Study of Protein Secondary Structure Prediction Using Support Vector Machine

Md. Nazrul Islam Mondal, Md. Al Mamun, Shaju Saha

Abstract— **Prediction of secondary structure of protein is important problem in bioinformatics, because the tertiary structure of protein can be determinant from the folds that are found in the secondary structure. Knowing the tertiary structure of protein can help us to find the function of protein. Moreover knowing the function of protein help to create of the antibody of protein and their work in human body. Protein secondary structure prediction mostly depends on the information stored in the primary amino acid sequence. Support Vector Machine (SVM) has shown special ability of predicting in a number of application areas including secondary structure prediction. The objective of this paper is to find out the protein secondary structure prediction using Support Vector Machine (SVM). However we introduce six binary classifiers as an almost new technique. We distinguish between the classes helices (H) strand (E), and coil (C). In this paper, we predict secondary structure using Gaussian kernel with a fixed a parameter at γ=0.1 and varying cost parameter C within the range [0.1, 5]. The goal of our approach is to propose a time efficient method for checking accuracy using different tests. Our results show the prediction accuracies are in the range 62-72%. More specifically, our results for H/~H, E/~E, C/~C, H/E, E/C and H/C are respectively 66.25%, 72.28%, 62.58%, 65.33%, 68.56% and 70.85%. The highest accuracy of 72% for OAtest is observed for the One-Against-All approach while the highest accuracy of about 70% is observed in One-Against-One approach on OAtest. We say that our approach is simple with better time complexity in comparison to Tsilo's work.**

Keywords— Secondary Structure; Support Vector Machine; Binary Classifiers; OAtest.

I. Introduction

Protein secondary structure is closely related to the protein tertiary structure, which determines the characteristically behavior of the proteins. Many researches have been done over the decades to study and predict the protein structure. Till date, the total number of experimentally determined structures is less than twenty thousand (Protein Data Bank) whereas there are over a million known protein sequences. It is therefore becoming increasingly important to predict protein structure from its amino acid sequence, using insight obtained from already known structures [6].

There are about 5 million protein sequences available from http://www.ebi.ac.uk and about fifty thousand protein known sequences that are available in http://www.rcsb.org/pdb/ [1]. Different techniques have been developed that can predict secondary structure of proteins from their amino acids sequences. They are based on different algorithms, such as Statistical Analysis (Chou and Fasman, 1974),Information theory, Bayesian Statistics and Evolutionary Information (Sen, Jernigan, Garnier, and Kloczkowski, 2005), Neural Networks (Holley and Kurplus, 1989; Qian and Sejnowski, 1988), Nearest Neighbour Methods (Salamov and Salovyev, 1995), a combination of multiple alignment and Neural Networks (Rost and Sander, 1993). For these approaches, the accuracy levels are in the range 65–80% [1].

II. SUPPORT VECTOR MACHINE

The SVM method is a comparatively new learning system which has mostly been used in pattern recognition problems. Support Vector Machines are machine learning algorithms implemented for classification and regression [5]. For classification, Support Vector Machines operate by finding a separating hyper plane in the space of possible inputs. This hyper plane attempts to split the positive examples from the negative examples. The split will be chosen to have the largest distance from the hyper plane to the nearest of the positive and negative examples. Data points that are at the margin are called Support Vectors. These data points are very important in the theory of Support Vector Machines because they can be used to summarize information contained in the dataset. [5]. The hyper plane with a maximum margin allows more accurate classification of new points. However, not all problems can be linearly separated by a hyper plane. For such problems, the resulting algorithm is formally similar, except that a nonlinear transformation of the data into a feature space is performed. This allows the algorithm to fit the maximum margin hyper plane in the transformed feature space. Kernels are used to perform the mapping [1].

III. Protein Secondary Structure Prediction

Protein structure can be used to infer bio-chemical and biological functional information, and in identification of amino acids that are involved in active site. The functional properties of proteins depend upon their three dimensional structures [1]. To understand the biological functions of proteins, the structure of a protein from the amino acid sequence should be known beforehand. Protein Secondary Structure Prediction is accomplished in five major steps. They are as follows:

Manuscript received October 08, 2015.

Md. Nazrul Islam Mondal, Dept. of CSE, RUET, Rajshahi, Bangladesh, Mobile No. +8801912744327 (e-mail: mondal@ruet.ac.bd).
Md. Al Mamun, Dept. of CSE, RUET, Rajshahi, Bangladesh, Mobile No. +8801962143416 (e-mail: cse_mamun@yahoo.com).
Saju Saha, Dept. of CSE, RUET, Rajshahi, Bangladesh, Phone No. +880721750838, (e-mail: ujas093017@gmail.com).

- Sliding Window Scheme,
- Orthogonal Input Profile,
- Reshape the Orthogonal Input,
- Learn Input Using SVM,
- Secondary structure assignment.

A. Sliding Window Scheme

To train the SVM with protein sequence and structural information, a sliding window scheme is sued [6]. In this sliding scheme, a window becomes one training pattern for predicting the structure of the residue at the center of the window. And in this training pattern, the information about the local interactions among neighboring residues is embedded. Fig.1 shows an example of this scheme with window size of 5. Here, to predict the structure of amino acid 'N', the sequence 'AKNLK' goes together as one input pattern. To predict structure of 'L', the next amino acid, the window slides down to the next group of sequence, 'KNLKQ' and so on.

Input Protein sequence…N A T A A K N L K Q D A T K S E R VA
Output Secondary structure…H H H H H H H C E C C H H H C C H H H

Input pattern i: …N A T A A K N L K Q D A T K S E R V
Input pattern i+1: …N A T A A K N L K Q D A T K S E R V
Input pattern i+2:… N A T A A K N L K Q D A T K S E R V

Sliding window

Fig.1. Sliding window scheme with window length of 5 [6].

B. Orthogonal Input profile

The feature value of each amino acid residue in a window means the weight (costs) of each residue in a pattern. In this study all weight assignment schemes are not tested. The test results from Hua's method are used to select the best scheme [5]. In Hu's study orthogonal method is used as reference for comparison with different encoding scheme [5]. Among the different schemes explained in the previous studies, the first simplest way is to use the orthogonal encoding which assigns a unique binary vector to each residue, such as (1, 0, 0, …), (0, 1, 0, …), (0, 0, 1 …) and so on [6]. In this method, the weights of all the residues in a window are assigned to 1 equally. The method is explained as follows here for simplicity only single window encoding scheme is explained. For multiple windows the same technique is followed for all the elements within the window and target is the center element of the middle window [6]. Fig.2 shows the sample training data to which orthogonal encoding method is applied as an input profile with window size 5. In this figure, the value of the first column, {-1,+1}, are target values of each binary classifier. For example, if the binary classifier is the one which classifies helix or not, and if the structure of the residue from the training data is helix, (i.e. the center value has H, G or I corresponding to its position) the target value becomes +1[6].

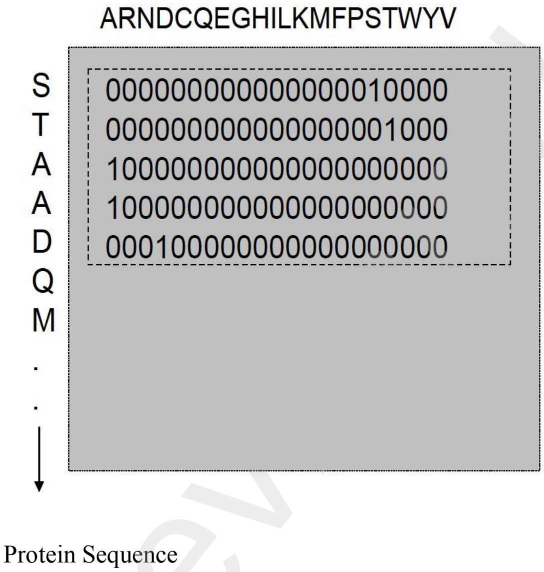

Fig. 2. An Example of Orthogonal Vector Profile [6].

C. Reshape the Orthogonal Input

The multidimensional orthogonal matrix can be converted into one dimensional matrix. A binary encoding scheme is therefore used to assign numerical values to letters. For each binary vector, there are 21 positions: 20 positions for the letters of amino acids and 21st for a null input denoted by _. For the window size of 13, the input pattern contains 13 input residues. Each residue will be assigned 1 depending on its position while the other positions will be assigned 0's. However, prediction will only be made for the central residues. In that way, there will be 21 × 13 input groups for which 13 of them will have values 1 and the rest is 0. For the window length of 13 obtained from 2n + 1 for n = 6, the dimension of the samples is (2n + 1) × 21 = 273. We convert the 21×13 input groups into 1×273 input groups.

D. Learn Input Using SVM

The soft margin classifier is an extension of linear SVM. The kernel method is a scheme to find the nonlinear boundaries. The concept of the kernel method is transformation of the vector space to a higher dimensional space. For SVM, we use a Gaussian kernel with parameter fixed at $\gamma = 0.1$ and varying cost parameters C in the range [0.1,5]. The mapping function is represented in forms of a positive definite kernel function, k (x, x'), which is easier to calculate the inner product in the feature space specified by following from[1]:

$$k(x, x') = (\Phi(x), \Phi(x')) \qquad (1)$$

kernel function $K(x_i, y_i)$ was written as:

$$K(x_i, y_i) = \exp(-\gamma \| x_i - x_j \|^2) \qquad (2)$$

SVM has two parameters: the kernel γ and the cost parameters C.

E. Secondary Structure Assignment

There are various methods that can be used to perform secondary structure assignment. The assignments of secondary structure in this study are based on the DSSP method [7]. It classifies the secondary structures into 8 classes. The 8 secondary structures are reduced to the known 3 structures: α−helices, β−strand and coils residues neither helices nor strand. Table 6.3 shows the 8 secondary structural classes defined by DSSP, their standard abbreviations and the mappings to three structural classes. Rest (−) defines secondary structures which do not belong to any DSSP defined categories [3].

Table 3: Secondary structure assignment [2]

DSSP class	*Abbreviation*	*3 state classes*
α− helix	H	H
3_{10}helix	G	H
β− strand	E	E
isolated β- ridge	B	E
Л− helix	I	C
Turn	T	C
Bend	S	C
Rest	-	C

For Support Vector Machines however, target coding is different: +1 represents target belonging to class 1, i.e. alpha while −1 denotes targets belonging to class 2, i.e. no alpha. These sections also gave a discussion of how the inputs and targets are to be created. The next section outlines formulation of both NN and SVM and describes how the results are going to be evaluated [2].

IV. Results

Six SVM binary classifier including three one-versus-rest classifier ('one': positive class, 'rest': negative class) names H/~H (helix vs no helix), E/~E (beta vs no beta) and C/~C (coil vs no coil) and three one-versus-one classifier named H/E (helix vs beta), E/C (beta vs coil), C/H (coil vs helix) were constructed. The results of SVM application in predicting the secondary structure of proteins based on OAtest are given in Table I.

Table I. SVM: Optimal classifiers for OAtrain and OAtest.

Quantity	*OAtrain*	*OAtest*
H/ ~H	1	66.25
E/~E	1	72.28
C/~C	1	62.58
H/E	1	65.33
E/C	1	68.56
C/H	1	70.85

The following figure (Fig.3) shows the result of H vs E classifier accuracy where cost perameter are plotted in X axis and accuracy are plotted in Y axis.

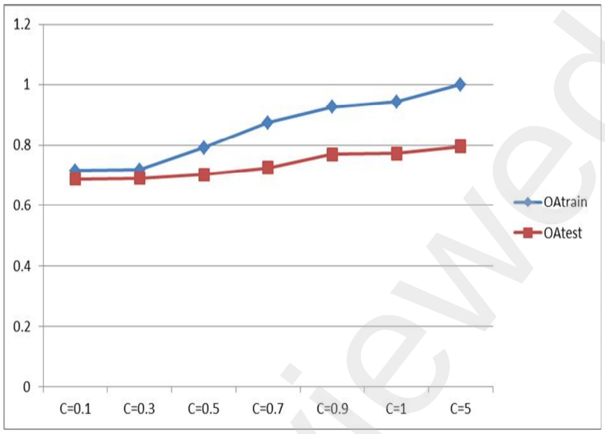

Fig. 3. Cost parameters vs. accuracy (H/E).

All the results were obtained by averaging 10 experiments for each cost parameter in SVM. The results are presented as follows:

In the following Table (Table II) alpha vs no alpha accuracy are represented:

TABLE II: SVM: ALPHA VS NO ALPHA

Experiment	SVMopt	C=0.1	C=0.3	C=0.5	C=0.7	C=0.9	C=1	C=5
OAtrain	1	0.7170	0.7199	0.7911	0.8742	0.9261	0.9434	1
OAtest	0.7397	0.6824	0.6895	0.7123	0.7197	0.7232	0.7341	0.7397

The following figure (Fig.4) shows the result of H vs ~H classifier accuracy where cost parameters are plotted in X axis and accuracy are plotted in Y axis:

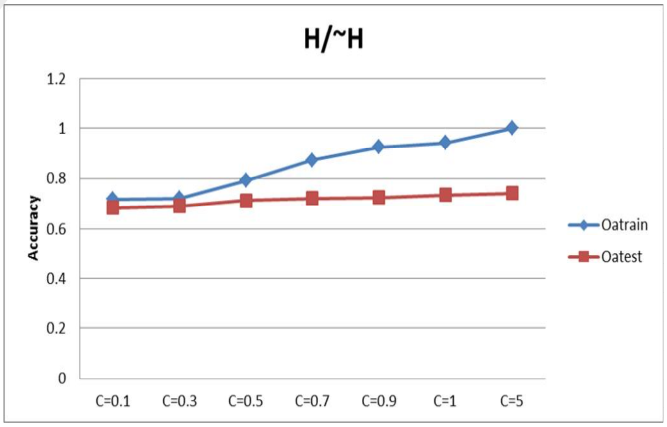

Fig 4: Cost parameters vs. accuracy (H/~H).

The following table (Table III) represents the accuracy between beta vs no beta :

TABLE III: SVM: BETA VS NO BETA.

Experiment	SVMopt	C=0.1	C=0.3	C=0.5	C=0.7	C=0.9	C=1	C=5
OAtrain	1	0.7037	0.7165	0.8021	0.8824	0.9329	0.9531	1
OAtest	0.7567	0.6834	0.6870	0.6931	0.7033	0.7326	0.7379	0.7567

The following figure shows the result of E vs ~E classifier accuracy where cost parameters are plotted in X axis and accuracy are plotted in Y axis:

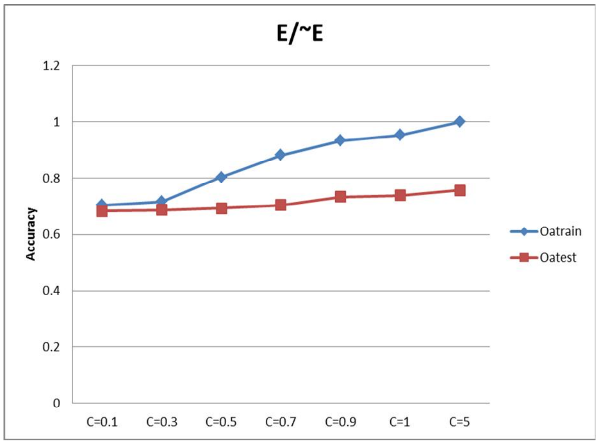

Fig. 5. Cost parameters vs. accuracy (E/~E).

The following table represents a comparison among the Tsilo's work[6] and our work in terms of accuracy:

TABLE I. COMPARISON OF OUR APPROACH AND TSILO APPROACH

Binary Classifiers	Our Accuracy (%)	Tsilo Accuracy (%)
H/ ~H	66.25	73.74
E/~E	72.28	80.32
C/~C	62.58	68.31
H/E	65.33	71.75
E/C	68.56	72.73
C/H	70.85	75.35

The following figure (Fig.5) shows a graphical representation of exposing a comparison between the Tsilo's work and our work in terms of Accuracy. Although our accuracy is slightly near or equal of Tsilo's accuracy, however our method is simple with better time complexity which is omitted for page limitation in this paper.

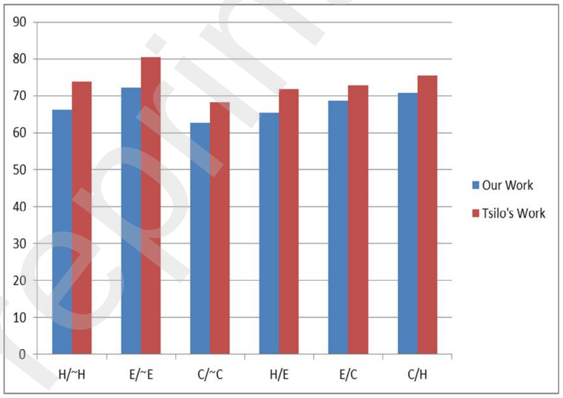

Fig. 5. Comparison between Our Work and Tsilo's Work

V. CONCLUSION

The main goal of this paper is to propose a simple method with better time complexity for comparing accuracy based on SVM. By our method, we predict the protein secondary structure using support vector machine and also earn the best accuracy in the result. Although our results show a slightly less or equal accuracy in comparison to Tsilo's work [6], however our method is simple with better time complexity. The SVM method is a comparatively new learning system which has mostly been used in pattern recognition problems. Protein structure prediction is an important step towards predicting the tertiary structure of proteins (and Quaternary structure). The reason is that knowing the tertiary structure of proteins can help to determine their functions. The main aim of this paper is to use support Vector Machine (SVM) predicting the secondary structure of proteins from their amino acid sequences.

REFERENCES

[1] Reyaz-Ahmed, Anjum B., "Protein Secondary Structure Prediction Using Support Vector Machines, Nueral Networks and Genetic Algorithms" (2007). Computer Science Theses. Paper 43.

[2] Rost, B. and Sander, C." Improved prediction of protein secondary structure by use of sequence profile and neural networks." Proc Natl Acad Sci U S A 90, 7558-62 (1993).

[3] Kabsch, W. & Sander, C. (1983). *"Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen Bonded and Geometrical Features."* Biopolymers, 22: pp. 2577-2637.

[4] Salamov, A.A. & Solovyev, V.V. "Protein secondary structure prediction using local Alignments" J. Mol. Biol, 268, pp. 31-36. (1997).

[5] Hua, S. and Sun, Z. "A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach." J. Mol. Biol. 308, 397-407 (2001).

[6] Lipontseng Cecilia Tsilo*"* A thesis submitted to Rhodes University in partial fulfillment of the requirements for the degree of Master of Science." Department of Statistics, February 2008.

[7] Nguyen, M.N. & Rajapakse, J.C (2003). *"*Multi-Class Support Vector Machines for Protein Secondary Structure Prediction. Genome Informatics*"* 14: pp. 218-227.

[8] V.Vapnik and C. Corter. *"*Support vector networks . Machine Learning.*"* Vol. 20 pp. 273-293, 1995.

[9] Voet, D., Voet, G. & Pratt, W.C. (2006*)*. *"*Fundamentals of biochemistry: life at the molecular level*"*. New York : Wiley.

M. N. I Mondal received BE degree from the Department of Electrical & Electronic Engineering, RUET, Bangladesh in 2000, ME degree from the Department of Information & Communication Technologies, Asian Institute of Technology, Thailand in 2008 and Ph.D. degree from the Department of Information Engineering, Hiroshima University, Japan in 2012. He joined as a lecturer to the Department of Computer Science & Engineering, RUET in 2001. In 2004 and 2013, he became Assistant Professor and Associate Professor in the same Department respectively. He also joined as a Professor in 2015 at the same Department. From March, 2012 to September, 2012 and from May, 2013 to April, 2014, respectively he was Visiting Research Scholar and Specially Appointed Assistant Professor in the Department of Information Engineering, Hiroshima University, Japan. He has been serving as a CISCO instructor since 2006. He has published his contributions extensively in journals, conference proceedings. He served as an Organizing Chair, a PC member, reviewer and sub-reviewer for many Journals and Conferences such as Journal of Foundation of Computer Science, Journal of Communication and Computer, International Journal of Networking and Computing, IEICE, APDCM, PDP, ICNC, CANDAR, IJPEDS, ICPP and so on. He is a Fellow of Institution of Engineers, Bangladesh and IEEE Member. His research interest includes FPGA-based Reconfigurable Computing, Parallel Computing, Algorithms and Architectures, Image Processing, DSP and Computer Networks and Data Communications.

Dr. Md. Al Mamun have 8 long years of teaching experience in the various fields of computer science and engineering. Graduated from Rajshahi University of Engineering & Technology, Bangladesh Mr. Mamun got his first teaching opportunity in the same university to take courses like Computer Programming, Database Management System, Computer Architecture, Computer Graphics, Object Oriented Programming, Digital Image Processing and many more. In 2009, he got teaching assistantship in the University of New South Wales, Australia. This was the opportunity, which he got when he was doing his PHD in the same university. He was responsible for lecturing various computer science courses like Object oriented programming (Java), Computer Graphics (Game Simulation-Alice) etc in UNSW@ADFA, Australia . Now he is serving as associate professor in RUET. He is a Fellow of Institution of Engineers, Bangladesh and IEEE Member. His research interest includes Satellite Image Processing, Data Mining, Computer Vision.

Shaju Saha received BE degree from the Department of Computer Science & Engineering, RUET, Bangladesh in 2013. He is now working in famous IT industry in Bangladesh. He was an excellent organizer of many Int'l Conferences in Bangladesh. He is a member of Institution of Engineers, Bangladesh. His research interest includes FPGA-based Reconfigurable Computing, Parallel Computing, Algorithms and Architectures, Image Processing, DSP and Computer Networks and Data Communications.