단백질의 세포내 소 기관별 분포 예측을 위한 서열 기반의 특징 추출 방법

김종경° 최승진
포항공과대학교 컴퓨터공학과
{blkimjk°, seungjin}@postech.ac.kr

Sequence driven features for prediction of subcellular localization of proteins

Jong Kyoung Kim° Seungjin Choi
Department of Computer Science, Pohang University of Science and Technology

요 약

Predicting the cellular location of an unknown protein gives valuable information for inferring the possible function of the protein. For more accurate prediction system, we need a good feature extraction method that transforms the raw sequence data into the numerical feature vector, minimizing information loss. In this paper, we propose new methods of extracting underlying features only from the sequence data by computing pairwise sequence alignment scores. In addition, we use composition based features to improve prediction accuracy. To construct an SVM ensemble from separately trained SVM classifiers, we propose specificity based weighted majority voting. The overall prediction accuracy evaluated by the 5-fold cross-validation reached 88.53% for the eukaryotic animal data set. By comparing the prediction accuracy of various feature extraction methods, we could get the biological insight on the location of targeting information. Our numerical experiments confirm that our new feature extraction methods are very useful forpredicting subcellular localization of proteins.

1. INTRODUCTION

Predicting the cellular location of an unknown protein gives valuable information for inferring the possible function of the protein. To achieve a good prediction result, we need a good feature extraction method that transforms the raw sequence data into the numerical feature vector, minimizing information loss. Although much work has been done on improving the prediction accuracy of subcellular localization, little research has been conducted on feature extraction methods relying solely on amino acid sequence properties. In this paper, we propose new sequence driven feature extraction methods to predict subcellular localization. To this end, we introduce various pairwise sequence alignment methods so that a protein sequence is represented as a numerical vector of pairwise sequence alignment scores. Additionally, we use amino acids composition based features to improve prediction accuracy. For classification, we use a SVM ensemble to combine mixed types of features. Our numerical experiments confirm that our proposed methods considerably improve the prediction accuracy and give biological insight into the position of targeting information in the protein sequence.

2. FEATURE EXTRACTION

We proposed, in our previous work, a new feature extraction method representing a protein sequence as the scores of dynamic global sequence alignment [1]. Despite its very high prediction accuracy, its time complexity was relatively higher than that of composition based method. Additionally, its location coverage was limited to some proteins whose signal sequences are located at the N-terminus. To overcome these limitations, we present sequence driven feature extraction methods. We also use two composition based methods to improve the prediction accuracy. We select, in this study, representative sequences in the training set in order to decrease the time complexity. For this purpose, we clustered all sequences in the training set by using a constructed phylogenetic tree. To represent the N-terminal sequence families for each cluster, we built profile Hidden Markov models (HMMs) [2]. The N-terminal truncated sequence is then converted into the corresponding feature vector by computing the log-odds scores between the sequence and all the constructed profile HMMs. Instead of using all the sequences in the training set, selecting representative sequences reduces the time complexity of calculating the scores of pairwise sequence alignment. For this purpose, we randomly select a protein sequence from each cluster. The minimum length of the chosen sequences is 80 and the first residue should be methionine, which means the first synthesized residue from the start codon.

The processed sequence is then converted into the corresponding feature vector by computing the scores of the Needleman-Wunsch algorithm between the sequence and all the selected representative sequences [3]. So far we have assumed that the signal sequences are located at the N-terminal, and that we are looking for the global match between two N-terminal regions. A much more common situation is that the signal sequences or the targeting information are located anywhere in the protein. In this situation, the most sensitive way to detect the internal targeting signals is to use the algorithm for finding optimal local alignments, which are referred to as the Smith-Waterman algorithm [4]. The general procedures of this feature extraction method is almost same to the above one using the Needleman-Wunsch algorithm. The dipeptide composition is the extension of amino acid composition adding the information on the local order of amino acids. In practice, it is proved that the predictive power of dipeptide composition is superior to amino acid composition. It is generally thought that the factors determining the cellular destination are physico-chemical properties such as hydrophobicity or the position of charged amino acids since the signal sequences are not well conserved. The 121 physico-chemical properties were used to represent a protein sequence as a 121 dimensional feature vector based on amino acid composition.

3. CLASSIFICATION

At the classification step, the feature vector is used as input to (M-1)M/2 binary SVM classifiers for M classes. In this pairwise classification, the feature vector is classified into the class getting the highest number of votes. The kernel function used in this study is the *radial basis function* (RBF) kernel with one parameter γ. Since we trained several independent SVM classifiers for each feature, we need to aggregate them in an appropriate manner. The majority voting is the simplest and widely used aggregation method. This voting scheme treats all SVM classifiers with equal weights. Since the prediction errors of the classifiers are often different, it is more realistic to give different weights in proportional to their prediction performance. The overall schematic diagram of our prediction system is illustrated in Fig. 1. In our system, four major steps are needed to get the final decision. In the first preprocessing step, a test protein sequence is truncated after first 40 or 80 residues to get the N-terminal regions. The truncated N-terminal sequences, in the next feature extraction step, are converted into the two feature vectors by computing the scores of pairwise sequence alignments based on the profile HMM and the Needleman-Wunsch algorithm. The full sequence is also converted into the three feature vectors by computing the scores of the Smith-Waterman algorithm, or by calculating the compositional fractions of all dipeptides and the average values of the 121 physico-chemical properties. After these feature extraction steps, we obtain the fixed-length feature vectors. At the classification step, each of the five feature vectors is used as the input to (M-1)M/2 binary SVM classifiers for M classes. In this pairwise classification, the feature vector is assigned to the class associated with the highest value in the majority voting. After that, in the weighted majority voting step, the final predicted class label is decided based on the five predicted class labels.

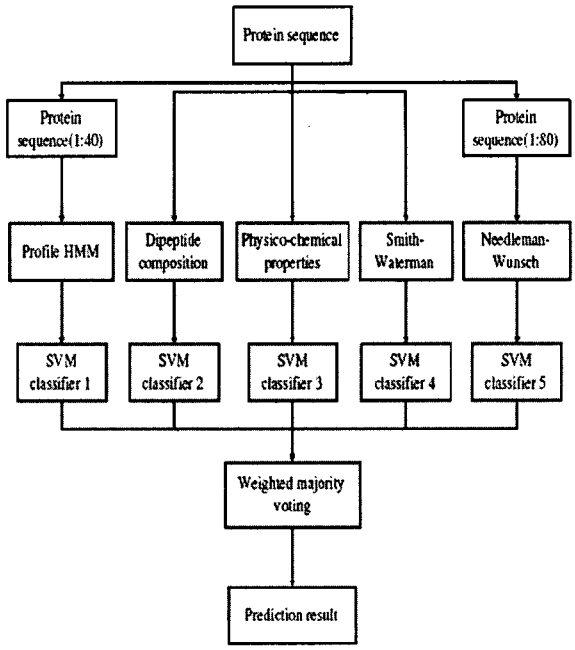

Figure 1. The schematic diagram of the proposed prediction system is illustrated.

4. NUMERICAL EXPERIMENTS AND RESULTS

We used the animal data set generated by [5] for training and evaluating our prediction system. The data set consists of 11688 eukaryotic animal proteins with 9 cellular locations: cytoplasm,

endoplasmic reticulum (ER), extracellular, golgi, lysosome, mitochondrion, nucleus, plasm membrane, and peroxisome. The performance of our prediction system was evaluated using the 5-fold cross-validation. The N-terminal profile HMM based prediction method showed the overall accuracy of 83.62%. In the case of the N-terminal Needleman-Wunsch algorithm based method, the prediction accuracy reached 83.25% which is slightly lower that that of the profile HMM based method. The two methods, which are specialized for extracting targeting information from the N-terminal signal sequences, showed that similar prediction patterns. The results of the methods based on the full sequence were considerably different from the above N-terminal based methods. The Smith-Waterman algorithm based method showed the prediction accuracy of 83.23%. The overall accuracy of dipeptide composition and physico-chemical properties based methods reached 85.82% and 82.29%, respectively. The SVM ensemble combining all the five SVM classifiers showed the overall accuracy of 88.53%. The prediction accuracy of our developed SVM ensemble is nearly 10% higher than the previous methods relying solely on amino acid sequence properties [6, 7]. The proteins targeted to extracellular, mitochondrion, and nucleus were predicted by the N-terminal based methods with higher accuracy. This result, in the case of nuclear proteins, are not well matched with the fact that the nuclear localization signals can be located anywhere. But, from this result, we could logically assume that most nuclear localization signals are located at the N-terminus. For the proteins whose targeting information are not restricted to the N-terminal region, the full sequence based methods showed better sensitivity.

5. CONCLUDING REMARKS

In this paper we have proposed various sequence driven feature extraction methods for prediction of subcellular localization. Taking into account the improved prediction performance, we conclude that our feature extraction methods using pairwise sequence alignment are well fitted to this classification problem. This study also has shown that the specificity based majority voting scheme is very effective for constructing the SVM ensemble from separately trained SVM classifiers. From the comparative study of the performance of the several SVM ensembles, we have shown that the performance of the feature extraction methods based on pairwise sequence alignment is significantly improved by combining the composition based methods. By comparing the average sensitivity of the N-terminal and full sequence based methods, we could get the biological insight on the location of targeting information. There are, however, a main problem that remain to be explored. In the case of proteins whose targeting information is not restricted to the N-terminal region, the sensitivity is considerably low. Therefore, more research is needed to resolve this low sensitivity problem. We hope this study will serve as a platform from which studies on developing feature extraction methods based on amino acid sequence property may be undertaken with greater depth and specificity.

6. REFERENCES

[1] J. K. Kim, G. P. S. Raghava, K. S. Kim, S. Y. Bang, and S. Choi. Prediction of subcellular localization of proteins using pairwise sequence alignment and support vector machine. In Proc. 3rd annual conference for the Korean society for bioinformatics, pages 158--166, Seoul, Korea, 2004.

[2] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge, 1998.

[3] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol., 48:443--453, 1970.

[4] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. J. Mol. Biol., 147:195--197, 1981.

[5] Z. Lu, D. Szafron, R. Greiner, P. Lu, D. S. Wishart, B. Poulin, J. Anvik, C. Macdonell, and R. Eisner. Predicting subcellular localization of proteins using machine-learned classifiers. Bioinformatics, 20:547--556, 2004.

[6] Y. D. Cai, X. J. Liu, X. B. Xu, and K. C. Chou. Supprot vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. J. Cell Biochem., 84:343--348, 2002.

[7] K. Park and M. Kanehisa. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics, 19:1656--1663, 2003.