

# Hybrid SVM kernels for protein secondary structure prediction

Gulsah Altun, Hae-Jin Hu, Dumitru Brinza, Robert W. Harrison, Alex Zelikovsky, Yi Pan

**Abstract**—The Support Vector Machine is a powerful methodology for solving problems in nonlinear classification, function estimation and density estimation. When data are not linearly separable, data are mapped to a high dimensional feature space using a nonlinear function which can be computed through a positive definite kernel in the input space. Using a suitable kernel function for a particular problem and input data can change the prediction results remarkably and improve the accuracy. The goal of this work is to find the best kernel functions that can be applied to different types of data and problems. In this paper, we propose two hybrid kernels  $SVM_{SM+RBF}$  and  $SVM_{EDIT+RBF}$ .  $SVM_{SM+RBF}$  is designed by combining the best performed RBF kernel with substitution matrix (SM) based kernel developed by Vanschoenwinkel and Manderick.  $SVM_{EDIT+RBF}$  kernel combines the RBF kernel and the edit kernel devised by Li and Jiang. We tested these two hybrid kernels on one of the widely studied problems in bioinformatics which is the protein secondary structure prediction problem. For the protein secondary structure problem, our results were 91% accuracy on H/E binary classifier.

**Index Terms**— kernel method, protein secondary structure prediction, support vector machine.

## I. INTRODUCTION

MANY methods are proposed for classification problems in bioinformatics and support vector machines (SVM) is one of them that has attracted a lot of attention recently [5][6][7]. SVM is a powerful methodology for solving problems in nonlinear classification, function estimation and density estimation. When data are not linearly separable, data are mapped to a high dimensional feature space using a nonlinear function, which can be computed through a positive definite kernel in the input space. Different kernel functions can change the prediction results remarkably. The goal of this work is to find the best kernel function that can be applied to different types of problems and application domains. We propose two hybrid kernels  $SVM_{SM+RBF}$  and  $SVM_{EDIT+RBF}$  [19].  $SVM_{SM+RBF}$  is designed by combining the best performed radial basis function (RBF) kernel with substitution matrix (SM) based kernel developed by Vanschoenwinkel and Manderick.  $SVM_{EDIT+RBF}$  is designed by

combining the edit kernel devised by Li and Jiang [14] with the RBF kernel.

## II. BACKGROUND

### A. Protein Structure Prediction

Proteins are linear polymers formed by linking amino acids with a peptide bond connecting the carboxylic acid from one amino acid with the amino group of the next. The relative conformations that the polymer of amino acids or peptide chain can adopt are restricted by the physical characteristics of the peptide bond and the steric interactions between the atoms of the amino acids. Protein structure can be described by a hierarchy of terms: primary structure refers to the amino acid sequence; secondary structure to the pattern of peptide backbone conformations; tertiary structure to the folding of the peptide chain into compact domains; and quaternary structure to the association of these domains into higher order structures. This paper studies the prediction of secondary structure as a first step in the prediction of protein structure from amino acid sequence data.

The prediction of secondary structure has been proposed as an intermediate step in the prediction of the full three-dimensional tertiary structure because the factors that control it are thought to be simpler than those that control folding. Patterns of secondary structure may also give clues to the structural class of a protein from sequence data.

Recently, there have been many approaches to reveal the protein secondary structure from the primary sequence information [3][15][16][17][18]. Among those, machine learning approaches such as neural networks or support vector machines have shown successful results [9][10][11]. In recent years, support vector machine became the common machine learning tool for structure prediction based on its outstanding features such as effective avoidance of over-fitting, the ability to handle large feature spaces, information condensing of the given data set [4][10].

### B. Support Vector Machines

SVM is a modern learning system designed by Vapnik and his colleagues [20]. Based on statistical learning theory which explains the learning process from a statistical point of view, the SVM algorithm creates a hyperplane that separates the data into two classes with the maximum margin. Originally it was a linear classifier based on the optimal hyperplane algorithm. However, by applying the kernel method to maximum-margin hyperplane, Vapnik and his colleagues

proposed a method to build a non-linear classifier. In 1995, Cortes and Vapnik suggested a soft margin classifier which is a modified maximum margin classifier that allows for misclassified data. If there is no hyperplane that can separate the data into two classes, the soft margin classifier selects a hyperplane that separates the data as cleanly as possible with maximum margin.

SVM learning is related to recognize the pattern from the training data [6]. Namely, we estimate a function  $f: \mathbb{R}^N \rightarrow \{\pm 1\}$ , based on the training data which have  $N$ -dimensional pattern  $x_i$  and class labels  $y_i$ . By imposing the restriction called structural risk minimization (SRM) on this function, it will correctly classify the new data  $(x, y)$  which has the same probability distribution  $P(x, y)$  as the training data. SRM is to find the learning machine that yields a good trade-off between low empirical risk (mean error over the training data) and small capacity (a set of functions that can be implemented by the learning machine).

In the linear soft margin SVM which allows some misclassified points, the optimal hyperplane can be found by solving the following constrained quadratic optimization problem.

$$\min_{w, b, \varepsilon} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \varepsilon_i \quad (1)$$

$$s.t. \quad y_i (w \bullet x_i + b) \geq 1 - \varepsilon_i \quad \varepsilon_i > 0 \quad i=1, \dots, l$$

Where,  $x_i$  is an input vector,  $y_i = +1$  or  $-1$  based on whether  $x_i$  is in positive class or negative class, ' $l$ ' is the number of training data, ' $w$ ' is a weight vector perpendicular to the hyperplane and ' $b$ ' is a bias which moves the hyperplane parallel to itself. Also ' $C$ ' is a cost factor (penalty for misclassified data) and  $\varepsilon$  is a slack variable for misclassified points. The resulting hyperplane decision function is

$$f(x) = \text{sign} \left( \sum_{i=1}^{SV} \alpha_i y_i (x \bullet x_i) + b \right) \quad (2)$$

where,  $\alpha_i$  is a Lagrange multiplier for each training data. The points  $\alpha_i > 0$  are lie on the boundary of the hyperplane and are called 'support vectors'. In Eq. (1) and (2), it is observed that both the optimization problem and the decision function rely on the dot products between each pattern.

In the non-linear SVM, the algorithm first map the data into high-dimensional feature space ( $F$ ) via kernel function  $\phi(\bullet): X \rightarrow F$  and construct the optimal separating hyperplane there using the linear algorithm. According to Mercer's theorem, any symmetric positive definite matrix can be regarded as a kernel function. The positive definite kernel is defined as follows [6]:

Definition 1. Let  $X$  be a nonempty set. A function  $k(\bullet, \bullet):$

$X \times X \rightarrow \mathbb{R}$  is called a positive definite kernel if  $k(\bullet, \bullet)$  is symmetric and for all  $n \in \mathbb{N}$ ,  $x_1, \dots, x_n \in X$  and  $a_1, \dots, a_n \in \mathbb{R}$ .

The traditional positive definite kernel functions are the following:

$$K(x, y) = (x \bullet y + 1)^p \quad (3)$$

$$K(x, y) = e^{-\gamma \|x - y\|^2} \quad (4)$$

$$K(x, y) = \tanh(kx \bullet y - \delta) \quad (5)$$

Eq. (3) is a polynomial, Eq. (4) is a Gaussian radial basis

function (RBF), and Eq. (5) is a two-layer sigmoidal neural network kernel. Based on one of the above kernel functions, the final non-linear decision function has the form

$$f(x) = \text{sign} \left( \sum_{i=1}^{SV} \alpha_i y_i K(x \bullet x_i) + b \right) \quad (6)$$

The choice of proper kernel is critical to the success of the SVM. In the previous protein secondary structure prediction studies, a radial basis function worked best [9][10]. Therefore, this work combines different distance measures in the Gaussian kernel in equation 4.

### III. METHODS

In our approach, two hybrid kernels are devised by combining the best performed RBF kernel with substitution matrix (SM) based kernel [19] and with edit kernel [14].

#### A. Hybrid kernel: $SVM_{SM+RBF}$

The SM based kernel is developed by Vanschoenwinkel and Manderick [19]. The authors introduced a pseudo inner product (PI) between amino acid sequences based on the Blosum62 substitution matrix values [8]. PI is defined in [19] as follows:

Definition 2. Let  $M$  be a  $20 \times 20$  symmetric substitution matrix with entries  $M(a_i, a_j) = m_{ij}$  where  $a_i, a_j$  are components of the 20-tuple  $A = (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y) = (a_1, \dots, a_{20})$ . Then for two amino acid sequences  $x, x' \in \sum^n$  with  $x = (a_{i1}, \dots, a_{in})$  and  $x' = (a_{j1}, \dots, a_{jn})$ , with  $a_{ik}, a_{jk} \in A$ ,  $i, j \in \{1, \dots, 20\}$  and  $k = 1, \dots, n$ , their inner product is defined as:

$$\langle x | x' \rangle = \sum_{k=1}^n M(a_{ik}, a_{jk}) \quad (7)$$

Based on the PI above, substitution matrix based distance function between amino acid sequences is defined in [19] as follows:

Definition 3. Let  $x, x' \in \sum^n$  be two amino acid sequences with  $x = (a_{i1}, \dots, a_{in})$  and  $x' = (a_{j1}, \dots, a_{jn})$  and let  $\langle x | x' \rangle$  be the inner product as defined in (7) [19], then the substitution distance  $d_{\text{sub}}$  between  $x$  and  $x'$  is defined as:

$$d_{\text{sub}}(x, x') = \sqrt{\langle x | x \rangle - 2 \langle x | x' \rangle + \langle x' | x' \rangle} \quad (8)$$

In our approach, we combined the SM kernel with the RBF kernel. An example and the algorithm of  $SVM_{SM+RBF}$  is given in Fig. 1, which shows how we used a sequence segment when it is used in the hybrid kernel for finding the distances with different kernel functions.

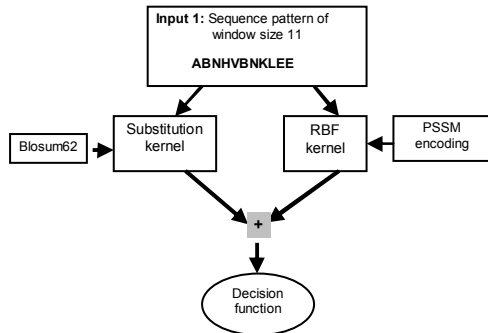


Fig 1. SVM<sub>SM+RBF</sub> algorithm

The data encoding given to the SVM<sub>SM+RBF</sub> is as shown in detail in Fig. 1. The data input for each sequence is the position specific scoring matrix (PSSM) [1][2][13] encoding of the sequence and the sequence combined together. The same data encoding is used for SVM<sub>EDIT+RBF</sub>.

### B. Hybrid kernel: SVM<sub>EDIT+RBF</sub>

The edit kernel is devised by Li and Jiang [14] to predict translation initiation sites in Eukaryotic mRNAs with SVM. It is based on string edit distance which contains biological and probabilistic information. The edit distance is the minimum number of edit operations (insertion, deletion, and substitution) that transform one sequence to the other. These edit operations can be considered as a series of evolutionary events. In nature, the evolutionary events happen with different probabilities. The authors [14] defined the edit kernel as follows:

$$K(x, y) = e^{-\gamma \cdot \text{edit}(x, y)} \quad (9)$$

$$\text{edit}(x, y) = -\frac{1}{2} \left( \sum_i \log P(x_i | y_i) + \sum_i \log P(y_i | x_i) \right) \quad (10)$$

where edit distance is the average of the negative log probability of mutating  $x$  into  $y$  and that of mutating  $y$  into  $x$ . The authors modified the 1-PAM matrix to get the asymmetric substitution cost matrix (SCM) for the edit kernel above. In our approach, we combined the edit kernel with the RBF kernel. An example of SVM<sub>EDIT+RBF</sub> is given in Fig. 2, which shows how a sequence segment is used in the hybrid kernel for finding the distances.

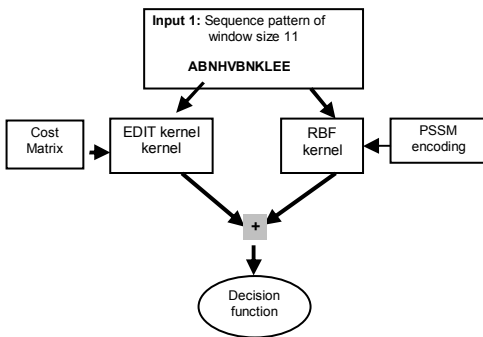


Fig 2. SVM<sub>EDIT+RBF</sub> algorithm

## IV. EXPERIMENTAL RESULTS

### A. Hybrid kernels applied to Protein Secondary Structure Prediction problem

The dataset used in this work includes 126 protein sequences obtained from Rost and Sander [18]. The sliding windows with eleven successive residues are generated from protein sequences. Each window is represented by a vector of 20x11. Twenty represents 20 amino acids and eleven represents each position of the sliding window. In Table 1, we show the results of the binary classifiers of the 6-fold cross validation test for the protein secondary structure prediction. SVM<sub>freq</sub> are from Hua and Sun [10] and the SVM<sub>psi</sub> results are obtained by PSI-BLAST profiles from Kim and Park [11]. SVM<sub>RBF</sub> is the profile which adopts the PSSM by Hu (2004) [9]. As Hu's result show, since PSSM encoding achieves the best results in the previous studies, we adopted the PSSM encoding scheme for the RBF kernel part of our hybrid kernel approaches.

TABLE I  
6-FOLD CROSS VALIDATION OF THE BINARY CLASSIFIERS

Binary	RS126		
Classifier	SVM <sub>freq</sub>	SVM <sub>psi</sub>	SVM <sub>RBF</sub>
H/~H	80.4	87.5	87.4
E/~E	81.3	86.3	86.8
C/~C	73.2	77.9	77.5
H/E	80.9	90.2	91.1
E/C	76.7	81.9	82.4
C/H	77.6	85.0	85.1

In Table 2, 6-fold cross validation results of the binary classifiers obtained by using different kernels in SVM are shown. The hybrid SVM method SVM<sub>SM+RBF</sub> proposed in this paper shows results that are almost similar as SVM<sub>RBF</sub>. This is because the data encoded for the RBF part in SVM<sub>SM+RBF</sub> uses PSSM encoding which is same as in SVM<sub>RBF</sub>. These results indicate combining SM to the RBF kernel can not improve the accuracy of RBF kernel used alone. This means that the additional distance information from SM part was not helpful to make the final decision. As alternatives, instead of adding the distance functions together, we have also tried different approaches such as taking the maximum of the two distances returned by the two kernels, or giving different weight to each distance before sending it to the decision function. However, all these methods gave the similar or worse results than those obtained by just adding the distance functions together. SVM<sub>EDIT+RBF</sub> could not achieve the results that SVM<sub>SM+RBF</sub> achieved. This suggests that for the protein secondary structure problem, SVM<sub>SM+RBF</sub> is a more suitable kernel.

TABLE 2  
COMPARISON RESULTS FOR THE BINARY CLASSIFIERS

Binary Classifier	RS126				
	SVM <sub>RBF</sub>	SVM <sub>SM</sub>	SVM <sub>EDIT</sub>	SVM <sub>SM+RBF</sub>	SVM <sub>EDIT+RBF</sub>
H/~H	87.4	75.18	68.2	87.4	74.0
E/~E	88.2	78.44	40.0	86.8	76.7
C/~C	79.4	69.83	52.5	77.9	64.0
H/E	91.7	73.32	48.8	91.0	79.2
E/C	83.6	75.36	41.8	82.5	71.8
C/H	85.3	73.48	48.9	85.0	71.1

## V. CONCLUSION

In this paper, we propose two hybrid kernels SVM<sub>SM+RBF</sub> and SVM<sub>EDIT+RBF</sub>. We tested these two hybrid kernels on one of the widely studied problems in bioinformatics, which is the protein secondary structure prediction problem. For the protein secondary structure problem, our results were 91% accuracy on H/E binary classifier. In this case, the information in the substitution matrix reinforces the information in the RBF on PSSM profiles. However, this is not true with the edit distance. These results show us that the data are consistent when substitution matrix is used and not consistent when edit distance is used. The edit distance kernel gives good results in [14], but not when used with our dataset in this work. Our results show that it is critically important to use mutually consistent data when merging different distance measures in support vector machines.

## ACKNOWLEDGMENT

The authors would like to thank Prof. T. Joachims for making SVMlight software available and Rost and Sander for providing the RS126 dataset. This research was supported in part by the U.S. National Institutes of Health under grants R01 GM34766-17S1, and P20 GM065762-01A1, and the U.S. National Science Foundation under grants ECS-0196569, and ECS-0334813. This work was also supported by the Georgia Cancer Coalition (GCC) and the Georgia Research Alliance. Dumitru Brinza and Hae-Jin Hu are supported by Georgia State University Molecular Basis of Disease Fellowship. Dr. Harrison is a GCC distinguished scholar.

## REFERENCES

- [1] F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, no.17, pp. 3389-3402, Sep 1, 1997.
- [2] S. F. Altschul and E. V. Koonin, "Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases," *Trends in Biochemical Sciences*, vol. 23, no. 11, pp. 444-447, November 1, 1998.
- [3] J. M. Berg, J. L. Tymoczko and L. Stryer, *Biochemistry*. 5th edit, W.H. Freeman and Company New York, 2002.
- [4] C. Berge, J.P.R. Christensen and P. Ressel *Harmonic Analysis on Semigroups*, Springer Verlag, 1984.
- [5] J. Casbon, "Protein secondary structure prediction with support vector machines," M.Sc. thesis, Univ. Sussex, Brighton, U.K., 2002.
- [6] N. Cristianini and J. Shawe-Taylor, *An introduction to Support Vector Machines*, Cambridge University Press 2000.

- [7] Burges, C. and J. C., "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery* 2, 121-167, 1998.
- [8] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks", vol. 89, no. 22, pp. 10915–10919, November 15, 1992.
- [9] H. Hu, Yi Pan, R. Harrison, and P. C. Tai, "Improved Protein Secondary Structure Prediction Using Support Vector Machine with a New Encoding Scheme and an Advanced Tertiary Classifier" *IEEE Transactions on NanoBioscience*, vol. 3, no. 4, Dec. pp. 265- 271, 2004.
- [10] Hua, S. & Sun, Z. "A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach", *J. Mol. Biol.* vol. 308, pp. 397-407, 2001.
- [11] H. Kim and H. Park, "Protein secondary structure prediction based on an improved support vector machines approach", *Protein Engineering* vol. 16 no. 8 pp. 553-560, 2003.
- [12] J. Joachims, SVMlight Support Vector Machine, Department of Computer Science.
- [13] Jones D., "Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices", *Journal of Molecular Biology*, vol. 292, pp. 195-202, 1999.
- [14] H. Li, & T. Jiang, A Class of Edit Kernels for SVMs to Predict Translation Initiation Sites in Eukaryotic mRNAs. *Journal of Computational Biology* 12, pp. 702-718, 2004.
- [15] N. Qian and T.J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models", *Journal of Molecular Biology*, vol. 202, pp. 865-884, 1988.
- [16] S. K. Riis, A. Krogh, "Improving Prediction of Protein Secondary Structure Using Structured Neural Networks and Multiple Sequence Alignments", *Journal of Computational Biology*, vol. 3, no.1, pp. 163-184, 1996.
- [17] B. Rost, and C. Sander, "Prediction of protein secondary structure at better than 70% accuracy," *J. Mol. Biol.*, vol. 232, pp. 584-599, 1993.
- [18] B. Rost, C. Sander and R. Schneider, "Evolution and neural networks - protein secondary structure prediction above 71% accuracy", 27th Hawaii International Conference on System Sciences, vol. 5, pp. 385-394, Wailea, Hawaii, U.S.A. Los Alamitos, CA, 1994.
- [19] B. Vanschoenwinkel and B. Manderick, "Substitution Matrix based Kernel Functions for Protein Secondary Structure Prediction", *In the proceedings of ICMLA*, 2004.
- [20] V. Vapnik and C. Cortes, "Support vector networks", *Machine Learning* vol 20, no 3, pp. 273-293, 1995.