

A Class of New Kernels Based on High-Scored Pairs of k -Peptides for SVMs and Its Application for Prediction of Protein Subcellular Localization

Zhengdeng Lei and Yang Dai*

Department of Bioengineering (MC063),
University of Illinois at Chicago,
851, South Morgan Street, Chicago, IL 60607, USA
{zlei2, yangdai}@uic.edu

Abstract. A class of new kernels has been developed for vectors derived from a coding scheme of the k -peptide composition for protein sequences. Each kernel defines the biological similarity for two mapped k -peptide coding vectors. The mapping transforms a k -peptide coding vector into a new vector based on a matrix formed by high BLOSUM scores associated with pairs of k -peptides. In conjunction with the use of support vector machines, the effectiveness of the new kernels is evaluated against the conventional coding scheme of k -peptide ($k \leq 3$) for the prediction of subcellular localizations of proteins in Gram-negative bacteria. It is demonstrated that the new method outperforms all the other methods in a 5-fold cross-validation.

Keywords: Protein subcellular localization, BLOSUM matrix, kernel, support vector machine, Gram-negative bacteria.

1 Introduction

Advances in genome sequencing and proteomics are generating enormous numbers of genes and proteins. Accordingly, the development of automated systems for the annotation of protein structure and function has become extremely important. Since many cellular functions are compartmentalized in specific regions of a cell, subcellular localization of a protein is biologically highlighted as a key element in understanding its function. Specific knowledge of the subcellular location can direct further experimental study of proteins.

Methods and systems have been developed during the last decade for the predictive task of protein localization. Machine learning methods such as Artificial Neural Networks, the k -nearest neighbor method, and Support Vector Machines (SVMs) have been utilized in conjunction with various methods of feature extraction for protein sequences. Most of the early approaches employed

* Corresponding author.

the amino acid and di-peptide compositions [7,12,27,28] to represent sequences. These methods may miss information on sequence order and inter-relationships among amino acids. In order to overcome these shortcomings, it has been shown that motifs, frequent-subsequences, functional domains, and other useful features, which are obtained from various databases (SMART, InterPro, PROSITE) or extracted using Hidden Markov Models, Fourier Transform, and other data mining techniques, can be used for the representation of protein sequences for the prediction of subcellular localizations [2,3,6,15,29,30]. Methods have also been developed based on the use of the N-terminal sorting signals [1,5,10,21,24,25,26] and sequence homology searching [23].

Most robust methods adopt an integrative approach by combining several methods, each of which may be a suitable predictor for a specific localization or a generic predictor for all localizations. PSORT is an example of such successful system. Developed by Nakai and Kanehisa [25], PSORT, recently upgraded to PSORT II [11,24], is an expert system that can distinguish between different subcellular localizations in eukaryotic cells. It also has a dedicated subsystem PSORT-B for bacterial sequences [8].

Several recent studies [19,31], however, have indicated that a predicting system based on the use of a generalized k -peptide composition or sequence homology could obtain similar or better performance compared to that of the integrated system PSORT-B. The outcome from our work supports these findings.

In this study, a new similarity measurement for protein sequences has been developed based on the use of high-scored pairs of k -peptides. It is the extension of the concept used in our previous work [16] for a fixed k value ($k = 3$). More specifically, each pair of k -peptides is assigned a score based on a BLOSUM matrix. A small portion of pairs with high scores is selected to retain their original scores in order to reduce noise and computational time. The remaining pairs are given zero scores. The reassigned score associated with each pair of k -peptides is then considered as an entry in a matrix D_k , which is named as the matrix of high-scored pairs of k -peptides. When $k = 1$, this matrix is the same as the BLOSUM matrix, except that the entries with negative values are replaced by zeroes. When $k \geq 2$, each entry is the BLOSUM score corresponding to a pair of k -peptides with negative value being replaced by zero. Each protein sequence is first coded by its k -peptide composition. Then each k -peptide coding vector \mathbf{x}_k is mapped to another vector $D_k \mathbf{x}_k$, and the similarity between the sequences is measured by those mapped vectors. That is, the kernel is defined based on these mapped vectors.

The new kernels combined with SVMs are evaluated against the conventional coding scheme of k -peptide ($k \leq 3$) composition for the prediction of subcellular localizations for proteins obtained from Gram-negative bacteria [8]. It is demonstrated by the result of a 5-fold cross-validation that the new kernel method outperforms significantly the coding methods based on the conventional k -peptide composition.

2 Method

This section introduces a new kernel for the coding vectors derived from the k -peptide compositions of protein sequences. This coding scheme based on the k -peptide composition for $k \leq 2$ has been used for the prediction of subcellular localizations [12,27,31], but has never been directly evaluated for $k = 3$. Below a short description of SVMs is presented.

2.1 Support Vector Machines

Suppose that a set of m training points \mathbf{x}_i ($1 \leq i \leq m$) in an n -dimensional space is given. Each point \mathbf{x}_i is labeled by $y_i \in \{1, -1\}$ denoting the membership of the point. An SVM is a learning method for binary classification. Using a nonlinear transformation ϕ , it maps the data to a high dimensional feature space in which a linear classification is performed. It is equivalent to solving the quadratic optimization problem:

$$\begin{aligned} \min_{\mathbf{w}, b, \xi_1, \dots, \xi_m} \quad & \frac{1}{2} \mathbf{w} \cdot \mathbf{w} + C \sum_{i=1}^m \xi_i \\ \text{subject to} \quad & y_i (\phi(\mathbf{x}_i) \cdot \mathbf{w} + b) \geq 1 - \xi_i \quad (i = 1, \dots, m), \\ & \xi_i \geq 0 \quad (i = 1, \dots, m), \end{aligned} \quad (1)$$

where C is a parameter. The decision function is defined as $f(\mathbf{x}) = \text{sign}(\phi(\mathbf{x}) \cdot \mathbf{w} + b)$, where $\mathbf{w} = \sum_{i=1}^m \alpha_i \phi(\mathbf{x}_i)$ and α_i ($i = 1, \dots, m$) are constants determined by the dual problem of the optimization defined above.

For any pair of mappings $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_j)$, the kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ is defined as a dot product of $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_j)$, i.e.,

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j). \quad (2)$$

The kernel function is essentially a measurement of similarity for the mapped points in terms of their inner products. The matrix $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ is called the kernel matrix. The decision function can be represented by using the kernel function:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^m \alpha_i \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}) + b\right) = \text{sign}\left(\sum_{i=1}^m \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b\right). \quad (3)$$

Typical kernel functions are, for example, polynomial kernel $(\mathbf{x}_i \cdot \mathbf{x}_j + a)^d$ ($d \geq 1$) and the radial basis kernel $\exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$. In most of these cases, the corresponding nonlinear mappings ϕ are not known explicitly, although their existence is guaranteed. For other details of SVMs refer to [4].

2.2 Sequence Coding Schemes and a Class of New Kernels Based on High-Scored Pairs of k -Peptides

The effectiveness of the coding schemes for protein sequences based on the k -peptide compositions or their variations has been demonstrated in the prediction

of subcellular localizations, combining with machine learning tools such as neural networks and support vector machines [12,23,27,31]. If $k = 1$, the k -peptide composition reduces to the amino acid composition, and if $k = 2$, the k -peptide composition gives the di-peptide composition. When k becomes larger, the k -peptide composition will encompass more global sequence information, but at the same time, such a coding scheme becomes less attractive from the computational viewpoint.

In order to code a sequence, a window with a length of k is moved along the sequence from the first amino acid to the k th amino acid from the end. Every k -letter pattern that appears in the window is recorded with an increment of 1 in the corresponding entry of the vector. The final vector is normalized by dividing the number of window positions associated with that sequence. Upon the termination of this procedure, the vector provides the k -peptide composition of the sequence. Since the symbol "X" may appear in some sequences, it is added to the set of the original 20 symbols of the amino acids to give a total of 21. Therefore, vectors of 21, $21^2 = 441$ and $21^3 = 9261$ dimensions are required, respectively, for $k = 1, 2$, and 3 in this coding scheme.

However, a more sensitive and biologically relevant coding method would allow some degree of mismatch of amino acids in the k -peptide representation for $k \neq 1$. That is, the similarity should be large if two sequences share many similar k -peptides. This idea has been explored by Leslie *et al.* [17] for protein homology detection, and a set of mismatch kernels was developed. In their paper, the coding vector represents the occurrence of the corresponding k -peptides and its mismatched peptides in a protein sequence. In our work, the concept of mismatch kernel is explored in an implicit and different way. The similarity of two k -peptides is measured by the sum of BLOSUM scores between two residues at the same position.

In order to define the new kernel, we introduce a matrix in which each entry corresponds to the pairwise score of two k -peptides. For example, the scores are 12 for an AAA-AAA pair, 11 for an AAY-ACY pair, and 6 for a TVW-TVR pair, if the BLOSUM62 matrix is used. Since the majority of all possible pairs is associated with lower scores, the elimination of those pairs can reduce noise that may confuse the prediction. In addition, this procedure also reduces training time. Accordingly, only a very small portion of the entries corresponding to high-scored pairs is kept given a proper threshold, and the other entries are replaced by 0 in the matrix. The resulting matrix is called *the matrix of high-scored pairs of k -peptides*, and is denoted as D_k . The new kernel $k(\cdot, \cdot)$ is then defined as

$$k(\mathbf{x}_k^i, \mathbf{x}_k^j) = \exp(-\gamma \|D_k \mathbf{x}_k^i - D_k \mathbf{x}_k^j\|^2) \quad (4)$$

for the radial basis functions; or

$$k(\mathbf{x}_k^i, \mathbf{x}_k^j) = (D_k \mathbf{x}_k^i \cdot D_k \mathbf{x}_k^j + a)^d, \quad d \geq 1 \quad (5)$$

for polynomial functions. Basically, the similarity is measured between the transformed vectors $D_k \mathbf{x}_k^i$ and $D_k \mathbf{x}_k^j$, instead of that between the original k -peptide coding vectors \mathbf{x}_k^i and \mathbf{x}_k^j .

The example in Fig. 1 describes the coding vectors obtained from the two methods for two short amino acids sequences AAACY and AACCY: \mathbf{x}_3^1 and \mathbf{x}_3^2 are based on the tri-peptide composition; and $D_3\mathbf{x}_3^1$ and $D_3\mathbf{x}_3^2$. For the tri-peptide composition, the vectors \mathbf{x}_3^1 and \mathbf{x}_3^2 share one common tri-peptide “AAC”, which is the entry 2 in the vectors. However, the transformed vectors $D_3\mathbf{x}_3^1$ and $D_3\mathbf{x}_3^2$ have many non-zero common entries, such as 2, 16, 23, 24, 26, 28, etc (see boldfaced numbers in Fig.1). This implies that the transformation can capture similarity even if the two sequences do not share many exactly matched tri-peptides.

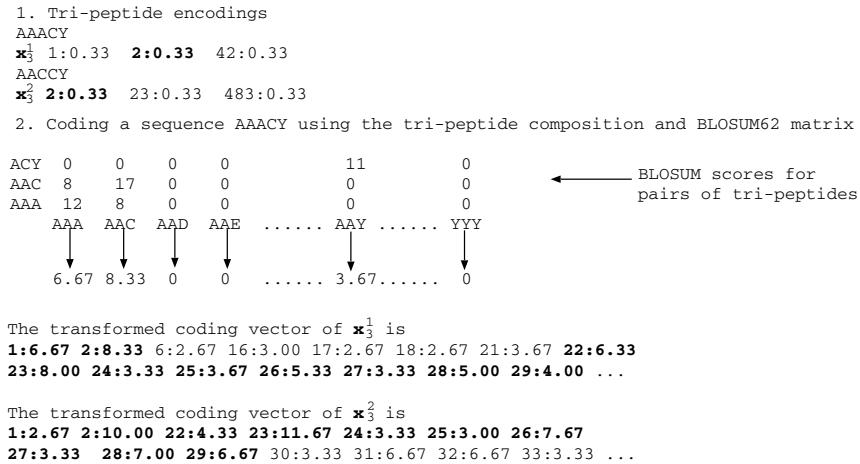


Fig. 1. The coding vectors for sequences AAACY and AACCY based on the tri-peptide composition and the transformed vectors based on high-scored pairs of tri-peptides. The representation of coding vectors follows the sparse format of SVMlight [14], i.e., the numbers appearing in the format of **vector index : score**. The shared elements between two sequences are boldfaced.

It is noted that the size of the matrix D_k for $k = 3$ is 9261×9261 . However, after score thresholding, very few non-zero entries in the matrix are kept. Therefore, the matrix is represented using a sparse data structure to ensure the efficiency of computation. The selection of the high-scored pairs of k -peptides is virtually filtering the k -peptides sharing more residues in common. In addition, the procedure also retains those pairs with high similarity BLOSUM scores between the residues.

3 Experimental Results and Discussion

In order to evaluate the performance of our new kernels on the prediction of protein subcellular localization for different values of $k = 1, 2, 3$, a set of proteins from Gram-negative bacteria was used. In addition, the computation with

the conventional k -peptide ($k = 1, 2, 3$) coding scheme was also performed for comparison.

3.1 Dataset

The set of proteins from Gram-negative bacteria used in the evaluation of PSORT-B [8] was considered (available at <http://www.psort.org/>) in this experiment. It consists of 1443 proteins with experimentally determined localizations. The dataset comprises 1302 proteins resident at a single localization site: 248 cytoplasmic, 268 inner membrane, 244 periplasmic, 352 outer membrane, and 190 extra cellular; it additionally contains a set of 141 proteins resident at multiple localization sites: 14 cytoplasmic/inner membrane, 50 inner membrane/periplasmic, and 77 outer membrane/extracellular. In our experiment, we considered only the 1302 proteins possessing a single localization.

3.2 Experiments and Results

The BLOSUM62 matrix was used for the assignment of scores to pairs of k -peptides. The threshold for high-scored pairs was 0 for $k = 1, 2$; and 8 for $k = 3$. The nonzero entries account for about 1.3% of the entries in matrix D_3 . In order to ease the computational burden, the 2000 top scored entries from a transformed vector $D_3\mathbf{x}_3$ were further selected to form the input vector for SVMs. The threshold 8 and the number 2000 were determined empirically from the preliminary study to ensure good performance and fast training.

The experiment was carried out with a 5-fold cross-validation (CV) for each specific localization. Each time, the relevant dataset consisting of the proteins with the specific localizations was designated as the positive set; and the remainder of the proteins was denoted as the negative set. The radial basis (4) and polynomial kernel (5) (degree ranged from 1 to 6) functions were used for the SVMs. Since the polynomial kernels did not generate good results, we only present the results obtained from the radial basis kernel.

As the sizes of the positive and negative sets are substantially different, the performance of SVMs was evaluated for precision, defined as $tp/(tp + fp)$; and recall, defined as $tp/(tp + fn)$, where tp , tn , fp , and fn are the numbers of predicted true positive, true negative, false positive and false negative, respectively. In addition, the F-score combining the precision and recall:

$$F\text{-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}, \quad (6)$$

was also evaluated. The reported values of precision, recall, and F-score are the averages from the 5-fold CVs.

The generalization performance of an SVM is controlled by the following parameters:

- (1) C : the trade-off between the training error and the class separation;
- (2) γ : the parameter in the radial basis function $\exp(-\gamma\|D_k\mathbf{x}_k^i - D\mathbf{x}_k^j\|^2)$;
- (3) J : the biased penalty for errors from positive and negative training points.

The penalty term $C\sum_{i=1}^m \xi_i$ in SVM is split into two terms [22]:

$$C\sum_{i=1}^m \xi_i \Rightarrow C\sum_{\{i:y_i=1\}} \xi_i + CJ\sum_{\{i:y_i=-1\}} \xi_i. \quad (7)$$

The choices of the parameters in this experiment are given as follows: for the new kernels,

- C : from 1 to 40 with an incremental size of 3;
 γ : from 0.001 to 1 with an incremental size of 0.003;
 J : from 0.1 to 3.0 with an incremental size of 0.4;

and for the conventional k -peptides compositions,

- C : from 1 to 150 with an incremental size of 10;
 γ : from 1 to 100 with an incremental size of 10;
 J : from 0.1 to 3.0 with an incremental size of 0.2.

The SVMLight package was used as the SVM solver [14]. The values of precision and recall of a 5-fold CV were computed for each triplet (C, γ, J) . The best values of precision, recall and the corresponding F-score for each method are reported. The symbols P, R and F used in Tables 1 and 2 stand respectively for precision, recall, and F-score.

From Table 1, it can be seen that the performance is sensitive to the value of k . With $k = 2$, the new kernel achieves the best performance in terms of precision, recall, and F-score. Specifically, the recall (85.73) is about 10% higher compared with that (75.76) obtained when $k = 3$, while maintaining a similar level of precision; the precision (90.07) is about 8% higher than that (81.93) obtained when $k = 1$; while keeping almost the same recall value.

The results of prediction with the conventional k -peptide composition scheme for the same data set are reported in Table 2. It is readily seen from the table that the three coding methods do not show significant difference in their performance, although the coding with composition ($k = 1$) achieves a slightly better level

Table 1. Results obtained from the new kernel method with different matrices for the proteins from Gram-negative bacteria

Method	D1			D2			D3		
	P	R	F	P	R	F	P	R	F
Cytoplasmic	76.74	87.05	81.46	88.12	84.53	86.24	77.38	73.48	75.38
Inner membrane	95.30	84.95	89.69	95.39	90.73	92.90	97.29	85.27	90.88
Periplasmic	76.43	79.69	77.88	80.44	82.55	81.36	85.98	68.45	76.22
Outer membrane	84.92	90.72	87.63	95.20	92.83	93.95	96.25	86.73	91.24
Extra cellular	76.26	83.73	79.73	91.22	78.00	83.85	92.11	64.86	76.12
Average	81.93	85.23	83.28	90.07	85.73	87.66	89.80	75.76	81.94

Table 2. Results obtained from the conventional k -peptide coding method for the proteins from Gram-negative bacteria

Method	composition			di-peptide			tri-peptide		
Localization	P	R	F	P	R	F	P	R	F
Cytoplasmic	80.09	70.77	74.66	81.12	57.69	66.09	83.43	45.00	55.09
Inner membrane	98.52	82.27	89.54	98.15	81.51	88.80	99.52	80.75	89.01
Periplasmic	94.12	55.17	68.38	91.80	54.14	65.77	90.37	50.34	63.11
Outer membrane	87.86	84.23	85.74	90.12	79.76	84.00	93.15	83.29	87.79
Extra cellular	88.38	53.68	66.05	89.71	53.68	66.27	92.57	50.53	64.63
Average	89.79	69.23	76.87	90.18	65.36	74.18	93.17	64.80	74.62

of recall. In this comparison it is clear the new kernel method demonstrates superior performance over the conventional k -peptide coding method. The recall (85.73) produced by the new method with $k = 2$ shows substantial improvement from 69.23 (composition), 65.36 (di-peptide), and 64.80 (tri-peptide); the F-score is likewise improved to a level of 87.66, from 76.87 (composition), 74.18 (di-peptide), and 74.62 (tri-peptide); while a similar level of precision is maintained.

The performance of the new kernel method also compares favorably with SCL-BLAST [23], a BLAST-search based predictor for all localizations. The new method improves recall from 60.40 to 85.73 and F-score from 74.36 to 87.66, while having a lower precision (90.07) compared to that (96.70) of SCL-BLAST.

It is worth noting that the new method ($k = 2$) yields a similar overall performance compared with the latest version of PSORT-B (v.2.0) [9], which gives a precision of 95.88, a recall of 82.6 and an F-score of 88.7. As the PSORT-B comprises several modules designed for the prediction of specific localization sites, it is surprising that our single module can match the performance of this integrative predictor.

4 Discussion

Kernel-based learning algorithms, such as SVMs, are among the most advanced machine learning methods. The success largely depends on the choice of kernel functions. In general, the more that prior knowledges is incorporated into the kernel function, the better the performance of the SVMs. Several successful approaches have focused on the design of new kernels reflecting higher levels of biological knowledge. This includes the mismatch kernel for protein fold recognition [17], the Fisher kernel for the detection of remote protein homologies [13], a class of edit kernels for the prediction of translation initiation sites in eukaryotic mRNAs [18], and an oligo kernel for the prediction of prokaryotic translation initiation sites [20]. The approach most relevant to our study is the mismatch kernel. In that work, each protein sequence is coded by a vector with each entry representing the number of occurrences of a k -peptide including its mismatched partners, namely, those that have a limited number of mutated amino acids in reference to the original k -peptide. Then, a linear kernel is essentially a weighted sum of numbers of shared mismatched k -peptides between two sequences. The class of new kernels proposed in this study can be considered as a generalization

of the mismatch kernel. The similarity between two k -peptides is measured not only by the number of mismatched residues, but also by the evolution distances between the residues based on their BLOSUM scores. This is concluded that these features are the basis of the improved performance of the new kernels that is revealed in the comparison with the conventional k -peptide coding scheme.

Although the class of the new kernels proposed in this study is general for any k -peptides, the implementation presents a particular difficulty when $k > 3$. This is why the experiments in this work were performed with $k \leq 3$. A clever data structure, such as the one used in [17], is needed for fast computation. This issue is currently under investigation.

5 Conclusions

This work has introduced a class of novel kernels based on matrices formed by the BLOSUM scores assigned to pairs of k -peptides of protein sequences. Through a linear mapping defined by the matrix, this method generalized the conventional k -peptide coding method to allow the measurement of similarity between mismatched k -peptides based on BLOSUM scores. The kernels have been used in support vector machines for the prediction of subcellular localizations. The performance of the new kernels was evaluated on a set of proteins with experimentally determined localizations from Gram-negative bacteria. Compared with other coding systems using k -peptide compositions, the experimental results demonstrate that the new kernel exhibited superior overall performance for the predictions. The method also achieved a similar level of overall performance comparing with that of the integrated system PSORT-B.

Acknowledgments

This research is partially supported by National Science Foundation (EIA-022-0301) and Naval Research Laboratory (N00173-03-1-G016). The authors are thankful for Deepa Vijayraghavan for the assistant with computing environment.

References

1. Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., Miyano, S.: Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics* **18** (2002) 298-305
2. Cai, Y.D., Chou, K.C.: Predicting subcellular localization of proteins in a hybridization space. *Bioinformatics* **20** (2003) 1151-1156
3. 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) 45765-4576
4. Cristianini, N., Shawe-Taylor, J.: *An Introduction to Support Vector Machines*, Cambridge University Press (2000)
5. 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

6. Emanuelsson, O.: Predicting protein subcellular localisation from amino acid sequence information. *Brief. Bioinform.* **3** (2002) 361-376
7. Feng, Z.P.: Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. *Biopolymers* **58** (2001) 491-99
8. Gardy, J.L. *et al.*: PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res.* **31** (2003) 3613-3617
9. Gardy, J.L. *et al.*: PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* **21** (2005) 617-623
10. von Heijne, G.: Signals for protein targeting into and across membranes. *Subcell. Biochem.* **22** (1994) 1-19
11. Horton, P., Nakai, K.: PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* **24** (1999) 34-36
12. Hua, S., Sun, Z.: Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* **17** (2001) 721-728
13. Jaakkola, T., Diekhans, M., Haussler, D.: Using the Fisher kernel method to detect remote protein homologies. *Proc. of the Seventh International Conference on Intelligent Systems for Molecular Biology* (1999) 149 - 158
14. Joachims, T.: Making Large Scale SVM Learning Practical. *Advances in Kernel Methods-Support Vector Learning*. MIT Press, Cambridge (1999)
15. Lei, Z, Dai, Y.: A novel approach for prediction of protein subcellular localization from sequence using Fourier analysis and support vector machines. *Proc. of the Fourth ACM SIGKDD Workshop on Data Mining in Bioinformatics* (2004) 11-17
16. Lei, Z, Dai, Y.: A new kernel based on high-scored pairs of tri-peptides and its application in prediction of protein subcellular localization. *Proc. of International Conference on Computational Science (ICCS 2005)*, LNCS **3515** (2005) 903-910
17. Leslie, C., Eskin, E., Cohen, A., Weston, J., Noble, W.: Mismatch string kernels for discriminative protein classification. *Bioinformatics* **20** (2004) 467-476
18. Li, H., Jiang, T.: A class of edit kernels for SVMs to predict translation initiation sites in eukaryotic mRNAs. *Proc. of the Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB)* (2004) 262-271
19. Lu, Z., Szafron, D., Greiner, R., Lu, P., Wishart, D.S, Poulin, B., Anvik, J., Macdonell, C., Eisner, R.: Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics* **20** (2004) 547-556
20. Meinicke, P., Tech, M., Morgenstern, B., Merkl, R.: Oligo kernels for datamining on biological sequences: a case study on prokaryotic translation initiation sites. *BMC Bioinformatics* **5** (2004) 169
21. Menne, K. M. L., Hermjakob, H., Apweiler, R.: A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics* **16** (2000) 741-742
22. Morik, K., Brockhausen, P., Joachims, T.: Combining statistical learning with a knowledge-based approach - A case study in intensive care monitoring. *Proc. of the Sixteenth International Conference on Machine Learning* (1999) 268-277
23. Nair, R., Rost, B.: Sequence conserved for subcellular localization. *Protein Sci.* **11** (2002) 2836-2847
24. Nakai, K.: Protein sorting signals and prediction of subcellular localization. *Adv. Protein. Chem.* **54** (2000) 277-344
25. Nakai, K., Kanehisa, M.: Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins* **11** (1991) 95-110

26. Nielsen, H., Engelbrecht, J., Brunak, S., von Heijne, G.: A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural Syst.* **8** (1997) 581-599
27. Park, K., 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
28. Reinhardt, A., Hubbard, T.: Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.* **26** (1998) 2230-2236
29. Tusnady, G.E., Simon, I.: Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.* **283** (1998) 489-506
30. Tusnady, G.E., Simon, I.: The HMMTOP transmembrane topology prediction server. *Bioinformatics* **17** (2001) 849-850
31. Yu, C.S., Lin, C.J., Hwang, J.K.: Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on *n*-peptide compositions. *Protein Sci.* **13** (2004) 1402-1406