

A New Kernel Based on High-Scored Pairs of Tri-peptides and Its Application in 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 new kernel has been developed for vectors derived from a coding scheme of the tri-peptide composition for protein sequences. This kernel defines the sequence similarity through a mapping that transforms a tri-peptide coding vector into a new vector based on a matrix formed by the high BLOSUM scores associated with pairs of tri-peptides. In conjunction with the use of support vector machines, the effectiveness of the new kernel is evaluated against the conventional coding schemes 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, Gram-negative bacteria, BLOSUM matrix, kernel, support vector machine.

1 Introduction

Advances in proteomics and genome sequencing are generating enormous numbers of genes and proteins. 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 the cell, subcellular localization of a protein is biologically highlighted as a key element in understanding its function. Specific knowledge of 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 (SVM) have been utilized in conjunction with various methods of feature

* 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.

** Corresponding author.

extraction from protein sequences. Most of the early approaches employed the amino acid composition and the di-peptide composition [7, 11, 20] to represent sequences. This method may miss the information on sequence order and the inter-relationships between the amino acids. In order to overcome this shortcoming, it has been shown that motifs, frequent-subsequences, and functional domains, which are obtained from various databases (SMART, InterPro, PROSITE) or extracted using Hidden Markov Models and data mining techniques, can be used for the representation of protein sequences for the prediction of subcellular localizations [2, 3, 6, 21, 22]. Methods have also been developed based on the use of the N-terminal sorting signals [1, 5, 9, 15, 17, 18, 19] and sequence homology searching [16].

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 a successful system. Developed by Nakai and Kanehisa [18], PSORT, recently upgraded to PSORT II [10, 17], 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 [14, 23], however, have indicated that a predicting system based on the use of generalized k -peptide compositions or sequence homology could obtain similar or better performance compared to that of the integrated system PSORT-B. The outcome from our work also supports these findings.

In this study, a new similarity measurement for protein sequences has been developed based on the use of a matrix derived from high-scored pairs of tri-peptides. Each protein sequence is first coded by its tri-peptide composition. Since the repeating of the same tri-peptide is relatively lower comparing to that of di-peptides, the tri-peptide coding is more faithful in retaining the order of amino acids. Each pair of tri-peptides is then assigned with 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 the computational time. The rest of pairs are given zero scores. The reassigned score associated with each pair of tri-peptides is then considered as an entry of an imaginary matrix D , which is named as the matrix of high-scored pairs of tri-peptides. It is obvious that pairs with more than two amino acids in common or sharing residues with high BLOSUM scores usually receive higher scores. Then each tri-peptide coding vector \mathbf{x} is mapped to another vector $D\mathbf{x}$, 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 method is evaluated against the coding schemes of k -peptide ($k \leq 3$) compositions 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 method outperforms the coding methods based on the k -peptide compositions.

2 Method

This section introduces the new kernel for the coding vectors derived from the tri-peptide compositions for protein sequences. The coding scheme based on the tri-peptide composition has been used in protein fold recognition [13], but has never been evaluated for the prediction of subcellular localizations. First a short description of support vector machines, the machine learning method used in this study will be 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:

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

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. Define a dot product $k(\mathbf{x}_i, \mathbf{x}_j)$ for any pair of mappings $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_j)$, i.e., $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$. This is called kernel function. The matrix $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ is called kernel matrix. The decision function can be represented as $f(\mathbf{x}) = \text{sign}(\sum_{i=1}^m \alpha_i \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}) + b) = \text{sign}(\sum_{i=1}^m \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b)$. The typical kernel functions are, for example, polynomial kernel $(\mathbf{x}_i \cdot \mathbf{x}_j)^d$ ($d \geq 1$) and Gaussian kernel $\exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$. For other details of SVMs refer to [4].

2.2 Sequence Coding Schemes and a New Kernel Based on High-Scored Pairs of Tri-Peptides

The coding schemes of protein sequences based on k -peptide compositions or their variations have been demonstrated effective in the predictions of protein folds and subcellular localizations, in conjunction with the use of machine learning tools such as neural networks and support vector machines [16, 23]. If $k = 1$, then 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 compositions will cover more global sequence information, but at the same time, such a coding scheme becomes less attractive from a 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 increment of 1 in the corresponding entry of the vector. Upon the termination of this procedure, the vector provides the k -peptide composition of the sequence. The final vector is normalized by dividing the number of window positions associated with that 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. Each entry of the vector is associated with a possible permutation of k amino acids.

Since there is only a small collection of the possible 3-letter patterns appearing in each protein sequence, the dot product $(\mathbf{x}_i \cdot \mathbf{x}_j)$ in the linear kernel for the tri-peptide composition calculates a value proportional to the number of tri-peptides coincide in two sequences \mathbf{x}_i and \mathbf{x}_j . The efficacy of the tri-peptide coding scheme in prediction of protein folds and subcellular localization is essentially due to the successful capture of local similarity by the coding scheme.

However, a more sensitive and biologically realistic coding method would allow some degree of mismatching in the tri-peptide representation. That is, the similarity should be large if the two sequences share many similar tri-peptides. This idea has been proposed and explored by Leslie *et al.* [13] for protein homology detection, and a set of spectral kernels was developed. In this work, the concept of mismatch kernel is explored in an implicit and different way. Here the discussion is restricted to case $k = 3$ for the simplicity of presentation, but the idea can be generalized to cases $k > 3$. In order to define the new kernel, we introduce a matrix in which each entry corresponds to the pairwise score of any two tri-peptides. For example, 12 for AAA-AAA pair, 11 for AAY-ACY pair, and 6 for TVW-TVR pair, if the BLOSUM62 matrix is used. The size of the matrix is 9261×9261 , however, the matrix is only for the description and is never explicitly used in computation. Since majority of these pairs are associated with lower scores, the elimination of those pairs can reduce noise that may hinder the prediction. In addition, this also reduces training time. Accordingly, only a very small portion of the entries corresponding to high-scored pairs are kept, and other entries are replaced by 0 in the matrix. The matrix is called *the matrix of high-scored pairs of tri-peptides*, and is denoted as D . The new kernel $k(\cdot, \cdot)$ is defined as

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|D\mathbf{x}_i - D\mathbf{x}_j\|^2).$$

The term $\|D\mathbf{x}_i - D\mathbf{x}_j\|^2$ can be considered as a new sequence similarity for the two coding vectors of tri-peptide compositions. The similarity is measured between the transformed vectors $D\mathbf{x}_i$ and $D\mathbf{x}_j$, instead of the similarity between the original tri-peptide coding vectors \mathbf{x}_i and \mathbf{x}_j . The example in Fig. 1 describes the coding vectors \mathbf{x}_1 and \mathbf{x}_2 based on tri-peptide compositions and the transformed vectors $D\mathbf{x}_1$ and $D\mathbf{x}_2$ for the two short sequences of amino acids AAACY and ADCCY.

The selection of the high-scored pairs of tri-peptides is virtually filtering the tri-peptides sharing more than two residues in common. The concept of the mis-

Tri-peptide encodings

AAACY
x1 1:1 2:1 42:1
ADCCY
x2 44:1 483:1 905:1

Coding a sequence AAACY using the tri-peptide composition and BLOSUM62 matrix

ACY	0	0	0	0	11	0		
AAC	8	17	0	0	0	0	← BLOSUM scores for	
AAA	12	8	0	0	0	0	pairs of tri-peptides	
	AAA	AAC	AAD	AAE	AAY	YYY
	↓	↓	↓	↓		↓		↓
	6.67	8.33	0	0	3.67.....		0

The transformed coding vector of x1 is
1:6.67 **2:8.33** 6:2.67 **16:3.00** 17:2.67 18:2.67 21:3.67 22:6.33
23:8.00 **24:3.33** 25:3.67 **26:5.33** 27:3.33 **28:5.00** 29:4.00...

The transformed coding vector of x2 is
2:3.67 3:4.67 4:3.33 12:3.00 14:2.67 **16:2.67** **23:3.33** **24:3.33**
26:3.33 **28:3.00** 40:3.00 42:4.67 43:3.33 44:6.33 45:2.67 47:2.67...

Fig. 1. The coding vectors for sequences AAACY and ADCCY based on the tri-peptide compositions and the transformed vectors based on the matrix of high-scored pairs of tri-peptides. The representation of coding vectors follows the sparse format of SVM-Light [12], i.e., the numbers appeared in the format of **vector index : score**. The shared elements between x_1 and x_2 are boldfaced

match string is explored, since only those mis-matched tri-peptides can yield high scores and survive the selection.

3 Experimental Results and Discussion

We employed the SVMs in conjunction with the coding vectors extracted by the method described above for training and testing. The evaluation of the methods was conducted on the following dataset.

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.

Experiment and Results. We have compared the performance of the new kernel with that of the coding schemes based on the conventional k -peptide compositions using the above data set. The pairs of tri-peptides with scores

greater than a cut-off value 8 were selected to form the nonzero entries the matrix D of high-scored pairs of tri-peptides. This accounts for about 1.3% of the entries in matrix D . To ease the computational burden, the 2000 top scored entries from a transformed vector were further selected to form the input vector for SVMs. The cut-off value 8 and the number 2000 were determined empirically from the preliminary study to achieve the good predicting performance and fast training. The BLOSUM62 matrix was used for the assignment of scores to pairs of tri-peptides.

The experiment was carried out using a 5-fold cross-validation 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 function was chosen as the kernel function for the SVMs, since a preliminary experiment indicated that this kernel exhibited better performance.

As the sizes of the positive and negative sets are substantially different, the performance of SVMs was evaluated for precision (or sensitivity), defined as $tp/(tp + fp)$, and recall (or positive prediction value), 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. The F-score combining the precision and recall was also provided: $F\text{-score} = 2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$. The precision, recall, and F-score of the 5-fold cross-validation was computed respectively as the average of the values from 5 folds.

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

- (1) C : the trade-off between the training error and class separation;
- (2) γ : the parameter in the radial basis function $\exp(-\gamma \|D\mathbf{x}_i - D\mathbf{x}_j\|^2)$;
- (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 splitted into two terms:

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

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

- 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 rest of the methods

- 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 [12]. The values of precision and recall of a 5-fold cross-validation are computed for each triplet (C, γ, J) . The best values of precision, recall and the corresponding F-score for

Table 1. Results obtained from four different methods for the proteins from Gram-negative bacteria

Method	composition			di-peptide			tri-peptide			new method		
Localization	P	R	F	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	77.38	73.48	75.38
Inner membrane	98.52	82.27	89.54	98.15	81.51	88.80	99.52	80.75	89.01	97.29	85.27	90.88
Periplasmic	94.12	55.17	68.38	91.80	54.14	65.77	90.37	50.34	63.11	85.98	68.45	76.22
Outer membrane	87.86	84.23	85.74	90.12	79.76	84.00	93.15	83.29	87.79	96.25	86.73	91.24
Extra cellular	88.38	53.68	66.05	89.71	53.68	66.27	92.57	50.53	64.63	92.11	64.86	76.12
Average	89.79	69.23	76.87	90.18	65.36	74.18	93.17	64.80	74.62	89.80	75.76	81.97

Composition, di-peptide and tri-peptide represents the method using the coding vector of the amino acid composition, di-peptide composition and tri-peptide composition, respectively. The symbols P, R and F stand respectively for precision, recall and F-score.

each method are given in Table 1. The new kernel based method demonstrated superior performance over the other three methods. The recall is improved substantially to a level of 75.76, from 69.23 (Composition), 65.36 (di-peptide), and 64.80 (tri-peptide).

The performance of the new kernel method also compares favorably with SCL-BLAST [16], a BLAST search based predictor for all localizations. The new method improves recall from 60.40 to 75.76 and F-score from 74.36 to 81.97, while having a lower precision (89.80) compared to that 96.70 of *SCL-BLAST*.

It is worth noting that the new method yields a similar overall performance comparing with PSORT-B, which gives precision 95.88, recall 73.20 and F-score 82.59. The PSORT-B comprises six 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 Conclusions

This work has introduced a novel kernel based on a matrix formed by the BLOSUM scores assigned to pairs of mis-matched tri-peptides of protein sequences. This kernel has been used in support vector machines for the prediction of subcellular localizations. The performance of the new kernel was empirically evaluated on a set of proteins with experimentally determined localizations from Gram-negative bacteria. Compared with the other coding systems using k -peptide compositions, the experimental results demonstrated that the new kernel exhibited superior overall performance for the prediction. The method also achieved a similar level of overall performance comparing with that of the integrated system PSORT-B.

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) 491–499
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. von Heijne, G.: Signals for protein targeting into and across membranes. *Subcell. Biochem.* **22** (1994) 1–19
10. 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
11. Hua, S., Sun, Z.: Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* **17** (2001) 721–728
12. Joachims, T.: *Making Large Scale SVM Learning Practical*. Advances in Kernel Methods-Support vector learning. MIT Press, Cambridge (1999)
13. Leslie, C., Eskin, E., Cohen, A., Weston, J., Noble, W.: Mismatch string kernels for discriminative protein classification. *Bioinformatics* **20** (2004) 467–476
14. 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
15. 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
16. Nair, R., Rost, B.: Sequence conserved for subcellular localization. *Protein Sci.* **11** (2002) 2836–2847
17. Nakai, K. (2000) Protein sorting signals and prediction of subcellular localization. *Adv. Protein. Chem.*, **54**, 277-344
18. Nakai, K., Kanehisa, M.: Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins*, **11** (1991) 95–110
19. 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
20. Reinhardt, A., Hubbard, T.: Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.* **26** (1998) 2230–2236
21. 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
22. Tusnady, G.E., Simon, I.: The HMMTOP transmembrane topology prediction server. *Bioinformatics* **17** (2001) 849–850
23. 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