A New Ensemble Scheme for Predicting Human Proteins Subcellular Locations

Abdul Majid[1, 2] and Tae-Sun Choi[1]

[1]Department of Mechatronics, Gwangju Institue of Science and Technology
[2]Department of Information and Computer Sciences, Pakistan Institute of Engineering and Applied Sciences
{abdulmajiid; tschoi}@gist.ac.kr

Abstract

Predicting subcellular localizations of human proteins become crucial, when new unknown proteins sequences do not have significant homology to proteins of known subcellular locations. In this paper, we present a novel approach to develop CE-Hum-PLoc system. Individual classifiers are created by selecting a fixed learning algorithm from a pool of base learners and then trained by varying feature dimensions of Amphiphilic Pseudo Amino Acid Composition. The output of combined ensemble is obtained by fusing the predictions of individual classifiers. Our approach is based on the utilization of diversity in feature and decision spaces. As a demonstration, the predictive performance was evaluated for a benchmark dataset of 12 human proteins subcellular locations. The overall accuracies reach upto 80.83% and 86.69% in jackknife and independent dataset tests, respectively. Our method has given an improved prediction as compared to existing methods for this dataset. Our CE-Hum-PLoc system can also be a used as a useful tool for prediction of other subcellular locations.

Keywords: *Subcellular location, ensemble classifier, individual classifier, Amphiphilic Pseudo Amino Acid Composition*

1. Introduction

The function of a protein is closely correlated with its specific location in the cell. It means precise protein function requires to locate properly its subcellular location, otherwise there is danger that protein may lose its function [1]. Information about subcellular location give understanding about their engagement in specific metabolic pathways [2]. The locations of proteins with known function help in understanding its biological function [3] and proteins interaction [4]. Newly synthesized proteins are localized to the appropriate subcellular spaces to perform their biological functions. Therefore, in large-scale genome analysis, demand to develop more accurate and reliable predictor is increasing [5].

In the literature, research related to accurately predict human proteins into various subcellular localizations has gained much importance. Researchers have proposed both individual and fusion of classifier strategies. Early attempts were based on the decision of a single learner. Covariant Discriminant Classifier (CDC) was attempted using different feature extraction techniques [5-10]. Support Vector Machines (SVM) classifier was tried with Functional Domain Composition [11] features. A SVM based prediction model was developed by constructing new Amino Acid Composition (AAC) distribution features [12]. The prediction of a single classifier is limited due to large variation in length and order of protein sequences. Therefore, researchers have also proposed fusion of classifiers strategies

[4, 13-15]. In this way, fusion of diverse types of classifiers often yield better prediction than the individual ones [16]. An ensemble of CDC classifiers is developed using Amphiphilic Pseudo Amino Acid Composition (PseAAC) [13]. An ensemble of KNN classifiers was built by fusing individual KNN classifiers to develop *Hum-PLoc* system [14]. In this work, individual classifiers are trained on the hybridized features of Gene Ontology and Amphiphilic PseAAC.

The above predicator schemes do not combine classifiers that are individually trained on different feature spaces and at the same time posses different learning mechanisms. In this paper, we propose a novel ensemble approach called *CE-Hum-PLoc*. The main idea of this scheme is based on the utilization of diversity in feature and decision spaces simultaneously. In the work, comparative analysis shows improved prediction accuracy than the existing approaches.

2. Materials and Methods

In this work, four learning mechanisms are selected as base learners: 1) 1-Nearest Neighbor (NN), 2) Probabilistic Neural Network (PNN), 3) SVM and 4) CDC. All learning mechanisms, except SVM, are inherently based on proximity. SVM is a margin based binary classifier that constructs a separation boundary to classify data samples. CDC and NN classifiers are commonly used for predicting protein sequences. The prediction of CDC is found by exploiting the variation in the PseAA features of protein sequence [13]. NN is reported to perform well on classification tasks regarding protein sequences [13, 17, 18]. PNN classifier is based on the Bayes theory to estimates the likelihood of a sample being part of a learned class [19].

The detail of benchmark datasets is provided in Table 1. This dataset was developed by Chou and Elrod [7]. Later on, researchers adopted this dataset to compare the results of their proposed methods. To reduce redundancy and homology bias, this datasets was passed through window screening.

2.1. Proposed method

Individual ensembles (IEs) are produced by exploiting diversity in feature spaces. Combined ensemble (CE) is then developed by fusing predictions of IEs classifiers. CE classifier is expected to be more effective as compared to IEs classifiers. Suppose we have N proteins feature vectors (**P$_1$, P$_2$ . . . P$_N$**) derived from protein dataset. Each **P$_i$** belong to one of V classes with labels $Q_1, Q_2,..,Q_V$. A k^{th} subcellular protein feature vector from a class v can be expressed as:

$$\mathbf{P}_v^k = \left[p_{v,1}^k \; p_{v,2}^k \cdots p_{v,20}^k \cdots p_{v,\Phi}^k\right]^{\mathbf{T}} \qquad \textbf{(1)}$$

where $p_{v,1}, p_{v,2}, ..., p_{v,20}$ are the frequencies of occurrence of 20 amino acid sequences. The elements $p_{v,21}, p_{v,22}, ..., p_{v,\Phi}$ are the 1st-tier to (ξ-1)-tier correlation factors of an amino acid sequence in the protein chain based on two indices of hydrophobicity and hydrophilicity. In order to develop IEs, first, *PseAA* composition with varying dimensions ranging from 20 to 62 is utilized, i.e. $\Phi=20+2(i-1)$, where $i=1,2,..., \xi$. Here, ξ =22, represents the number of *IE* classifiers. The individual predictions R$_i$ of IE classifiers can be expressed as:

$$\{R_1, R_2, R_3 \cdots, R_\xi\} \in \{Q_1, Q_2, Q_3, \cdots, Q_V\} \qquad \textbf{(2)}$$

Now IE based voting mechanism for a protein can be formulated as:

$$Z_j^{IE} = \sum_{i=1}^{\xi} w_i \Delta(R_{i,} Q_j), \quad j = 1,2,\ldots V \tag{3}$$

where w_i represents weight factor. Here, for simplicity, its value is set to unity and function $\Delta(R_{i,} Q_j)$ is defined as: $\Delta(R_{i,} Q_j) = \begin{cases} 1, & \textit{if } R_i \in Q_j \\ 0, & \textit{otherwise} \end{cases}$.

Finally, the query protein is assigned the class γ that obtains maximum votes:

Table 1. Number of proteins sequences in each subcellular location

Sr. no.	Subcellular locations	Dataset	
		Jackknife test	Independent test
1	Chloroplast	145	112
2	Cytoplasm	571	761
3	Cytoskeletons	34	19
4	Endo. Reticulum	49	106
5	Extracell	224	95
6	Golgi Apparatus	25	4
7	Lysosome	37	31
8	Mitochondria	84	163
9	Nucleus	272	418
10	Peroxisome	27	23
11	Plasma Memb.	699	762
12	Vacuole	24	---
	Total	2,191	2,494

$$Z_\gamma^{IE} = Max\left\{Z_1^{IE}, Z_2^{IE}, \ldots Z_V^{IE}\right\} \tag{4}$$

In the second step, the aim was to combine the diverse decision spaces generated by *IE* classifiers. In this way, the shortcoming of one classifier can be overcome by the advantage of others. Let *l=1,2,3,..., L* represents the number of different base learners in the entire-pool voting. We compute the votes of each class for *CE* as:

$$Z_j^{CE} = \sum_{i=1}^{L*\xi} w_i \Delta(R_{i,} Q_j), \quad j = 1,2,\ldots V \tag{5}$$

The predicted class τ by the *CE* classifier will be decided by using the *Max* function:

$$Z_\tau^{CE} = Max\left\{Z_1^{CE}, Z_2^{CE}, \ldots Z_V^{CE}\right\} \tag{6}$$

In jackknife test, if a tie occurs for a query protein; then decision of the highest performing IE^{SVM} classifier is taken. If the highest performing ensemble also delivers a tie, then the vote of the 2nd highest performing IE^{NN} ensemble is considered. Results reported in the Table 2 justify this action.

2.2. Evaluation methods

In the literature of Bioinformatics, both independent and jackknife tests are used by the leading investigators to evaluate the performance of their prediction methods [4-10, 12-15, 20]. The jackknife test is considered as the most rigorous and objective [21]. This test is conducted for the cross validation based performance analysis. During this test, each protein sequence is singled out as a test sample and remaining samples are used to train. The overall percent accuracy (*acc*.%) is calculated as:

$$acc.\% = \frac{\sum_{i=1}^{V} p(i)}{N} \times 100 \tag{7}$$

where, V is the 12 class number, $p(i)$ is the number of correctly predicted sequences of location i.

The numerical value of Q-statistic indicates the independency of component classifiers [22]. For any two base classifiers C_i and C_j, the Q-statistic is defined as:

$$Q_{i,j} = \frac{ad - bc}{ad + bc} \tag{8}$$

where, a and d represent the frequency of both classifiers making correct and incorrect predictions, respectively. However, b shows the frequency when first classifier is correct and second is incorrect; c is the frequency of second classifier being correct and first incorrect. The average value of Q-statistic among all pairs of L base classifiers in CE ensemble is calculated as:

$$Q_{avg} = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{k=i+1}^{L} Q_{i,k} \tag{9}$$

The positive value of Q_{avg} shows that classifiers recognize the same objects correctly. The positive value of Q_{avg} (<1) shows the diversity level of base classifiers in the CE.

3. Results and Discussions

In this section, the overall accuracy and Q-statistic of IEs and CE are computed and results are given in the Table 2. This table indicates average Q-statistics of IEs and CE, in jackknife test, are in the range of 0.94-0.74 and 0.63, respectively. In case of independent test, Q-values of IEs and CE are in the range of 0.96-0.69 and 0.86, respectively. This highlights sufficient diversity in IEs and CE. This diversity in individual learners is accumulated that result improvement in CE. Further, in jackknife test, this table shows better prediction accuracy of IE^{SVM} among other IEs. IE^{SVM} predicts correctly 1738 out of 2191 sequences. Thus it gives an overall accuracy of 79.32%. Therefore, if SVM based learning mechanism is incorporated as a base classifier, then chance of CE improved enhances. IE^{NN} predicts correctly 1665 out of 2191 sequences and it gives overall accuracy of 76.01%. This predicted accuracy of IE^{NN} is comparable with IE^{PNN}. The prediction accuracies of IEs are also investigated on independent dataset containing 2494 protein sequences. The results show IE^{NN} correctly predicts 2115 protein sequences and gives an overall accuracy of 84.85%. In this case, the prediction accuracy of IE^{PNN} is lower than IE^{NN}. For independent dataset test, overall prediction of IE^{SVM} is not appreciable (69.52%). However, to improve SVM prediction, there is a need to find optimal kernel parameters.

Table 2. Performance comparison of IEs vs. CE

IEs/CE	Jackknife test			Independent test		
	Correct Prediction	Acc. %	Avg. Q statistics	Correct Prediction	Acc. %	Avg. Q statistics
IE^{SVM}	1738	79.32	0.94	1734	69.52	0.96
IE^{CDC}	1600	73.60	0.74	1965	78.79	0.69
IE^{NN}	1665	76.01	0.92	2115	84.85	0.93
IE^{PNN}	1686	76.99	0.85	2057	82.48	0.93
CE	1771	80.83	0.63	2162	86.65	0.86

Table 3. Summary of comparative analysis

Prediction methods		Jackknife test	Independent test	Ref
Input features form	Prediction algorithm	Acc. %	Acc. %	
28 PseAAC component	Aug. CDC	1590/2191=72.6	1865/2494=74.8	[20]
PseAAC formation via three types filters	Aug. CDC	1532/2191=69.9	----------	[6]
Simple AAC	CDC	1492/2191=68.	1888/2494=75.7	[7]
PseAAC	CDC	1600/2191=73	2017/2494=80.9	[8]
PseAAC generated by DSP	Aug. CDC	1483/2191=67.68	1842/2494=73.86	[9]
Lempel-Ziv complexity	Aug.CDC	1612/2191=73.6	1990/2494=79.8	[10]
Quasi-sequence-order	Aug. CDC	1588/2191=72.5	1985/2494=79.6	[5]
Functional Domain Composition	SVM	1461/2191=66.7	2037/2494=81.7	[11]
AAC distribution	SVM	1800/2191=82.15	2132/2494=85.49	[12]
PseAAC	CE-classifier	1771/2191=80.83	2162/2494=86.69	Our method

In Table 3, we summarize the results of existing approaches and then compare with our method. The basic reason for this comparison was; these researchers have used the same human protein dataset and estimation tests to evaluate their prediction algorithm, except Shi et al. [12]. This table indicates that our CE classifier, in jackknife test, correctly classifies 1771 protein sequences out of 2191 giving an overall accuracy of 80.83%. However, for independent test, CE correctly classifies 2162 protein sequences out of 2494 to give an overall accuracy of 86.69%. This shows an improved in prediction accuracy of our approach as compared to the existing approaches proposed, except [12].

By comparing our results with Shi et al., we obtained a comparable performance. Average accuracies of both methods come out to be nearly equal, i.e. 83.74% and 83.82%. Currently, we have utilized a simple feature extraction strategy and exhaustive single-one-out cross validation test. However, Shi et al. have developed a complex feature extraction and prediction results are estimated using 5-fold cross validation.

4. Conclusion

In this paper, we have developed a new ensemble in predicting human protein into 12 subcellular localizations. The proposed CE-Hum-PLoc system delivers more accurate predictions than existing approaches, except [12]. This improvement was made possible by exploiting diversity in feature and decision spaces simultaneously. Currently, we have attempted with four base classifiers and one feature extraction strategy. However, by adding more base learners, further improvement is possible.

Acknowledgement

This research was supported by the Bio Imaging Research Center at Gwangju Institute of Science and Technology (GIST), South Korea.

References

[1] P. Bjorses, M. Halonen, J. J. Palvimo, M. Kolmer, J. Aaltonen, P. Ellonen, J. Perheentupa, I. Ulmanen, and L. Peltonen, "Mutations in the AIRE gene: Effects on subcellular location and transactivation function of the autoimmune polyendocrinopathy-candidiasis - Ectodermal dystrophy protein," American Journal of Human Genetics, vol. 66, 2000, pp. 378-392.

[2] A. Garg, M. Bhasin, and G. P. S. Raghava, "Support Vector Machine-based Method for Subcellular Localization of Human Proteins Using Amino Acid Compositions, Their Order, and Similarity Search," J. Biol. Chem., vol. 280, 2005, pp. 14427-14432.

[3] A. Reinhardt and T. Hubbard, "Using neural networks for prediction of the subcellular location of proteins," Nucleic Acids Research, vol. 26, 1998, pp. 2230-2236.

[4] Y. Shen and G. Burger, "'Unite and conquer': enhanced prediction of protein subcellular localization by integrating multiple specialized tools," BMC Bioinformatics, vol. 8, 2007, p. 420.

[5] K. C. Chou, "Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Order Effect," Biochemical and Biophysical Research Communications, vol. 278, 2000, pp. 477-483.

[6] Y. Gao, S. Shao, X. Xiao, Y. Ding, Y. Huang, Z. Huang, and K. C. Chou, "Using pseudo amino acid composition to predict protein subcellular location: Approached with Lyapunov index, Bessel function, and Chebyshev filter," Amino Acids, vol. 28, 2005, pp. 373-376.

[7] K. C. Chou and D. W. Elrod, "Protein subcellular location prediction," Protein Eng, vol. 12, 1999, pp. 107 - 118.

[8] C. Kuo-Chen, "Prediction of protein cellular attributes using pseudo-amino acid composition," Proteins: Structure, Function, and Genetics, vol. 43, 2001, pp. 246-255.

[9] Y. X. Pan, Z. Z. Zhang, Z. M. Guo, G. Y. Feng, Z. D. Huang, and L. He, "Application of Pseudo Amino Acid Composition for Predicting Protein Subcellular Location: Stochastic Signal Processing Approach," Journal of Protein Chemistry, vol. 22, 2003, pp. 395-402.

[10] X. Xiao, S. Shao, Y. Ding, Z. Huang, Y. Huang, and K. C. Chou, "Using complexity measure factor to predict protein subcellular location," Amino Acids, vol. 28, 2005, pp. 57-61.

[11] K. C. Chou and Y. D. Cai, "Using functional domain composition and support vector mchines for prediction of protein subcellular location," J. Biol. Chem., vol. 277, 2002, pp. 45765-45769.

[12] J. Y. Shi, S. W. Zhang, Q. Pan, and G. P. Zhou, "Using pseudo amino acid composition to predict protein subcellular location: approached with amino acid composition distribution," Amino Acids, vol. 35, 2008, pp. 321-327.

[13] K. C. Chou and H. B. Shen, "Predicting protein subcellular location by fusing multiple classifiers," Journal of Cellular Biochemistry, vol. 99, 2006, pp. 517 - 527.

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

[15] H. B. Shen and K. C. Chou, "Virus-PLoc: A fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells," Biopolymers, vol. 85, 2007, pp. 233-240.

[16] S. Dzeroski and B. Zenko, "Is combining classifiers with stacking better than selecting the best one?," Machine Learning, vol. 54, 2004, pp. 255-273.

[17] P. Jia, Z. Qian, Z. Zeng, Y. Cai, and Y. Li, "Prediction of subcellular protein localization based on functional domain composition," Biochemical and Biophysical Research Communications, vol. 357, 2007, pp. 366-370.

[18] A. Khan, M. Fayyaz, and T. S. Choi, "Proximity based GPCRs prediction in transform domain," Biochemical and Biophysical Research Communications, vol. 371, 2008, pp. 411-415.

[19] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: John Wiley & Son s, Inc., 2001.

[20] X. Xiao, S. Shao, Y. Ding, Z. Huang, and K. C. Chou, "Using cellular automata images and pseudo amino acid composition to predict protein subcellular location," Amino Acids, vol. 30, 2006, pp. 49-54.

[21] K. C. Chou and C. T. Zhang, "Prediction of Protein Structural Classes," Critical Reviews in Biochemistry and Molecular Biology., vol. 30, 1995, pp. 275 - 349.

[22] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," MachLearn, vol. 51, 2003, pp. 181-207.

Authors

ABDUL MAJID received his M.Sc. degree in Electronics from Quaid-i-Azam University, Islamabad, Pakistan in 1991. He received his M.S. and Ph.D. degrees in Computer Systems Engineering from Ghulam Ishaq Khan Institute of Engineering Sciences and Technology (GIK Institute), Topi, Pakistan, in 2003 and 2006, respectively. He has more than 13 years of research experience and is working as Assistant Professor in Department of Computer and Information Sciences at PIEAS. His research areas include, Pattern Recognition, Image Processing, Machine Learning and Computational Materials Science.

TAE-SUN CHOI received his B.S. degree in Electrical Engineering from the Seoul National University, Seoul, Korea, in 1976, and his M.S. degree in Electrical Engineering from the Korea Advanced Institute of Science and Technology, Seoul, Korea, in 1979, and his Ph.D. degree in electrical engineering from the State University of New York at Stony Brook, in 1993. He is currently a Professor in the Department of Mechatronics at Gwangju Institute of Science and Technology, Gwangju, Korea. His research interests include Image Processing, Machine/Robot