Protein subcellular localization classification from multiple subsets of amino acid pair compositions

Thai Quang Tung[1], Doheon Lee[2*], Jong Tae Lim[1], Kwang Hyung Lee[2]

[1] Department of Computer Engineering, Kongju National University, Kongju, Korea

[2] Department of BioSystems, KAIST, Daejeon, Korea

[*]To whom correspondence should be addressed.E-mail: dhlee@bisl.kaist.ac.kr

Abstract

Subcellular localization is a key functional char acteristic of proteins. With the number of sequences entering databanks rapidly increasing, the importance of developing a powerful tool to identify protein subcellular location has become self -evident. In this paper, we introduce a novel method for predic ting protein subcellular locations from protein sequences. The main idea was motivated from the observation that amino acid pair composition data is redundant. By classifying from multiple feature subsets and using many kinds of amino acid pair composition s, we forced the classifiers to make uncorrelated errors. Therefore when we combined the predictors using a voting scheme, the prediction accuracy c ould be improved. Experiment was conducted on several data sets and significant improvement has been achieve d in a jackknife test.

Introduction

The localization of a protein in a cell is closely correlated with its biological function. Experimental determination of subcellular location is time-consuming and costly. With the number of sequences entering databa nks rapidly increasing, the importance of developing a powerful tool to identify protein sub-cellular location automatically has become self -evident.

Several methods and systems have been developed. Most of them fall into two categories: one is based on pr ediction of individual sorting signals; the other is based on global properties of protein sequences such as amino acid composition, dipeptide frequenc y. The former approach has a clear biological implication because newly synthesized proteins *in vivo* are governed by an intrinsic sequence to their destination, whether they are to pass through a membrane into a particular organelle, to become integrated into the membrane, or to be exported out of the cell [23]. Nakai and Kanehisa [10] were the first who proposed to predict the subcellular location of proteins based on N -terminal sorting signals. This approach was integrated eventually into the PSORT prediction system. Nielsen [17][18] worked extensively on identifying individual sorting signals using neural networks. Then they combined these individual predictions into a integrated system - TargetP [7] for subcellular loca tion prediction. Chou [4] provided a comprehensive review of predicting protein signal sequences. However, as pointed out by Reinhardt and Hubbard [19], "in large genome analysis projects, genes are u sually automatically assigned and these assignments are often unreliable for the 5' -region. This can lead to the leader sequences being missing or partially included, thereby causing the problems for prediction algorithms depending on them" Therefore, most of recent research focused on the second approach.

Cedano [3] pointed out in his paper that subcellular location is correlated with amino acid composition. Several machine learning methods were then employed in conjunction wit h the amino acid composition of protein sequences. Reinhardt and Hubburt [19] used neural networks while Sun and Hua used SVMs [8] and their overall accuracies were 66% and 79% respectively on a datas et of 2427 eukaryotic proteins in four locations. Park and Kanehisa [9] also used SVMs but they tried on different sequence features represented by the amino acid, amino acid pair compositions and their accuracy was 78.2% on a dataset of 7589 eukaryotic proteins classified in 12 locations. Recently, Ying [21] employed a fuzzy k-NN algorithm based on amino acid dipeptides and his result was 80.1% on a new dataset which contained 7203 proteins classified in 11 locations.

In this paper, we introduce a new classification method based on multiple subsets of amino acid pair compositions. The idea comes from the observation that amino acid pair composition data is quite redundant. Building nearest neighbor classifiers based on randomly selected subsets, we force the classifiers to make different and hopefully uncorrelated errors. Therefore by applying a voting scheme the prediction accuracy can be improved.

Material and method

Dataset

In this research, for training the classifier, we used PLOC dataset [9]. All sequences in this dataset were collected from SWISS-PROT database release 39.0. Eukaryotic proteins with specific subcellular locations are identified according to the annotation information the CC (comments or notes) and OC (organism classification) fields of SWISS-PROT. Several keywords that specified subcellular locations were identified to search against the categorization of subcellular locations (-!- SUBCELLULAR LOCATION) in the CC field). There were totally 12 categories of protein locations: choloroplast, cytoplasmic, cytoskeleton, endoplasmic reticulum, extracellular, Golgi apparatus, lysosomal, mitochondrial, nuclear, peroxisomal, plasma membrane and vacuolar proteins. Proteins annotated with two or more subcellular locations were not included in the dataset.

Collected sequences with a high degree of similarity to the others were removed by all-to-all sequence similarity searching using the program ALIGN [11], which produces an optimal global alignment between two protein sequences. Sequences with full length matches of 80% similarity were placed in one group. Then for each group, a sequence was randomly selected as the representative entry. The final dataset is summarized in Table 1.

Subcellular location	Number of entries
Choloroplast	671
Cytoplasmic	1245
Cytoskeleton	41
Endoplasmic reticulum	114
Extracellular	862
Golgi apparatus	48
Lysosomal	93
Mitochondrial	727
Nuclear	1932
Peroxisomal	125
Plasma membrane	1677
Vacuolar	54
Total	**7589**

Table 1 Number of proteins used in the data set

Amino acid pair compositions

In order to keep as much as possible the sequence order property, protein sequences are represented by amino acid dipeptide, gapped amino acid pair and double gapped amino acid pair compositions. We expected that these different representations can detect different sequence features. Each protein in the training data set is characterized by a vector x_i (i=1..N) where N is the number of features, together with a label specifying the category of subcellular location. With 20 amino acids, there are 400 features for each kinds of amino acid pair composition. Therefore we need 1200 (N=1200) features to represent one protein sequence. In this research, we did not use amino acid composition as features because having only 20 features made it not suitable for subset selection (see the algorithm).

Algorithms

It has been shown in previous works that fuzzy k-NN worked very well in subcellular locations prediction problem. In this paper we used nearest neighbor algorithm for classification followed by a voting scheme. Nearest neighbor (NN) algorithm is a simple non-parametric classification algorithm [6]. Despite its simplicity, it can give competitive performance compared to many other methods. Given a test sample of unknown label, it finds the nearest neighbor entry in the training set and assigns a label to the test sample according to label of the neighbor.

Due to the fact that amino acid pair composition data is quite redundant. Since when we did classifying with only half of 400 amino acid dipeptides, the overall accuracy did not decrease much in comparison with that of classification with 400 features. The same thing happened to other kinds of amino acid pair composition data. Therefore, instead of building classifier using the whole features, we did classification from multiple feature subsets. The algorithm for classification from multiple feature subsets (MFS) was proposed by Stephen Bay [20]. It is simple and can be stated as:

Using simple voting, combine the outputs from multiple NN classifiers, each having access only to a random subset of features.

```
Input:
    n: number of subsets
    m: number of features for each subsets
    x: pattern for classifying
Output:
    y: pattern's category

Begin
Initialize values for category hit tablecount
for (1:n)
    for each type of amino acid pair composition data
        - select m features randomly
        - do NN classification on those features
        - update category hit tablecount
    end for
end for
y = argmax(count[y])
End
```

Figure 1 **Pseudo code for classification from multiple feature subsets**

For each type of amino acid pair, we randomly select n subsets of features, and each subset has m features; m and n are parameters of the algorithm. We build $3n$ NN classifiers based on these subsets. Whenever a pattern is presented, it is classified by $3n$ classifiers then a voting scheme is applied for determining the class of the pattern. Figure 1 shows the pseudo code for classifying on multiple feature subsets.

By selecting different feature subsets for classifying, we attempt to force the NN classifiers to make different and hopefully uncorrelated errors in order to improve the classification accuracy. Although k-nearest neighbor and its variations usually work better than simple nearest neighbor algorithm, we used nearest neighbor classifiers because we expected that they would create different uncorrelated errors making the voter work more effectively.

Measurement accuracy

To evaluate the algorithm performance, jackknife test was employed for cross-validation. According to Mardia [11], the jackknife test is thought to be more rigorous and reliable in comparison to subsampling test or independent data set test. A comprehensive discussion about this problem was provided by Chou and Zhang [5]. In the jackknife test process, each protein is singled out in turn as a test sample, the remaining proteins are used as training set to calculate test sample nearest neighbor and predict the class. The prediction quality was evaluated by the overall prediction accuracy and prediction accuracy for each location as defined below:

$$\text{Overall accuracy} = \frac{\sum_{s=1}^{k} p(s)}{N}$$

$$\text{Accuracy}(s) = \frac{p(s)}{obs(s)}$$

Here N is the total number of sequences, k is the class number, obs(s) is the number of sequences observed in location s and p(s) is the number of correctly predicted sequences in locations.

The other measure of prediction accuracy is Matthew's correlation coefficients (MCC) [11] between the observed and predicted locations over a data set as given by following equation:

$$\text{MCC}(s) = \frac{p(s)n(s) - u(s)o(s)}{\sqrt{(p(s)+u(s))(p(s)+o(s))(n(s)+u(s))(n(s)+o(s))}}$$

Here, *p(s)* is the number of properly predicted proteins in location *s*, *n(s)* is the number of correctly predicted proteins not in location *s*, *u(s)* is the number of under-predicted and *o(s)* is the number of over-predicted sequences.

Results and discussion

Parameters selection and prediction accuracy of MFS method

Test has been done with various values of the number of classifiers and the number of features used in each classifier. We set the subset size parameter based on one-leave-out cross-validation accuracy estimated on the training set. At first the number of classifiers was fixed to n=15 (it means that totally 45 classifiers were used) and varied the subset size from 150 to 300 with. We found that the number of features in the range of 200 to 250 gave almost the same accuracy. Therefore we selected 200, the smallest value, as the subset size.

Location **(No of entries)**	**Our method** *MCC*	 Accuracy	**PLoc** Accuracy
Chloroplast (671)	*0.84*	88.5	72.3
Cytoplasmic (1245)	*0.74*	83.2	72.2
Cytoskeleton (41)	*0.89*	82.5	58.5
ER (114)	*0.86*	78.0	46.5
Extracellular (862)	*0.89*	88.9	78.0
Golgi apparatus (48)	*0.72*	61.7	14.6
Lysosomal (93)	*0.84*	77.4	61.8
Mitochondrial (727)	*0.69*	61.4	57.4
Nuclear (1932)	*0.83*	92.5	89.6
Peroxisomal (125)	*0.69*	54.4	25.2
Plasma membrane (1677)	*0.92*	93.8	92.2
Vacuolar (54)	*0.74*	61.1	25.0
Overall	---	**85.8**	**78.2**

Table 2 The accuracy of our method in comparison with PLOC.

After setting the subset size value, we varied the number of classifiers for each kind of amino acid pair composition data from 1 to 30. The larger the value of *m*, the better the accuracy we got, but the slower the algorithm worked. When *m*>15, the accuracy does not improve much. Finally we set the value of *m* to 15 as a reasonable trade-off between computational expense and accuracy. With these algorithm parameters, the jackknife testing result is listed in Table 2. The overall predictive accuracy of our method reached 85.8%.

Confusion matrix analysis

For detail analysis, we constructed a confusion matrix according to the result of jackknife test as shown in Table 3. We can see from Table 2 and Table 3 that predictive accuracy varies substantially with subcellular locations. Proteins in major classes which have large number of data entries such as nuclear proteins, plasma membrane proteins can be inferred more reliably than other classes. Among the major classes, mitochondrial protein prediction accuracy is the worst. The poor result had been achieved in other methods based on sequence global features. It seems that mitochondrial proteins should be treated specially by incorporating with other methods based on sorting signal.

In contrast to the major classes, prediction accuracies on most minor classes which have few data entries (Golgi, endoplasmic reticulum, vacuole, peroxisome and lysosome) are still low. Many of them are misclassified as major class proteins. To overcome this problem, perhaps a more sophisticated voting scheme that takes into account sizes of the classes should be considered. As the same time, more data entries for minor classes should be added to into the dataset from updated databases.

Predicted *Actual*	Plas	Ext	Cytop	Chl	Nuc	Mit	Endo	Cytos	Gol	Lyso	Pero	Vac	SUM
Plasma	**1571**	9	30	7	50	4	1	0	1	1	0	0	1674
Ext	10	**749**	41	12	35	3	1	0	2	6	0	2	861
Cytop	11	11	**1033**	31	112	34	2	1	2	0	4	0	1241
Chl	5	0	32	**594**	23	16	0	0	0	0	1	0	671
Nuc	22	10	83	13	**1788**	15	1	0	0	0	0	0	1932
Mit	16	11	115	46	89	**446**	0	0	0	0	4	0	727
Endo	3	3	8	2	6	3	**89**	0	0	0	0	0	114
Cytos	0	0	0	0	6	1	0	**33**	0	0	0	0	40
Gol	1	2	7	1	7	0	0	0	**29**	0	0	0	47
Lyso	8	4	2	0	4	1	0	0	0	**72**	0	2	93
Pero	11	0	22	8	9	7	0	0	0	0	**68**	0	125
Vac	0	1	10	0	10	0	0	0	0	0	0	**33**	54
SUM	1658	800	1383	714	2139	530	94	34	34	79	77	37	7579

Table 3 Confusion matrix for prediction results of PLOC data set

Improvement by voting

Our voting scheme involves 15 classifiers for each kind of amino acid pair information. We estimate the average and standard deviation of the accuracies for all classifiers in every subcellular location classes. Table 4 shows the prediction accuracies on three kinds of amino acid pair information and the accuracy achieved by voting.

Location	**Amino acid pair**	**1 gaped amino acid pair**	**2 gapped amino acid pair**	**Vote**
Plasma	84.17 ± *0.56*	82.54 ± *0.76*	81.25 ± *0.62*	93.8
Ext	78.56 ± *0.49*	78.56 ± *0.53*	76.28 ± *0.85*	88.9
Cytop	73.78 ± *0.55*	73.15 ± *0.59*	72.52 ± *0.25*	83.2
Chl	82.47 ± *0.79*	82.98 ± *0.82*	82.15 ± *0.75*	88.5
Nuc	80.71 ± *0.40*	81.00 ± *0.37*	79.81 ± *1.03*	92.5
Mit	56.70 ± *0.90*	56.18 ± *1.88*	54.22 ± *0.82*	61.4
Endo	73.68 ± *0.72*	73.86 ± *1.58*	72.28 ± *0.76*	78.0
Cytos	81.50 ± *0.71*	81.50 ± *1.15*	77.50 ± *0.91*	82.5
Gol	44.26 ± *2.38*	48.94 ± *3.20*	47.23 ± *0.92*	61.7
Lyso	73.33 ± *1.44*	73.55 ± *1.28*	70.32 ± *1.95*	77.4
Pero	52.00 ± *1.13*	52.32 ± *0.75*	53.76 ± *1.32*	54.4
Vac	56.30 ± *1.45*	56.67 ± *0.86*	60.37 ± *1.86*	61.1
Overall	76.88 ± *0.18*	76.53 ± *0.42*	75.28 ± *0.15*	85.8

Table 4 Prediction accuracies of different amino acid pair composition information and prediction accuracy achieved by voting

As we expected, using three kinds of amino acid pair information and selecting feature subsets randomly, the NN classifiers can create different uncorrelated errors making the voter work efficiently. There was apparent improvement in all protein classes. However the voting scheme did not work very well on protein classes which have few data entries. It is due to the fact that simple voting can only improve accuracy if the classifiers select the correct class more often than any other class, but proteins in minor classes tend to be misclassified to major class because of their lack of data more often. A more sophisticated voting scheme may work better than simple voting.

Comparison with other methods

Park and Kanehisa [9] used a set of binary SVM classifiers based on amino acid, amino acid pair compositions and a voter to solve the multi-class classification problem. Their method had the overall accuracy of 78.2% on PLOC dataset. Although we used only three kinds of amino acid pair information while Park used four kinds of that together with amino acid composition, our method produced better results in all subcellular location categories. We also tried to use more kinds of amino acid pair information but no improvement has made. It seems that three is a reasonable number of amino acid pair information. There are two reasons that make our algorithm work much better than Park's algorithms: 1) in this classification problem, NN algorithm is superior to SVMs in term of balance. SVMs showed very poor result in categories which have few training sequences while the NN classifier worked much better. That makes Park's followed voter not effective; 2) using different feature subsets, the NN classifiers tend to make different location predictions as we expected. That makes our voter work very well as discussed above.

In order to prove the robustness of our method, we applied the algorithms to two other data sets. The first one is Fuzzy_Loc data set [21]. This data set contained 7203 eukaryotic proteins located in 11 subcellular locations. Membrane proteins are excluded because it has been predicted with very high accuracy by other methods. Ying [21] used a fuzzy k-NN algorithm based on amino acid dipeptide and got 80.1% accurate. Our method is somewhat similar to his method because both are instance based classification algorithms. However in our approach, we added more sequence information by using three kinds of amino acid pair composition while Ying used only amino acid dipeptide composition. Furthermore, as shown in the experiment performed by Stephan [20], MFS usually worked better than fuzzy k-NN in term of predictive accuracy. We applied our method to Fuzzy_Loc dataset and overall accuracy retrieved was 85.2%. Table 5 shows the comparison between MFS and fuzzy k-NN on Fuzzy_Loc data set. We made significant improvement in most location categories except Golgi apparatus. The predictive accuracy for Golgi apparatus class is low because the accuracy of the individual classifier is so low that simple voting does increase the expected errors.

Location (No of entries)	Fuzzy_Loc	Our method
Extracellular (2134)	93.7	94.4
Nuclear (2149)	81.9	90.0
Mitochondrial (692)	59.0	68.2
Cytoplasmic (1251)	70.2	75.9
ER (82)	57.3	63.4
Chloroplast (645)	84.7	88.1
Cytoskeleton (10)	40.0	50.0
Peroxisomal (81)	56.8	64.2
Golgi apparatus (31)	16.1	12.9
Lysosomal (83)	67.5	75.9
Vacuolar (41)	34.1	36.6
Overall	**80.1**	**85.2**

Table 5 Comparison with Fuzzy_Loc

We also applied the algorithms to Reinhardt's dataset [1] which contains 2427 eukaryotic proteins in 4 locations. This dataset has been used in many other researches. Table 2 shows the result of our method in comparison with the previous method. Among many approaches, MFS shows the best result.

Locations	Neural Network	Markov Model	SVMs	Fuzzy k-NN	Our method
Cytop	55	78.1	76.9	86.7	**90.8**
Extra	75	62.2	80.0	83.7	**86.5**
Nuc	72	74.1	87.4	**92.0**	90.0
Mito	61	69.2	56.7	60.4	66.7
Overall accuracy	66	73.0	79.4	85.2	**88.5**

Table 6 Comparison with other methods on Reinhardt's dataset

Conclusion

In this paper, a nearest neighbor classification from multiple subsets of amino acid pair compositions algorithm was proposed for protein subcellular locations prediction. This method takes advantage of sequence order effect and the redundancy of amino acid pair compositions. We have applied it to several data sets and high predictive accuracies have been achieved using a jackknife test. This method is simple and it just needs raw sequence data, so we can predict protein which has only sequence information. In the future we will use this method to annotate protein database.

Beside the significant improvement, the predictor still shows low predictive accuracy for some localization categories. Perhaps further improvement can be obtained by preparing a higher quality data set. It should be possible to increase the number of protein entries, especially for small groups such as Golgi apparatus, peroxisomal, vacuolar and lysosomal. In additional, we should consider protein groups that belong to multiple locations, such as those that move between cytoplasm and nucleus under different conditions. For system analysis of great amounts of genome data, this method should be integrated with other existing methods based on sorting signals.

Acknowledgement

The authors would like to thank Dr Park K.J., Dr A. Reinhardt for providing datasets. We thank Sammy Tan for his valuable comments on the manuscripts. We are also grateful to Chung Moon Soul Center for BioInformation and BioElectronics for supporting this work. This research was funded by the Brain Korea 21 research grant.

References

[1] Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J, Basic local alignment search tool. J. Mol. Biol., **215**,403–410, 1990.

[2] Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M., The SWISSPROT protein knowledge base and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370, 2003.

[3] Cadeno J, Alloy P, P'erezPons JA, Querol E., Relation between amino acid composition and cellular location of proteins,*J Mol Biol* **266**:594-600, 1997.

[4] Chou, K.C., Prediction of protein signal sequences. *Curr. Protein Peptide Sci.*, **3**, 615-622, 2002.

[5] Chou, K.C. and Zhang,C.T., Review: Prediction of protein structure classes,*Crit. Rev. Biochem. Mol. Biol.*, **30**, 275-349, 1995.

[6] Duda,R.O., Hart,P.E. and Stork,D.G., Pattern Classification, 2nd edn. Wiley, New York, 2000.

[7] Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G., Predicting subcellular localzation of proteins based on their N-terminal amino acid sequence,*J. Mol. Biol.*, **300**, 1005–1016, 2000.

[8] Hua S, Sun Z, Support vector machines approach for protein subcellular localization prediction,*Bioinformatics*, **17**, 721-728, 2001.

[9] Keun-Joon Park and Minoru Kanehisa, Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs, *Bioinformatics*, **19**, 1656-1663, 2003.

[10] Keller J.M., Gray M.R. and Givens J.A., A fuzzy k-nearest neighbor algorithm. *IEEE Trans. Syst. Man Cybern.*, **15**, 580–585, 1985.

[11] Matthew,B.W., Comparison ofpredicted and observed secondary structure of T4 phage lysozym, *Biochem. Biophys. Acta.*, **405**, 442-451, 1975.

[12] Madria,K.V., Ken,J.T. and Bibby,J.M., Multivariate Analysis, *Academic Press*, London, 322-381, 1979.

[13] Myer, E.W. and Miller, W., Optimal alignment in linear space, *CABIOS*, **4**, 11-17, 1988.

[14] Nakai,K. and Kanehisa,M., A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomic*, **14**, 897-911, 1992.

[15] Nakai, K. and Horton, P., PSORT: a program for detecting sorting signals in protein and predict their protein subcellular localization, *Trend Biochem. Sci.*, **24**, 34-36, 1999.

[16] Nakai,K. ,Protein sorting signals and prediction of subcellular localization. *Adv. Protein Chem.*, **54**, 277–344, 2000.

[17] Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G., Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, **10**, 1–6, 1997.

[18] Nielsen, H., Brunak, S. and von Heijne, G., Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng*., **12**, 3–9, 1999

[19] Reinhardt A., Hubbard T., Using neural networks for prediction of the subcellular location of proteins, *Nucleic Acids Res*, **26**, 2230-2236, 1998.

[20] Stephen, B., Nearest neighbor classification from multiple feature subsets,*Intelligent Data Analysis*, **3**, 191-209, 1999.

[21] Ying Huang, Prediction of protein subcellular locations using fuzzy k-NN method, *Bioinformatics*, **20**, 21-28, 2004

[22] Yuan,Z., Prediction of protein subcellular locations using Markov chain models.*FEBS Lett.*, **451**, 23–2, 1999.

[23] http://nobelprize.org/medicine/laureates/1999/press.htm