2007－BIO－8 (1)
2007／3／5

アラインメントとアミノ酸構成比に基づいた サポートベクターマシンによるタンパク質の細胞内局在予測

田村武幸　　阿久津達也
京都大学 化学研究所 バイオインフォマティクスセンター

概要 タンパク質の細胞内局在予測とは、タンパク質のアミノ酸配列が入力として与えられた時に、そのタンパク質が細胞内のどの部位に移送されていくかを予測する問題である。本稿では、文字列アラインメントとアミノ酸構成比に基づく特徴ベクトルを組み合わせる方法を紹介する。我々はこの方法をサポートベクターマシンと TargetP データベースから入手した植物に関するデータ集合を用いて実装した。Fivefold cross validation test によって得られた overall accuracy と average MCC はそれぞれ 0.8915 と 0.8363 であり、アミノ酸配列の情報だけを用いる既存の予測法よりも良い値であった。本稿における提案手法は一般的かつ簡素であるため、バイオインフォマティクスにおける他の問題にも応用可能であると考えられる。

Subcellular Location Prediction of Proteins Using Support Vector Machines with Alignment and Amino Acid Composition

Takeyuki Tamura and Tatsuya Akutsu
Bioinformatics Center, Institute for Chemical Research, Kyoto University

Abstract Subcellular location prediction of proteins is a problem of predicting which part in a cell a given protein is transported to, where an amino acid sequence of the protein is given as an input. In this report, we introduce a novel and general predicting method by combining techniques for sequence alignment and feature vectors based on amino acid composition. We implemented this method with support vector machines on plant data sets extracted from the TargetP database. Through fivefold cross validation tests, obtained overall accuracy and average MCC were 0.8915 and 0.8363 respectively. These values are higher than existing sequence-based predictors which use only sequence information. Our predictor is considered to be applicable to other problems in bioinformatics since our method is simple and general.

1 Introduction

Bioinformatics is one of the important fields for application of intelligent systems and technologies. Though there exist many important problems in bioinformatics for which intelligent technology can be applied, prediction of subcellular location of proteins is one of the most studied problems. This is a problem of predicting which part (e.g., Mitochondria, Chloroplast, etc.) in a cell a given protein is transported to, where an amino acid sequence (i.e., string data) of the protein is given as an input as shown in Fig. 1. This problem is becoming more important because information on subcellular location is helpful for annotation of proteins and genes and the number of complete genomes is rapidly increasing.

For the protein subcellular location problem, many methods have been proposed using various intelligent techniques. Furthermore, many web-based prediction systems have been developed based on these proposed methods. PSORT [14, 20], which is historically the first subcellular location predictor, uses various sequence-derived features such as the presence of sequence motifs and amino acid compositions. Although there are many predicting methods, they can be roughly classified into two groups. One is the N-terminal based method and the other is based on amino acid composition. TargetP [11] requires the

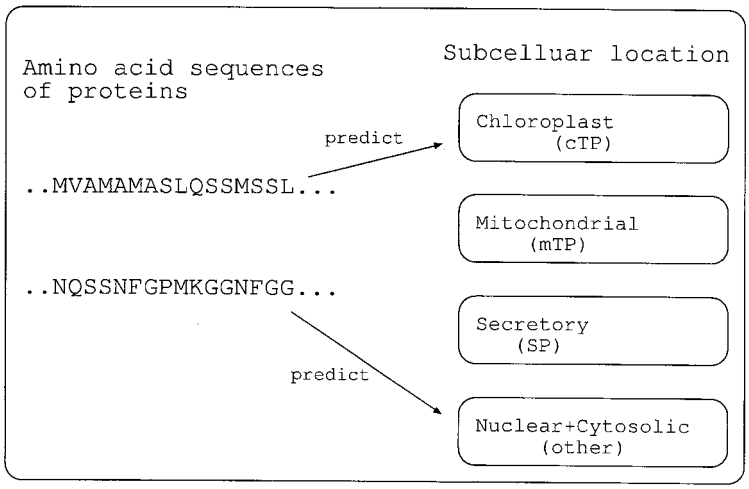

Figure 1: Subcellular location prediction of proteins is a problem of predicting which part in a cell a given protein is transported to, where an amino acid sequence of the protein is given as an input.

N-terminal sequence as an input into two layers of artificial neural networks (ANN) with utilizing the earlier binary predictors, SignalP [21] and ChloroP [12]. Reczko and Hatzigeorgiou [24] used a bidirectional recurrent neural network with the first 90 residues in the N-terminal sequence.

ProtLock [4] is based on the amino acid composition and the least Mahalanobis distance algorithm. Chou and Elrod [9, 10] used the covariant discriminant algorithm besides amino acid composition. NNPSL [25] is an ANN-based method using the amino acid composition. After the successful report in [25], application of machine learning techniques became popular in this field. A support vector machine (SVM) was implemented for SubLoc [15] instead of the ANN. Incorporating an amino acid order as well as the amino acid composition is expected to make it possible to improve prediction performance. Chou proposed the pseudo-amino acid composition to take the effect of the amino acid order into account [5]. Moreover, Cai and Chou [3] have recently developed an accurate method integrating the pseudo-amino acid composition, the functional domain composition [6, 8], and the information of gene ontology [7]. Park and Kanehisa [4] developed an SVM based method that incorporates compositions of dipeptides and gapped amino acid pairs besides the conventional amino acid composition. The concepts of the pseudo-amino acid and gapped amino acid pair compositions were merged in the residue-couple model proposed by Guo et al. [13].

Recently, Mastuda et al. [18] proposed a novel representation of protein sequences. That representation involves local amino acid compositions and twin amino acids, and local frequencies of distance between successive (basic, hydrophobic, and other) amino acids. Each sequence is split into three parts: N-terminal, middle, and C-terminal in order to calculate the local features. The N-terminal part is further divided into four regions for considering ambiguity in the length and position of signal sequences. It was combined with SVM for prediction of subcellular location of proteins. The results of computational experiments suggest that their method is one of the state-of-the-art methods. Though the prediction accuracy is high, the method is based on various heuristics. Furthermore, many of the heuristics are specific to the protein subcellular location problem. In this paper, we try to develop a less heuristic method for the protein subcellular location problem with keeping similar prediction accuracy. Development of such a method is important since it may be applied to other problems in bioinformatics. For example, the *spectrum kernel* [17], which is a simple and general kernel function for SVM, has been applied to various problems including remote homology detection [17], recognition of DNA-binding proteins [2], prediction of protein-protein interactions [1] and prediction of protein subcellular location.

To develop a general method, we combine two-techniques: sequence alignment and a feature vector

based on amino acid composition. It should be noted that amino acid composition-based feature vector is as spectrum kernel. Elements of our proposed kernel matrix are scores of alignment between sequences of subsequences of proteins. The alignment scores are calculated in accordance with amino acid composition-based feature vectors. Note that there is a possibility that the obtained matrix is not semi-definite since we use alignment. Fortunately, in most cases, our method can train SVM without additional operations in order to satisfy the kernel condition. To evaluate the efficiency of our method, we compared the prediction accuracy of subcellular location for TargetP plant data sets with existing methods through fivefold cross validation tests. Although our prediction method is less heuristic than existing predictors, the overall accuracy and average MCC, which are standard measures of prediction accuracy, are 0.8915 and 0.8363 respectively. They are higher than existing predictors.

2 Method

Contents of this section are as follows. First, our method is explained by using figures and examples intuitively. Second, the method is expressed mathematically.

Assume that *sequence*1 = $AAAAACCCCCDEFGHIIKKKLLLLL$ and *sequence*2 = $MMMMMC$ $CCCCAAAAACCCCCNNN$ are given as shown in Fig. 2 (a) and (b) respectively. We make sequences of subsequences as shown in Fig. 2 (c) in accordance with w and c, where w is the length of subsequences and c is the unit of distances of left ends of subsequences. Note that $w = 10$ and $c = 5$ are used in Fig. 2 (a) and (b). In Fig. 2 (b), "ϕ"s are assigned to the rightmost subsequence since there are not corresponding amino acids. Obtained sequences are aligned as shown in Fig. 2 (d). In our method, while both left ends of sequences must be used by the alignment, we do not have to use right ends of sequences. Note that subsequences $DEFGHIIKKK$ and $IIKKKLLLLL$ of *sequence*1, and a subsequence $CCCCCNNN\phi\phi$ of *sequence*2 are not used in Fig. 2 (d). One of the simplest methods for calculating pairing scores is to use inner products of vectors based on amino acid compositions. For example, the amino acid compositions for $AAAAACCCCC$, $DEFGHIIKKK$, and $CCCCCNNN\phi\phi$ are (A=0.5, C=0.5, the other=0), (D=0.1, E=0.1, F=0.1, G=0.1, H=0.1, I=0.2, K=0.3, the other=0), and (C=0.5, N=0.3, the other=0) respectively in our method. The score obtained by pairing $AAACCCCCDD$ and $DDDDDAACCC$ is $0.3 \times 0.2 + 0.5 \times 0.3 + 0.2 \times 0.5 = 0.31$. However, our implemented method is slightly different. $2 \cdot exp(-\gamma||\mathbf{b}_{x,j_x} - \mathbf{b}_{y,j_y}||^2) - 1$ is used as a pairing score where $\mathbf{b}_{x,j_x}$ and $\mathbf{b}_{x,j_x}$ are feature vectors for subsequences which are based on amino acid compositions. Since $2 \cdot exp(-\gamma||\mathbf{b}_{x,j_x} - \mathbf{b}_{y,j_y}||^2) - 1$ always takes $[-1, 1]$, pairing scores also take $[-1, 1]$ in our implemented method. Note that the pairing score takes a positive value when two subsequences are similar each other. On the other hand, the pairing score takes a negative value when two subsequences are not similar each other. Then, there is a possibility that higher alignment scores are obtained in the case where right ends of sequences are not used than in the case where right ends of sequences must be used. For example, in Fig. 2 (c) and (d), pairing $IIKKKLLLLL$ and $CCCCCNNN\phi\phi$ takes a negative value in our implemented method since these are not similar each other. Note that the inner product of the vectors of amino acid compositions is 0. However, $2 \cdot exp(-\gamma||\mathbf{b}_{x,j_x} - \mathbf{b}_{y,j_y}||^2) - 1$ takes a negative value by assigning appropriate γ. The optimal alignment of given two sequences is calculated by using these obtained pairing scores of subsequences. The optimal scores of alignments are used as elements of the kernel matrix. Finally, our predictor selects a location whose "discriminant" value is higher than any other location. Note that our "discriminant" values are calculated by slightly modifying values outputted by "gist-classify" [23].

Let c and w be some positive integers used as parameters. Let $S_i = s_{i,1}s_{i,2}\dots s_{i,n_i}$ $\phi\phi\phi\cdots$ be given protein sequences ($i = 1, 2, \cdots, m$), where m is the number of sequences and n_i is the length of S_i. Let $B_i = b_{i,1}b_{i,2}\dots b_{i,\max(\lceil (n_i-w)/c \rceil,0)+1}$ be a sequence of subsequences of S_i, where $b_{i,j} = s_{i,c(j-1)+1}s_{i,c(j-1)+2}\cdots s_{i,c(j-1)+w}$. Note that there is a possibility that $b_{i,j}$ includes ϕ.

Let x and y be integers which satisfy $1 \le x, y \le m$. Moreover, let j_x and j_y be integers which satisfy $1 \le j_x \le \max(\lceil (n_x - w)/c \rceil, 0) + 1$ and $1 \le j_y \le \max(\lceil (n_y - w)/c \rceil, 0) + 1$. Feature vectors (defined later) of b_{x,j_x} and b_{y,j_y} are denoted by $\mathbf{b}_{x,j_x}$ and $\mathbf{b}_{y,j_y}$ respectively. The kernel-like value between B_x and B_y, which is denoted by $K(B_x, B_y)$, is calculated by the following dynamic programming (DP) procedures:

$$K(B_x, B_y) = \max_{j_x, j_y} D(j_x, j_y),$$

Subcellular location	No. of sequences (plant)
Chloroplast(cTP)	141
Mitochondrial(mTP)	368
Secretory(SP)	269
Nuclear+cytosolic(other)	162
Total	940

Table 1: Number of sequences in each subcellular location of TargetP plant data sets

$$D(j_x, j_y) = \max \begin{cases} D(j_x - 1, j_y) - p \\ D(j_x, j_y - 1) - p \\ D(j_x - 1, j_y - 1) + f(j_x, j_y), \end{cases}$$

where $f(j_x, j_y) = 2 \cdot exp(-\gamma ||\mathbf{b}_{x,j_x} - \mathbf{b}_{y,j_y}||^2) - 1$, $D(0,0) = 0$, $D(j_x, 0) = -pj_x$, $D(0, j_y) = -pj_y$, γ is the parameter of RBF kernel, and p is the gap penalty of the alignment. Note that the value of $f(j_x, j_y)$ is on the interval $[-1, 1]$. When $\mathbf{b}_{x,j_x}$ and $\mathbf{b}_{y,j_y}$ are similar, $f(j_x, j_y)$ takes a positive value. On the other hand, when $\mathbf{b}_{x,j_x}$ and $\mathbf{b}_{y,j_y}$ are not similar, $f(j_x, j_y)$ takes a negative value.

The feature vector to represent a protein subsequence is expressed as follows: $\mathbf{b} = (r_1, r_2, \ldots, r_{20}, q_1, q_2, \ldots, q_{20}, z_1, z_2, \ldots, z_{18})^T$, where $r_1, r_2, \ldots, r_{20}$ indicate the composition of 20 amino acids. $q_1, q_2, \ldots, q_{20}$ are the composition of 20 twin amino acids (e.g., RR, KK). $z_1, \ldots, z_6$ represent the distance frequency [18] of basic amino acids (R, K, and H). To calculate distance frequencies, we defined six distance classes ($H = 1, 1 < H \leq 6, 6 < H \leq 11, 11 < H \leq 16, 16 < H \leq 21, 21 < H$). Similarly the distance frequencies for hydrophobic amino acids (I, V, L, F, M, A, G, W, and P) and the other amino acids (D, N, E, Q, Y, S, T, and C) are represented by $z_7, \ldots, z_{12}$ and $z_{13}, \ldots, z_{18}$ respectively.

In this work, the data sets were collected from plant proteins of TargetP [11] (See Table 1). In order to perform a fivefold cross-validation test, each data set was partitioned into five subsets that have exactly equal sizes. Note that it can be done since the number of sequences is 940. Before partitioning, we shuffled the sequences by using at least 1000 random numbers. One subset is regarded as test data and the remaining four subsets as training data. This procedure is repeated five times so that each subset is used as test data once.

Let $score(cTP)$, $score(mTP)$, $score(SP)$, and $score(other)$ be values of "discriminant" calculated for a protein sequence by gist-classify [23] when "cTP", "mTP", "SP", and "other" are positive locations respectively. Our predictor calculates $\max\{score(cTP) + constant, score(mTP) + constant, score(SP) + constant, score(other) + constant\}$ and chooses the corresponding location as an output. Note that "constant"s are different for each location as shown in Table 3.

3 Results

In order to implement SVM, we used the software GIST [23]. We evaluated the prediction performance of our method by calculating sensitivity, specificity, Matthew's correlation coefficient (MCC) [19], and overall accuracy for each subcellular location. The definitions of these measures are as follows:

$$Sensitivity(l) = \frac{tp(l)}{tp(l) + fn(l),} \qquad Specificity(l) = \frac{tp(l)}{tp(l) + fp(l),}$$

$$MCC(l) = \frac{tp(l) \cdot tn(l) - fp(l) \cdot fn(l)}{\sqrt{(tp(l) + fn(l))(tp(l) + fp(l))(tn(l) + fp(l))(tn(l) + fn(l))},}$$

$$Total \ \ accuracy = \frac{1}{m} \sum_{l=1}^{k} tp(l),$$

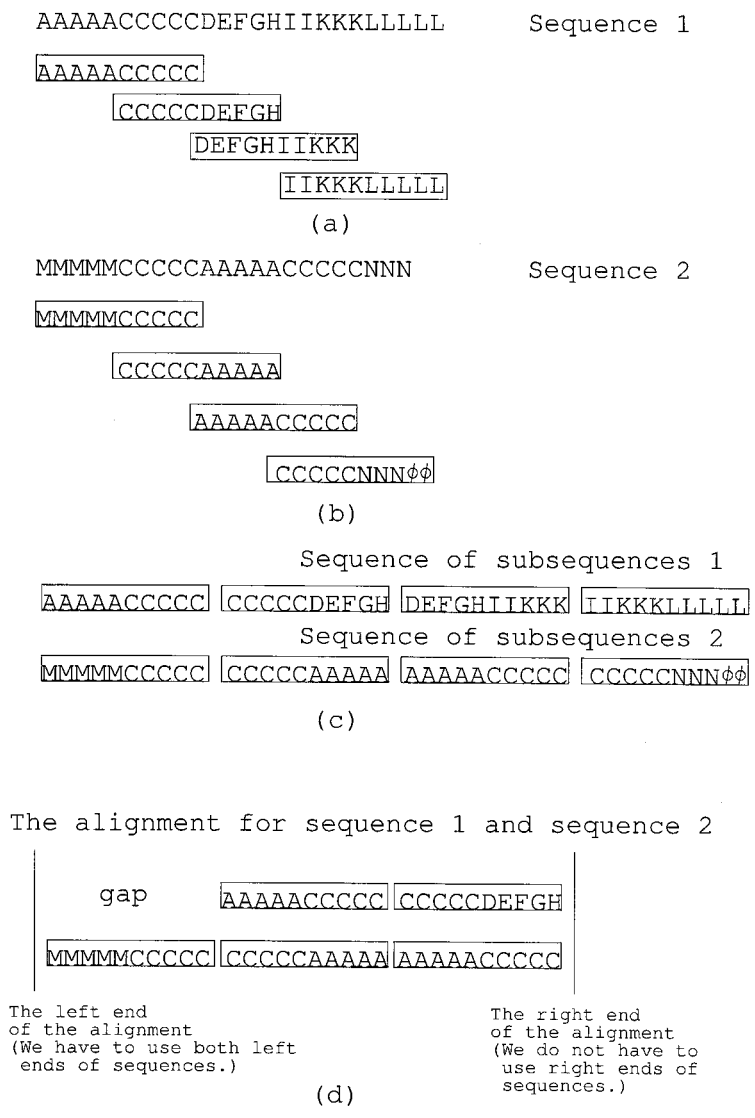

Figure 2: (a)(b)(c) Sequences of subsequences are obtained, where $w = 10$ is the length of subsequences and $c = 5$ is the unit of distances of left ends of subsequences. (d) Obtained sequences are aligned. While both left ends of sequences must be used by the alignment, we do not have to use right ends of sequences in our method.

Predictor	Location	Sensitivity	Specificity	MCC	Average MCC	Overall accuracy
Our method	cTP	0.8227	0.8169	0.7879	0.8363	0.8915
	mTP	0.9158	0.9158	0.8616		
	SP	0.9517	0.9209	0.9124		
	other	0.7963	0.8431	0.7831		
Matsuda et al. (2005)	cTP	0.7591	0.8474	0.7694	0.8244	0.8809
	mTP	0.9240	0.8652	0.8227		
	SP	0.9219	0.9326	0.8983		
	other	0.8210	0.8586	0.8070		
Kim et al. (2004)	cTP	0.6874	0.8435	0.7222	0.7791	0.8479
	mTP	0.8970	0.8392	0.7773		
	SP	0.8592	0.9428	0.8872		
	other	0.8027	0.7549	0.7296		
Emanuelsson et al. (2000)	cTP	0.85	0.69	0.72	0.79	0.853
	mTP	0.82	0.90	0.77		
	SP	0.91	0.95	0.90		
	other	0.85	0.78	0.77		

Table 2: Comparison of predictive accuracy for plant proteins in the TargetP data set. "cTP", "mTP", "SP", and "other" indicate proteins destined for chloroplast, mitochondria, secretory pathway, and other locations (nucleus and cytosol), respectively.

where m is the total number of protein sequences and k is the number of subcellular locations. $tp(l)$ is the number of correctly predicted sequences belonging to location l (true positive). $tn(l)$ is the number of correctly predicted sequences that do not belong to location l (true negative). $fp(l)$ is the number of overpredicted sequences in location l (false positive). $fn(l)$ is the number of underpredicted sequences in location l (false negative).

Results and used parameters are shown in Table 2 and 3 respectively. "posconstraint" is a parameter of GIST which sets an explicit upper bound on the magnitude of the weights for positive training examples. Similarly, "negconstraint" sets an explicit upper bound on the magnitude of the weights for negative training examples. "constant" is added to the obtained score when locations are predicted. The other parameters are explained above.

Table 2 shows the comparison of predictive accuracies with existing methods on the TargetP plant data sets. Although it is known that the overall accuracies of the predictor by Chou and Cai [7, 8] are remarkably high, the sensitivity, specificity, and MCC of their method are not given in their paper and their method uses the information of gene ontology and functional domain. Since the other methods require only sequence information, we cannot compare their method with the other methods directly.

In Table 2, our overall accuracy and average MCC are higher than any other predictor. Although our MCC for "other" is lower than Matsuda et al. (2005), our MCC for "cTP", "mTP" and "SP" are higher than any other predictor. Then, it can be said that the prediction accuracy of our method is higher than existing methods.

4 Conclusion and Future Works

In this paper, we introduced a novel subcellular location predicting method which is based on sequence alignment and amino acid composition. Through fivefold cross validation tests for TargetP plant data sets, we obtained the overall accuracy of 0.8915 and the average MCC of 0.8363. These values are higher than existing predictors which use only sequence information.

We are trying to improve the accuracy by optimizing some parameters and actually obtained better

Location	gap penalty	γ of RBF	posconstraint	negconstraint	c	w	constant
cTP	0.6	2	0.05	0.012	10	20	0
mTP	0.6	2	not limited	not limited	10	20	- 0.013
SP	0.6	2	0.05	0.019	10	20	0
other	0.6	2	0.05	0.014	10	20	+ 0.011

Table 3: Parameters which are used in our method in Table 2. Gap penalty is used in alignment. γ is the parameter of RBF kernel. "posconstraint" and "negconstraint" are parameters of GIST. "constant" is added to the obtained score when locations are predicted.

results although these are not shown in this paper. We are also developing the web-based prediction system based on our proposed method. Our predictor is considered to be applicable to other problems in bioinformatics since our method is simple and general.

References

[1] Ben-Hur, A., Noble, W. S.: Kernel methods for predicting protein-protein interactions. *Bioinformatics.* **21** (2006) i38-i46

[2] Bhasin, M., Reinherz, E. L., Reche, P. A.: Recognition and classification of histones using support vector machines. *Journal of Computational Biology.* **13** (2006) 102-112

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

[4] Cedano, J., Aloy, P., Perez-Pons, J.A., Querol, E.: Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol.* **266** (1997) 594-600

[5] Chou, K.-C.: Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins.* **43** (2001) 246-255

[6] Chou, K.-C., Cai Y.-D.: Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.* **277** (2002) 45765-45769

[7] Chou, K.-C., Cai Y.-D.: A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. *Biochem. Biophys. Res. Commun.* **311** (2003) 743-747

[8] Chou, K.-C., Cai Y.-D.: Predicting subcellular localization of proteins by hybridizing functional domain composition and pseudo-amino acid composition. *J. Cell. Biochem.* **91** (2004) 1197-1203

[9] Chou, K.-C., Elrod, D. W., Using discriminant function for prediction of subcellular location of prokaryotic proteins. *Biochem. Biophys. Res. Commun.* **252** (1998) 63-68

[10] Chou, K.-C., Elrod, D. W.: Protein subcellular location prediction. *Protein Eng.* **12** (1999) 107-118

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

[12] Emanuelsson, O., Nielsen, H., von Heijne, G.: ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci.* **8(5)** (1999) 978-984

[13] Guo, J., Lin, Y., Sun, Z.: A novel method for protein subcellular localization: Combining residue-couple model and SVM. *Proceedings of the 3rd Asia-Pacific Bioinformatics Conference.* (2005) 117-129

[14] Horton, P., Nakai, K.: Better prediction of protein cellular localization sites with the k nearest neighbors classifier. *Proc. Intelligent Systems for Molecular Biology.* **5** (1997) 147-152

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

[16] Kim, J. K., Raghava, G. P. S., Kim, K. S., Bang, S. Y., Choi, S.: Prediction of subcellular localization of proteins using pairwise sequence alignment and support vector machine. *Proc. The 3rd Annual Conference of the Korean Society for Bioinformatics.* (2004) 158-166

[17] Leslie, C., Eskin, E., Noble, W. S.: The spectrum kernel: a string kernel for SVM protein classification. *Proc. Pacific Symposium on Biocomputing.* (2002) 564-575

[18] Matsuda, S., Vert, J.-P., Saigo, H., Ueda, N., Toh, H., and Akutsu, T.: A novel representation of protein sequences for prediction of subcellular location using support vector machines. *Protein Science.* **14** (2005) 2804-2813

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

[20] Nakai, K., Kanehisa, M.: A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics.* **14** (1992) 897-911

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

[22] 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** (2003) 1656-1663

[23] Pavlidis, P., Wapinski, I., Noble, W. S.: Support vector machine classification on the web. *Bioinformatics.* **20(4)** (2004) 586-587

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

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