MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **73** (2015) 559-571

ISSN 0340 - 6253

A Novel Reduced Triplet Composition Based Method to Predict Apoptosis Protein Subcellular Localization

Shengli Zhang [a,1], **Yunyun Liang** [a], **Zhenguo Bai**

School of Mathematics and Statistics, Xidian University, Xi'an 710071, P. R. China

(Received May 1, 2014)

Abstract

Apoptosis, or programmed cell death, plays an important role in development of an organism. Obtaining information on subcellular location of apoptosis proteins is very helpful to understand the apoptosis mechanism. In this paper, based on the hydropathy characteristics, a novel feature extraction method with triplet composition features for reduced protein sequence is presented, and applied to apoptosis protein subcellular localization prediction associated with support vector machine. The experiment results show that the new feature extraction method is efficient to extract the information implicated in protein sequence and the method has reached a satisfied performance despite its simplicity. The overall prediction accuracy of our proposed method on dataset ZD98 reaches 96.9% in Jackknife test. For the dataset ZW225, the overall prediction accuracy achieves 89.8%. Comparison with other existing methods shows that RTC_SVM model is a very promising prediction model for apoptosis protein subcellular localization and may at least play an important complementary role to existing methods.

1 Introduction

Apoptosis, or programmed cell death, is a fundamental process controlling normal tissue homeostasis by regulating a balance between cell proliferation and death[1]. When

[1]Corresponding author. Tel./Fax:+86-29-88202860. E-mail:shengli0201@163.com
[a]Both authors contributed equally.

apoptosis malfunctions, a variety of formidable disease can ensue: blocking apoptosis is associated with cancer, and autoimmune disease, whereas unwanted apoptosis can possibly lead to ischemic damage or neurodegenerative disease[2]. Obtaining information about subcellular location of apoptosis proteins will help us to understand the apoptosis mechanism and functions of proteins. This is because subcellular location of proteins are strongly correlated to their function. Thus, the determination of apoptosis protein subcellular location remains a quite meaningful topic in bioinformatics and computational biology. However, it is both time-consuming and expensive to do so based solely on experimental techniques. Hence there exists a critical challenge to develop automated methods for fast and accurate determination of the locations of proteins. Therefore there is a need to develop reliable and effective computational methods for identifying the subcellular classes of newly found proteins based on their primary sequences.

During the last decade, many different algorithms and efforts have been made to address this problem. There are generally two aspects in the computational prediction: feature vector and classification algorithm. Various sequence features have been applied to represent protein sequences. Zhou and Doctor[3] 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. Bulashevska and Eils[4] used the same dataset with Zhou and Doctor[3] to predict subcellular location of apoptosis proteins by using hierarchical ensemble of Bayesian classifiers. Zhang et al.[5] 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. In addition, many methods were proposed using support vector machine. Chen and Li[6] combined the increment of diversity algorithm with support vector machine (ID_SVM) in the 317 protein sequences data set provided into six kinds of subcellular locations. Zhang et al.[7] proposed a novel approach (DF_SVM) by combining the distance frequency and support vector machine. Jian et al.[8] proposed a novel approach (2-BF_SVM) by the frequency of 2-blocks and pK value of the α-NH_3^+ group of 2-blocks and achieved the higher accuracy. 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[9–13] or sequence-order information[7, 14–19].

In this study, we develop a new computational method to predict apoptosis protein subcellular localization by using the triplet composition features of reduced protein sequences which is obtained by the coding of the amino acids. Finally, the support vector machine (SVM) classifier is employed to perform the prediction. Jackknife cross-validation tests on two widely used benchmark datasets show that our method presents high prediction accuracies in comparison with other existing methods.

2 Materials and methods

2.1 Datasets

To compare the accuracy of our prediction with those of existing prediction methods, the datasets constructed by Zhou[3] and Chen[20] are adopted in our work. Proteins in those datasets are extracted from Swiss-Prot[21]. The ZD98 consisted of 98 apoptosis protein sequences, 43 of which are cytoplasmic proteins(CY), 30 plasma membrane-bound proteins(ME), 13 mitochondrial proteins(MI) and 12 other proteins(OTHER). The accession numbers can be referred to [3, 20]. The other much larger dataset, ZW225, is also adopted to further test the prediction model. The ZW225 dataset included 225 apoptosis proteins in four subcellular localizations with 41 nuclear proteins(NU), 70 cytoplasmic proteins(CY), 25 mitochondrial proteins(MI) and 89 membrane proteins(ME). All the proteins in this dataset were selected from Swiss-Prot[21] using the same selection rule as ZD98. Employed benchmarks in this study and the number of proteins belonging to each class are shown in Table 1.

Table 1: The properties of ZD98 and ZW225 benchmarks.

Benchmark	CY	ME	MI	OTHER	Total
ZD98	43	30	13	12	98
Benchmark	NU	CY	MI	ME	Total
ZW225	41	70	25	89	225

2.2 Feature extraction methods

One of the key steps in developing a powerful predictor for a protein system is to formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the attribute to be predicted[22].

Structural similarity among proteins is reflected in the hydrophilic/hydrophobic distribution along the amino acids in the protein sequence. Similarities in the hydropathy distributions are obvious for homologous proteins within a protein family. They are also observed for proteins with related structures, even when sequence similarities are undetectable.

Mutations happen in a more or less random manner at the molecular level, while selections shape the direction of evolution. From the perspective of molecular evolution, k-word frequency may reflect both the results of random mutation and selective evolution. The triplet code is how the long strand of DNA is read. A copy of the DNA molecule is translated on the ribosomes into a polypeptide. The ribosomes "read" the DNA in groups of 3 bases at a time. Each 3 bases code for a certain amino acid that is incorporated into the polypeptide. In this section, we will explain briefly the reduced coding of amino acids and triplet composition feature extraction strategies.

2.2.1 Reduced coding of amino acids

Research shows that physical and chemical properties of amino acids have an important impact on the protein subcellular localization. And it has been confirmed that the hydrophilic/hydrophobic distribution of the amino acid sequence has an important role on the formation of protein structure. On the basis of the idea that structure determines function, we believe that the subcellular localization of proteins are closely related to the hydrophilic/hydrophobic distribution of the amino acid sequence. So we adopt a new reduced coding method of protein sequence based on the above assumption[23].

To obtain the hydropathy characteristics, the amino acids are divided into groups using their individual hydropathies according to the ranges of the hydropathy scale (Table 1). Each group is characterized by that hydropathy characteristic that is common for hydropathy of all amino acids in the group. Supposing that proteins composed of 20 amino acids with a single-letter code can be expressed as:

$$AA = A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, Y$$

Where A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V or Y represents one of 20 amino acids respectively. According to the hydrophilicity/hydrophobicity of the amino acid residues, they can be divided into strongly hydrophilic, highly hydrophobic, weakly

hydrophilic or weakly hydrophobic amino acids and other categories, Because the proline, glycine and cysteine have the special role on the protein structure, they are considered as their respective category. In view of this classification, 20 kinds of amino acids are divided into six categories which each category with a single letter, the six classifications are shown in Table 2. According to the rule in Table 2, the representation of protein sequences is characterized by the six letters R, B, W, P, G and C. For example, given a protein sequence $S = SCWMFKDGNAL$, we can get its reduced sequence $WCWBBRRGRBB$. In the following study, we extract the features based on the reduced protein sequence.

Table 2 Reduced coding of the amino acids

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

2.2.2 Triplet composition

Triplet composition(TC) is defined by three consecutive residue occurrence frequencies. In this study, the TC of the reduced protein sequence, which has been previously obtained according to the hydrophilic/hydrophobic distribution, is calculated. The method is as follows:

Let $N_{m\times n\times k}$ be the number of the consecutive triplets that appear in the amino acid sequence, then

$$N_{m\times n\times k} = \sum_{i=1}^{N-2} I_{m,n,k}(i, i+1, i+2), (m, n, k = 1, 2, \cdots, 6) \tag{2.1}$$

where $I_{m,n,k}(i, i+1, i+2)$ can be expressed as:

$$I_{m,n,k}(i, i+1, i+2) = \begin{cases} 1 & m, n, k \; represent \;\; i, i+1, i+2-th \; residue; \\ 0 & else. \end{cases} \tag{2.2}$$

Triplet composition(TC) is represented by three consecutive residue occurrence frequencies in the reduced protein sequence, that is to say, the reduced protein sequence is

mapped into a point on the $6 \times 6 \times 6 = 216$ dimensional Euclidean space, its reduced triplet composition vector(RTC) representation is expressed as follows:

$$RCT(S) = (p_1, p_2, p_3, \cdots, p_{216}), \tag{2.3}$$

where p_i can be calculated as follows:

$$p_i = \frac{N_{m\times n\times k}}{N-3+1}.(i = 1, 2, \cdots, 216) \tag{2.4}$$

2.3 Support vector machine

Support vector machine(SVM) is assumed to be a very powerful algorithm that often achieves superior classification performance in comparison with other classification algorithms [24]. The basic idea of SVM is to map data of samples into a high dimensional Hilbert space and to seek a separating hyperplane in this space. There are total four protein subcellular classes, and prediction of protein subcellular class is therefore a four-classification problem. So we adopt the multiclass prediction method, SVM using 'One-Versus-One' strategy. Usually, four kinds of kernel functions, i.e. linear function, polynomial function, sigmoid function and radial basis function(RBF), can be available to perform prediction. Empirical studies have demonstrated that the RBF outperforms the other three kinds of kernel functions [25, 26]. RBF kernel is defined as follows:

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma\|\mathbf{x}_i-\mathbf{x}_j\|^2}, \tag{2.5}$$

where γ is the kernel parameter, $\mathbf{x_i}$ and $\mathbf{x_j}$ are input feature vectors. In this study, we choose the RBF to perform prediction. The regularization parameter C and kernel parameter γ are optimized based on fifteen-fold cross-validation using a grid search strategy in the LIBSVM software [27].

Here, we determine the values of C and γ by aiming to achieve the highest overall prediction accuracy as possible. For this purpose, a simple grid search strategy is adopted, where C is allowed to take a value only between 2^{-5} to 2^{15} and γ only between 2^{-15} to 2^5 (fifteen-fold cross-validation). We use the dataset $ZD98$ to compute the overall prediction accuracies for different combinations of C and γ. By the above grid search, we find that the highest accuracy is obtained with the combination of C=1.4142 and γ=32. So we choose C=1.4142 and γ=32 in our experiments.

2.4 Prediction assessment

In statistical prediction, three cross-validation methods that are independent dataset test, sub-sampling test and jackknife test often can be used to examine a predictor for its effectiveness in practical application. Among these three methods, the jackknife test has been increasingly and widely used to examine the performance of various predictors [28–30]. Hence, the jackknife cross-validation is utilized to examine the power of our method. During the process of the jackknife test, each protein sequence in the dataset is singled out in turn as a test sample, and the predictor is trained by the remaining protein sequences.

The following measures are used to assess the performance of the classifiers used in this study. The definition is showed as follows:

$$\begin{aligned} Precision &= TP/(TP+FP)(P) \\ Recall\ or\ Sensitivity &= TP/(TP+FN)(R,\ Sn) \\ Specificity &= TN/(TN+FP)(Sp) \\ Overall\ accuracy &= (TP+TN)/(TP+FN+FP+TN)(OA) \end{aligned}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives and FN is the number of false negatives, respectively. OA is the overall accuracy for all the classes. The Matthews correlation coefficient (MCC) value takes account of both over- and under-predictions and is between -1 and 1.

3 Results and discussion

3.1 Prediction performances of our method

The prediction method is examined with two benchmark datasets including ZD98 and ZW225 by jackknife test and report the Accuracy, Sensitivity, Specificity and MCC for each protein subcellular class, as well as the OA. The results are shown in Table 3.

From Table 3, we can see that the overall accuracies for the two datasets are all above 89%. For the dataset of ZW225, the fourth class is nuclear proteins, while for the dataset ZD98, it is else proteins. Specifically, the overall accuracies of 96.9% and

Table 3 The prediction quality of our method on two test datasets.

Dataset	Class	Sensitivity(%)	Specificity(%)	MCC(%)
ZD98	CY	100	96.4	95.9
	MI	84.6	100	90.9
	OTHER	91.7	100	95.2
	ME	100	98.5	97.7
	OA	**96.9**		
ZW225	NU	87.8	97.3	85.1
	CY	91.4	96.1	87.6
	MI	92.1	95.6	87.9
	ME	80.0	97.0	75.7
	OA	**89.8**		

89.8% are achieved for the datasets ZD98 and ZW225, respectively. If comparing the four subcellular classes to each other, the predictions of the cytoplasmic proteins(CY) and plasma membrane-bound proteins(ME) in ZD98 dataset run up to the highest accuracies 100%. The Specificity values reach 100% both in the mitochondrial proteins(MI) and other proteins(OTHER). Furthermore, the MCC coefficients for all four protein subcellular classes are higher than 90%.

Referring to the ZW225 dataset, our method RTC_SVM also performs satisfactorily with prediction accuracies of 89.8%. Specifically, the overall accuracies of 87.8%, 91.4%, 92.1% and 89.8% are achieved for the classes nuclear proteins(NU), cytoplasmic proteins(CY), mitochondrial proteins(MI) and membrane proteins(ME), respectively. Also the Specificity values for all the four classes reach higher than 95%. Furthermore, the MCC coefficients for all four protein subcellular classes are higher than 75%. The prediction performance for two datasets are variant. This may be owing to the discrepancy of dataset traits, e.g. the size, sequence homologous and unbalance of subset. And the sequence homologous will greatly affect the prediction performance.

Hence, we may conclude that the triplet composition features which are extracted by the reduced protein sequences make their positive contributions to the overall predictions. In other words, as more effective features are involved in the prediction, the overall accuracy are shown to increase steadily. The results show that RTC_SVM is effective and helpful for prediction of apoptosis protein subcellular location.

3.2 Comparison with existing methods

To further evaluate the prediction performance of the current method objectively, we also make comparisons with some previously published methods studied on the same dataset by the jackknife tests. Among the compared methods, we denoted the model with dipeptide composition and increment of diversity algorithm as Dipep_Diver; model with EBGW and SVM as EBGW_SVM; model with the distance frequency and SVM as DF_SVM and model with the frequency of 2-blocks and SVM as 2-BF_SVM. The detailed results are shown in Tables 4 and 5.

Table 4 Performance comparison of different methods on ZD98 dataset.

Method	Prediction accuracy(%)				
	CY(43)	ME(30)	MI(13)	OTHER(12)	OA(%)
Covariant[a]	97.7	73.3	30.8	25.0	72.5
Dipep_Diver[b]	88.4	90.0	92.3	50.0	84.7
BC[c]	90.7	90.0	92.3	50.0	85.7
Hens_BC[d]	95.3	90.0	92.3	66.7	89.8
ID_SVM[e]	95.3	93.3	84.6	58.3	88.8
EBGW_SVM[f]	97.7	90.0	92.3	83.3	92.9
DF_SVM[g]	97.7	96.7	92.3	75.0	93.9
2-BF_SVM[h]	97.7	96.7	91.7	83.3	94.9
This paper	**100**	**100**	**84.6**	**91.7**	**96.9**

[a]This result is based on Covariant method[3].
[b]This result is based on Dipep_Diver method[20].
[c]This result is based on BC method[4].
[d]This result is based on Hens_BC method[4].
[e]This result is based on ID_SVM method[6].
[f]This result is based on EBGW_SVM method[5].
[g]This result is based on DF_SVM method[7].
[h]This result is based on 2-BF_SVM method[8].

Table 5 Performance comparison of different methods on ZW225 dataset.

Method	Prediction accuracy(%)				
	NU(41)	CY(70)	MI(25)	ME(89)	OA(%)
Dipep_Diver[a]	70.7	81.4	76.0	51.7	67.1
EBGW_SVM[b]	63.4	90.0	60.0	93.3	83.1
DF_SVM[c]	73.2	87.1	64.0	92.1	84.0
ID_SVM[d]	73.2	92.9	68.0	91.0	85.8
This paper	**87.8**	**91.4**	**92.1**	**80.0**	**89.8**

[a]This result is based on Dipep_Diver method[20].
[b]This result is based on EBGW_SVM method[5].
[c]This result is based on DF_SVM method[7].
[d]This result is based on ID_SVM method[6].

From Table 4, we can find that the prediction capacity of our method RTC_SVM is stronger than that of other existing models. The overall prediction accuracy of RTC_SVM is 8.1, 4.0, 3.0 and 2.0 percentile higher than that of ID_SVM, EBGW_SVM, DF_SVM and 2-BF_SVM, respectively. Especially for the other class proteins, the prediction accuracy of other models do not exceed 83.3%, while that of RTC_SVM achieves 91.7%. From Table 4, we also find that among the existing models, the prediction accuracy of RTC_SVM achieves the highest 100% for the cytoplasmic proteins(CY) and plasma membrane-bound proteins(ME).

For the dataset ZW225, the prediction results of three models are listed in Table 5. The overall prediction accuracy of RTC_SVM reaches 89.8%, which is 22.7, 6.7 and 4.0 percentile higher than that of Dipep_Diver, EBGW_SVM and ID_SVM. Especially for the nuclear proteins(NU) and the mitochondrial proteins(MI), the prediction accuracies of RTC_SVM are 24.4% and 32.1% higher than those of EBGW_SVM model; 14.6% and 24.1% higher than those of ID_SVM model. But we also notice that for the membrane proteins(ME), the prediction accuracy of RTC_SVM is lower than that of EBGW_SVM or ID_SVM model but higher than that of Dipep_Diver model. These results show that the prediction capacity of different models are complementary, if we can better joint them together, the prediction accuracies will be further improved. The hybrid model will be taken into consideration to further improve the prediction accuracy in our future work.

The different performances on the two datasets are presumably due to the different numbers of sequences in the dataset and possibly multiple localizations of these sequences. At present, our method does not deal with multiple subcellular localizations. However, it is straightforward to extend our approach to the case of multiple localizations, because SVM output is, in fact, a probability distribution of subcellular localization, we can set a proper probability threshold to determine the possible subcellular compartment candidates. In summary, the outstanding performance of the current method can be attributed to the effective usage of the triplet composition features that are extracted from the reduced protein sequence with the coding of amino acids as well as well-trained SVM.

4 Conclusions

Prediction of apoptosis protein subcellular localization can provide important information about their functionalities. The accuracy of predicting apoptosis protein subcellular lo-

calization is mainly depended on the representation vector of a protein sequence and the classifier for prediction. Though some of existing methods have shown the state-of-the-art performance, there is always room for improvement. Hydropathy is one of important physicochemical properties of amino acids, and is better conserved than protein sequences in evolution. In this paper, by introducing the reduced coding of the amino acids according to the hydrophobic/hydrophilic or other categories and triplet composition, a novel reduced triplet composition vector is proposed to predict subcellular localization. Then we apply it to predict apoptosis proteins with support vector machine method on two benchmark datasets ZD98 and ZW225. The results of the jackknife cross validation test using the standard datasets show that the proposed method can be used as an efficient approach for predicting apoptosis protein subcellular localization. Our method achieved an overall prediction accuracy of 96.9% and 89.8% for the two widely used datasets, ZD98 and ZW225, respectively. The experiment results show that RTC_SVM approach is convenient to calculate and provides an effective tool to extract valuable information from protein sequences, which may be a useful tool in other assignment problems in proteomics and genome research.

Acknowledgements: The authors thank the anonymous referees for many valuable suggestions that have improved this manuscript. This work was supported by the TianYuan Special Funds of the National Natural Science Foundation of China (Grant Nos. 11326201, 11326202), the Fundamental Research Funds for the Central Universities (No. JB140703), and the Natural Science Basic Research Plan in Shaanxi Province of China.

References

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

[2] S. H. Kaufmann, M. O. Hengarther, Programmed cell death: alive and well in the new millennium, *Trends Cell Biol.* **11** (2001) 526–534.

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

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

[5] 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.

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

[7] 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.

[8] G. Jian, Y. Zhang, P. Qian, Prediction of subcellular localization for apoptosis protein: Approached with a novel representation and support vector machine, *MATCH Commun. Math. Comput. Chem.* **67** (2012) 867–878.

[9] 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)

[10] L. Zou, C. Nan, F. Hu, Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles, *Bioinformatics* **29** (2013) 3135–3142.

[11] M. A. Singh, A. Navarro, A. Miguel, A novel approach for protein subcellular location prediction using amino acid exposure, *BMC Bioinformatics* **14** (2013) UNSP 342.

[12] P. Feng, W. Chen, H. Lin, K. C. Chou, iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition, *Anal. Biochem.* **442** (2013) 118–125.

[13] H. Chen, B. Liao, L. Cai, X. Chen, S. Liu, A novel numerical feature extraction method for protein subcellular localization, *J. Comput. Theor. Nanos.* **10** (2013) 2618–2625.

[14] W. Chen, P. Feng, H. Lin, K. C. Chou, iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, *Nucl. Acids Res.* **41** (2013) e68.

[15] W. Chen, H. Lin , P. Feng, C. Ding, Y. Zuo, K. C. Chou, iNuc-PhysChem: A sequence–based predictor for identifying nucleosomes via physicochemical properties, *PLoS ONE* **7** (2012) e47843.

[16] B. Liao, J. Jiang, Q. Zeng, W. Zhu, Predicting apoptosis protein subcellular location with PseAAC by incorporating tripeptide composition, *Protein Peptide Lett.* **18** (2011) 1086–1092.

[17] C. Huang, J. Yuan, Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites, *Biosystems* **113** (2013) 50–57.

[18] Y. Zeng, Y. Guo, R. Xiao, L. Yang, L. Yu, M. Li, Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach, *J. Theor. Biol.* **259** (2009) 366–372.

[19] C. Jia, Y. Zhang, Z. Wang, SulfoTyrP: A high accuracy predictor of protein sulfotyrosine sites, *MATCH Commun. Math. Comput. Chem.* **71** (2014) 227–240.

[20] Y. L. Chen, Q. Z. Li, Prediction of the subcellular location apoptosis proteins using the algorithm of measure of diversity, *Acta Sci. Natur. Univ. NeiMongol* **25** (2004) 413–417.

[21] A. Bairoch, R. Apeweiler, The Swiss-Prot protein sequence data bank and its supplement TrEMBL, *Nucleic Acids Res.* **25** (2000) 31–36.

[22] K. C. Chou, Z. C. Wu, X. Xiao, iLoc-Euk: A multi–label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic Proteins, *PLoS ONE* **6** (2011) e18258.

[23] J. Panek, 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, Bioinformatics* **58** (2005) 923–934.

[24] L. A. 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 Recognit.* **39** (2006) 2323–2343.

[25] Z. Yuan, B. Huang, Prediction of protein accessible surface areas by support vector regression, *Proteins* **57** (2004) 558–564.

[26] Z. Yuan, T. L. Bailey, R. D. Teasdak, Prediction of protein B-factor profiles, *Proteins* **58** (2005) 905–912.

[27] C. C. Chang, C. J. Lin, LIBSVM: a library for support vector machines, (2001) http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

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

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

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