PROTEIN SECONDARY STRUCTURE PREDICTION WITH HYDROPHOBICITY AND HYDROPHOBIC MOMENT

TZU-CHENG CHUANG
School of Electrical and Computer
Engineering, Purdue University,
West Lafayette, Indiana 47907

SAUL B. GELFAND
School of Electrical and Computer
Engineering, Purdue University,
West Lafayette, Indiana 47907

OKAN K. ERSOY
School of Electrical and Computer
Engineering, Purdue University,
West Lafayette, Indiana 47907

ABSTRACT

Protein secondary structure prediction has been satisfactorily performed by machine learning techniques such as support vector machines (SVM's). We discuss a special technique to include hyrophobicity information to further improve the classification results. Hydrophobicity or hydrophobic moment measure of each amino acid is included within a given window length in the protein secondary structure prediction using support vector machines. The input data is divided into two groups, which is subsequently classified by an SVM. By including hydrophobicity or hydrophobic moment, the classification accuracy is increased. Comparing the accuracy between using 1 SVM and 2 SVMs. 2 SVMs method has 3-9% higher accuracy than 1 SVM method.

INTRODUCTION

Protein structure prediction is a research topic of growing interest. The number of protein sequences deposited in the Protein Data Bank (PDB) grows faster than the numberof known protein structures. It is very time-consuming to crystallize each protein and use X-ray or nuclear magnetic resonance to analyze its structure. Higher accuracy in secondary structure prediction by using machine learning techniques may help predict tertiary structure more precisely.

Such research is usually initiated by using basic local alignment search tool (BLAST) or position-specific iterated BLAST (PSIBLAST) to find similar homologous proteins. Then, the sequences are aligned, and the position-specific scoring matrix (PSSM) is calculated. This is typically followed by a machine learning algorithm such as a neural network (NN) or a support vector machine (SVM) to do secondary structure prediction.

It is known that discriminating features may help improve classification accuracy. In this paper, we investigate how best to include new features to improve secondary structure prediction. We focus on several membrane proteins which have special distributions with respect to input features. In a previous paper on secondary structure prediction with support vector machines (Ibrikci et al., 2005), the protein structure was classified into 8 classes. In another paper (Rost et al., 2003), the protein structure was classified into 3 classes as alpha-helix, beta sheet and coil. In this paper, we first classify the data into 2 classes, such as alpha-helix and non-alpha-helix. Later by combining the results of three binary classifiers, the three-class classifier is obtained. The window size was investigated in a previous paper (Ibrikci et al., 2005). We pick window length equal to 7 in this paper.

Amino acids have different characteristics in terms of hydrophobicity and hydrophobic moment information. We first segment input data into two groups based on hydrophobicity or hydrophobic moment. Then, the data in each group is classified by an SVM. Below we refer to hydrophobicity unless otherwise specified since it tends to be more significant . A threshold is picked by maximizing the difference of the composition. The input pattern with hydrophobicity greater than the threshold is sent to SVM1. The input pattern with hydrophobicity smaller than the threshold is sent to SVM2. In this way, by segmenting the original dataset into 2 groups, each SVM is trained with similar composition of structural types, and the overall classification accuracy is expected to be higher. The overall structure of the learning algorithm is shown in Fig. 1. It is possible to include the hydrophobicity and/or hydrophobic moment information as part of the input vector. However, the method proposed here was more successful in increasing classification accuracy.

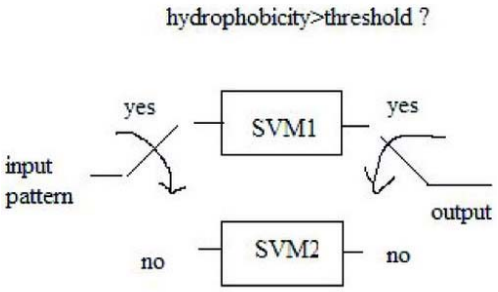

Figure 1. Hydrophobicity decision on which input pattern goes to which SVM.

DATASET

We focused upon membrane proteins since they are less understood than other proteins. The protein sequence and protein structure data for membrane proteins were obtained from the database of secondary structure assignments (DSSP) and the Protein Data Bank (PDB). The decision on which proteins are membrane proteins is obtained by using the information given in the Stephen White laboratory at UC Irvine (Stephen White laboratory). The protein id’s for these membrane proteins are 2PPS, 1JB0, 1FE1, 1S5L, 2AXT, 1IZL, 1VF5 and 1Q90. Some proteins have more than one chain, resulting in 119 chains.

In the DSSP data, there are 8 classes: alpha-helix (H), beta-sheet (E), residue in isolated beta-bridge (B), 3-helix (G), 5 helix (I), hydrogen bonded turn (T), bend (S) and Coil (C). We treat {H, G} as alpha-helix, {E, B} as beta-sheet and all the others as coil since it is more commonly used in other literature as well (Guo et al., 2004).

HYDROPHOBICITY AND HYDROPHOBIC MOMENT DATASET

Hydrophobicity and hydrophobic moment were previously used to classify membrane and surface protein sequences (Eisenberg et al., 1984), (Yang, 2005).

In Eisenberg hydrophobicity, the scale is normalized with mean value equal to zero and standard deviation equal to unity. The hydrophobic moment is calculated as follows :

$$\mu_H = \left\{ \left[\sum_{n=1}^{N} H_n \sin(\delta n) \right]^2 + \left[\sum_{n=1}^{N} H_n \cos(\delta n) \right]^2 \right\}^{1/2} \tag{1}$$

where μ_H is the hydrophobic moment, H_n is the hydrophobicity of the amino acid number *n*, *N* is the window length used (7 in this paper), and δ is the angle of the turn between two amino acids. For classical helices as discussed here, the angle was set to 100.

The following 2 figures show that alpha-helix, beta-sheet and coil have different distributions of hydrophobicity and hydrophobic moment. We included this information into our training and testing method.

Each amino acid has its own hydrophobicity. We calculated the sum of hydrophobicity scale in a given window length. We also calculated the hydrophobic moment in a given window length.

Based on this information, the threshold values were estimated. By including one or the other or both of these attributes, we decide which input pattern is classified by which SVM.

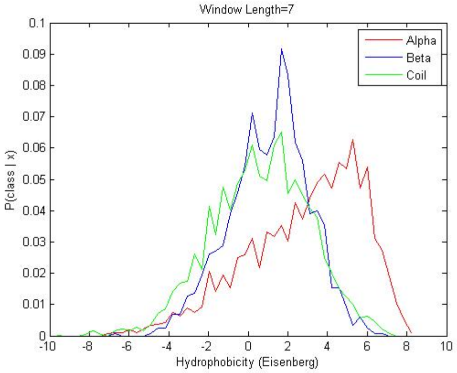

Figure 2: Hydrophobicity distribution for membrane photosystem protein with window length equal to 7.

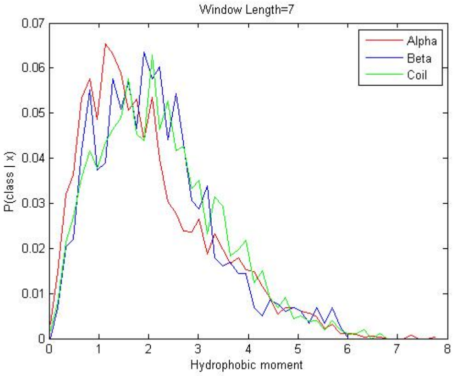

Figure 3: Hydrophobic moment distribution for membrane photosystem protein with window length equal to 7.

INPUT CODING METHOD

The class of an amino acid has some relation to its nearby neighbor. To take this into account, the window length was chosen as an odd number. For a reference amino acid, the amino acids which are half window length in front of the amino acid and behind it make up the input pattern in the window length. For example, for a window length equal to 7, the window includes the previous 3 and subsequent 3 positions of the amino acid. Each amino acid is represented as a 1x 20 vector in one-of-m representation like [0 1 0 ... 0]. If the window length is 7, the pattern within a window is represented as a 1x140 vector.

In DSSP files, some representations of amino acid are B, Z or X. B means the amino acid at that position could be Aspartic acid or Asparagine. Z means the amino acid at that position could be Glutamic acid or glutamine. X means the amino acid at that position is unknown or any amino acid. In such cases, we use 1 x 23 vector to represent an amino acid. In the experiments to be described, the 23-bit representation was used.

SUPPORT VECTOR MACHINES

Vapnik invented Support vector machines (SVM's) with a kernel function in the 1990s (Boser et al., 1992). This algorithm is initially designed for the two-class classification problem. One class output is marked as 1, and another class output is marked as -1. SVM learning involves optimization resulting in support vectors, which are vectors near the boundaries between two classes.

The algorithm trys to find the best separating hyperplane with the largest margin width. The hyperplane of the non-separable case is determined by solving the following equation:

$$\begin{aligned} &\min \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \\ &subject.to. \\ &y_i(x_i^T w + b) \geq 1 - \xi_i \\ &\xi_i \geq 0 \end{aligned} \tag{2}$$

where x_i is the ith data vector, y_i is the binary (-1 or 1) class label of the ith data vector, ξ_i is the slack variable, w is the weight vector normal to the hyperplane, C is the regularization parameter and b is the bias. It can be shown that the margin width is equal to 2/|w|.

Usually the original data is mapped by using a kernel function to a higher dimensional representation before classification. Some common kernel functions are linear, polynomial, radial basis and sigmoid functions. From previous research (Ibrikci et al., 2005), radial basis function kernel is known to give good results. Hence, we used this kernel function in the experiments. Its kernel function is given by

$$K(x, x_i) = C * \exp(-\gamma * |x - x_i|^2) \tag{3}$$

We chose γ equal to 1 and C equal to 1. SVM-light software package was used in the experiments (Joachims, 2004).

EXPERIMENTS AND RESULTS

In each experiment, we randomly picked some proteins for training and the others for testing so that the training set and the testing set do not overlap with each other. Since SVM is initially designed for 2 class classification problem, we classified the proteins into H/not H, E/ not E and C/not C. Later the results were combined to obtain 3-class classification. We set the hydrophobicity threshold equal to 0.42. We also tried several different thresholds near 0.42, and the accuracies were quite similar. Table 1 shows the results for H / not H classification.

We also investigated the distribution of hydrophobic moment to improve classification accuracy. In these experiments, we set the threshold equal to 1.7. The classification accuracy results are shown in Table 2.

In subsequent experiments, we combined three two-class SVM classifiers of H/ not H, E/ not E and C/ not C to obtain 3-class classification. We assumed the class belongs to the class which gives the largest output value. In these experiments, we used hydrophobicity to decide which input pattern goes to which SVM. The threshold is again chosen as 0.42.

The accuracy rate for the 3 classes is denoted as *Qtotal*, and is given by

$$Q_{total} = \frac{CorrectClasses}{N}\% \qquad (4)$$

where N is the total number of predicted residues and CorrectClasses is the number of correct classifications in which the prediction class is the same as the actual class. Table 3 shows the results.

Table 1. Comparison of the testing classification accuracy for H/ not H between one SVM only, and with two SVMs by considering hydrophobicity.

Experiment	1	2	3	4	5
Only 1 SVM	73.49%	81.19%	58.08%	71.69%	62.55%
SVM1	86.78%	88.63%	74.89%	82.61%	73.59%
SVM2	76.57%	83.52%	67.41%	81.22%	65.39%
2 SVMs overall	80.07%	85.29%	70.06%	81.72%	68.4%

Table 2. Comparison of the testing classification accuracy for H/ not H between one SVM only, and with two SVMs by considering hydromoment.

Experiment	1	2	3	4	5
Only 1 SVM	73.49%	81.19%	71.69%	62.55%	72.81%
SVM1	72.99%	82.19%	75.42%	60.57%	72.74%
SVM2	79.46%	83.79%	76.87%	60.64%	83.04%
2 SVMs overall	75.75%	82.87%	76.05%	60.6%	77.2%

Table 3. Comparison of Qtotal with one SVM only, and with two SVMs.

Experiment	1	2	3	4	5
Only 1 SVM	76.02%	54.09%	60.35%	54.09%	69.05%
2 SVMs overall	80.66%	63.04%	63.74%	63.04%	76.95%

SUMMARY

This method is most useful when the distributions of hydrophobicity of the three classes are significantly separable. From the hydromoment experiment, Table 2, we see that one of the experiments actually gives worse result with 2 SVMs than with one SVM. The differentiability of the three classes with respect to the characteristic information used is really important to improve classification accuracy. Our results suggest that, since a binary classifier is used, other characteristics which differentiate beta-sheet versus non-beta-sheet or coil versus non-coil can be used with three different thresholds for different characteristics within different binary classifiers. Each classifier is specialized with the characteristic assigned to it. By combining their outputs with a consensus method, classification results can be improved further.

In the experiments conducted, the ratio of support vectors was almost half of the training vectors. This means the patterns are really difficult to differentiate.

The distribution of membrane photosystem protein is quite different from the other proteins.With some other protein data, such as the CB513 dataset, the three classes of structure have slightly different distribution in hydrophobicity and in hydrophobic moment. In such cases, using hydrophobicity to classify E/not E and using hydrophobic moment to classify H/not H would give better results.

In this study, we showed a special approach to include hydrophobicity and/ or hydrophobic moment to improve classification accuracy in classifying the membrane proteins. We could also include other natural discriminating features in alpha-helix, beta-sheet and coil to further increase classification accuracy.

ACKNOWLEDGEMENT

This research was supported by NSF Grant MCB-9873139 and partly by NSF Grant #0325544.

REFERENCES

T.Ibrikci, Ayca Cakmak, Irem Ersoz and Okan K. Ersoy, ”Hemoglobin secondarystructure prediction with four kernels on support vector machines,” *Computational Intelligence Methods and Applications*, 2005 ICSC Congress Digital ObjectIdentifier 10.1109/CIMA.2005.1662310.

B Rost, G Yachdav and J Liu, ”The PredictProtein Server,” *Nucleic Acids Research* 32(Web Server issue), 2003 W321-W326.

Stephen White laboratory http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html#Latest

D. Eisenberg, E. Schwarz, M. Kmaromy and R. Wall, ”Analysis of membrane and surface protein sequences with the hydrophobic moment plot,” *Journal of Molecular Biology*, Volume 179, Issue 1, 15 October 1984, Pages 125-142.

J.Y. Yang, M.Q. Yang, O. Ersoy, ”Datamining and knowledge discovery from membrane proteins,” *Proceeding of IEEE Intelligence: Methods and Applications Conference*, Istanbul, June 2005.

Jian Guo, Hu Chen, Zhirong Sun, and Yuanlie Lin, “A novel method for protein secondary structure prediction ssing dual-layer SVM and profiles,” *PROTEINS: Structure, Function, and Bioinformatics*, Volume 54, 2004, Pages 738-743.

B. E. Boser, I. M. Guyon, and V. N. Vapnik, ”A training algorithm for optimal margin classifiers,” D. Haussler, editor, *5th Annual ACM Workshop on COLT*, pages 144-152, Pittsburgh, PA, 1992. ACM Press.

Thorsten Joachims, 2004, http://www.cs.cornell.edu/People/tj/svm_light/