or neurodegenerative disease [4]. Therefore, base on the primary sequence, obtaining information and the accurate prediction of subcellular location can be significantly improved our knowledge from apoptosis proteins, on the basis of these knowledge we can more clearly to understand the apoptosis mechanism and functions.

Zhou and Doctor [5], firstly investigated the prediction of subcellular location of apoptosis proteins. They explored amino acid composition and the covariant discrimination function to predict the four kinds of subcellular locations for 98 apoptosis proteins dataset. Through more than ten years of efforts, the prediction accuracy is improved, researchers have proposed many methods. Though the overall predictive accuracy have been improved for apoptosis proteins using existed methods, the representation of protein sequence was mainly by using the amino acid composition, or dipeptide composition.

Amino acid index has been used in wide-ranging bioinformatics research on protein sequences, such as predicting protein subcellular localization and protein structural classes prediction, the information encoded in the amino acid sequence ultimately determines the three-dimensional structure and biological function of a protein under physiological conditions [6]. A wide variety of properties of amino acids have been investigated through a large number of experiments and theoretical studies. An amino acid index is a set of 20 numerical values representing each of the different physicochemical properties of the 20 amino acids. Auto-correlation functions (ACF) based on the profile of the amino acid index along the primary sequence is normally used to predict the subcellular location of apoptosis proteins.

In order to comprehensively study the effection of amino acid index for apoptosis proteins subcellular localization prediction. In this paper, three dataset are applied to scan the AAindex database, AAindex databse contains 544 kinds of amino acids which can classified in six groups of A, B, C, H, P and O. Based on the ACF model and SVM classifier, we made a series of experiments which successful to mined the knowledge that using the amino acid index belongs group hydrophobicity (group H) it's more effective for apoptosis proteins subcellular location prediction.

II. MATERIALS

Three datasets were adopted in our work, proteins in those datasets were extracted from SWISS-PROT (version 49.5). The ZD98 dataset consists of 98 apoptosis protein sequences, which include 43 cytoplasmic proteins, 30 plasma membrane-bound proteins, 13 mitochondrial proteins and 12 other proteins[5]. The ZW225 dataset consists of 41 nuclear proteins, 70 cytoplasmic proteins, 25 mitochondrial proteins and 89 membrane proteins[7]. The dataset CL317 constructed by Chen and Li, consists of 112 cytoplasmic proteins, 55 membrane proteins, 34 mitochondrial proteins, 17 secreted proteins, 52 nuclear proteins and 47 endoplasmic reticulum proteins[8].

AAindex is a database of numerical indices representing various physicochemical and biochemical properties of amino acids and pairs of amino acids [9-11]. The AAindex database version is 9.1, currently contains 544 indices. In 1996, Tomii [12] defined the 402 indices classified in six groups, the 402 indices is contains in the 544 indices of AAindex database.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

The six groups are alpha and turn propensities (A), beta propensity (B), composition (C), hydrophobicity (H), physicochemical properties (P) and other properties (O), in this paper these six group of amino acid index are representation to A, B, C, H, P and O.

III. METHODS

A. Auto-correlation functions (ACF)

Auto-correlation functions based on the amino acid index propose by Feng [13] it was use to proteins structure classes prediction. The starting point of the ACF model algorithm is that each protein sequence is represented by an attribute vector on the basis of amino acid idnex.

For the ACF model, the attribute vector is defined as:

$$X = (r1, r2, ...ri...rm)^T, (i = 1, 2, ...m) \quad (1)$$

Where *ri* is the auto-correlation functions defined below, and *m* is an integer to be determined later.

To calculate the auto-correlation functions, fist is to replace each residue in the primary sequence by its amino acid index. Consequently, the replacement results in a numerical sequence:

$$h_1, h_2, ...h_N, \quad (2)$$

Where h_i is the amino acid index for the *i-th* residue and *N* is the number of residues of the protein. The auto-correlation function r_n is defined as:

$$r_n = \frac{1}{N-n}\sum_{i=1}^{N-n} h_i h_{i+n}, (n = 1, 2, ..., m) \quad (3)$$

Where h_i is the amino acid index for the *i-th* residue and *m (m<N)* is an integer. The key point of the present work is to replace the attribute vector defined in Equation (1) by the one defined in Equation (2) and Equation (3). From the above formula we know that each protein can derived an *m*-D feature vector through the ACF model.

Representation each protein sequence by a fixed-length feature vector is the primary tasks for apoptosis protein subcellular location prediction. On each dataset, the length of each sequence is different, so there have different *m* for each sequence. Therefore, we let the *m* is less-then or equal the length of the shortest sequence in the dataset, so in dataset ZD98 the *m* must be less-then 131, in dataset ZW225 *m* must less-then 77, in dataset CL317 *m* must less-then 88.

B. Novel feature selection strategy

This work apply the dimension reduction techniques principal component analysis (PCA) to selection feature. Principal component analysis is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. Principal components are guaranteed to be independent. Principal component analysis was invented by K. Pearson[14].

In this work, through PCA processing the original feartures set transform a new feartures set that according to the contribution rate arrangement from high to low, the new features set equal dimension with original feartures set and new feartures set each dimension is uncorrelated.We select the feature using a simple grid search strategy based on the jackknife test for three datasets.

C. Support vector machine

This work adopts Vapnik's support vector machine [15] to predict the subcellular location of apoptosis proteins. Prediction of protein subcellular location is a multi-classification problem. Therefore, we adopt the multi-class prediction method. Support vector machine using "one-against-others" strategy, given a test protein of unknown category, SVM first map the input vectors into one feature space. Then SVM finds an optimized linear division to solve two-class or multi-class problem in feature space. Finally, a prediction label is assigned to the test sample. In our study, the LIBSVM package is used to implement the SVM classifier. The radial basis function (RBF) is chosen as the kernel function. The regularization parameter c and the kernel width parameter r are optimized on the training set using a grid search strategy in the LIBSVM.

In this work, jackknife test is employed to evaluate the prediction performance of our method. To evaluate the performance of the test, the overall prediction accuracy OA is discussed, and OA calculated as follows:

$$OA = \frac{\sum TP_i}{N} \quad (4)$$

Where TP_i denotes the numbers of the *i*-th subcellular location correctly recognized positives, *N* is the number of all protein sequences in each dataset.

IV. RESULT AND DISCUSSION

Based on the ACF model and SVM classifier, we made a series of experiments to scan the AAindex database in three benchmark datasets. First we transform each apoptosis protein sequence into a fixed-length vector through the ACF model, and then these feature vectors are fed to the SVM classifier to perform the prediction.

In our work, we try to investigate the effect of variation of parameter *m* on the ACF model. According to the above test, we found out the best performance amino acid index and parameter *m*. Finally we use the optimal amino acid index and parameter *m* and try to investigate the affection of the novel feature selection technique based on principal component analysis (PCA). In this work, the best prediction accuracies in dataset ZD98, ZW225 and CL317 are reaches 92.9%, 82.2% and 89.9%, respectively.

A. Performance of three datasets in the AAindex database

On ACF model the parameter *m* have to selected, in this work we select five (10, 20, 30, 40, 50) of *m* value to test the performance. With the ACF model and SVM classifier to scan the AAindex database in datasets ZD98, ZW225 and CL317, we can mine that the best performance of amino acid index in each dataset. Table 1 show that when the *m* selected is 50 the

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

OA is best in three datasets, the best OAs in ZD98, ZW225 and CL317 are 92.9%, 81.8% and 89.0%, the accession number in AAindex database is NAKH900110, PONP930101 and CIDH920101, respectively.

More detailed analysis can be seen from Figure 1. Figure 1 shows that the most of index the *OAs* concentrate in 40%-80% in three datasets. In order to better evaluate the stand or fall of each experiments, here we define a variable that is the proportion of high performance index (*OA>80%*) in each experiment. Obviously, the more proportion of high performance index in each experiment, the more significance there is. From Figure 1 we can found out that with the increase of *m*, the proportion of high performance index (*OA>80%*) increases.

TABLE 1. THE RESULT OF SELECT PARAMETER *m* IN THREE DATASET

	The max OA in AAindex database (%)				
	m=10	m=20	m=30	m=40	m=50
ZD98	85.7	89.8	88.8	90.8	**92.9**
ZW225	77.3	80.4	81.8	81.8	**81.8**
CL317	87.3	89.5	88.8	89.0	**89.0**
	Accession number of the max OA in AAindex database				
	m=10	m=20	m=30	m=40	m=50
ZD98	SWER830101	SWER830101	EISD860103	HOPT810101	NAKH900110
ZW225	WOLS870101	SWER830101	CORJ870104	PONP930101	PONP930101
CL317	WOLS870101	JURD980101	COWR900101	COWR900101	CIDH920101
	The average of OA in AAindex database (%)				
	m=10	m=20	m=30	m=40	m=50
ZD98	60.7	65.5	67.5	68.5	**69.1**
ZW225	55.5	59.3	61.2	62.3	**62.8**
CL317	58.4	63.2	65.8	67.2	**67.4**

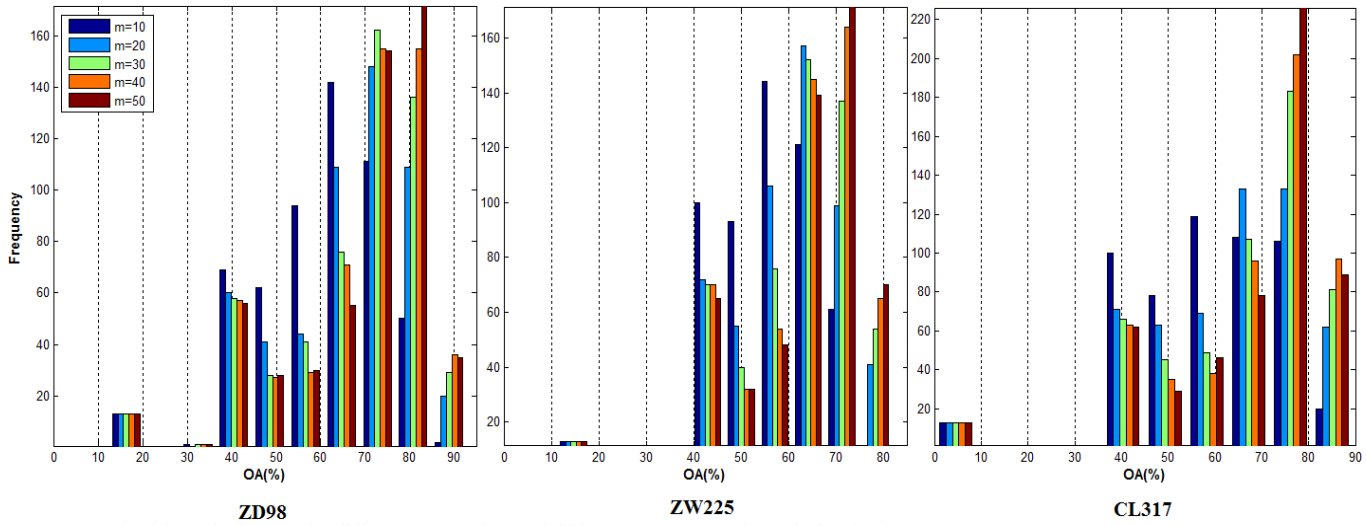


Figure 1. Result of three datasets as the different OA section and different parameter *m* in AAindex database

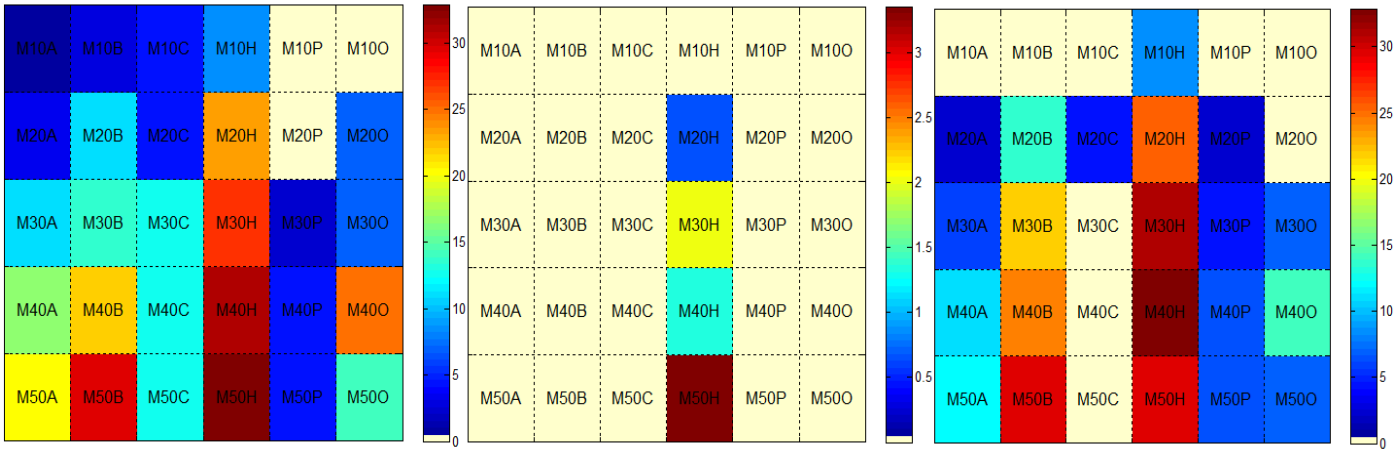
B. Experiment of the six groups on amino acid index

In 1996 Tomii defined the 402 indices classified in six groups A, B, C, H, P and O, number of amino acid index in each group is: A=118, B=37, C=24, H=149, P=46, O=28. On this basis we try to compare the performance among the six groups.

Figure 2 is the heat map representation of the result that combinations parameter *m* of each group of amino acid index in ZD98, ZW225 and CL317 dataset. The result shows that when the group selection is H the proportion of the high performance index is higher than other group, in ZD98 and ZW225 when the *m=50* and group is H the proportion of the

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

high performance index is highest, in CL317 when the *m=40* and group is H the proportion of the high performance index is highest.



ZD98 **ZW225** **CL317**

Figure 2. The different of parameter *m* in ACF model performance at six group of amino acid index, test in three dataset. Here M10, M20, M30, M40, M50 is represent the parameter *m =(10,20,30,40,50)* in ACF model, A, B, C, H, O and P is represent the six group of amino acid index, here M10A is record combination of *m=10* and the amino acid index group is A. The colored grid represent the performance of each combination of *m* and the group, the grid of color more close to red explain that the more proportion of the high performance index (*OA> 80%*) in that group.

According to the above test, we sought out the optimal parameter *m* is 50. With the ACF model and SVM classifier to scan the six groups of amino acid index, test in dataset ZD98, ZW225 and CL317, we can mine that the best performance of amino acid index in each dataset, the result is shown in Table2. When the *m* selects is 50 the best *OA* is in group H, the best *OA* in ZD98, ZW225 and CL317 dataset reaches 92.9%, 81.3% and 89.0%, respectively, and their accession numbers in AAindex database is NAKH900110, CIDH920104 and CIDH920101, respectively.

TABLE 2. THE PERFORMANCE IN THE SIX GROUPS OF AMINO ACID INDEX

	The max OA in six groups (%)					
	A	B	C	H	P	O
ZD98	87.6	84.7	83.7	**92.9**	82.7	83.7
ZW225	77.3	78.2	72.4	**81.3**	75.1	71.6
CL317	84.7	86.7	79.0	**89.0**	83.0	80.3
	Accession number of the max OA in six groups					
	A	B	C	H	P	O
ZD98	QIAN880133	OOBM850101	DAYM780101	NAKH900110	CHAM830105	CHAM830104
ZW225	RICJ880112	QIAN880120	NAKH900109	CIDH920104	FAUJ880106	CHAM830104
CL317	ROBB760108	LIFS790102	CHAM830108	CIDH920101	WOLS870102	ISOY800105

C. Effect of feature selection process

According to the above of experiment, we found out the optimal of amino acid index in each dataset and parameter *m=50* in ACF model. Finally we use the amino acid index NAKH90011, PONP930101, CIDH920101 in ZD98, ZW225 and CL317 experiment, respectively; the numerical values of three amino acid indexes are show in Table 4.

TABLE 3. OVERALL ACCURACY AND DIMENSION OF ORIGINAL FEATURES AND NEW FEATURES FOR THREE DATASETS

	Original features		**New features after PCA**	
	OA (%)	Dimension	OA (%)	Dimension
ZD98	92.9	50	**92.9**	**29**
ZW225	81.8	50	**82.2**	**32**
CL317	89.0	50	**89.9**	**46**

In feature selection process, first, we transform each apoptosis protein sequence into a fixed-length vector through the ACF model, and then principal component analysis applied to reduction the features dimension. Through the PCA

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

process the raw features set can transform a new features set that contains more information and less dimensions. Finally these new feature vectors are fed to the SVM classifier to perform the prediction. Table 3 shows the result of PCA process, through the PCA process the dimension are reduced to 29, 32 and 46 in ZD98, ZW225 and CL317, respectively. In ZW225 and CL317 the *OAs* improved to 82.2% and 89.9% and improved by 0.4% and 0.9%, respectively.

TABLE 4. THE NUMERICAL VALUE OF THREE SETS AMINO ACID INDEX IN 20 AMINO ACIDS

Amino acid	Symbol	NAKH90011	PONP930101	CIDH920101
Alanine	A	0.34	0.85	-0.45
Leucine	L	0.52	1.99	1.29
Arginine	R	-0.57	0.20	-0.24
Lysine	K	-0.75	-1.19	-0.36
Asparagine	N	-0.27	-0.48	-0.20
Methionine	M	0.47	1.42	1.37
Aspartate	D	-0.56	-1.10	-1.52
Phenylalanine	F	1.30	1.69	1.48
Cysteine	C	-0.32	2.10	0.79
Proline	P	-0.19	-1.14	-0.12
Glutamine	Q	-0.34	-0.42	-0.99
Serine	S	-0.20	-0.52	-0.98
Glutamate	E	-0.43	-0.79	-0.80
Threonine	T	-0.04	-0.08	-0.70
Glycine	G	0.48	0.00	-1.00
Tryptophan	W	0.77	1.76	1.38
Histidine	H	-0.19	0.22	1.07
Tyrosine	Y	0.07	1.37	1.49
Isoleucine	I	0.39	3.14	0.76
Valine	V	0.36	2.53	1.26

V. CONCLUSIONS

In this paper, three datasets are scanned the AAindex database, based on the ACF model and SVM classifier, and we made a series of experiments to mine the knowledge when we using amino acid index to predict apoptosis protein subcellular location.

Through this work, we successfully to mined the knowledge that the hydrophobicity of amino acid index is sensitive to apoptosis protein subcellular location prediction, and the amino acid indexes NAKH90011, PONP930101, CIDH920101 that belong to the hydrophobicity group have the best performances in ZD98, ZW225 and CL317, respectively. This knowledge can provide reference for other researchers in protein prediction task. In addition, we propose a novel feature selection technique based on principal component analysis, the result shows that this method can effectively reduce the feature dimension and it can even improve prediction accuracy. In future work, we will fuse the other information with the amino acid index to explore more effective schemes of apoptosis protein subcellular location prediction and improve the prediction accuracy.

ACKNOWLEDGMENT

We appreciate the financial support of this work that was provided by Zhejiang Provincial Natural Science Foundation of China (No. LY12F02043). This work was also partially supported by the National Natural Science Foundation of China (No. 61272312, 61170316, 61170110).

REFERENCES

[1] M.D. Jacobson, M. Weil, and M.C. Raff, Programmed cell death in animal development, Cell, 1997, 88(3): p. 347-354.

[2] J.M. Adams, and S. Cory, The Bcl-2 protein family: arbiters of cell survival, Science, 1998. 281(5381): p. 1322-1326.

[3] J.C. Reed, and G. Paternostro, Postmitochondrial regulation of apoptosis during heart failure, Proc. Natl. Acad. Sci. USA, 1999, 96(14): p. 7614-7616.

[4] J.B. Schulz, M. Weller, and M.A. Moskowitz, Caspases as treatment targets in stroke and neurodegenerative diseases, Ann. Neurol, 1999, 45(4): p. 421-429.

[5] G.P. Zhou, and K. Doctor, Subcellular location prediction of apoptosis proteins, Proteins, 2003, 50(1): p. 44-48.

[6] K. Nakai, A. Kidera, and M. Kanehisa, Cluster analysis of amino acid indices for prediction of protein structure and function, Protein. Eng, 1988, 2(2): p. 93-100.

[7] Z.H. Zhang, et al, A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine, FEBS. Lett, 2006, 580(26): p. 6169-6174.

[8] Y.L. Chen, and Q.Z. Li, Prediction of the subcellular location of apoptosis proteins, J. Theor. Biol, 2007, 245(4): p. 775-783.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

[9] S. Kawashima, H. Ogata, and M. Kanehisa, AAindex: Amino Acid Index Database, Nucleic. Acids. Res, 1999, 27(1): p. 368-369.

[10] S. Kawashima, and M. Kanehisa, AAindex: amino acid index database, Nucleic. Acids. Res, 2000, 28(1): p. 374.

[11] S. Kawashima, et al, AAindex: amino acid index database, progress report 2008, Nucleic. Acids. Res, 2008, 36(Database issue): p. D202-205.

[12] K. Tomii, and M. Kanehisa, Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins, Protein. Eng, 1996, 9(1): p. 27-36.

[13] W.S. Bu, et al, Prediction of protein (domain) structural classes based on amino-acid index, Eur. J. Biochem, 1999, 266(3): p. 1043-1049.

[14] K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space". in Philosophical. Magazine, 1901.

[15] V. Vapnik, The nature of statistical learning theory, Springer. Verlag, 2000.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE