MATCH
Communications in Mathematical
and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **67** (2012) 867-878

ISSN 0340 - 6253

Prediction of Subcellular Localization for Apoptosis Protein: Approached with a Novel Representation and Support Vector Machine

Guiqian Jian, Yusen Zhang*, Panpan Qian

School of Mathematics and Statistics, Shandong University at Weihai
Weihai 264209, China

(Received June 6, 2011)

Abstract. Apoptosis proteins play a crucial role in the development and homeostasis of an organism. Obtaining information about subcellular location of these proteins is very important to understand the mechanism of programmed cell death. In this paper, based on the hydropathy characteristics, we introduce the frequency of 2-blocks and pK value of the α-NH_3^+ group of 2-blocks. By using the new representation for apoptosis protein sequence and support vector machine, we predict subcellular location of 317 apoptosis proteins in jackknife test. The overall prediction accuracy is 91.80% which is higher than other existing algorithms. Furthermore, another dataset containing 98 apoptosis proteins is examined in the same method. The overall predicted successful rate is 94.85%. The promising results indicate that our method may play a complementary role for predicting subcellular location of apoptosis protein.

1 Introduction

Apoptosis proteins play an important role in the growth and dynamic equilibrium of an organism. It can regulate the balance between cell proliferation and death [1]. A range of diseases may outbreak when apoptosis proteins are out of tune, such as cancer [2, 3], autoimmune diseases [4] and neurodegenerative disease [5]. Since the function of an apoptosis protein is closely related with its subcellular location [5, 6], prediction of subcellular location is very necessary. However, for large protein sequences, it is both time consuming and costly to predict subcellular location by doing biochemical experiments.

*Corresponding author: zhangys@sdu.edu.cn

It demonstrates that using information of protein primary structure to predict subcellular location of apoptosis proteins will be more economic and fast.

Many efforts have been made to develop prediction methods of subcellular location of proteins. However, research for predicting subcellular location of apoptosis proteins appears relatively late [7–15]. With the number of known apoptosis proteins increasing, developing a faster and accurate prediction method is necessary. Recently, more and more scholars began to engage in research in this area. Zhou and Doctor [16] firstly provided a method (covariant discriminant algorithm) for predicting subcellular location of apoptosis proteins. Their data set only consisted of 98 protein sequences with four kinds of subcellular locations. The overall accuracy achieved 72.5% in jackknife test. Zhang et al. [17] proposed a novel approach (group weight coding method, EBGW_SVM) in the expanded 151 and 225 protein sequences data set with other four kinds of subcellular locations. The overall accuracies achieved 91.4% and 83.1% in jackknife test separately. Later, many methods were proposed using support vector machine. Chen and Li [18] combined the increment of diversity algorithm with support vector machine (ID_SVM) in the new 317 protein sequences data set provided into six kinds of subcellular locations. Their prediction accuracy was 84.2% by jackknife test. Zhang et al. [19] proposed a novel approach (DF_SVM) by combining the distance frequency and support vector machine using the same data set and test method with Chen and Li. They got an overall predictive accuracy of 88.0%. Lin et al. [20] utilized the Chou's pseudo amino acid composition and achieved the accuracy of 91.1%. Though the overall predictive accuracies have been improved for apoptosis proteins using existed methods, the representation of protein sequence was mainly composed of the amino acid frequency [21] or sequence-order information [19]. PseAAC can represent a protein sequence with a discrete model yet without completely losing its sequence order information, but the calculation is a little complicated.

In this paper, based on the hydropathy distribution information, the frequency and pK value of the α-NH_3^+ group of 2-blocks are proposed. With the novel representation including the frequencies and pK values of 20 native amino acids and the relative frequencies and pK values of 16 2-blocks, a protein sequence can be converted into fixed-dimensional feature vector and each element in it is calculated using the frequency multiplied by the corresponding pK value of the α-NH_3^+ group. Although there are lots of classifiers which can solve the protein classification problem [22,23] and here we choose the support vector machine which can get higher prediction accuracy [24–28]. The support vector machine is utilized to solve the multiple classification problem and the jackknife cross-validation is applied to examine the predictive ability of method. Two data sets, CL317 and ZD98 are used to examine our method. The overall prediction accuracies are improved, which imply that the proposed method is a simple but efficient model for predicting apoptosis protein subcellular location .

2 DataSet

The 317 apoptosis proteins (CL317) extracted from Swiss-Prot 49.0 can be classified into six subcellular locations: 112 cytoplasmic proteins, 55 membrane proteins, 34 mitochondrial proteins, 17 secreted proteins, 52 nuclear proteins and 47 endoplasmic reticulum proteins. The distribution of the sequence identity percentage is 40.1% with $\leq 40\%$ sequence identity, 15.5% with sequence identity from 41% to 80%, 18.9% with sequence identity from 81% to 90% and 25.6% with $\geq 91\%$ sequence identity [29].

In addition, the 98 apoptosis proteins (ZD98) extracted from SWISS-PROT data bank containing 43 cytoplasmic proteins, 30 plasma membrane–bound proteins, 13 mitochondrial proteins and 12 other proteins [16] is also used to estimate the effectiveness of the method.

3 Methods

The accuracy of predicting apoptosis protein subcellular localization is mainly depended on the following two aspects. The first is the representation vector of a protein sequence and the second is the classifier for prediction. Our new representation for apoptosis protein is a 36-dimensional vector. Each element in the first 20-dimensional vector contains the frequency and the pK value of the α-NH_3^+ group of the 20 native amino acids while each element in the last 16-dimensional vector contains the frequency and pK value of the α-NH_3^+ group of hydropathy blocks. Then, we apply the feature vector to predict apoptosis proteins with support vector machine on two datasets.

3.1 Representation of protein sequence

It was demonstrated that the patterns of hydrophobic and hydrophilic residues play an important part in the definition of global protein structure [30]. Among all the physicochemical properties of amino acids in protein sequences, such as polarity, solubility, hydropathy, and so forth, hydropathy (the patterns of hydrophilicity and hydrophobicity) is known to be well conserved during the evolution process. Furthermore, the hydropathy patterns presenting in the protein sequences are used to develop reduced amino acid alphabets for protein secondary structure prediction [31,32]. A protein sequence of length n can be defined as a linear succession of n symbols from the 20-letter amino acid alphabet {R, D, E, N, Q, K, H, L, I, V, A, M, F, S, Y, T, W, P, G, C}. According to the hydropathy scale, we divide the 20 basic amino acids into four groups [33]. Each group is denoted with a letter. Let L, B and W denote strongly hydrophilic amino acids, strongly hydrophobic amino acids and weakly hydrophilic or weakly hydrophobic amino acids, respectively. As the amino acids P, G and C have unique backbone properties, we divide them into a single group, denoted by P. In this way, a protein primary sequence can be converted into a four-letter sequence just like DNA sequence without considering the differences of representative letter. The four-letter sequence may be regarded as a simplification of the protein primary sequence which makes it easier to extract features from the sequence. We show the classifications of 20 amino acids in table 1.

Table 1: Classification of basic amino acids

Classification	Group	Amino acids
Strongly hydrophilic or polar	L	R, D, E, N, Q, K, H
Strongly hydrophobic	B	L, I, V, A, M, F
Weakly hydrophilic or weakly hydrophobic (ambiguous)	W	S, T, Y, W
Proline or Glycine or Cysteine	P	P, G, C

For a protein sequence S, suppose the size of L be L_L, the size of B be L_B, the size of W be L_W , and the size of P be L_P. Actually, from the above table, we can see that L_L is 7, L_B is 6, L_W is 4, L_P is 3 and $L_L + L_B + L_W + L_P$ is 20. Before simplifying a protein sequence, we can obtain the number and frequency of the 20 amino acids denoted by n_i^j and f_i^j, with

$$f_i^j = n_i^j/n$$

where i represents the class L, B, W or P, n_1^j or $f_1^j(j = 1, \ldots, L_L)$ represents the number or frequency of the jth amino acids of Class L occurring in sequence S, n_2^j or $f_2^j(j = 1, \ldots, L_B)$ represents the number or frequency of the jth amino acids of Class B occurring in sequence S, n_3^j or $f_3^j(j = 1, \ldots, L_W)$ represents the number or frequency of the jth amino acids of Class W occurring in sequence S, n_4^j or $f_4^j(j = 1, \ldots, L_P)$ represents the number or frequency of the jth amino acids of Class P occurring in sequence S. Also, we can define the the frequencies of L, B, W, P labeled as f_i with

$$f_i = \frac{1}{n}\left(\sum_{j=1}^{n_i} n_i^j\right)$$

where $i = 1, \ldots, 4, n_1 = L_L, n_2 = L_B, n_3 = L_W, n_4 = L_P$.

By considering the successive groups in a simplified four-letter protein sequence just like the dinucleotide in DNA [34–36], we can obtain sixteen binary groups (We call it 2-blocks){LL, LB, LW, LP, BL, BB, BW, BP, WL, WB, WW, WP, PL, PB, PW, PP}. We can count the number of the occurrences of linear succession 2-blocks by taking a 2-letter sliding window that is run through the sequences, from position 1 to n-1. Consequently, the number of occurrences of the 2-blocks in protein sequence S can be represented as:

$$\{n_{11}, n_{12}, n_{13}, n_{14}, n_{21}, n_{22}, n_{23}, n_{24}, n_{31}, n_{32}, n_{33}, n_{34}, n_{41}, n_{42}, n_{43}, n_{44}\}$$

Every elements in the above vector divided by n-1, we can get the frequency of 2-blocks as follows:

$$f_{ij} = n_{ij}/(n-1), i, j = 1, 2, 3, 4$$

where n-1 is the total number of 2-blocks in protein S.

The value of pK of the α-NH$_3^+$ group is one of the important physicochemical properties of amino acids. We use $pK_i(i = 1, \ldots, 20)$ to represent pK values of the α-NH$_3^+$ group of the 20 amino acids in the sequence R, D, E, N, Q, K, H, L, I, V, A, M, F, S, Y, T, W, P, G, C. We define the pK values of the α-NH$_3^+$ group of class L,B,P,W as follows:

$$pK_L = \sum_{i=1}^{L_L} pK_i/L_L, pK_B = \sum_{i=L_L+1}^{L_L+L_B} pK_i/L_B$$

$$pK_W = \sum_{i=L_L+L_B+1}^{L_L+L_B+L_W} pK_i/L_W, pK_P = \sum_{i=L_L+L_B+L_W+1}^{L_L+L_B+L_W+L_P} pK_i/L_P$$

where pK_L, pK_B, pK_W, pK_P represent the average pK values of the α-NH$_3^+$ group in class L,B,P,W separately. In the same way, we can get the pK values of the α-NH$_3^+$ group of the sixteen binary groups (2-blocks) as follows:

$$pK_{11} = (pK_L + pK_L)/2, pK_{12} = pK_{21} = (pK_L + pK_B)/2$$

$$pK_{22} = (pK_B + pK_B)/2, pK_{23} = pK_{32} = (pK_B + pK_W)/2$$

$$pK_{24} = pK_{42} = (pK_B + pK_P)/2, pK_{33} = (pK_W + pK_W)/2$$

$$pK_{34} = pK_{43} = (pK_W + pK_P)/2, pK_{44} = (pK_P + pK_P)/2$$

$$pK_{13} = pK_{31} = (pK_L + pK_W)/2, pK_{14} = pK_{41} = (pK_L + pK_P)/2$$

Consequently, the protein sequence S can be converted into a 36-dimensional vector:

$V = [pK_1 * f_1^1, ..., pK_7 * f_1^7, pK_8 * f_2^1, ..., pK_{13} * f_2^6, pK_{14} * f_3^1, ..., pK_{17} * f_3^4, pK_{18} * f_4^1, ..., pK_{20} * f_4^3, pK_{11} * f_{11}, ..., pK_{14} * f_{14}, pK_{21} * f_{21}, ..., pK_{24} * f_{24}, pK_{31} * f_{31}, ..., pK_{34} * f_{34}, pK_{41} * f_{41}, ..., pK_{44} * f_{44}]$

where every part of the fixed-dimensional vector is the pK value multiplied by the corresponding frequency.

Finally, we get a feature vector including 20 native amino acids frequency, 16 2-blocks frequency, pK values of the α-NH$_3^+$ group of 20 native amino acids and pK values of 16 2-blocks, which is simple but has enough information. In this way, all the proteins are 36-D and we need not consider the influence of the different lengths of the protein sequences.

3.2 Support vector machine

SVM is not only a kind of machine learning method based on statistical learning theory [37] but is also superior in practical applications. As a supervised machine learning technology, it has been successfully used in wide fields of bioinformatics by transforming the input vector into a high-dimension Hilbert space and to seek a separating hyperplane in this space.

In general, One-Versus-Rest (OVR) and One-Versus-One (OVO) are the most commonly used approach for solving multi-class problems by reducing a single multiclass problem into multiple binary problems. This paper use the OVO strategy. For a k-classification problem, the OVO strategy constructs $k \times (k-1)/2$ classifiers with each one trained with the data from two different classes. The software used to implement

SVM is LibSVM written by Lins lab and can be freely downloaded from: http://www.csie.ntu.edu.tw/*cjlin/libsvm [38].

There are some common kernel functions, including polynomial kernel, radial basis function, Gaussian radial basis function and sigmoid function. Here, the RBF is used for all our calculations. The regularization parameter C and the kernel parameter γ of the RBF must be determined in advance, which will be discussed in results and discussion section.

3.3 Assessment of prediction performance

In order to evaluate a predictive algorithm, selecting a test method is an important issue. There are three cross-validation tests often used in evaluating the prediction performance: independent dataset test, sub-sampling (such fivefold or tenfold sub-sampling) test, and jackknife test. Of these three examine method, the jackknife test is deemed the most rigorous and objective one [39]. The jackknife test reflects ability of the

extrapolation of a prediction method. For each test sequence, the rule parameters are extracted from the remaining sequences. Every time the tested sequence should be put into the data set and single out another one to test. Therefore, we explore the jackknife test to examine proposed method.

The individual sensitivity S_{in}, individual specificity S_{ip}, Matthew's correlation coefficient MCC_i, and overall prediction accuracy A_c are used to measure the prediction performance of our work. The definition is shown as follows:

$$S_{in} = TP_i/(TP_i + FN_i)$$

$$S_{ip} = TP_i/(TP_i + FP_i)$$

$$MCC_i = \frac{(TP_i \times TN_i) - (FP_i \times FN_i)}{\sqrt{(TP_i + FP_i) \times (TN_i + FN_i) \times (TP_i + FN_i) \times (TN_i + FP_i)}}$$

$$A_c = \sum_{i=1}^{n} TP_i/N$$

where TP_i denotes the number of true positives in ith subcellular location, TN_i denotes the number of true negatives in ith subcellular location, FP_i denotes the number of false positive, FN_i denotes the number of false negative in ith subcellular location and N is the number of all the protein sequences.

4 Results and discussion

Chen and Li constructed the 317 data set and predicted the subcellular location of the apoptosis proteins by increment of diversity method (ID) [29] and ID with support vector machine (ID_SVM) [18]. The overall success rates were 82.7 % and 84.2 % for jackknife test, respectively. Zhang et al. [19] achieved the overall accuracy of 88.0% with the distance frequency and support vector machine method. Lin et al. [20] got the accuracy of 91.1% by use of Chou's pseudo amino acid composition (PseAAC). In our method, we examine a great deal of parameters of SVM (C and γ) by using jackknife cross-validation. For the current study, we found that, when C=1000 and γ=0.99, the predicted successful rate is 91.8% which is the highest. The results on dataset CL317 are listed in Table 2. The compared results with other methods are shown in Table 3. Especially, for membrane protein, mitochondrial protein, nuclear protein and endoplasmic protein, sensitivities are higher than other methods. The results of comparison with other results on dataset ZD98 [16] are exhibited in Table 4. By lots of examination, we select C=30 and γ=0.29 for this prediction. However, one protein is missing now and we use the rest 97 proteins to test. The results show that the predictive successful rate of our method is 94.9%. The results of comparison with different methods on different datasets indicate that our our method is effective for predicting the subcellular localization of apoptosis protein.

Table 2: The prediction results on dataset CL317 in Jackknife test

Subbcellular location	Jackknife test		
	S_n(%)	S_p(%)	MCC
Cytoplasmic	92.9	92.0	0.88
Membrane	94.6	88.1	0.89
Mitochondrial	88.2	90.9	0.88
Secreted	70.6	92.3	0.80
Nuclear	92.3	92.3	0.91
Endoplasmic	95.7	95.7	0.95
A_c(%)	91.8		

SVM: C=1000, γ=0.99.

5 Conclusion

A good representation of protein and powerful classifier are important for predicting apoptosis protein subcellular localization. Hydropathy is one of important physicochemi-

Table 3: Comparison of predicting results on the 317 apoptosis proteins data set with a jackknife test

Algorithm	*Sn*%						
	Cyto	Memb	Mito	Secr	Nucl	Endo	Overall
ID[a]	81.3	81.8	85.3	88.2	82.7	83.0	82.7
ID_SVM[b]	91.1	89.1	79.4	58.8	73.1	87.2	84.2
FKNN[c]	92.0	89.1	85.3	76.5	92.3	93.7	90.2
FKNN[d]	93.8	92.7	82.4	76.5	90.4	93.6	90.9
PseAAC[e]	93.8	90.9	85.3	76.5	90.4	95.7	91.1
DF_SVM[f]	92.9	85.5	76.5	76.5	86.5	93.6	88.0
Our method	92.9	94.6	88.2	70.6	92.3	95.7	91.8

[a] Comes from [29].
[b] Comes from [18].
[c] Comes from [40].
[d] Comes from [41].
[e] Comes from [20].
[f] Comes from [19].

Table 4: Comparison of predicting results on the 98 apoptosis proteins data set with a jackknife test

Algorithm	Sn(%)				
	Cyto	Memo	Mito	Other	Overall
Covariant [a]	$42/43 = 97.7$	$22/30 = 73.3$	$4/13 = 30.8$	$3/12 = 25.0$	$71/98 = 72.5$
20 sqrt-AAC [b]	$37/43 = 86.0$	$27/30 = 90.0$	$13/13 = 100$	$12/12 = 100$	$89/98 = 90.8$
EBGW_SVM [c]	$42/43 = 97.7$	$27/30 = 90.0$	$12/13 = 92.3$	$10/12 = 83.3$	$91/98 = 92.9$
BC[d]	$39/43 = 90.7$	$27/30 = 90.0$	$12/13 = 92.3$	$6/12 = 50.0$	$84/98 = 85.7$
HensBC[e]	$41/43 = 95.3$	$27/30 = 90.0$	$12/13 = 92.3$	$8/12 = 66.7$	$88/98 = 89.8$
Dual-layer [f]	$41/43 = 95.3$	$29/30 = 96.7$	$12/13 = 92.3$	$11/12 = 91.7$	$93/98 = 94.9$
ID[g]	$39/43 = 90.7$	$27/30 = 90.0$	$12/13 = 92.3$	$11/12 = 91.7$	$89/98 = 90.8$
ID_SVM[h]	$41/43 = 95.3$	$28/30 = 93.3$	$11/13 = 84.6$	$7/12 = 58.3$	$87/98 = 88.8$
HHT[i]	$41/43 = 95.3$	$29/30 = 96.7$	$12/13 = 92.3$	$9/12 = 75.7$	$91/98 = 92.9$
FKNN[j]	$41/43 = 95.3$	$29/30 = 96.7$	$13/13 = 100$	$11/12 = 91.7$	$94/98 = 95.9$
PseAAC[k]	$41/43 = 95.3$	$28/30 = 93.3$	$12/13 = 92.3$	$10/12 = 83.3$	$91/98 = 92.9$
DF_SVM[l]	$42/43 = 97.7$	$29/30 = 96.7$	$12/13 = 92.3$	$9/12 = 75.0$	$92/98 = 93.9$
Our method	$42/43 = 97.7$	$29/30 = 96.7$	$11/12 = 91.7$	$10/12 = 83.3$	$92/97 = 94.9$

SVM: C=30, γ=0.29. One protein sequence is missing now.

[a] Comes from [16], by using covariant discriminant function.
[b] Comes from [28], by using 20 sqrt-amino acid composition and SVM.
[c] Comes from [17], by using group weight coding method.
[d] Comes from [42], by using single Bayesian classifier.
[e] Comes from [42], by using hierarchical ensemble of Bayesian classifiers.
[f] Comes from [43], by using Dual-layer SVM.
[g] Comes from [29], by using increment of diversity method.
[h] Comes from [18], by using increment of diversity combined with support vector machine.
[i] Comes from [44], by using Hilbert Huang transform.
[j] Comes from [41], by using fuzzy K-nearest neighbor classifier.
[k] Comes from [20], by using Chou's pseudo amino acid composition (PseAAC) and SVM.
[l] Comes from [19], by using the distance frequency and support vector machine method.

cal properties of amino acids, and is better conserved than protein sequences in evolution. Based on it, we introduce the frequency of 2-blocks and pK value of the α-NH_3^+ group of 2-blocks. Combining them with the frequency and pK value of native amino acids, a novel representative for protein is proposed to predict subcellular location. Using the new feature extraction method and support vector machine, we can reduce dimension of inputting vector, improve calculating efficiency and extract important classify information. Compared with other existed approaches in two datasets, we can see that our method is convenient, effective and powerful in improving the overall predicting accuracy.

Acknowledgments

This work was supported in part by the Shandong Natural Science Foundation (ZR2010AM020).

References

[1] M. D. Jacobson, M. Weil, M. C. Raff, Programmed cell death in animal development, *Cell* **88** (1997) 347–354.

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

[3] G. Evan, T. Littlewood, A matter of life and cell death, *Science* **281** (1998) 1317–1322.

[4] J. C. Reed, G. Paternostro, Postmitochondrial regulation of apoptosis during heart failure, *Proc. Natl. Acad. Sci. USA* **96** (1999) 7614–7616.

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

[6] M. Suzuki, R. J. Youle, N. Tjandra, Structure of Bax: coregulation of dimmer formation and intracellular location, *Cell* **103** (2000) 645–654.

[7] K. C. Chou, D. Elrod, Protein subcellualr location prediction, *Protein Eng.* **12** (1999) 107–118.

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

[9] K. J. Park, M. Kanehisa, Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs, *Bioinformatics* **19** (2003) 1656–1663.

[10] Q. B. Gao, Z. Z. Wang, C. Yan, Prediction of protein subcellular location using a combined feature of sequence, *FEBS Lett.* **579** (2005) 3444–3448.

[11] S. Matsuda, J. P. Vert, N. Ueda, H. Toh, T. Akutsu, A novel repersentation of protein sequences for prediction of subcellular location using sopport vector machines, *Protein Sci.* **14** (2005) 2804–2813.

[12] K. C. Chou, H. B. Shen, Predicting protein subcellular locaton by using multiple classifiers, *J. Cell. Biochem.* **99** (2006) 517–527.

[13] K. C. Chou, H. B. Shen, Large–scale predictions of Gram–negative bacterrial protein subcellular locations, *J. Proteome Res.* **5** (2006) 3420–3428.

[14] K. C. Chou, H. B. Shen, Large–scale plant protein subcellular location prediction, *J. Cell. Biochem.* **100** (2007a) 665–678.

[15] J. Y. Shi, S. W. Zhang, Q. Pan, G. P. Zhou, Using pseudo amino acid composition to predict protein subcellularlocation: approach with amino acid composition distribution, *Amino Acids* **35** (2008) 321–327.

[16] G. P. Zhou, K. Doctor, Subcellular location prediction of apoptosis proteins, *Proteins Struct. Funct. Genet.* **50** (2003) 44–48.

[17] Z. H. Zhang, Z. H. Wang, Z. R. Zhang, Y. X. Wang, A novel method for apoptosis protein subcellular localization prediction combining encoding based on group weight and support vector machine, *FEBS Lett.* **580** (2006) 6169–6174.

[18] Y. L. Chen, Q. Z. Li, Prediction of apoptosis proetin subcellular location using improved hybrid approach and pseudo amino acid composion, *J. Theor. Biol.* **248** (2007b) 377–381.

[19] L. Zhang, B. Liao, D. C. Li, W. Zhu, A novel representation for apoptosis protein subcellular localization prediction using support vector machine, *J. Theor. Biol.* **259** (2009) 361–365.

[20] H. Lin, H. Wang, H. Ding, Y. L. Chen, Q. Z. Li, Prediction of subcellular localization of apoptosis protein using Chou's pseudo amino acid composition, *Acta Biotheor.* **57** (2009) 321–330.

[21] K. C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, *Proteins Struct. Funct. Genet.* **43** (2001) 246-255. (Erratum: K. C. Chou **44** (2001) 60.)

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

[23] H. B. Shen, K. C. Chou, Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram–positive bacterial proteins, *Protein Eng. Des. Sel.* **20** (2007) 39–46.

[24] S. Hua, Z. R. Sun, Support vector machine approach for protein subcellular localization prediction, *Bioinformatics* **17** (2001) 721–728.

[25] C. W. Hsu, C. J. A. Lin, Comparison of methods for multiclass supportvector machines, *IEEE Trans. Neural Networks* **13** (2002) 415–425.

[26] M. Bhasin, G. P. S. Raghava, ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST, *Nucl. Acids Res.* **32** (2004) W414–W419.

[27] A. Garg, M. Bhasin, G. P. Raghava, SVM-based method for subcellular localization of human proteins using amino acid compositions, their order and similarity search, *J. Biol. Chem.* **280** (2005) 14427–14432.

[28] J. Huang, F. Shi, Support vector machines for predicting apoptosis proteins types, *Acta Biotheor.* **53** (2005) 39–47.

[29] Y. L. Chen, Q. Z. Li, Prediction of the subcellular location of apoptosis proteins, *J. Theor. Biol.* **245** (2007) 775–783.

[30] D. P. Goldenberg, Finding the right fold, *Nat. Struct. Biol.* **6** (1999) 987–990.

[31] T. Klingler, D. Brutlag, Discovering structural correlations in a-helices, *Protein Sci.* **3** (1994) 1847-1857.

[32] S. C. Schmidler, J. S. Liu, D. L. Brutlag, Bayesian segmentation of protein secondary structure, *J. Comput. Biol.* **7** (2000) 233–248.

[33] J. Pánek, I. Eidhammer, R. Aasland, A new method for identification of protein (sub)families in a set of proteins based on hydropathy distribution in proteins, *Proteins Struct. Funct. Bioinf.* **58** (2005) 923–934.

[34] Y. S. Zhang, A simple method to construct the similarity matrices of DNA sequences, *MATCH Commun. Math. Comput. Chem.* **60** (2008) 313–324.

[35] W. Chen, Y. S. Zhang, Comparisons of DNA sequences based on dinucleotide, *MATCH Commun. Math. Comput. Chem.* **61** (2009) 533-540.

[36] Y. S. Zhang, W. Chen, A new measure for similarity searching in DNA sequences, *MATCH Commun. Math. Comput. Chem.* **65** (2011) 477–488.

[37] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.

[38] C. C. Chang, C. J. Lin, LIBSVM: a library for support vector machines, 2001.

[39] K. C. Chou, C. T. Zhang, Prediction of protein structural classes, *Crit. Rev. Biochem. Mol. Biol.* **30** (1995) 275–349.

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

[41] Y. S. Ding, T. L. Zhang, Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm–based ensemble classifier, *Pattern Recognit. Lett.* **29** (2008) 1887–1892.

[42] A. Bulashevska, R. Eils, Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains, *BMC Bioinf.* **7** (2006) 298.

[43] X. B. Zhou, C. Chen, Z. C. Li, X. Y. Zou, Improved prediction of subcellular location for apoptosis proteins by the dual–layer support vector machine, *Amino Acids* **35** (2008) 383-388.

[44] F. Shi, Q. J. Chen, N. N. Li, Hilbert Huang transform for predicting proteins subcellular location, *J. Biomed. Sci. Eng.* **1** (2008) 59–63.