Scientific Research Publishing

Using position specific scoring matrix and auto covariance to predict protein subnuclear localization

Rong-Quan Xiao[1], Yan-Zhi Guo[2], Yu-Hong Zeng[2], Hai-Feng Tan[1], Xue-Mei Pu[2], Meng-Long Li[1,2*]

[1]College of Life Sciences, Sichuan University, Chengdu 610064. [2]College of Chemistry, Sichuan University, Chengdu 610064, P.R. China. Correspondence should be addressed to Meng-Long Li(liml@scu.edu.cn). Tel: +86 28 89005151; Fax: +86 28 85412356.

Received September 8th, 2008; revised November 13th, 2008; accepted November 20th, 2008

ABSTRACT

The knowledge of subnuclear localization in eukaryotic cells is indispensable for understanding the biological function of nucleus, genome regulation and drug discovery. In this study, a new feature representation was proposed by combining position specific scoring matrix (PSSM) and auto covariance (AC). The AC variables describe the neighboring effect between two amino acids, so that they incorporate the sequence-order information; PSSM describes the information of biological evolution of proteins. Based on this new descriptor, a support vector machine (SVM) classifier was built to predict subnuclear localization. To evaluate the power of our predictor, the benchmark dataset that contains 714 proteins localized in nine subnuclear compartments was utilized. The total jackknife cross validation accuracy of our method is 76.5%, that is higher than those of the Nuc-PLoc (67.4%), the OET-KNN (55.6%), AAC based SVM (48.9%) and ProtLoc (36.6%). The prediction software used in this article and the details of the SVM parameters are freely available at http://chemlab.scu.edu.cn/predict_SubNL/index.htm and the dataset used in our study is from Shen and Chou's work by downloading at http://chou.med.harvard.edu/bioinf/Nuc-PLoc/Data.htm.

Keywords: Position Specific Scoring Matrix; Auto Covariance; Support Vector Machine; Protein Subnuclear Localization Prediction

1. INTRODUCTION

The cell nucleus is complex, important subcellular organelle in eukaryotes cell. It organizes the comprehensive assembly of our genes and their corresponding regulatory factors [1]. Meanwhile, it also reflects various intricate biological activities, and controls various kinds of biologic processes [2]. Many proteins, from outside a nuclear, trend to be localized into specific subnuclear locations of the nucleus [3]. If proteins can not be correctly localized into its specific subnuclear locations in human, it will lead to genetic disease [4], cancer [5] or virally infected cells [6]. Thus, it's desirable to get the knowledge of protein subnuclear localization for in-depth understanding cell biological processes and genomic regulation. However, it is costly and time-consuming to assay the subnuclear localization of proteins by biology experiments [7]. The number of protein sequences is increasing more rapidly than that of identified proteins [7]. So it is of great practical significance to develop computational approaches for identifying the protein subnuclear localizations in cell nucleus. At the same time, many lines of evidences have indicated that computational approaches, such as structural bioinformatics [8], molecular docking [9], pharmacophore modelling [10], QSAR [11,12,13], protein subcellular location prediction [7,14], identification of membrane proteins and their types [15], identification of enzymes and their functional classes [16], identification of proteases and their types [17], protein cleavage site prediction [18,19], and signal peptide prediction [20,21] can provide very useful information for both basic research and drug discovery in a timely manner. The present study is devoted to develop a new method for predicting protein subnuclear localization in hope to stimulate the development of the relevant areas.

Recently, many algorithms have already been developed for predicting protein subcellular localizations [22,23,24,25,26,27,28,29,30,31,32,33], as reviewed by Chou [7]. Even several web severs have been constructed for predicting subcellular localization of various organisms [14,34,35,36,37]. However, there are only a few computational methods for predicting protein subnuclear localization [38,39,40,41], such as OET-KNN [42], ProLoc [43], Nuc-PLoc [44], and AdaBoost classifiers [45].

Compared to the conventional amino acid composition (AAC), pseudo amino acid (PseAA) composition [46], originally introduced by Chou [47,48], can include the sequence-order information of sequences. Similarly, the PsePSSM was also proposed by Shen and Chou in order to incorporate the evolution information of proteins [44]. They built a new web server called Nuc-PLoc for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM with a promising prediction result. In this study, we developed a new method by fusing position specific scoring matrix (PSSM) and auto covariance (AC), so that this method can incorporate sequence-order information by AC and the evolutionary information by PSSM. A classifier based on SVM was constructed to predict protein subnuclear localization using jackknife test. The result indicates that our method has successfully enhanced accuracies of the existing methods for predicting protein subnuclear localization.

2. MATERIALS AND METHODS

2.1. Data Sets

In this paper, our dataset is obtained from article by Shen and Chou [44]. And anyone can freely download it at this page (http://chou.med.harvard.edu/bioinf/Nuc-PLoc/Data.htm). This dataset consists of nine classes and 714 proteins in total. Details of this benchmark dataset are shown in **Table 1.** S$_i$ (i=1, 2… 9) is used to represent each of nine subsets and S represents the total dataset.

2.2. Feature Representations

2.2.1. Auto Covariance (AC)

We selected three common physicochemical properties, hydrophobicity [49], volumes of side chains of amino acids [50], and polarity [51], to represent the structure and function [52], the stereospecific blockade [53] and the electronic property [54] of residues in a protein respectively. These original values were taken from Guo *et al*. [55] and were first normalized to zero mean value and unit standard deviation (SD) by Equation (1):

$$P'_{i,j} = \frac{P_{i,j} - \overline{P}_j}{S_j} \tag{1}$$

(i=1, 2, 3; j=1, 2, 3…, 20.)

Where $P_{i,j}$ is the *i-th* descriptor value for *j-th* amino acid, $\overline{P}_j$ is the mean of the *j-th* descriptor of the 20 amino acids and S_j is the value of SD. So each protein sequence was translated into three vectors with each amino acid represented by the normalized values.

There are many approaches to convert the protein sequences into numerical order sequences, including autocorrelations and auto covariance (AC). Autocorrelations, quite similar to AC, has been used in the prediction of secondary structure content [56,57,58] and structural class [59,60,61,62]; however, AC as a statistical tool for analyzing sequences of vectors has also been successfully adopted by our research group for protein classifications [55,63] from primary sequence. So in our study, AC was selected to transform these numerical vectors into uniform matrices in order to take the neighboring effect of the sequences into account. Here, *lag* is the distance between one residue and its neighbour, a certain number of residues away. The AC variables are calculated by the Equation (2) [55].

Table 1. The benchmark dataset consists of 714 nuclear proteins classified into nine subnuclear localizations

Subnuclear localization	Subset	No. of proteins
Chromatin	S$_1$	99
Heterochromatin	S$_2$	22
Nuclear envelope	S$_3$	61
Nuclear matrix	S$_4$	29
Nuclear pore complex	S$_5$	79
Nuclear speckle	S$_6$	67
Nucleolus	S$_7$	307
Nucleoplasm	S$_8$	37
Nuclear PML body	S$_9$	13
Total	S	714

$$AC_{lag,j} = \frac{1}{L-lag}\sum_{i=1}^{L-lag}\left(P_{i,j} - \frac{1}{L}\sum_{i=1}^{L}P_{i,j}\right)\times\left(P_{(i+lag),j} - \frac{1}{L}\sum_{i=1}^{L}P_{i,j}\right) \tag{2}$$

Where i is the position in the sequence P, j is one descriptor, L is the length of the sequence P and *lag* is the value of the lag.

In this way, the number of AC variables, D, can be calculated according to Equation (3) [55].

$$D = lg \times p \tag{3}$$

Where lg is the maximum *lag* (*lag*=1, 2, 3…, *lg*) and p represents the number of descriptors.

2.2.2. Position Specific Scoring Matrix (PSSM)

A PSSM is a Position Specific Scoring Matrix and is a commonly used representation of motifs (patterns) in biological sequences [64]. So far, this method has been used for predicting protein subcellular localization [65] and subnuclear localization [40,44].

For a protein sequence P with L amino acid residues, PSSM is obtained according to the following Equation [44].

$$P_{PSSM} = \begin{bmatrix} P_{1\to1} & P_{1\to2} & \cdots & P_{1\to j} & \cdots & P_{1\to20} \\ P_{2\to1} & P_{2\to2} & \cdots & P_{2\to j} & \cdots & P_{2\to20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{i\to1} & P_{i\to2} & \cdots & P_{i\to j} & \cdots & P_{i\to20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{L\to1} & P_{L\to2} & \cdots & P_{L\to j} & \cdots & P_{L\to20} \end{bmatrix} \tag{4}$$

In Equation (4), where *i*→*j* describes *i-th* amino acid residue of the protein sequence P being mutated to amino acid type j in the biology evolution process, $P_{i\to j}$ is the score of this mutation and L is the length of the sequence P. Here we used the numerical codes 1, 2, 3…20 to represent the single character of ordered 20 native amino acid types in Equation (4). To get the $L\times20$ scores of the P_{PSSM} in the Equation (4), we used three iterations of PSI-BLAST [66] with default threshold (the default E-value is 0.001) to search the Swiss-Prot database (version 54.4, released on 25 Oct. 2007) for multiple sequence alignment against the protein P. Then, the value of $P_{i\to j}$ is standardized by Equation (5), as given below.

$$P_{i\to j} = \frac{P^{o}_{i\to j} - \frac{1}{20}\sum_{j=1}^{20}P^{o}_{i\to j}}{\max(P^{o}_{i\to j}) - \min(P^{o}_{i\to j})} \tag{5}$$

$$(i=1, 2, 3\ldots L; j=1, 2, 3\ldots 20)$$

Where $\boldsymbol{P}^{o}_{i\to j}$ is the original scores generated by PSI-BLAST, $\boldsymbol{P}_{i\to j}$ is a zero mean value over the 20 native amino acids and the value is between -1 and 1. However, because of proteins with different lengths L, the matrices of the PSSM descriptor in Equation (4) have different numbers of rows. To gain the uniform matrix for protein sequences of different lengths, we converted the PSSM of protein P to a uniform vector through the Equation (6) [44].

$$\overline{P}_{PSSM}=\left[\overline{P}_{1}\quad \overline{P}_{2}\quad \cdots\quad \overline{P}_{j}\quad \cdots\quad \overline{P}_{20}\right]^{\mathrm{T}}\ (j=1,2,\cdots,20) \qquad (6)$$

Where T is the transpose operator, $\overline{P}_{j}$ is the average score over *j-th* column in Equation (4).

Finally, the $\overline{P}_{PSSM}$ describes the evolutionary information of a protein sample, and AC variables contain the interaction information between two amino acid residues of a sequence. So each protein sequence was converted into a numerical vector by concatenating PSSM and AC. Here, each AC variable was appended a weight factor of 0.05.

2.2.3. Accuracy and Matthew's Correlation Coefficient (MCC)

To evaluate the performance of this method, two parameters, accuracy and Matthew's correlation coefficient (MCC), were selected in this article. They are calculated by Equation (7) and Equation (8), respectively.

$$Accuracy=\frac{TP}{TP+FN} \qquad (7)$$

$$MCC=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}} \qquad (8)$$

Where TP represents the true positive; TN, the true negative; FP, the false positive and FN, the false negative.

3. RESULTS AND DISCUSSION

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling test, and jackknife test [67]. However, as elucidated in [14] and demonstrated by Eq.50 of [7], among the three cross-validation methods, the jackknife test is deemed the most objective that can always yield a unique result for a given benchmark dataset, and hence has been increasingly used by tigators to examine the accuracy of various predictors (see, e.g., [7,33,68,69,70,71,72,73,74,75,76,77,78,79,80, 81,82]). So in this paper, the jackknife test was chosen to validate the current algorithm. Because the benchmark dataset used has nine subsets, the one-to-one multiclass classification system led to 9*(9-1)/2=36 SVM models for one single encoding methods. Meanwhile, for AC variables, the value of *lg* was optimized as 13 through a series of control experiments, and the value of *p* is 3. So, the number of AC variables, *D*, is 39 ($D=lg\times p=13\times 3=39$) according to Equation (3).

Amino acid composition (AAC) has been widely used for predicting subcellular localizations [7,14,22,23,24, 25,26,27,28,30,31,32,34,35,36,37,83,84,85], so it was also used as a substitution model in our study. And thus, three SVM models based on AAC, AC and PSSM, were respectively constructed.

The results according to jackknife test are listed in **Table 2**. As can be seen from **Table 2**, the prediction accuracy of PSSM based model is nearly equal to that of AAC based model. However, AC based model gives the lower accuracy of 64.13%. Then we constructed models by fusing the three substitution models, so four fused classifier were built. **Table 2** shows that the accuracies of the four fused models are higher than those of the three anterior models. Among those four fused models, the accuracy of the model combining PSSM, AAC and AC is lower than that of PSSM and AC based model that obtains the best performance with an accuracy of 76.45%. So the final SVM model was built based on PSSM and AC. The kernel function of SVM is radio basis function (rbf), and the parameters of C and γ are listed in the table by downloading at http://chemlab.scu.edu.cn/predict_SubNL/index.htm.

In order to further examine the prediction power of the current classifier, the performance of this method was also compared with those of the existing methods on the same training dataset. The results obtained by several algorithms with different substitution models were summarized in **Table 3**. From **Table 3**, we can see that the accuracy obtained by Nuc-PLoc [44] is much higher than those of ProtLoc [43], AAC based SVM and OET-KNN [42]. When compared to Nuc-PLoc, our method obtains a better performance with the accuracy of 76.5%. It means our method is successful in predicting protein subnuclear localization only using primary sequences of proteins

Table 2. Overall accuracies by jackknife tests with different substitution models on the benchmark dataset of Table 1

Substitution Model	AAC^{a}	AC^{b}	PSSM^{c}	AAC+AC^{d}	PSSM+AAC	PSSM+AC^{d}	PSSM+AAC+AC^{d}
Accuracy	73.82%	64.13%	73.85%	74.05%	75.97%	76.45%	75.99%

a: Amino acid composition
b: Auto covariance
c: Position specific scoring matrix
d: While fused models were constructed, a weight factor added on AC is 0.05.

Table 3. Overall accuracy by jackknife tests with different algorithms on the benchmark dataset of Table 1

Algorithm	Protein sample descriptor	Overall accuracy
ProtLoc[a,d]	Amino acid composition	261/714=36.6%
SVM[d]	Amino acid composition	349/714=48.9%
OET-KNN[b,d]	PseAA Composition	397/714=55.6%
Nuc-PLoc[c,d]	Fusion of PsePSSM and PseAA Composition	481/714=67.4%
Our method	Combination of PSSM and AC	546/714=76.5%

a: See Cedano *et al*. (1997)[86]
b: See Shen and Chou (2005)[42]
c: See Shen and Chou (2007)[44]
d: The results were from Shen and Chou (2007)[44], and the original data could been seen in that article.

Table 4. The MCC values obtained by the jackknife tests with Nuc-PLoc and our method on the benchmark dataset of Table 1

Subnuclear localization	Matthew's correlation coefficient	
	Nuc-PLoc[a]	Our method[b]
Chromatin S_1	0.60	0.55
Heterochromatin S_2	0.52	0.58
Nuclear envelope S_3	0.53	0.65
Nuclear matrix S_4	0.52	0.61
Nuclear pore complex S_5	0.70	0.72
Nuclear speckle S_6	0.43	0.57
Nucleolus S_7	0.57	0.57
Nucleoplasm S_8	0.31	0.54
Nuclear PML body S_9	0.32	0.51

a: The results were from Shen and Chou (2007)[44], and the original data could be seen in that article.
b: The classifier fused PSSM and AC.

In addition, to evaluate the stability of our method, the values of the MCC for the nine subsets were compared based on Nuc-PLoc and our current predictor, respectively, as seen in **Table 4**. For nine subsets, our method yields a higher MCC than Nuc-PLoc, except the subset $S_{1.}$ So, compared to the existing methods, our classifier combined with PSSM and AC has further improved the prediction accuracy of protein subnuclear localization.

4. CONCLUSION

In this paper, a new classifier was developed by fusing PSSM and AC for predicting protein subnuclear localization only using the primary sequences of nuclear proteins. The SVM predictor was constructed based on PSSM and AC. AC variables represent the interactions between amino acids in protein sequences; PSSM describes the evolutionary information. So the method incorporated not only the evolution information, but also the sequence-order information. Compared with the current methods, this method successfully raises the prediction accuracy. Hence, it may be a good supplementary tool for protein function studies.

ACKNOWLEDGEMENT

The authors gratefully thank Shen and Chou for sharing the benchmark dataset. The work was funded by the National Natural Science Foundation of China (No. 20775052).

REFERENCES

[1] G. S. Stein, J. B. Lian, W. A. van, J. L. Stein, A. Javed, G. Barnes, L. Gerstenfeld, D. Vradii, S. K. Zaidi, and J. Pratap, *et al.* (2006) Organization of transcriptional regulatory machinery in nuclear microenvironments: Implications for biological control and cancer. Cancer Treatment Reviews, 32, 13−13.

[2] A. I. Lamond, and W. C. Earnshaw, (1998) Structure and function in the nucleus. Science, 280, 547−553.

[3] J. M. Bridger, and W. A. Bickmore, (1998) Putting the genome on the map. Trends in Genetics, 14, 403−409.

[4] H. G. Sutherland, G. K. Mumford, K. Newton, L. V. Ford, R. Farrall, G. Dellaire, J. F. Caceres, and W. A. Bickmore, (2001) Large-scale identification of mammalian proteins localized to nuclear sub-compartments. Human Molecular Genetics, 10, 1995−2011.

[5] S. K. Zaidi, D. W. Young, A. Javed, J. Pratap, M. Montecino, W. A. van, J. B. Lian, J. L. Stein, and G. S. Stein, (2007) Nuclear microenvironments in biological control and cancer. Nature Reviews Cancer, 7, 454−463.

[6] R. D. Phair, and T. Misteli, (2000) High mobility of proteins in the mammalian cell nucleus. Nature, 404, 604−609.

[7] K. C. Chou, and H. B. Shen, (2007) Recent progress in protein subcellular location prediction. Analytical Biochemistry, 370, 1−16.

[8] K. C. Chou, (2004) Structural bioinformatics and its impact to biomedical science. Current Medicinal Chemistry, 11, 2105−2134.

[9] K. C. Chou, D. Q. Wei, and W. Z. Zhong, (2003) Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. Biochemical and Biophysical Research Communications, 308, 148−151.

[10] Sirois, S., Wei, D. Q., Du, Q. S. and Chou, K. C. (2004) Virtual screening for SARS-CoV protease based on KZ7088 pharmacophore pointst. Journal of Chemical Information and Computer Sciences, 44, 1111−1122.

[11] Du, Q. S., Mezey, P. G. and Chou, K. C. (2005) Heuristic molecular lipophilicity potential (HMLP): A 2D-QSAR study to LADH of molecular family pyrazole and derivatives. Journal of Computational Chemistry, 26, 461−470.

[12] Du, Q. S., Huang, R. B., Wei, Y. T., Du, L. Q. and Chou, K. C. (2008) Multiple field three dimensional quantitative structure-activity relationship (MF-3D-QSAR). Journal of Computational Chemistry, 29, 211−219.

[13] Prado-Prado, F. J., Gonzalez-Diaz, H., de, V. O., Ubeira, F. M. and Chou, K. C. (2008) Unified QSAR approach to antimicrobials. Part 3: First multi-tasking QSAR model for Input-Coded prediction, structural back-projection, and complex networks clustering of antiprotozoal compounds. Bioorganic and Medicinal Chemistry, 16, 5871−5880.

[14] Chou, K. C. and Shen, H. B. (2008) Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nature Protocols, 3, 153−162.

[15] Chou, K. C. and Shen, H. B. (2007) MemType-2L: a web server

for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochemical and Biophysical Research Communications, 360, 339−345.

[16] Shen, H. B. and Chou, K. C. (2007) EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochemical and Biophysical Research Communications, 364, 53−59.

[17] Chou, K. C. and Shen, H. B. (2008) ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information. Biochemical and Biophysical Research Communications, 376, 321−325.

[18] Chou, K. C. (1996) Prediction of human immunodeficiency virus protease cleavage sites in proteins. Analytical Biochemistry, 233, 1−14.

[19] Shen, H. B. and Chou, K. C. (2008) HIVcleave: a web-server for predicting human immunodeficiency virus protease cleavage sites in proteins. Analytical Biochemistry, 375, 388−390.

[20] Chou, K. C. and Shen, H. B. (2007) Signal-CF: A subsite-coupled and window-fusing approach for predicting signal peptides. Biochemical and Biophysical Research Communications, 357, 633−640.

[21] Shen, H. B. and Chou, K. C. (2007) Signal-3L: A 3-layer approach for predicting signal peptides. Biochemical and Biophysical Research Communications, 363, 297−303.

[22] Chou, K. C. (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochemical and Biophysical Research Communications, 278, 477−483.

[23] Chou, K. C. and Cai, Y. D. (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. Journal of Biological Chemistry, 277, 45765−45769.

[24] Cai, Y. D., Liu, X. J., Xu, X. B. and Chou, K. C. (2002) Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. Journal of Cellular Biochemistry, 84, 343−348.

[25] Cai, Y. D., Liu, X. J. and Chou, K. C. (2002) Artificial neural network model for predicting protein subcellular location. Computers and Chemistry, 26, 179−182.

[26] Chou, K. C. and Cai, Y. D. (2004) Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. Journal of Cellular Biochemistry, 90, 1250−1260.

[27] Chou, K. C. and Cai, Y. D. (2004) Prediction of protein subcellular locations by GO-FunD-PseAA predictor. Biochemical and Biophysical Research Communications, 320, 1236−1239.

[28] Gao, Q. B., Wang, Z. Z., Yan, C. and Du, Y. H. (2005) Prediction of protein subcellular location using a combined feature of sequence. Febs Letters, 579, 3444−3448.

[29] Zhang, T. L., Ding, Y. S. and Chou, K. C. (2006) Prediction of protein subcellular location using hydrophobic patterns of amino acid sequence. Computational Biology and Chemistry, 30, 367−371.

[30] Chou, K. C. and Shen, H. B. (2006) Predicting protein subcellular location by fusing multiple classifiers. Journal of Cellular Biochemistry, 99, 517−527.

[31] Chou, K. C. and Shen, H. B. (2006) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-Nearest Neighbor classifiers. Journal of Proteome Research, 5, 1888−1897.

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

[33] Shi, F., Chen, Q. J. and Li, N. N. (2008) Hilbert Huang transform for predicting proteins subcellular location. Journal of Biomedical Science and Engineering, 1, 59−63.

[34] Chou, K. C. and Shen, H. B. (2006) Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization. Biochemical and Biophysical Research Communications, 347, 150−157.

[35] Shen, H. B., Yang, J. and Chou, K. C. (2007) Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. Amino Acids, 33, 57−67.

[36] Shen, H. B. and Chou, K. C. (2007) Hum-mPLoc: An ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochemical and Biophysical Research Communications, 355, 1006−1011.

[37] Chou, K. C. and Shen, H. B. (2007) Euk-mPLoc: A fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. Journal of Proteome Research, 6, 1728−1734.

[38] Lei, Z. D. and Dai, Y. (2005) An SVM-based system for predicting protein subnuclear localizations. BMC Bioinformatics, 6, 291−298.

[39] Lei, Z. D. and Dai, Y. (2006) Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. BMC Bioinformatics, 7, 491−500.

[40] Mundra, P., Kumar, M., Kumar, K. K., Jayaraman, V. K. and Kulkarni, B. D. (2007) Using pseudo amino acid composition to predict protein subnuclear localization: Approached with PSSM. Pattern Recognition Letters, 28, 1610−1615.

[41] Li, F. M. and Li, Q. Z. (2008) Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. Amino Acids, 34, 119−125.

[42] Shen, H. B. and Chou, K. C. (2005) Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. Biochemical and Biophysical Research Communications, 337, 752−756.

[43] Huang, W. L., Tung, C. W., Huang, H. L., Hwang, S. F. and Ho, S. Y. (2007) ProLoc: Prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features. Biosystems, 90, 573−581.

[44] Shen, H. B. and Chou, K. C. (2007) Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Engineering Design and Selection, 20, 561−567.

[45] Jiang, X. Y., Wei, R., Zhao, Y. J. and Zhang, T. L. (2008) Using Chou's pseudo amino acid composition based on approximate entropy and an ensemble of AdaBoost classifiers to predict protein subnuclear location. Amino Acids, 34, 669−675.

[46] Shen, H. B. and Chou, K. C. (2008) PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition. Analytical Biochemistry, 373, 386−388.

[47] Chou, K. C. (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins-Structure Function and Genetics, 43, 246−255.

[48] Chou, K. C. (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics, 21, 10−19.

[49] Tanford, C. (1962) Contribution of Hydrophobic Interactions to the Stability of the Globular Conformation of Proteins. Journal of the American Chemical Society, 84, 4240−4247.

[50] Krigbaum, W. R. and Komoriya, A. (1979) Local interactions as a structure determinant for protein molecules: II. Biochimica et Biophysica Acta, 576, 204−248.

[51] Grantham, R. (1974) Amino acid difference formula to help explain protein evolution. Science, 185, 862−864.

[52] Guo, Y. Z., Li, M., Lu, M., Wen, Z., Wang, K., Li, G. and Wu, J. (2006) Classifying G protein-coupled receptors and nuclear receptors on the basis of protein power spectrum from fast Fourier transform. AMINO ACIDS, 30, 397−402.

[53] Hackel, M., Hinz, H. J. and Hedwig, G. R. (1999) Partial molar volumes of proteins: amino acid side-chain contributions derived from the partial molar volumes of some tripeptides over the temperature range 10−90 degrees C. BIOPHYSICAL CHEMISTRY, 82, 35−50.

[54] Guo, Y. Z., Li, M. L., Wang, K. L., Wen, Z. N., Lu, M. C., Liu, L. X. and Jiang, L. (2006) Fast fourier transform-based support vector machine for prediction of G-protein coupled receptor subfamilies. (Vol 37, pg 759, 2005). ACTA BIOCHIMICA ET BIOPHYSICA SINICA, 38, 456−456.

[55] Guo, Y.Z., Yu, L.Z., Wen, Z.N. and Li, M.L. (2008) Using sup-

port vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Research, 36, 3025–3030.

[56] Lin, Z. and Pan, X. M. (2001) Accurate prediction of protein secondary structural content. Journal Of Protein Chemistry, 20, 217–220.

[57] Zhang, C. T., Lin, Z. S., Zhang, Z. D. and Yan, M. (1998) Prediction of the helix/strand content of globular proteins based on their primary sequences. Protein Engineering, 11, 971–979.

[58] Zhang, Z. D., Sun, Z. R. and Zhang, C. T. (2001) A new approach to predict the helix/strand content of globular proteins. Journal Of Theoretical Biology, 208, 65–78.

[59] Kedarisetti, K. D., Kurgan, L. and Dick, S. (2006) Classifier ensembles for protein structural class prediction with varying homology. Biochemical And Biophysical Research Communications, 348, 981–988.

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

[61] Li, X. and Pan, X. M. (2001) New method for accurate prediction of solvent accessibility from protein sequence. Proteins-Structure Function And Genetics, 42, 1–5.

[62] Kurgan, L. and Chen, K. (2007) Prediction of protein structural class for the twilight zone sequences. Biochemical And Biophysical Research Communications, 357, 453–460.

[63] Guo, Y. Z., Li, M. L., Lu, M. C., Wen, Z.N. and Huang, Z. T. (2006) Predicting G-protein coupled receptors-G-protein coupling specificity based on autocross-covariance transform. Proteins, 65, 55–60.

[64] Ben-Gal, I., Shani, A., Gohr, A., Grau, J., S, A., Shmilovici, A., Posch, S. and Grosse, I. (2005) Identification of transcription factor binding sites with variable-order Bayesian networks. Bioinformatics, 21, 2657–2666.

[65] Xie, D., Li, A., Wang, M. H., Fan, Z. W. and Feng, H. Q. (2005) LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. Nucleic Acids Research, 33, 105–110.

[66] Schaffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I., Koonin, E. V. and Altschul, S. F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Research, 29, 2994–3005.

[67] Chou, K. C. and Zhang, C. T. (1995) PREDICTION OF PROTEIN STRUCTURAL CLASSES. Critical Reviews in Biochemistry and Molecular Biology, 30, 275–349.

[68] Chen, Y. L. and Li, Q. Z. (2007) Prediction of the subcellular location of apoptosis proteins. Journal of Theoretical Biology, 245, 775–783.

[69] Chen, Y. L. and Li, Q. Z. (2007) Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo-amino acid composition. Journal of Theoretical Biology, 248, 377–381.

[70] Zhou, X. B., Chen, C., Li, Z. C. and Zou, X. Y. (2007) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. Journal of Theoretical Biology, 248, 546–551.

[71] Chen, C., Chen, L. X., Zou, X. Y. and Cai, P. X. (2008) Predicting protein structural class based on multi-features fusion. Journal of Theoretical Biology, 253, 388–392.

[72] Chen, K., Kurgan, L. A. and Ruan, J. S. (2008) Prediction of protein structural class using novel evolutionary collocation-based sequence representation. Journal of Computational Chemistry, 29, 1596–1604.

[73] Du, P. F. and Li, Y. D. (2008) Prediction of C-to-U RNA editing sites in plant mitochondria using both biochemical and evolutionary information. Journal of Theoretical Biology, 253, 579–586.

[74] Jiang, X. Y., Wei, R., Zhang, T. L. and Gu, Q. (2008) Using the concept of Chou's Pseudo Amino Acid composition to predict apoptosis proteins subcellular location: An approach by approximate entropy. Protein and Peptide Letters, 15, 392–396.

[75] Jin, Y. H., Niu, B., Feng, K. Y., Lu, W. C., Cai, Y. D. and Li, G. Z. (2008) Predicting subcellular localization with AdaBoost Learner. Protein and Peptide Letters, 15, 286–289.

[76] Li, F. M. and Li, Q. Z. (2008) Protein Subcellular Location Using Chou's Pseudo Amino Acid Composition and Improved Hybrid Approach. Protein and Peptide Letters, 15, 612–616.

[77] Lin, H. (2008) The modified Mahalanobis Discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. Journal of Theoretical Biology, 252, 350–356.

[78] Lin, H., Ding, H., Guo, F. B., Zhang, A. Y. and Huang, J. (2008) Predicting Subcellular Localization of Mycobacterial Proteins by Using Chou's Pseudo Amino Acid Composition. Protein and Peptide Letters, 15, 739–744.

[79] Niu, B., Jin, Y. H., Feng, K. Y., Liu, L., Lu, W. C., Cai, Y. D. and Li, G. Z. (2008) Predicting membrane protein types with bragging learner. Protein and Peptide Letters, 1**5**, 590–594.

[80] Wang, T., Yang, J., Shen, H. B. and Chou, K. C. (2008) Predicting membrane protein types by the LLDA algorithm. Protein and Peptide Letters, 15, 915–921.

[81] Wu, G. and Yan, S. M. (2008) Prediction of mutations in H3N2 hemagglutinins of influenza A virus from North America based on different datasets. Protein and Peptide Letters, 15, 144–152.

[82] Zhang, G. Y. and Fang, B. S. (2008) Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo-amino acid composition. Journal of Theoretical Biology, 253, 310–315.

[83] Cai, Y. D. and Chou, K. C. (2003) Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. Biochemical and Biophysical Research Communications, 305, 407–411.

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

[85] Chou, K. C. and Cai, Y. D. (2004) Predicting subcellular localization of proteins by hybridizing functional domain composition and pseudo-amino acid composition. Journal of Cellular Biochemistry, 91, 1197–1203.

[86] Cedano, J., Aloy, P., PerezPons, J. A. and Querol, E. (1997) Relation between amino acid composition and cellular location of proteins. Journal of Molecular Biology, 266, 594–600.