Prediction of Protein Subcellular Locations by Combining K-Local Hyperplane Distance Nearest Neighbor*

Hong Liu, Haodi Feng, and Daming Zhu

School of Computer Science and Technology, Shandong University, Jinan 250061, Shan-dong Province, People's Republic of China
{Hong-liu, Fenghaodi, Dmzhu}@sdu.edu.cn

Abstract. A huge number of protein sequences have been generated and collected. However, the functions of most of them are still unknown. Protein subcellular localization is important to elucidate protein function. It would be worthwhile to develop a method to predict the subcellular location for a given protein when only the amino acid sequence of the protein is known. Although many efforts have been done to accomplish such a task, there is the need for further research to improve the accuracy of prediction. In this paper, with K-local Hyperplane Distance Nearest Neighbor algorithm (HKNN) as base classifier, an ensemble classifier is proposed to predict the subcellular locations of proteins in eukaryotic cells. Each basic HKNN classifiers are constructed from a separated feature set, and finally combined with majority voting scheme. Results obtained through 5-fold cross-validation test on the same protein dataset showed an improvement in pre-diction accuracy over existing algorithms.

Keywords: Protein, Subcellular Location, Ensemble Classifier.

1 Introduction

As a result of the Human Genome Project and related efforts, protein sequence data accumulate at an accelerating rate. This raises the challenge of understanding the functions of proteins from high throughput sequencing projects. Protein subcellular localization is a key functional characteristic of proteins [1] and correct prediction of protein subcellular localization will greatly help in understanding its functions. However, experimental determination of subcellular location is time-consuming and costly. Therefore, a reliable and efficient computational method is highly required to construct prediction systems to predict the subcellular location for a given protein when only the amino acid sequence of the protein is known.

Several machine learning techniques have been applied to construct such prediction systems, for example, Support Vector Machines (SVM) [2-4], Neural Network [5,6], Naïve Bayesian [7] and Fuzzy KNN [8], using different sets of

* Supported by National Science Foundation of China under grant No. 60603007 and Science and Technology Development Foundation of Shandong Province, China under grant No. 2006GG2201005.

R. Alhajj et al. (Eds.): ADMA 2007, LNAI 4632, pp. 345–351, 2007.

features extracted from amino acid composition [2,3,6,9], amino acid pair composition[9], gapped amino acid composition [2], pseudo amino acid composition [10], evolutionary and structural information [11] and motif information [12]. While SVM, Neural Network, Naïve Bayesian need a separate training procedure, Fuzzy KNN not. Given the various sets of features, one approach to make use of them is combining (or ensembling) different classifiers constructed from different set of features. By this means, reduction in variance caused by the peculiarity of a single feature set and consequently more reliable and stable prediction system could be obtained.

In this paper, we present a new method based on an ensemble of K-local Hyperplane Distance Nearest Neighbor algorithm (HKNN) [13] for the prediction of protein subcellular locations. Experimental results obtained through 5-fold cross-validation tests on the same protein dataset showed an improvement in prediction accuracy over existing algorithms.

The rest of this paper is organized as follows. Section 2 presents the methods proposed. Experimental results and comparison with existing methods are presented in Section 3. Conclusion is given in Section 4

2 Methods

2.1 Feature Presentation

Protein sequences are composed of amino acids, which are denoted by letters from the alphabet {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, W, V, Y}. In order to be able to perform computation on these sequences, non-numerical amino acids should be represented by numerical values.

From amino acid composition, we extracted the first set of features. In detail, for each protein sequence, the occurrence frequency of each amino acid (letter) was calculated and normalized (i.e. divided by the length of the sequence minus one). Each of the number obtained corresponds to an element of a 20-dimension vector (see Fig. 1). That is, a protein sequence was mapped as a point in a feature space with dimension 20.

Prediction based on only amino acid composition features would lose sequence order information. Thus, to capture this kind of information, amino acid pair composition and gapped amino acid composition [2] were considered. While the former corresponds to two adjacent amino acids, the latter corresponds to two amino acids separated by one or more intervening residue positions. In fact, amino acid pair composition can be seen as a special case of gapped amino acid composition with zero gaps. For four different gap values (i.e. 0, 1, 2, 3), we extracted four different set of features separately. In detail, since there are 20 different amino acids, we considered $20 \times 20 = 400$ amino acid pairs for each gap value. For each protein sequence, the occurrence frequency of each gapped pair was calculated and normalized (i.e. divided by the length of the sequence minus 3). Thus, a protein sequence was converted to other four different 400-dimension vectors one for each

gap value (see Fig. 1) or, in other words, was mapped as a point in other four different feature spaces with dimension 400.

Since amino acids have different biochemical and physical properties that influence their relative replace-ability in evolution, we re-substituted the 20-letter amino acids by 9-letter amino acids according to their physicochemical properties, as illustrated in Table 1[14]. Based on this new encoding scheme, each protein sequence was converted to other five different vectors using the similar process as in 20-letter encoding scheme case (see Fig. 1).

So far, for each protein sequence, we got ten feature vectors. In other words, each protein sequence was mapped as a point in ten different feature spaces.

Table 1. The 9-letter encoding scheme for the 20 amino acids based on their physical-chemical proper-ties

Group	Residues	Description
1	C	Highly conserved
2	M	Hydrophobic
3	N, Q	Amides, polar
4	D, E	Acids, positive, polar
5	S, T	Alcohols
6	P, A, G	Aliphatic, small
7	I, V, L	Aliphatic
8	F, Y, W	Aromatic
9	H, K, R	Bases, charged

2.2 K-Local Hyperplane Distance Nearest Neighbor Algorithm (HKNN)

HKNN is a modified k-nearest neighbor algorithm (KNN). By building a (non-linear) decision surface, separating different classes of the data, directly in the original feature space, it is intended to improve the classification performance of the conventional KNN to a level of SVM.

Suppose the number of different classes in the training set is c, HKNN computes distances of a test point $\mathbf{x}$ to c local hyperplanes, where each hyperplane is composed of k nearest neighbors of $\mathbf{x}$, belonging to the same class, in the training set. Then the test point $\mathbf{x}$ is assigned to the class whose hyperplane is closest to $\mathbf{x}$ (see [13] for details).

HKNN has two parameters, k and λ, a penalty term introduced to find the hyperplane.

2.3 Voting Scheme

In each of the ten feature space described above, a HKNN classifier was constructed. To combine the prediction results of all the ten classifiers, majority voting scheme [15] was used, in which the final prediction class was the most voting one. Ties were randomly resolved.

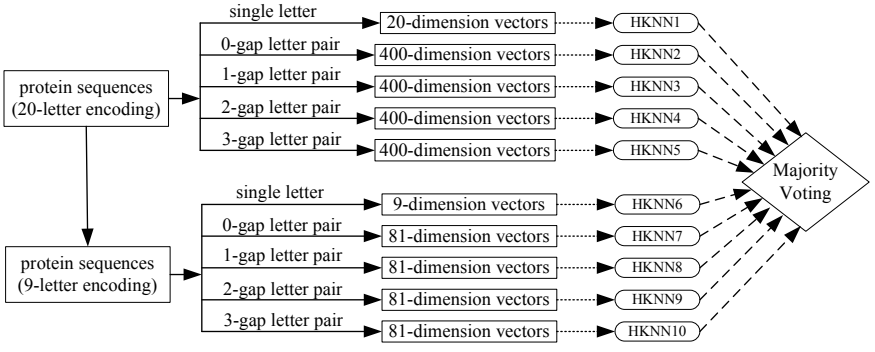

Fig. 1. Framework of the method used in this paper

3 Experimental Results and Discussion

3.1 Protein Dataset

For comparison, the protein dataset (downloadable at "http://web.kuicr.kyoto-u.ac.jp/~park/Seqdata/") studied in previous investigations [2] were used in this study. In this dataset, all protein sequences were collected from the SWISS-PROT database release 39.0. Totally 7579 protein sequences of eukaryotic cells for 12 subcellular locations were contained in this dataset. The number of protein sequences in each location is shown in Table 2.

3.2 Performance Measurement

The prediction performance was evaluated by 5-fold cross-validation test. In detail, proteins in the dataset were separated into five balanced sets. Each of these sets contained almost the same number of proteins. In each round of cross-validation, four sets were used to construct the ensemble of HKNNs while one set was set aside for evaluating the method. This procedure was repeated five times, once for each set. In order to evaluate the prediction performance of our method, two measures were used. The first measure, total accuracy (TA), is defined as

$$TA = \frac{\sum_{i=1}^{c} T_i}{N} \tag{1}$$

The second measure, local accuracy (LA), is defined as

$$LA = \frac{\sum_{i=1}^{c} P_i}{c} \tag{2}$$

In Equation (1) and (2), c is the number of subcellular locations, N is the total number of proteins in the dataset, T_i is the number of proteins correctly predicted in location i, and $P_i=T_i/n_i$, where n_i is the number of proteins in location i. These two measures were also used in [2].

Table 2. The 9-letter encoding scheme for the 20 amino acids based on their physical-chemical proper-ties

Subcellular Location	Number of protein sequences
Chloroplast	671
Cytoplasmic	1241
Cytoskeleton	40
Endoplasmic reticulum	114
Extracellular	861
Golgi apparatus	47
Lysosomal	93
Mitochondrial	727
Nuclear	1932
Peroxisomal	125
Plasma membrane	1674
Vacuolar	54
Total	7579

Table 3. Prediction performance for the 12 subcellular locations

Subcellular Location	Accuracy (%)
Chloroplast	78.1
Cytoplasmic	73.2
Cytoskeleton	67.5
Endoplasmic reticulum	71.9
Extracellular	76.1
Golgi apparatus	42.6
Lysosomal	71.0
Mitochondrial	55.7
Nuclear	92.2
Peroxisomal	44.0
Plasma membrane	94.5
Vacuolar	46.3
LA	67.8
TA	80.9

3.3 Results

In this paper, ten HKNNs with parameters k=4 and λ=0.8 were ensembled and test on the dataset. The total accuracy (TA) and location accuracy (LA) were calculated by 5-fold cross-validation. The prediction accuracy for each subcellular location is shown in Table 3.

3.4 Comparison with Previous Methods

In order to examine the performance of our method, we made comparison with previous methods, especially the methods by Park and Kanehisa[2] who had used the

same dataset and 5-fold cross-validation test. The comparison results are shown in Table 4. Our method improved the LA significantly, from 57.9% to 67.8%, together with a small increase in the TA, from 78.2% to 80.9%. Our method is more balanced than their method, that is, our method performs better than theirs for all small groups, such as Cytoskeleton (58.5% vs. 67.5%), Endoplasmic reticulum (46.5% vs. 71.9%), Golgi apparatus (14.6% vs. 42.6%), Lysosomal (61.8% vs. 71.0%), Peroxisomal (25.2% vs. 44.0%) and Vacuolar (25.0% vs. 46.3%).

Table 4. Comparison our method with a previous method

Subcellular Location	Park and Kanehisa[2]	Our Method
Chloroplast	72.3	**78.1**
Cytoplasmic	72.2	**73.2**
Cytoskeleton	58.5	**67.5**
Endoplasmic reticulum	46.5	**71.9**
Extracellular	78.0	76.1
Golgi apparatus	14.6	**42.6**
Lysosomal	61.8	**71.0**
Mitochondrial	57.4	55.7
Nuclear	89.6	**92.2**
Peroxisomal	25.2	**44.0**
Plasma membrane	92.2	**94.5**
Vacuolar	25.0	**46.3**
LA	57.9	**67.8**
TA	78.2	**80.9**

4 Conclusion and Future Work

In this paper, a new method based on an ensemble of K-local Hyperplane Distance Nearest Neighbor algorithm is proposed to predict protein subcellular locations of eukaryotic cells. The experimental results on the same dataset showed an improvement in prediction accuracy over existing algorithms. Other advantages of our method include: (1) it has relatively fewer adjustable parameters compared with SVM and Neural Network and (2) like KNN, it does not need a training process.

In the future, we will develop the ensembling classifier as a web service for public usage.

References

1. Chou, K.C.: Review: prediction of protein structural classes and subcellular locations. Current Protein and Peptide Science 1, 171–208 (2000)
2. Park, K.J., Kanehisa, M.: Prediction of Protein Subcellular Locations by Support Vector Machines using Compositions of Amino Acids and Amino Acid Pairs. Bioinformatics 19(13), 1656–1663 (2003)
3. Hua, S.J., Sun, Z.R.: Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17(8), 721–728 (2001)

4. Matsuda, S., et al.: A novel representation of protein sequences for prediction of subcellular location using support vector machines. Protein Science 14, 2804–2813 (2005)
5. Cai, Y.D., et al.: Artificial neural network model for predicting protein subcellular location. Computers and Chemistry 26, 179–182 (2002)
6. Emanuelsson, O., et al.: Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. Journal of Molecular Biology 300(4), 1005–1016 (2000)
7. Lu, Z., et al.: Predicting subcellular localization of proteins using machine-learned classifiers. Bioinformatics 20(4), 547–556 (2004)
8. Huang, Y., Li, Y.: Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics 20(1), 21–28 (2004)
9. Nakashima, H., Nishikawa, K.: Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. Journal of Molecular Biology 238(1), 54–61 (1994)
10. Chou, K.C.: Prediction of protein cellular attributes using pseudo-amino-acid-composition. Proteins 43(3), 246–255 (2001)
11. Nair, R., Rost, B.: Better prediction of sub-cellular localization by combining evolutionary and structural information. Proteins 53, 917–930 (2003)
12. Cai, Y.D., et al.: Support vector machines for predicting membrane protein types by using functional domain composition. Biophysical Journal 84(5), 3257–3263 (2003)
13. Vincent, P., Bengio, Y.: K-local hyperplane and convex distance nearest neighbor algorithms. In: Dietterich, T.G., Becker, S., Ghahramani, Z. (eds.) NIPS. Advances in Neural Information Processing Systems, vol. 14, pp. 985–992. MIT Press, Cambridge (2002)
14. Yang, M.Q., Yang, J.Y.: Identification of Intrinsically Unstructured Regions in Proteins Using Primary Structure. In: Arabnia, H.R., Valafar, H. (eds.) BIOCOMP'06. Proceedings of the 2006 International Conference on Bioinformatics & Computational Biology, pp. 303–309. CSREA Press (2006)
15. Freund, Y.: Boosting a weak learning algorithm by majority. Information and computation 121(2), 256–285 (1995)