BMB
reports

A novel method for predicting protein subcellular localization based on pseudo amino acid composition

Junwei Ma & Hong Gu*
School of Control Science and Engineering, Dalian University of Technology, Dalian 116024, China

In this paper, a novel approach, ELM-PCA, is introduced for the first time to predict protein subcellular localization. Firstly, Protein Samples are represented by the pseudo amino acid composition (PseAAC). Secondly, the principal component analysis (PCA) is employed to extract essential features. Finally, the Elman Recurrent Neural Network (RNN) is used as a classifier to identify the protein sequences. The results demonstrate that the proposed approach is effective and practical. [BMB reports 2010; 43(10): 670-676]

INTRODUCTION

Subcellular localization is a key functional attribute of a protein. Studies on protein subcellular localization are of great help in illuminating disease mechanisms and new drug targets (1). Although subcellular localization can be determined by experimental methods such as cell fractionation, electron microscopy and fluorescence microscopy, they are time-consuming and expensive (2). Therefore, it is very urgent to develop an automatic and reliable prediction system for protein subcellular localization.

Nakashima and Nishikawa proposed a computational method to predict the subcellular localization based on the amino acid composition (3). Using the Mahalanobis distance (4), Cedano *et al*. (5) proposed an algorithm named ProtLock. Shen and Chou (6) developed the algorithm, an OET-KNN classifier, for prediction of 370 nuclear proteins, and the overall accuracy of jackknife test achieved about 64%. Lei and Dai (7) proposed a support vector machine (SVM) system using protein sequence information for prediction of six localization sites. Huang *et al*. (8) developed a method named PGAC involving the informative Gene Ontology terms associated with amino acid composition features.

Although there already exist so many computational prediction methods, there is still room for improvement. This is due to the fact that the protein sorting process is very complex and not yet well understood. Only a small portion of proteins have clearly identifiable sorting signals in their primary sequence. A further challenge is how to deal with proteins presenting in multiple locations. For a comprehensive description in this area, readers can refer to three recent papers (9-11).

Actually, among all these methods, most of them were based on the amino acid composition (AAC), where the sample of a protein is represented by 20 discrete numbers, with each representing the occurrence frequency of 20 different constituent native amino acids. Clearly, if one uses the conventional AAC to represent the sample of a protein, all effects of sequence order and length will be lost. In order to incorporate the sequence-order information for statistical prediction, Chou (12) proposed a PseAAC, which consists of $20+\lambda$ components. The first 20 components are the same as those in the AAC, and the remaining components represent λ sequence-order correlation factors of different ranks. It is these additional λ factors that approximately incorporate the sequence order effects. So it has been adopted to improve the prediction quality of protein subcellular localization by many investigators (13-15).

However, the PseAAC with a different value λ will result in a different outcome. To overcome such a problem, Shen and Chou (16, 17) introduced an ensemble classifier, by which the pseudo amino acid compositions with a set of different values can be automatically fused into one prediction system. In this paper, a completely different approach, ELM-PCA, combing the feature extraction technology of PCA with higher computational ability of the Elman network, is designed to solve the problem. Firstly, the protein is represented by the PseAAC, where the value of λ is large enough to contain more sequence order information. Secondly, the principle component analysis (18), one of the most popular linear dimension reduction methods, is employed to extract key features from the high-dimensional space. Lastly, the Elman RNN (19, 20), which is characterized by higher computing ability than the BP network, is applied to classify the protein data. Experimental results show that ELM-PCA significantly improves the robustness and prediction reliability.

*Corresponding author. Tel: 86-0411-84749451; Fax: 86-0411-84667461; E-mail: junweima@yahoo.com.cn
DOI 10.5483/BMBRep.2010.43.10.670

Received 24 May 2010, Accepted 12 August 2010

Keywords: Elman neural network, Five-fold cross validation, Principal component analysis, Protein subcellular localization, Pseudo amino acid composition

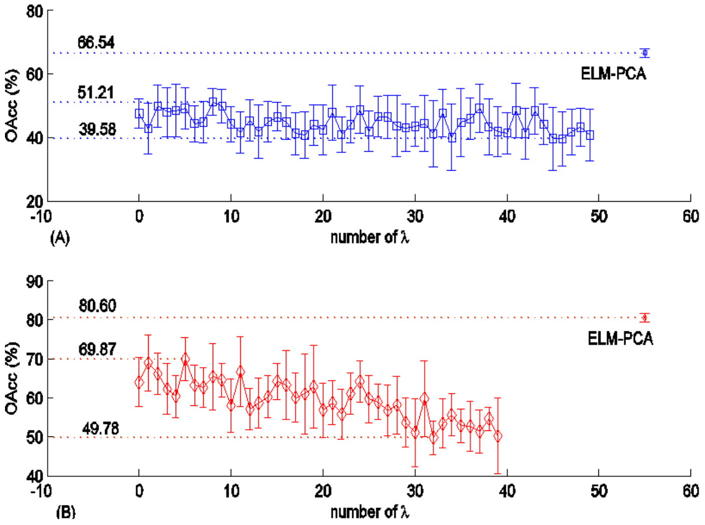

Fig. 1. Predicted results using Elman RNN on the two datasets: (A) On the MA629 dataset (B) On the GA541 dataset.

RESULTS

The overall prediction accuracy

First, the overall prediction accuracy, which can be regarded as the most important measurement, is used for assessment of the prediction system, and given by:

$$OAcc = \frac{\sum_{i=1}^{i=k} TP_i}{N} \tag{1}$$

where *N* is the total number of sequences in a data set, TP_i is the true positive number of subcellular localization *i*.

By making use of the Elman RNN, Fig. 1 (Error bar) shows the relations between the number of λ and prediction accuracy. Meanwhile, the accuracy by ELM-PCA is also given in this figure. It can be seen that prediction accuracy varies substantially with the number of, from 39.58% (λ= 46) to 51.21% (λ = 8) on the MA629 dataset and 49.78% (λ = 32) to 69.87% (λ = 5) on the GA541 dataset. When λ >44 for the MA629 or λ > 31 for the GA541, the accuracy is relatively low, which shows the correctness of the theoretical analysis of PseAAC. However, when ELM-PCA is adopted here, the accuracies reach 66.54% and 80.60%, about 15% and 11% higher, respectively, than the result at λ= 8 on the first dataset and at λ = 5 on the second dataset. Moreover, the standard deviations are much smaller than others. This indicates that the prediction quality can be remarkably improved by means of the PCA, and the performance of Elman RNN becomes more stable.

Other evaluation criteria

Evaluating a classifier, the success prediction rate and the prediction reliability should be considered together. Here, we report other standard performance measures over each of subcellular localization sites, including precision, sensitivity (recall or accuracy), specificity and Matthew's correlation coefficient (MCC). The precision (Pr) for class C_i, is defined as the fraction of proteins predicted to be in class C_i that are correct predictions; the sensitivity (Sn) for class C_i, is defined as the fraction of proteins belonging to class C_i that are correctly predicted; the specificity (Sp) for class C_i, is defined as the fraction of proteins not in class C_i that are correctly predicted; MCC provides a single measure of evaluating specificity and sensitivity together, and takes range from −1 to 1, where *MCC* = 1 indicates a perfect prediction; *MCC* = 0 indicates a completely random assignment; and *MCC* = −1 indicates a perfectly reverse correlation. The formula for each measurement is given below:

$$\mathrm{Pr} = \frac{TP}{TP+FP} \tag{2}$$

$$Se = \frac{TP}{TP+FN} \tag{3}$$

$$Sp = \frac{TN}{TN+FP} \tag{4}$$

$$MCC = \frac{(TP)(TN)-(FP)(FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{5}$$

where TP, TN, FP and FN are the number of true positives, true negatives, false positives and false negatives.

As mentioned above, the accuracy by only using Elman RNN reaches the maximum at λ = 8 on the MA629 dataset or at λ = 5 on the GA541 dataset. Therefore, ELM-PCA is compared with these two cases (Table 1, 2). It can be seen that all of the predicted results using ELM-PCA are less conservative.

Table 1. Comparison between ELM-PCA and Elman on the MA629 dataset

Location	ELM-PCA				Elman (λ = 8)			
	Pr (%)	Se (%)	Sp (%)	MCC	Pr (%)	Se (%)	Sp (%)	MCC
Cyt (152)	64.14	83.55	85.12	0.63	48.48	63.16	78.62	0.39
Ext (76)	57.89	14.47	98.55	0.25	33.33	7.89	97.83	0.11
Inn (186)	84.41	84.41	93.45	0.78	67.14	76.88	84.20	0.59
Out (103)	52.68	57.28	89.92	0.46	35.78	37.86	86.69	0.24
Pep (112)	57.02	58.04	90.52	0.48	41.76	33.93	89.75	0.26

Table 2. Comparison between ELM-PCA and Elman on the GA541 dataset

Location	ELM-PCA				Elman (λ = 5)			
	Pr (%)	Se (%)	Sp (%)	MCC	Pr (%)	Se (%)	Sp (%)	MCC
Cyt (194)	75.86	90.72	83.86	0.72	69.96	84.02	79.83	0.62
Cyt mem (103)	92.86	75.73	98.63	0.81	86.25	66.99	97.49	0.71
Cel (61)	77.78	68.85	97.50	0.70	63.89	37.70	97.29	0.44
Ext (183)	81.87	76.50	91.34	0.69	64.06	67.21	80.73	0.47

Table 3. The comparison of predictive performances of different approaches on another independent dataset

Classifier	Protein sample descriptor	Accuracy of five localization sites (%)					Overall (%)
		Cytoplasm	Extracell	Inner	Outer	Periplasm	
Randomtree	PseAAC (λ = 28)	50.00	25.68	45.00	52.63	55.00	42.76
J48	PseAAC (λ = 28)	55.00	41.89	61.43	**68.42**	46.38	56.24
Naivebayes	PseAAC (λ = 28)	89.29	39.19	60.36	28.95	37.68	59.90
PSLpred	-	89.29	40.54	60.71	47.37	60.87	64.06
PSORTb	-	83.57	31.08	65.36	57.89	69.57	65.39
Cello	-	85.71	54.05	61.79	52.63	66.67	66.39
PA	-	90.00	41.89	66.07	63.16	65.22	68.39
SOSUI	-	88.58	58.11	**70.00**	63.16	57.97	71.05
ELM-PCA	PseAAC (λ = 28)	**91.43**	**64.86**	67.86	52.63	**72.46**	**72.55**

In Table 1, the precision (Pr) using ELM-PCA increases by around 15% for each of the locations. MCC is 0.25 for the extracell using the current method, which is relatively low, but still 0.14 higher than the result at λ = 8. In Table 2, MCCs obtained by ELM-PCA are also much higher than the result at λ = 5, which shows the good classification performance of this method in subcellular localization prediction.

Comparison with other classifiers

In this work, the ELM-PCA predictor is compared with other classification methods such as Randomtree (21), J48 (22), Naivebayes (23) and five popular web-servers PSLpred (24), PSORTb (25), Cello (26), Proteome Analyst (PA) (27) and SOSUI (28). Also, we construct another independent dataset containing 601 proteins, of which 140 are of the cytoplasm, 74 of the extracellular locations, 280 of the inner membrane, 38 of the outer membrane, and 69 of the periplasm, to assess effectiveness of our method. The results for each of the subcellular localization sites are summarized in Table 3.

It can be seen that the prediction accuracy obtained by ELM-PCA is the highest for the cytoplam, the extracell and the periplasm, respectively, which shows that our method has better performance in the three subcellualr localizations. Moreover, the secretion pathways of extracellular proteins are at least seven pathways, so these complicated pathways result in the diversity of target signals and the low prediction accuracy for the popular online predictors. SOSUI has been proved to be superior to other systems in predicting the extracellular proteins (28), but the accuracy is 58.11% in this dataset, about 7% lower than that of our method. On the other hand, the prediction accuracy for the inner membrane using ELM-PCA is 67.86%, slightly smaller than the accuracy of SOSUI. And the result for the outer membrane is relatively low, only 52.63%, which means the learning effect is not good. However, the outer membrane sample number is 38, which is the smallest of the five subcellular localization sites. When PCA is adopted, it has not clearly portrayed the feature of the samples. It is expected that the correct rate for the outer membrane can be further enhanced by adding more new samples to the subset. Anyway, the overall prediction accuracy obtained by our method is the highest, about 72.55%, which indicates that ELM-PCA is indeed very useful in dealing with the complicated biological problem of predicting protein subcellular localization. It is also worth noting that ELM-PCA is different from the predictors PSORTb and Cello because PSORTb makes use of different modules and input information tuned up for specific localization sites, and Cello combines forty SVM classifiers to predict the sites, while our method only employs a single module approach and one classifier. Therefore the method proposed in this paper is simpler and more practical.

DISCUSSION

In order to contain more sequence order information, the dimension of input vector is set to $20 + L - 1$ before using PCA. Here, the first 20 numbers represent the classic AAC, each reflecting the occurrence frequency of one of the 20 native amino acids in a protein. So the increase of one of the variables will result in the decrease of others, and vice versa. The next $L - 1$ discrete numbers reflect the sequence length and the order effect along a protein chain. Therefore, all of the variables may possess some kind of correlation. And the higher the correlation, the greater the effect of PCA is. On the other hand, the Elman neural network, which adds a feedback function to Bp network, is more capable in computing and has relatively better global network stability. Further, the self-culture capabilities can be enhanced with the application of PCA so as to improve the network operation efficiency. So the new predictor of Elman-PCA can fit nonlinear prediction better. This may be one of the reasons why this classifier can well predict subcellular localization.

Here, 90% of the total variance was used to extract the key features for all. Clearly, this may be not the best choice. However, it can not only improve the predicting precision, but also avoid the low efficiency arising from looking for the optical value of λ. Generally speaking, the variance equal to or larger than 85% is appropriate. The future work is to optimize the parameter to improve the adaptability of PCA. In addition, PCA is a linear transformation in essence. Another work is to plan to use other nonlinear transformations, such as Kernel PCA (29) and Kernel independent component analysis (KICA) (30) for feature extraction. It is expected that it is a useful attempt in protein science and bioinformatics.

In conclusion, the novel classifier Elman-PCA is proposed to predict the subcellular localization. Further analysis suggests that the new method is not only a good complement for the existing methods, but also a potential alternative for some prediction tools arising from a machine learning approach (31).

The software in Matlab is available freely by request. Since user-friendly and publicly accessible web-servers represent the future direction, we shall make efforts in our future work to provide a web-server for the method presented in this paper.

MATERIALS AND METHODS

Datasets

To evaluate the effectiveness of our method, two datasets were adopted in the study. The first one (32) was obtained from Ma, which contained 629 protein sequences, of which 152 were located in the cytoplasm, 76 in the extracell, 186 in the inner membrane, 103 in the outer membrane, 112 in the periplasm. These sequences were extracted from SWISSPROT release 50.0 and only those sequences that appeared complete and had reliable experimental annotations for localization were included. In this dataset, no two sequences had more than 25% identity. The second (25) was generated by Gardy, which contained 541 sequences, of which 194 were in the cytoplasm, 103 in the cytoplasmic membrane, 61 in the cell wall and 183 in the extracell. Also, these sequences have been experimentally verified and reported in the literature, and were annotated for only one location. For convenience, these two datasets were referred to as the MA629 and GA541 datasets.

The PseAAC model

The PseAAC was proposed by Chou, which is a more advanced model than the traditional amino acid composition. According to the PseAAC discrete model, the protein P with L amino acid residues

$$S_1S_2S_3S_4S_5S_6 \cdots S_L$$

where S_1 represents the residue at the sequence position 1, S_2 at position 2, and so forth, can be formulated as

$$P= [P_1,\ P_2,\ P_3,\ \cdots, P_{20},\ P_{21},\ \cdots,\ P_{20+\lambda}]^T, (\lambda < L) \quad (6)$$

where the 20 + λ components are given by

$$P_u = \begin{cases} \dfrac{f_u}{\sum\limits_{i=1}^{20} f_i + w \sum\limits_{k=1}^{\lambda} \tau_k}, & 1 \le u \le 20 \\ \dfrac{w\tau_{u-20}}{\sum\limits_{i=1}^{20} f_i + w \sum\limits_{k=1}^{\lambda} \tau_k}, & 20+1 \le u \le 20+\lambda \end{cases} \quad (7)$$

where f_u is the normalized occurrence frequency of the 20 amino acids, τ_k is the k-tier sequence correlation factor for the protein P, and w is the weight factor for the sequence order effect. In the current study, we choose w = 0.05. One can easily convert protein sequences through the free server at the web site http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/.

As we can see, the first 20 components reflect the effect of the amino acid composition, whereas the components from 20 + 1 to 20 + λ reflect the effect of sequence order. Generally speaking, the larger the number of the λ components, the more the sequence-order effects incorporated. However, the number λ cannot exceed the length of a protein i.e. the number of its total residues (12). Moreover, if the number of λ is too large, the PseAAC will correspond to a very high dimensional vector, which may cause many problems in statistical prediction, such as the "dimension disaster", over-fitting and redundancy (33). Therefore, for different training data sets, λ may have different optimal values. However, when the number of amino acid residues of the shortest protein chain in the data set is very large, it is difficult to determine the value of λ.

ELM-PCA architecture

Here, a novel predictor, ELM-PCA, combing the feature extraction technology of PCA with higher computational ability of the Elman network, is designed to cope with this situation.

The PCA theory: The most common derivation of PCA is in terms of a standardized linear projection, which maximizes the variance in the projected space (34). For a given *p*-dimensional data set *X*, the *m* principal axes T_1, T_2, $\cdots$ T_m, where, $1 \leq m \leq p$, are orthonormal axes onto which the retained variance is maximum in the projected space. Generally, T_1, T_2, $\cdots$ T_m can be given by the *m* leading eigenvectors of the sample covariance matrix $S=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T$, where x_i, $\in X$, μ is the sample mean, and *N* is the number of samples, so that

$$ST_i = \lambda_i T_i,\ i\in 1,\ \cdots,\ m, \tag{8}$$

where λ_i is the *i* th largest eigenvalue of *S*. The *m* principal components of a given observation vector $x_i \in X$ are given by

$$y=[y_1,\ y_2,\ \cdots y_m]=T^T x. \tag{9}$$

The *m* principal components of *x* are decorrelated in the projected space. In multi-class problems, the variations are determined on a global basis, that is, the principal axes are derived from a global covariance matrix:

$$S=\frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{Ni}(x_{ij}-\mu)(x_{ij}-\mu)^T, \tag{10}$$

where μ is the global mean of all the samples, *K* is the number of classes, N_i is the number of samples in class *i*, *N* is equal to $\sum_{i=1}^{k} N_i$, and x_{ij} represents the *j* th observation from class *i*. Finally, the principal axes T_1, T_2, $\cdots T_m$ are the *m* leading eigenvectors of *S*:

$$ST_i = \lambda_i T_i,\ i\in 1,\ \cdots,\ m, \tag{11}$$

where λ_i is the th largest eigenvalue of *S*. An assumption made for feature extraction and dimensionality reduction by PCA is that most information of the observation vectors is contained in the subspace spanned by the first *m* principal axes, where $m < p$. Therefore, each original data vector can be represented by its principal component vector with dimensionality *m*.

The Elman RNN classifier: In developing a method for predicting protein subcellular location, how to choose a proper machine learning algorithm is another important problem. Here, The Elman RNN is taken as a basic classifier, which was developed by J. L. Elman (19). This network type consists of an input layer, a hidden layer, and an output layer. In this way, it resembles a three layer feedforward neural network. However, it also has a context layer, in which the neurons hold a copy of the output of the hidden neurons. The value of each context neuron is used one time step later as an extra input signal for all the neurons in the hidden layer. The addition of interior feedback network increases the capability of processing dynamic information of the network itself, and therefore makes the system have the ability to adapt to time-varying characteristics.

Suppose there are *r* inputs, *m* outputs and *n* neurons, respectively, in the hidden layer and in the context layer. *u(k-1)* represents the inputs of Elman network; *x(k)* represents the outputs of the hidden layer; x_c *(k)* represents the outputs of the context layer, and *y(k)* represents the outputs of Elman network. Then, its nonlinear state-space expression is

$$\begin{cases} y(k)=g(w_2 x(k)), \\ x(k)=f(w_3 x_c(k)+w_1(u(k-1))), \\ x_c(k)=x(k-1), \end{cases} \tag{12}$$

where w_1 is the weight from input layer to hidden layer, w_2 from hidden layer to output layer and w_3 from context layer to hidden layer. *g* represents the transfer function of the output layer, which is usually a linear function. *f* represents of the hidden layer, *S* type function is commonly used and can be defined as

$$f(x)=(1+e^{-x})^{-1} \tag{13}$$

Bp algorithm with momentum of variable learning rate is used here to modify the weight values and the error of the network is

$$E=\sum_{i=1}^{m}(t_i-y_i)^2, \tag{14}$$

in which t_i *(i=*1, 2, $\cdots$*m)* are the output vectors of the object.

Parameters setting

As mentioned above, λ must be smaller than the number of amino acid residues of the shortest protein chain in the dataset. The shortest length of the MA629 dataset is 50 and the GA541 is 40, so the number of λ is set to 49 and 39, respectively, in order to contain the more sequence order information for ELM-PCA. Thus, the input neurons correspond to 49 + 20 = 69 and 39 + 20 = 59. Likewise, for the independent dataset, λ is set to 29 − 1 = 28, and the input neurons correspond to 28 + 20 = 40. Eight neurons are used for the hidden layer. During the training process, the generalization error is estimated in each epoch on a validation set. If the error does not change in five consecutive epochs, the training of the network is terminated in order to avoid overfitting.

The use of PCA for dimensionality reduction is based on saving only a small number of eigenvectors to represent the data. Here, 90% of the total data variance is applied to preserve energy and extract the most important features.

We use five-fold cross-validation for training and evaluating the prediction performance, in which a data set is divided into five subsets of approximately equal size. This means that the data is partitioned into training and test data in five different ways. After training a classifier with a collection of four subsets, the performance of the classifier is tested against the fifth subset. This process is repeated five times so that every subset is once used as the test data. In the tests, it is run by five times with intent to ensure the rationality of results, because Bp algorithm over multilayer networks is only guaranteed to converge toward some local minimum and not necessarily to the global minimum error (35).

Acknowledgements

This work was supported by Special Public Sector Research Funds for Earthquake Study (no.200808075).

REFERENCES

1. Glory, E. and Murphy, R. (2007) Automated subcellular location determination and high-throughput microscopy. *Dev. Cell* **12**, 7-16.
2. Chou, K. C. and Shen, H. B. (2008) Cell-ploc: a package of web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* **3**, 153-162.
3. Nakashima, H. and Nishikawa, K. (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Biol.* **238**, 54-61.
4. Chou, K. C. (1995) A novel approach to predicting protein structural classes in a (20-1)-d amino acid composition space. *Proteins: Struct. Funct. Genet.* **21**, 319-344
5. Cedano, J. Aloy, P. Perez-Pons, J. and Querol, E. (1997) Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol.* **266**, 594-600.
6. Shen, H. B. and Chou, K. C. (2005) Predicting protein subnuclear location with optimized evidence-theoretic k-nearest classifier and pseudo amino acid composition. *Biochem. Biophys. Res. Commun.* **337**, 752-756.
7. Lei, Z. and Dai, Y. (2005) An SVM-based system for predicting protein subnuclear localizations. *BMC Bioinformatics* **6**, 291-298.
8. Huang, W. Tung, C., Huang, H. and Ho, S. (2009) Predicting protein subnuclear localization using GO-amino-acid composition features. *Biosystems* **98**, 73-79.
9. Glory, E. and Murphy, R. F. (2007) Automated subcellular location determination and high-throughput microscopy. *Dev. Cell* **12**, 7-16.
10. Shen, H. B. and Chou, K. C. (2009) A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. *Anal. Biochem.* **394**, 269-274.
11. Chou, K. C. and Shen, H. B. (2008) Cell-PLoc: a package of web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* **3**, 153-162.
12. Chou, K. C. (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Struct. Funct. Genet.* **43**, 246-255.
13. Ding, Y. S. and Zhang, T. L. (2008) 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**, 1887-1892.
14. Shen, H. B. and Chou, K. C. (2007) PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.* **373**, 386-388.
15. Zeng, Y., Guo, Y., Xiao, R., Yang, L., Yu, L. and Li, M. (2009) Using the augmented chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. *J. Theor. Biol.* **259**, 366-372.
16. 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 Eng. Des. Sel.* **20**, 561-567.
17. Shen, H. B. and Chou, K. C. (2006) Ensemble classifier for protein fold pattern recognition. *Bioinformatics* **22**, 1717-1722.
18. Jolliffe, I. (2002) Principal component analysis. pp. 29-43, Springer-Verlag, Second Edition, New York, USA
19. Elman, J. (1990) Finding structure in time. *Cog. Sci.* **14**, 179-211.
20. Shi, X. H., Liang, Y. C., Lee, H. P., Lin, W. Z., Xu, X. and Lim, S. P. (2004) Improved elman networks and applications for controlling ultrasonic motors. *Appl. Artif. Intell.* **18**, 603-629.
21. Dehling, H., Fleurke, S. and Klske, C. (2008) Parking on a random tree. *J. Stat. Phys.* **133**, 151-157.
22. Witten, I. and Frank, E. (2005) Data Mining: practical machine learning tools and techniques. pp.189-283, Morgan Kaufmann Publishers, Second Edition, San Francisco, USA.
23. Yousef, M., Jung, S., Kossenkov, A., Showe, L. S. and Showe, M. (2007) Naive Bayes for microRNA target predictions- machine learning for microRNA targets. *Bioinformatics* **23**, 2987-2992.
24. Bhasin, M., Garg, A. and Raghava, G. P. S. (2005) PSLpred: prediction of subcellular localization of bacterial proteins. *Bioinformatics* **21**, 2522-2524.
25. Gardy, J., Laird, M., Chen, F., Rey, S., Walsh, C., Ester, M. and Brinkman, F. (2005) Psortb v. 2.0: expanded pre-

diction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* **21**, 617-623.

26. Yu, C. S., Lin, C. J. and Hwang, J. K. (2004) Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Sci.* **13**, 1402-1406.
27. Szafron, D., Lu, P., Greiner, R., Wishart, D., Poulin, B., Eisner, R., Lu, Z., Anvik, J., Macdonell, C., Fyshe, A. and Meeuwis, D. (2004) Proteome Analyst: custom predictions with explanations in a web-based tool for high-throughput proteome annotations. *Nucleic Acids Res.* **32**, 365-371.
28. Imai, K., Asakawa, N., Tsuji, T., Akazawa, F., Ino, A., Sonoyama, M. and Mitaku, S. (2008) SOSUI-GramN: high performance prediction for sub-cellular localization of proteins in Gram-negative bacteria. *Bioinformatics* **2**, 417-421.
29. Hoffmann, H. (2007) Kernel pca for novelty detection. *Pattern Recognition* **40**, 863-874.
30. Yang J., Gao, X., Zhang D. and Yang, J. Y. (2005) Kernel ICA: an alternative formulation and its application to face recognition. *Pattern Recognition* **38**, 1784-1787.
31. Yu, U., Lee, S. H., Kim, Y. J. and Kim, S. (2004) Bioinformatics in the post-genome era. *BMB Rep.* **37**, 75-82.
32. Ma, J. W., Liu, W. Q. and Gu, H. (2009) Predicting protein subcellular locations for gram-negative bacteria using neural networks ensemble. Proceedings of the 6th Annual IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, pp.114-120, Tennessee, USA.
33. Hinton, G. and Salakhutdinov, R. (2006) Reducing the dimensionality of data with neural networks. *Science* **313**, 504-507.
34. Yeung, K. Y. and Ruzzo, W. L. (2001) Principal component analysis for clustering gene expression data. *Bioinformatics* **17**, 763-774.
35. Bishop, C. (2006) Pattern recognition and machine learning. pp. 225-284. Springer, New York, USA.