Prediction of subcellular localization of proteins using pairwise sequence alignment and support vector machine

Jong Kyoung Kim [1], G. P. S. Raghava[2], Kwang S. Kim[3], Sung Yang Bang[1] and Seungjin Choi[1,*]

[1] Department of Computer Science and Engineering, Pohang University of Science and Tec hnology, San 31, Hyojadong, Namgu, Pohang, Korea
[2] Bioinformatics Centre, Institute of Microbial Technology, Sector 39A, Chandigarh, India
[3] National Creative Research In itiative Center of Superfunctional Materials, Department of Chemistry, Division of Molecular and Life Sciences, Pohang University of Science and Technology, San 31, Hyojadong, Namgu, Pohang 790-784, Korea

*To whom correspondence s hould be addressed. E-mail: seungjin@postech.ac.kr

Abstract

Predicting the destination of a protein in a cell gives valuable information for annotating the function of the protein. Recent technological breakthroughs have led us to develop more accurate methods for predicting the subcellular localization of proteins. The most important factor in determining the accuracy of these methods, is a way of extracting useful features from protein sequences. We propose a new method for extracting appropriate features only from the sequence data by computing pairwise sequence alignment scores. As a classifier, support vector machine (SVM) is used. The overall prediction accuracy evaluated by the jackknife validation technique reach 94.70% for the eukaryotic non-plant data set and 92.10% for the eukaryotic plant data set, which show the highest prediction accuracy among methods reported so far with such data sets. Our numerical experimental results confirm that our feature extraction method based on pairwise sequence alignment, is useful for this classification problem.

Introduction

Cellular organelles in a eukaryotic cell require a continuous supply of appropriate proteins to make and maintain themselves. Proteins encoded in the nuclear genome are synthesized on ribosomes in the cytosol and delivered to the organelles in which they are required. Here, we do not consider the proteins that are synthesized on ribosomes inside the mitochondria and chloroplasts because they are not delivered to other organelles. Some proteins imported to ER are secreted from the cell and other proteins imported to organelles replace organelle proteins that have been degraded. The delivery of a protein to an appropriate organelle depends on an N-terminal signal sequence. Since the signal sequence specifying the same organelle is not well conserved, it is generally thought that the factors in determining the destination are physico-chemical properties such as hydrophobicity or the position of charged amino acids [1].

Predicting the destination of an unknown protein can give a valuable hint for guessing the possible function of the protein. Therefore, in recent years, numerous methods in computational

biology have been developed for more accurate prediction. In fact, this is a classification problem that has been extensively studied in machine learning and statistics communities, because class labels related to subcellular locations are already known in a set of training data. Various classifiers such as artificial neural networks (ANN), support vector machines (SVM), or k-nearest neighbor algorithms (k-NN) have been applied to this classification problem. However, a critical factor in determining the classification or prediction accuracy, lies in a way of feature extraction. Most of prediction methods can be divided into two classes, depending on their ways of feature extraction: (1) features based on protein sequences data; (2) features based on ontology data. In the protein sequence-based approach, two different feature extraction methods are popular. These involve the recognition of N-terminal sorting signals or the detection of amino acids compositions from an entire sequence. The former has the strong biological implication because the signal sequence specifying the cellular location of a protein is located in the N-terminal region [2, 3]. However, it is difficult to recognize underlying features from a highly diverse signal sequence and to vectorize those features. The latter approach partially overcomes these difficulties but lose the information regarding the context stored in the sequence data [4, 5, 6]. The ontology-based approach has received much attention recently because of its high prediction accuracy [7, 8]. This approach extracts text information of homologous sequences of a target sequence by searching biological databases and vectorizes the information. It is not surprising for this approach to show good performance because it utilizes various extra information derived from several sources.

In this paper, we propose a new method of extracting underlying features only from the sequence data in predicting cellular locations of proteins. To this end, we introduce a pairwise sequence alignment score so that a protein sequence is presented to a SVM classifier as a vector containing pairwise sequence alignment scores. Our numerical experimental results confirm that our proposed method considerably improve the prediction accuracy.

Systems and Methods

Data sets

We used two data sets for training and evaluating our prediction system. These data sets were generated by Emanuelsson *et al.* (2000). All sequences in the two data sets were extracted from SWISS-PROT release 36, 37, or 38, and their subcellular locations were chosen by referring annotations in FT or CC field. In the preprocessing step, all sequences containing ambiguous amino acids such as B, Z, or X were excluded, and sequences with high similarities were removed for redundancy reduction. As shown in Table 1, these data sets consist of 940 eukaryotic plant sequences with four classes (chloroplast, mitochondrion, extracellular, and other) and 2738 eukaryotic non-plant sequences with three classes (mitochondrion, extracelluar, and other).

Pairwise sequence alignment as a feature extractor

Representing a protein sequence by the scores of pairwise sequence alignments (SA) was already applied to the SVM-pairwise for detecting remote structural and evolutionary relationships [9]. In many ways SVM-pairwise is directly analogous to our prediction system. In the feature extraction step, the SVM-pairwise vectorizes a protein sequence by computing pairwise sequence similarity scores between the target sequence and all sequences in the training set. The resulting vectors are then used as input to SVM for classification. Yet locality makes a distinction between the two methods. The SVM-pairwise uses the Smith-Waterman algorithm for finding the optimal local alignment because the global SA of two very highly diverged sequences is not possible. In contrast to the SVM-pairwise, our prediction system uses the Needleman-Wunsch algorithm for obtaining the optimal global alignment [10]. In order to consider only N-terminal signal sequences, all sequences were truncated after first 90 residues, and then have the same length. Additionally, it can be thought that the whole N-terminal sequences are important in determining subcellular locations. Therefore, it is reasonable to use the global dynamic programming algorithm.

For the global dynamic programming algorithm, we used Matlab functions that are available at her[1]. A d-dimensional feature vector $\mathbf{x}_k$ for the kth protein sequence has the form

$$\boldsymbol{x}_k = [x_{k1}, x_{k2}, \ldots, x_{kd}]^T \qquad (1)$$

where T denotes the matrix or vector transpose operator and x_{ki} is the score of Needleman-Wunsch algorithm between sequence k and the i th sequence in the training set. Note that d is equal to the total number of sequences in the training set. The gap penalty is -3 and the substitution matrix is BLOSUM 50.

[1]http://www.cs.cornell.edu/courses/cs321/2001fa/matlab_examples.html

Table 1. Number of sequences in each subcellular localization category of eukaryotic plant and non-plant data sets. (Emmanuelsson *et al.*, 2000)

Species	Subcellular localization	Number of sequences
Eukaryotic Plant	Chloroplast (cTP)	141
	Mitochondirial (mTP)	368
	Extracellular (SP)	269
	Cytoplasmic + Nuclear (Other)	162
Eukaryotic Non-plant	Mitochondirial (mTP)	371
	Extracellular (SP)	715
	Cytoplasmic + Nuclear (Other)	1652

Support vector machine as a classifier

SVM classifiers receive their popularity from the fact that they are based on the concept of statistical learning theory, or VC (Vapnik-Chervonenkis) theory, and they can achieve high performance in practical applications [11, 12]. SVM classifiers are basically kernel-based learning algorithms and find the optimal hyperplane decision boundary in the feature space. In kernel-based algorithms, a kernel trick leads us to process the data in a higher-dimensional feature space constructed by a nonlinear mapping, without the explicit knowledge of the nonlinear mapping. In a view of statistics, the high dimensionality of the feature space can cause the curse of dimensionality. However, the optimal

separating hyperplane with a maximal margin in the feature space, can relieve this problem. In statistical learning theory, we can minimize the complexity term of the upper bound of the expected risk by maximizing the margin of the separating hyperplane. The minimization of the upper bound can be viewed as relieving the over-fitting problem [13]. The maximization of the margin can be formulated as a quadratic optimization program so that a global solution can be easily obtained.

In the present study, we used OSU SVM Matlab toolbox 3.00 for the SVM classifier that is freely available at here[2]. The prediction of subcellular localization is a multi-class classification problem but the SVM classifier can only deal with the binary classification problem. Therefore, we need to construct a set of binary classifier for multi-class classification. We constructed (M-1)M/2 binary classifiers for M classes. In this pairwise classification, each possible pair of classes is considered and a test pattern is classified by the majority voting. This approach has two advantages over the *one versus the rest* method. The weak point of the latter approach is that it compares the real values in outputs of M binary classifiers directly. Because each binary classifier is trained on different binary classification problems, their real values in outputs of the classifiers may not be suitable for comparison. In addition, in the 'one versus the rest' approach, the numbers of positive and negative training data points are not symmetric. These two weak points can be solved by the pairwise classification [14]. The kernel function used in this study is the *radial basis function* (RBF) kernel with one parameter γ:

$$k(x, y) = e^{-\gamma\|x-y\|^2} \quad (2)$$

During the training and testing, only the RBF kernel parameter γ and the regularization parameter C were considered and the remaining parameters were kept constant.

The proposed prediction system

The overall schematic diagram of our prediction system is illustrated in Fig. 1. The target protein sequence is truncated after first 90 residues, in order to take only N-terminal signal sequence into account. The processed target sequence is then converted into the corresponding feature vector by computing the scores of Needleman-Wunsch algorithm between the processed target sequence and all other sequences in the training set. Here, all sequences in the training set are also truncated after first 90 residues. The training set can be divided into two parts which are positive and negative vectorization set. The positive vectorization set means all sequences of this set belong to the same class with the target sequence. The negative vectorization set denotes the opposite case. Therefore, the discriminative power of the feature vector is expected to increase since it contains the information of positive and negative examples. After this feature extraction step, we obtain the fixed-length feature vector. Note that the fixed dimension of the feature vector is equal to the total number of the whole training set. At the classification step, the feature vector 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 voting.

[2] http://www.ece.osu.edu/~maj/osu_svm

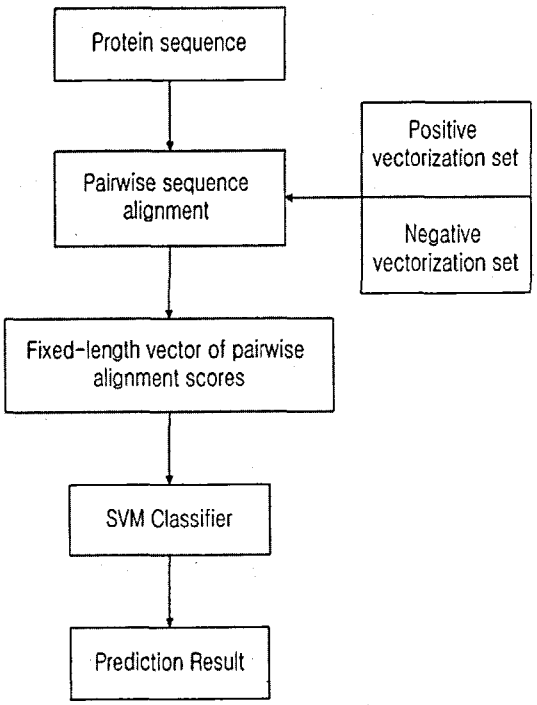

Fig. 1. The schematic diagram of our proposed prediction system, is illustrated. Through the pairwise sequence alignment, each protein sequence is converted into the corresponding feature vector, by computing Needleman-Wunsch scores between the protein sequence and the whole sequences in the training data set. The SVM classifier predict an appropriate class, given a protein sequence.

Evaluation of the prediction system

The performance of our prediction system was evaluated using the 5-fold cross-validation and jackknife validation techniques. In the 5-fold cross-validation, the whole data set was partitioned into five exclusive subsets, and in turn one subset was used for the test data and the remaining sets were used for the training data. In this study, the 5-fold cross-validation was just used for comparing the results obtained by this validation technique. For more objective and rigorous evaluation, we used the jackknife validation. In this technique, one protein sequence was left out in turn for the test data and the rest was used for the training data. In our prediction system, the dimension of the feature vector depends on the validation technique because the dimension is equal to the number of the training data. To measure the performance, sensitivity, specificity and Matthew's correlation coefficient (MCC) [15] and overall accuracy were calculated using the following equations:

$$Sensitivity(i) = \frac{tp(i)}{tp(i) + fn(i)}, \quad (3)$$

$$Specificity(i) = \frac{tp(i)}{tp(i) + fp(i)}, \quad (4)$$

$$MCC(i) = \frac{tp(i) \times tn(i) - fp(i) \times fn(i)}{\sqrt{(tp(i) + fn(i))(tp(i) + fp(i))(tn(i) + fp(i))(tn(i) + fn(i))}}, \quad (5)$$

$$Overall\ accuracy = \frac{\sum_{i=1}^{k} tp(i)}{N}, \quad (6)$$

where N is the total number of sequences, k is the number of class, tp(i) (true positive) is the number of correctly predicted sequences of class i, tn(i) (true negative) is the number of correctly predicted sequences which is not in class i, fp(i) (false positive) is the number of over predicted sequences of class i and fn(i) (false negative) is the number of under predicted sequences of class i.

Results

The performance measured for the eukaryotic plant and non-plant data is shown in Tables 2 and 3. To select the appropriate parameter values, we tested various parameter values of the RBF kernel parameter γ and the regularization parameter C through the 5-fold cross-validation. Table 2 shows

Table 2. Performance comparison of different subcellular localization predictions on the eukaryotic plant data set

Method	Category	Sensitivity	Specificity	MCC	Overall accuracy	Reference
	cTP	0.8511	0.8163	0.8003		this work
5-fold cross	mTP	0.8886	0.9355	0.8536	0.8957	
validation	SP	0.9375	0.9836	0.9435		
	other	0.8839	0.7654	0.7814		
	cTP	0.8794	0.8794	0.8562		this work
Jackknife	mTP	0.9136	0.9535	0.8898	0.9210	
validation	SP	0.9492	0.9918	0.9581		
	other	0.9290	0.7956	0.8278		
	cTP	0.85	0.69	0.72		
5-fold cross	mTP	0.82	0.90	0.77	0.853	Emanuelsson et al.
validation	SP	0.91	0.95	0.90		(2000)
	other	0.85	0.78	0.77		
Jackknife					0.861	Cai & Chou (2004)
validation						

Table 3. Performance comparison of different subcellular localization pre-dictions on the eukaryotic non-plant data set.

Method	Category	Sensitity	Specificity	MCC	Overall accuracy	Reference
	mTP	0.8702	0.8824	0.8565		this work
5-fold cross	SP	0.9216	0.9478	0.9116	0.9399	
validation	other	0.9632	0.9492	0.8859		
	mTP	0.8785	0.8908	0.8662		this work
Jackknife	SP	0.9390	0.9557	0.9287	0.9470	
validation	other	0.9656	0.9557	0.8981		
	mTP	0.89	0.67	0.73		
5-fold cross	SP	0.96	0.92	0.92	0.900	Emanuelsson et al.
validation	other	0.88	0.97	0.82		(2000)
	mTP	0.78	0.82	0.77		Reczko and
5-fold cross	SP	0.93	0.91	0.89	0.913	Hatzigeorgiou
valldation	other	0.93	0.94	0.84		(2004)
Jackknife					0.912	Cai and Chou
validation						(2004)

Table 4. Performances of our prediction system for various dimensions of the feature vector on the eukaryotic non-plant data set.

Dimension of feature vector	Category	Sensitivity	Specificity	MCC	Overall accuracy
	mTP	0.7044	0.8333	0.7311	
75	SP	0.8636	0.8763	0.8221	0.8829
	other	0.9307	0.8945	0.7668	
	mTP	0.7707	0.8506	0.7805	
150	SP	0.9115	0.9101	0.8781	0.9120
	other	0.9436	0.9248	0.8276	
	mTP	0.7983	0.8731	0.8096	
300	SP	0.9245	0.9286	0.8999	0.9250
	other	0.9534	0.9339	0.8526	
	mTP	0.8315	0.8750	0.8300	
600	SP	0.9376	0.9376	0.9150	0.9336
	other	0.9546	0.9442	0.8689	
	mTP	0.8702	0.8824	0.8565	
Full	SP	0.9216	0.9478	0.9116	0.9399
	other	0.9632	0.9492	0.8859	

the results for the eukaryotic plant data through the 5-fold cross-validation and the jackknife validation. The overall prediction accuracy (γ = 0.008 and C = 10) evaluated by the 5-fold cross-validation and the jackknife validation reached 89.57% and 92.10%, respectively. The accuracy measure by the jackknife validation was about 6~7% higher than those by other prediction methods. The sensitivity, specificity and MCC for each class were also improved considerably.

The results for the eukaryotic non-plant data are shown in Table 3. The overall accuracy (γ = 0.005 and C = 7) evaluated by the jackknife validation was 94.70% and the accuracy is about 3~4% higher than those by other prediction methods. The MCC for each class are improved significantly.

In this study, we evaluated the performance of our prediction system through two validation techniques. In general, the jackknife validation is more rigorous and the 5-fold cross-validation is more likely to overestimate. However, our results were the opposite. The reason is already mentioned above. Because the dimension of the jackknife validation is higher than that of the 5-fold cross-validation, the performance of the jackknife validation becomes higher. The dependency of the performance on the dimension of the feature vector is shown in Table 4. As the dimension increases, the overall accuracy (γ = 0.005 and C = 7) was improved. The results of Table 4 were measured by the 5-fold cross-validation for the eukaryotic non-plant data.

CONCLUDING REMARKS

Our proposed prediction system showed high performance capability in predicting subcellular localization of proteins. Taking into account the significantly improved prediction accuracy, we can conclude that our proposed feature extraction step is well fitted to this classification problem. To sum up, the advantages of our prediction system are: (1) the discriminative power of the feature vector is expected to increase since it contains the information of positive and negative data; (2) our prediction system has the strong

biological implication because it considers only N-terminal signal sequences; and (3) it is easy to understand and implement our prediction system. Despite these advantages, there remain two basic limitations inherent in this approach. First, it takes a long time to vectorize protein sequences because of the dynamic programming algorithm. Second, our prediction system is not suitable to discriminate between cytoplasmic and nuclear proteins since the sorting signals of these protein sequences are not located at the N-terminal region. Therefore, what remains to be improved by future research is to extend our prediction system to circumvent these limitations.

Acknowledgements

This work was supported by Korea Ministry of Science and Technology under Creative Research Initiative programm, by POSTECH Research Fund, and by BK 21 in POSTECH.

References

[1] Alberts,B., Bray,D., Johnson,A., Lewis,J., Raff,M., Roberts,K. and Walter,P. (1998) *Essential cell biology: an introduction to the molecular biology of the cell.* Garland Pub., New York.

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

[3] Reczko,M. and Hatzigerrorgiou,A. (2004) Prediction of the subcellular localization of eukaryotic proteins using sequence signals and composition. *Proteomics*, ***4,*** 1591-1596.

[4] Bhasin,M. and Raghava,G.P. (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.*, ***32,*** W414-9.

[5] Hua,S. and Sun,Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* ***17,*** 721-728.

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

[7] Cai,Y.D. and Chou,K.C. (2004) Predicting subcellular localization of proteins in a hybridization space. *Bioinformatics*, ***20,*** 1151-1156.

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

[9] Liao,L. and Noble,W.S. (2003) Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *J. Comput. Biol.* ***10,*** 857-868.

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

[11] Hearst,M.A., Dumais,S.T., Osman,E., Platt,J. and Schölkopf,B. (1998) Support vector machines. *IEEE Intelligent Systems*, **13**, 18-28.

[12] Cristianini,N. and Shawe-Taylor,J. (2000) *An introduction to support vector machines : and other kernel-based learning methods.* Cambridge, New York.

[13] Müller,K.R., Mika,S., Ratsch,G., Tsuda,K. and Schölkopf,B. (2001) An introduction to kernel-based learning algorithms. *IEEE Trans. on Neural Networks*, **12**, 181-202.

[14] Schölkopf,B. and Smola,A.J. (2002) Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT Press, Cambridge.

[15] Matthews,B.W. (1975) Comparison of predicted and observed secondary structure of T4 phase lysozyme. *Biochim. Biophys. Acta.*, **405**, 442-451.