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

Jong Kyoung Kim [a], G. P. S. Raghava [b,*], Sung-Yang Bang [a], Seungjin Choi [a,*]

[a] *Department of Computer Science*
*Pohang University of Science and Technology*
*San 31 Hyoja-dong, Nam-gu, Pohang 790-784, Korea*

[b] *Bioinformatics Centre*
*Institute of Microbial Technology*
*Sector 39A,Chandigarh, India*

**Abstract**

Predicting the destination of a protein in a cell is important for annotating the function of the protein. Recent advances have allowed us to develop more accurate methods for predicting the subcellular localization of proteins. One of the most important factors for improving the accuracy of these methods is related to the introduction of new useful features for protein sequences. In this paper we present a new method for extracting appropriate features 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 reached 94.70% for the eukaryotic non-plant data set and 92.10% for the eukaryotic plant data set, which is the highest prediction accuracy among the methods reported so far with such data sets. Our experimental results confirm that our feature extraction method based on pairwise sequence alignment is useful for this classification problem.

*Key words:* Pairwise sequence alignment, Subcellular localization, Support vector machine

* Corresponding author. Tel.: +82-54-279-2259; Fax: +82-54-279-2299
*Email:* `raghava@imtech.res.in` (G. Raghava),
`seungjin@postech.ac.kr` (S. Choi)
*URL:* `http://www.postech.ac.kr/∼seungjin` (S. Choi)

## 1 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 are 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. The delivery of a protein to the ER, mitochondria, and chloroplasts, depends on the N-terminal signal sequence. The N-terminal signal sequence, located at the N-terminus, is a continuous stretch of amino acid sequence which determines the proper cellular location. Since the signal sequence specifying the same destination is not well conserved, it is generally thought that the factors determining the destination are physico-chemical properties such as hydrophobicity or the position of charged amino acids (Alberts et al., 1998).

Predicting the destination of an unknown protein is important for inferring the possible function of the protein. Therefore, in recent years, numerous methods in computational biology have been developed to improve the prediction accuracy. In fact, this is a classification problem that has been extensively studied in machine learning, pattern recognition, and statistics communities, since class labels related to cellular locations are already available in a set of training data. Various classifiers including artificial neural networks (ANN), support vector machines (SVM), and k-nearest neighbor algorithm (k-NN), have been applied to this classification problem. However, one of the most critical factors for improving the prediction accuracy, involves a way of feature extraction. Most of prediction methods can be divided into two approaches, depending on their ways of feature extraction: (1) features based on protein sequence data; (2) features based on ontology data.

In the protein sequence-based approach, two popular feature extraction methods include: (1) methods involving the recognition of N-terminal signal sequences; (2) methods involving 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 at the N-terminus (Emanuelsson et al., 2000; Reczko and Hatzigerrorgiou, 2004). However, it is difficult to recognize underlying features from a highly diverged signal sequence and to vectorize those features. The latter approach partially overcomes these difficulties, but loses the information regarding the context stored in the sequence data (Bhasin and Raghava, 2004; Hua and Sun, 2001; Reinhardt and Hubbard, 1998). The ontology-based approach has received much attention recently because of its high prediction accuracy (Cai and Chou, 2004; Lu et al., 2004). This approach extracts the text information of homologous sequences of a target sequence by searching biological databases, and vectorizes this 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 for extracting appropriate features from the sequence data to predict cellular locations of proteins. To this end, we introduce a pairwise sequence alignment score such that a protein sequence is presented to a SVM classifier as a vector. Our experimental results confirm that our feature extraction method considerably improves the prediction accuracy.

## 2 Systems and methods

### *2.1 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 cellular 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).

### *2.2 Pairwise sequence alignment as a feature extractor*

Representation of a protein sequence by the scores of pairwise sequence alignments (SA) was already used in the SVM-pairwise for detecting remote structural and evolutionary relationships (Liao and Noble, 2003). In some aspects, the 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 the input to SVM for classification.

The main distinction between the SVM-pairwise and our method, is in the locality of the pairwise-sequence alignment. The SVM-pairwise uses the Smith-Waterman algorithm (Smith and Waterman, 1981) for finding the optimal *local alignment* because the global SA of two very highly diverged sequences is not possible. In contrast, our prediction system uses the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) for obtaining the optimal *global alignment*. In order to consider only N-terminal signal sequences, all sequences were truncated after first 90 residues such that every sequence has the same length. It is believed that the whole N-terminal sequences are important in determining cellular locations. Therefore, it is desirable to use the global dynamic programming algorithm.

For the global dynamic programming algorithm, we used Matlab functions that are available at http://www.cs.cornell.edu/courses/cs321/2001fa/matlab_examples.html. A $d$-dimensional feature vector $\boldsymbol{x}_k$ for the $k$th protein sequence has the form

$$\boldsymbol{x}_k = [x_{k1}, x_{k2}, \cdots, x_{kd}]^\top, \tag{1}$$

where $\top$ denotes the matrix or vector transpose operator and $x_{ki}$ represents the score of the Needleman-Wunsch algorithm between the $k$th sequence 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.

Table 1
The number of sequences in each cellular location of eukaryotic plant and non-plant data sets. (Emmanuelsson *et al.*, 2000)

| Species | Cellular location | # of sequences |
|---|---|---|
| Eukaryotic | Chloroplast (cTP) | 141 |
| Plant | Mitochondirial (mTP) | 368 |
| | Extracellular (SP) | 269 |
| | Cytoplasmic + Nuclear (Other) | 162 |
| Eukaryotic | Mitochondirial (mTP) | 371 |
| Non-plant | Extracellular (SP) | 715 |
| | Cytoplasmic + Nuclear (Other) | 1652 |

### *2.3 Support vector machine as a classifier*

SVM classifiers have recently been used as popular and powerful tools for classification, due to their strong theoretical origin at statistical learning theory as well as their high performance in practical applications(Hearst et al., 1998; Cristianini and Shawe-Taylor, 2000). SVM classifiers are kernel-based learning algorithms, determining the optimal hyperplane decision boundary in the feature space. In kernel-based algorithms, a kernel trick leads us to process the data in a feature space without the explicit knowledge of a nonlinear mapping from the data space to a feature space. The high dimensionality of a feature space might 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 (Müller et al., 2001). 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 http://www.ece.osu.edu/∼maj/osu_svm. The prediction of the 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 classifiers 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 (Schölkopf and Smola, 2002). The kernel function used in this study is the *radial basis function* (RBF) kernel with one parameter $\gamma$:

$$k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left\{-\gamma\|\boldsymbol{x} - \boldsymbol{y}\|^2\right\}. \tag{2}$$

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

### *2.4 The proposed prediction system*

The overall schematic diagram of our prediction system is illustrated in Fig. 1. Every protein sequence in consideration (including target sequence and all sequences in the training set), is truncated after first 90 residues, such that only N-terminal signal sequence is taken into account. A target sequence is converted into an associated feature vector by computing the scores of the Needleman-Wunsch algorithm between the target sequence and every sequence in the training set. The training set can be divided into two parts which are positive and negative vectorization set. The positive vectorization set means all sequences in 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 about positive as well as negative data. 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.

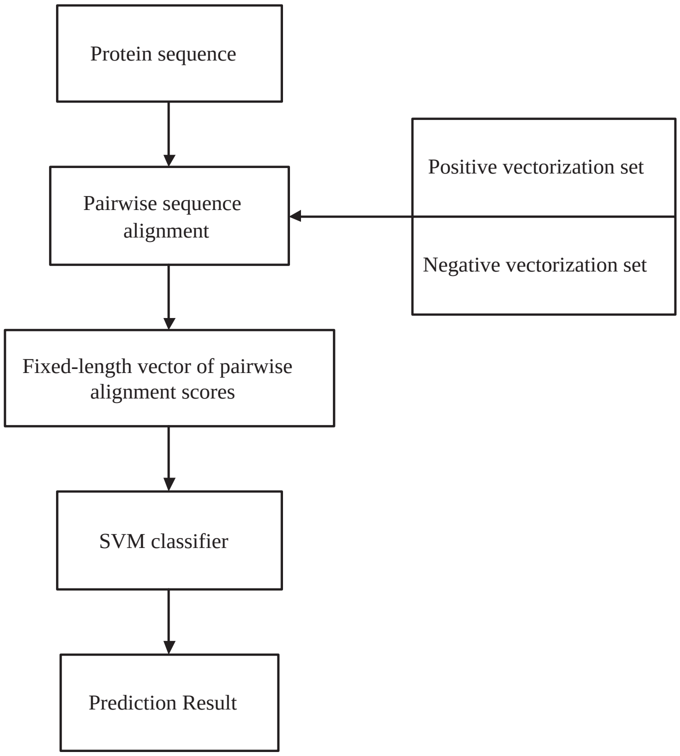
Protein sequence
Pairwise sequence alignment
Positive vectorization set
Negative vectorization set
Fixed-length vector of pairwise alignment scores
SVM classifier
Prediction Result

Fig. 1. The schematic diagram of our proposed system is illustrated. A target protein sequence is converted into the corresponding feature vector by computing the scores of the Needleman-Wunsch algorithm between the protein sequence and the whole sequences in the training data set. The SVM classifier predicts an appropriate class of the protein sequence.

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, Matthew's correlation coefficient (MCC) (Matthews, 1975), and overall accuracy were calculated using the following equations:

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

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

$$\text{MCC}(i) = \frac{\text{tp}(i)\text{tn}(i) - \text{fp}(i)\text{fn}(i)}{\sqrt{\text{de}(i)}}, \tag{5}$$

$$\text{Overall accuracy} = \frac{\sum_{i=1}^{M} \text{tp}(i)}{N}, \tag{6}$$

where

$$\begin{aligned}\text{de}(i) = {} & (\text{tp}(i) + \text{fn}(i))\,(\text{tp}(i) + \text{fp}(i)) \\ & (\text{tn}(i) + \text{fp}(i))\,(\text{tn}(i) + \text{fn}(i))\,,\end{aligned} \tag{7}$$

and $N$ is the total number of sequences, $M$ 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 $i$.

Table 2
Performance comparison of different subcellular localization prediction methods on the eukaryotic plant data set. In the table, we used 5-fold CV for 5-fold cross validation, Jackknife for Jackknife validation, categ for catergory, sensit for sesitivity, and accur for overall accuracy.

| method | categ | sensit | specif | MCC | accur | reference |
|---|---|---|---|---|---|---|
| 5-fold CV | cTP | 0.8511 | 0.8163 | 0.8003 | 0.8957 | our method |
| | mTP | 0.8886 | 0.9355 | 0.8536 | | |
| | SP | 0.9375 | 0.9836 | 0.9435 | | |
| | other | 0.8839 | 0.7654 | 0.7814 | | |
| Jackknife | cTP | 0.8794 | 0.8794 | 0.8562 | 0.9210 | our method |
| | mTP | 0.9136 | 0.9535 | 0.8898 | | |
| | SP | 0.9492 | 0.9918 | 0.9581 | | |
| | other | 0.9290 | 0.7956 | 0.8278 | | |
| 5-fold CV | cTP | 0.85 | 0.69 | 0.72 | 0.853 | Emanuelsson *et al.* (2000) |
| | mTP | 0.82 | 0.90 | 0.77 | | |
| | SP | 0.91 | 0.95 | 0.90 | | |
| | other | 0.85 | 0.78 | 0.77 | | |
| Jackknife | | | | | 0.861 | Cai-Chou (2004) |

## 3 Results

Prediction results with comparison to some other methods, are summarized in Table 2 and 3 for the eukaryotic plant and non-plant data, respectively. Parameters in the SVM classifier, including the kernel width $\gamma$ and the regularization parameter $C$, were selected through the 5-fold cross-validation. Table 2 shows the results for the eukaryotic plant data through the 5-fold cross-validation and the jackknife validation. The overall prediction accuracy ($\gamma = 0.008$ and$C = 10$) evaluated by the 5-fold cross-validation and the jackknife validation reached 89.57% and 92.10%, respectively. The accuracy measured 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 ($\gamma = 0.005$ and$C = 7$) evaluated by the jackknife validation was 94.70% and the accuracy was about 3∼4% higher than those by other prediction methods. The MCC for each class was improved significantly.

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

| method | categ | sensit | specif | MCC | accur | reference |
|---|---|---|---|---|---|---|
| 5-fold CV | mTP | 0.8702 | 0.8824 | 0.8565 | 0.9399 | our method |
| | SP | 0.9216 | 0.9478 | 0.9116 | | |
| | other | 0.9632 | 0.9492 | 0.8859 | | |
| Jackknife | mTP | 0.8785 | 0.8908 | 0.8662 | 0.9470 | our method |
| | SP | 0.9390 | 0.9557 | 0.9287 | | |
| | other | 0.9656 | 0.9557 | 0.8981 | | |
| 5-fold CV | mTP | 0.89 | 0.67 | 0.73 | 0.900 | Emanue- |
| | SP | 0.96 | 0.92 | 0.92 | | lsson *et al.* |
| | other | 0.88 | 0.97 | 0.82 | | (2000) |
| 5-fold CV | mTP | 0.78 | 0.82 | 0.77 | 0.913 | Reczko and |
| | SP | 0.93 | 0.91 | 0.89 | | Hatzi-georgiou |
| | other | 0.93 | 0.94 | 0.84 | | (2004) |
| Jackknife | | | | | 0.912 | Cai-Chou (2004) |

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 ($\gamma = 0.005$ and $C = 7$) was improved. The results of Table 4 were measured by the 5-fold cross-validation for the eukaryotic non-plant data.

In general, the high dimension of the feature vector can cause the over-fitting problem. Therefore, the high performance of our system may be the over-fitted result. To relieve this problem, we used SVM as a classifier which can be viewed as minimizing the complexity term of the upper bound of the expected risk. In addition, we tested our system through the rigorous validation technique. Finally, the high positive correlation between the dimension of the feature vector and the prediction accuracy supports that the high performance of our system is not the over-fitted result.

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

| dimension | categ | sensit | specif | MCC | accur |
|---|---|---|---|---|---|
| 75 | mTP | 0.7044 | 0.8333 | 0.7311 | 0.8829 |
| | SP | 0.8636 | 0.8763 | 0.8221 | |
| | other | 0.9307 | 0.8945 | 0.7668 | |
| 150 | mTP | 0.7707 | 0.8506 | 0.7805 | 0.9120 |
| | SP | 0.9115 | 0.9101 | 0.8781 | |
| | other | 0.9436 | 0.9248 | 0.8276 | |
| 300 | mTP | 0.7983 | 0.8731 | 0.8096 | 0.9250 |
| | SP | 0.9245 | 0.9286 | 0.8999 | |
| | other | 0.9534 | 0.9339 | 0.8526 | |
| 600 | mTP | 0.8315 | 0.8750 | 0.8300 | 0.9336 |
| | SP | 0.9376 | 0.9376 | 0.9150 | |
| | other | 0.9546 | 0.9442 | 0.8689 | |
| Full | mTP | 0.8702 | 0.8824 | 0.8565 | 0.9399 |
| | SP | 0.9216 | 0.9478 | 0.9116 | |
| | other | 0.9632 | 0.9492 | 0.8859 | |

## 4 Concluding remarks

We have presented a method for predicting the subcellular localization of proteins, where features associated with protein sequences were constructed by the scores of the global alignment (Needleman-Wunsch algorithm) between only N-terminal signal sequences. The high prediction performance of our method was verified, using the eukaryotic plant and non-plant data sets, through 5-fold cross validation as well as the Jackknife validation. The advantages of our prediction system are: (1) the discriminative power of the feature vector is expected to increase, since it contains the information on positive as well as negative data; (2) our prediction system has important biological implications because it considers only N-terminal signal sequences; and (3) the system is easy to understand and implement. Despite these advantages, there remain two basic limitations inherent in this approach. First, the vectorization of protein sequences is computationally ”expensive”, because it is based on a dynamic programming algorithm. Second, our prediction system is not suitable for the discrimination between cytoplasmic and nuclear proteins, since the sorting signals of these protein sequences are not located at the N-terminus. Therefore, what remains to be done in the future research is to extend the proposed system to circumvent these limitations.

## 5 Acknowledgments

This work was supported by National Core Research Center for Systems Bio-Dynamics.

## References

Alberts, B., Bray, D., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., 1998. Essential Cell Biology: An Introduction to the Molecular Biology of the Cell. Garland, New York.

Bhasin, M., 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–W419.

Cai, Y. D., Chou, K. C., 2004. Predicting subcellular localization of proteins in a hybridization space. Bioinformatics 20, 1151–1156.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge, New York.

Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G., 2000. Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. J. Mol. Biol. 300, 1005–1016.

Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., Schölkopf, B., 1998. Support vector machines. IEEE Intelligent Systems 13, 18–28.

Hua, S., Sun, Z., 2001. Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17, 721–728.

Liao, L., 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.

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

Matthews, B. W., 1975. Comparison of predicted and observed secondary structure of t4 phase lysozyme. Biochim. Biophys. Acta. 405, 442–451.

Müller, K. R., Mika, S., Ratsch, G., Tsuda, K., Schölkopf, B., 2001. An introduction to kernel-based learning algorithms. IEEE Trans. on Neural Networks 12, 181–202.

Needleman, S. B., 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.

Reczko, M., Hatzigerrorgiou, A., 2004. Prediction of the subcellular localization of eukaryotic proteins using sequence signals and composition. Proteomics 4, 1591–1596.

Reinhardt, A., Hubbard, T., 1998. Using neural networks for prediction of the subcellular location of proteins. Nucelic Acids Res. 26, 2230–2236.

Schölkopf, B., Smola, A. J., 2002. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge.

Smith, T. F., Waterman, M. S., 1981. Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197.