Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 10, October-2019

Application of Support Vector Machine in Prediction Secondary Structure Protein

Nahla .I.Jabbar
Chemical Engineering
Babylon University.
Babylon, Iraq

Rafah Ibraheem Jabbar
Metallurgical Engineering
Babylon University
Babylon, Iraq

***Abstract*— This paper studying the predication of secondary structure protein from primary structure protein using support vector machine (SVM).We classify 64 types of proteins in three types : Helices(H) ,Strand(E) and Coil (C).In our SVM ,we use a Gaussian kernel with parameter =0.1 and costing parameter c between [0.1,5].The results of support vector machine have different varying accuracies of three types of proteins .**

Keywords*— Support Vector Machine; Amino Acid Sequences and Secondary Protein Structure

INTRODUCTION

Support Vector Machine (SVM) was known in 1992, introduced by Boser, Guyon, and Vapnik[1]. Support vector machines (SVMs) are a set of related supervised learning methods used for classification and prediction tool. It have been successfully applied in a variety of biological application for example in automatic classification of microarray gene expression profiles [2]. It is also being used for many applications, such as hand writing analysis, face analysis. SVMs were developed to solve many problems but recently they have been extended to solve bioinformatic application .Support vector machine have been successfully applied in a prediction problem for example in protein secondary structure Protein structure prediction by using bioinformatics can involve sequence similarity searches, multiple sequence alignments, identification and characterization of domains, secondary structure prediction, A central part of a typical protein structure prediction is the identification of a suitable structural target from which to extrapolate three-dimensional information for a query sequence .The main objective of this paper is to predict the secondary structure of protein using support vector machine .This will be done by measuring the performance of the algorithm using their accuracies of prediction performance. All experiments are done by matlab software

I. SUPPORT VECTOR MACHINE

Support Vector Machine or SVM is related to Structural Risk Minimization (SRM). At the first place, SVM is an initial form for binary classification, but now it can also be used for multiclass classification. SVM method does mapping form input space to a higher dimensional space to support nonlinear classification problems where a maximal separating hyperplane is constructed. Hyperplane is a linear pattern whose maximum margin gives the maximum separation between the decision classes[3]. given data set $\{x_i, y_i\}^N$ where N is known as the number of samples, $x_i \in R^D$ is known as feature vectors from sample-i, where D is the number of feature (dimension), and y_i is known as class labels. For two class classification problem $y_i \in \{-1, +1\}$, however for the multiclass classification problem $y_i \in 1,2,...,k$ where k is the number of class. The main purpose of SVM is to find the best hyperplane:

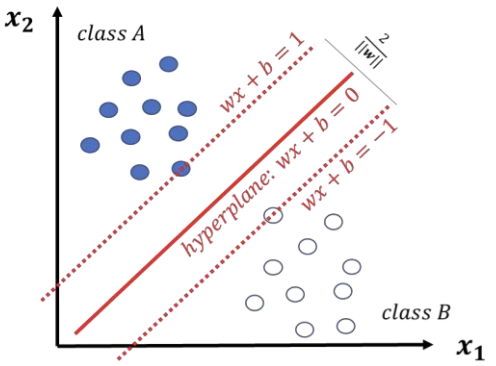


Fig. 1. SVM is trying to find the best hyperplane that separates the two classes, class A and B

Distance of closest point on hyperplane to origin can be found by maximizing the x as x is on the hyper plane. Similarly for the other side points we have a similar .This solving and subtracting the two distances we get the summed distance from the separating hyperplane to nearest points. Maximum Margin = M = 2 / ||w||

II. PROTEIN STRUCTURE PREDICATION

Proteins are made of simple building blocks called amino acids. There are 20 different amino acids that can occur in proteins. Their names are abbreviated in a three-letter code or a one letter code. The amino acids and their letter codes are given in Table [1]

Glycine	Gly	G	Tyrosine	Try	Y
Alanine	Ala	A	Methionine	Mer	M
Serine	ser	S	Tryptophan	Trp	W
Threonine	Thr	T	Asparagine	Asn	N

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 10, October-2019

Cysteine	Cys	C	Glutamine	Gln	Q
Valine	Val	V	Histidine	His	D
Isoleucine	Ile	I	Aspartic Acid	Asp	D
Leucine	Leu	L	Glutamic Acid	Glu	E
Proline	Pro	P	Lysine	Lys	K
Phenylalanine	Phe	F	Arginine	Arg	R

III PROTEIN STRUCTURE

There are four different structure types of proteins, namely Primary, Secondary, Tertiary and Quartenary structures. Primary structure refers to the amino acid sequence of a protein. It provides the foundation of all the other types of structures. Secondary structure refers to the arrangement of connections within the amino acid groups to form local structures. α helix, β. strand are some examples of structures that form the local structure. Tertiary structure is the three dimensional folding of secondary structures of a polypeptide chain. Quartenary structure is formed from interactions of several independent polypeptide chains. The four structures of proteins are shown in Figure (2)

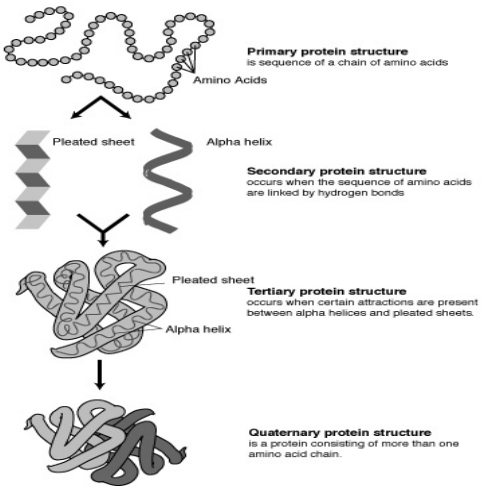
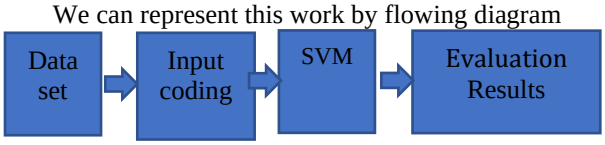


Fig. 2. Protein structure

There exists a relationship between protein structure and function. Proteins with similar sequences and structures have similar functions. Moreover, similar sequences in proteins imply that they also have similar structures. However, similar structures in proteins may have different sequences and different functions .The primary structure of proteins can be used to predict its tertiary structure. It is through the tertiary structure of the protein that we can derive its properties as well as how they function in an organism. Secondary structure prediction means predicting the secondary structure of a protein from its primary sequence. It is important because knowledge of secondary structure helps in the prediction of tertiary structure. This is very interesting for proteins whose sequences do not show any similarities with the sequences of proteins in the database.

IV.METHODLOGY OF THE WORK

We can represent this work by flowing diagram



A. *Data set*

The data are collected from Data bank ,it includes 62 types of proteins he data to be used consists of 62 proteins from Rost and Sander (1983) database available from [5] It contains a protein name, its primary and secondary sequences. The data are defined in rows :protein names and amino acid sequence for example

Acprotease
GVGTVPMTDYGNDVEYYGQVTIGTPGKSFNLNFDTGSSNLW
VGSVQCQASGCKGGRDKFNPSDGSTFKATGYDASIGYGDGSA
SGVLGYDTVQVGGIDVTGGPQIQLAQRLGGGGFPGDNDGLLG
LGFDTLSITPQSSTNAFQDVSAQGKVIQPVFVVYLAASNISDG
DFTMPGWIDNKYGGTLLNTNIDAGEGYWALNVTGATADST
YLGAIFQAILDTGTSLLILPDEAAVGNLVGFAGAQDAALGGFV
IACTSAGFKSIPWSIYSAIFEIITALGNAEDDSGCTSGIGASSLG
EAILGDQFLKQQYVVFDRDNGIRLAPVA.

For the two methods, the same partitioning of the data into training set, validation set and test set was used. 10-fold cross validation as described in was used; the data was randomly divided into 10 parts, one used as a test set and the rest for training. However, training data and used as validation set. The window size was fixed at 13.

B. *Input coding*

The data of the 20 amino acid residues into letters .The purpose of input coding was converted these letters into numbers, that's coding are done by orthogonal method in Figure(3) and Table[2].

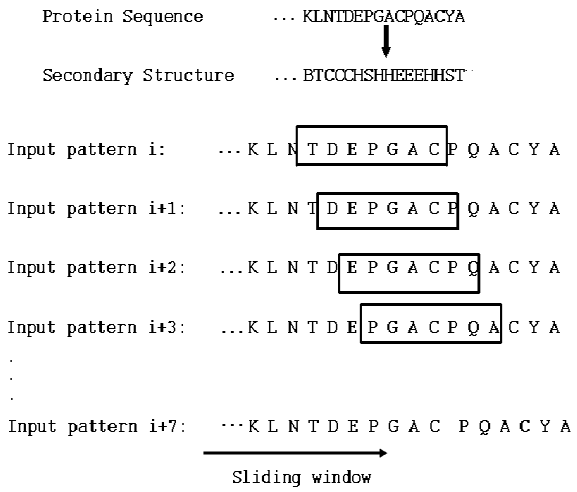


Fig.3. A sliding window of length 7

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 10, October-2019

Table 2

K	L	N	T	D	E	T	G	A	C	P	Q	A	C	Y	A
0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1
0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

The orthogonal coding for each residue. The function of input coding the letters for each sequence and the outputs are the unit vectors (orthogonal coding) for each residue in table[3].

Table 3

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Therefore, the orthogonal coding for a sequence KLNTDETGACPQACYA, is given in Table [1]From this table a unique binary vector is assigned for each residue.

C. Support Vector Machine (SVM)

The aim of this script is to produce the prediction overall training data and on test data using Support Vector Machines. In this approach, the following six binary classifiers are created :H/ ~ H, E/ ~ E, C/ ~ C, H/E, E/C and C/H were constructed as in SVM. Kernel parameter γ was fixed for γ = 0.1 [24]. The cost parameter C was however varied over the following values: 0.1,0.3,0.5,0.7,0.9,1,5. The best results are in window size 13

D. Evaluation of result

The objective of this step is to compare the performance Support Vector Machines. This approach was applied on 62 globular proteins. The results of multi class SVM show that H/ ~ H has the highest accuracy while C/ ~ C has the lowest prediction accuracy

Table 5 Protein secondary structure prediction with Support Vector Machines

Quality index	% Accuracy
Overall accuracy	73.5

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 10, October-2019

H/ ~ H	79.36
E/ ~ E	79.15
C/ ~ C	67.10
H/E	72.02
E/C	72.10
C/H	74.66

The effect of variations in cost parameter on prediction accuracy and the relationship between the values of cost parameters and time required for training are also discussed. Figure(4) shows different cost parameters and the time taken to train the six binary classifiers. Positive relationship exists between time and the cost parameters because an increase in the number of cost parameters increases the time required to train the classifiers. The longest time it to train a classifier was about 57 minutes, and that occurred at C = 5 for C/ ~ C. The shortest time to train a classifier was about 13 minutes, which occurred at C = 0.1 for H/E. Furthermore, training time increases rapidly between C = 1

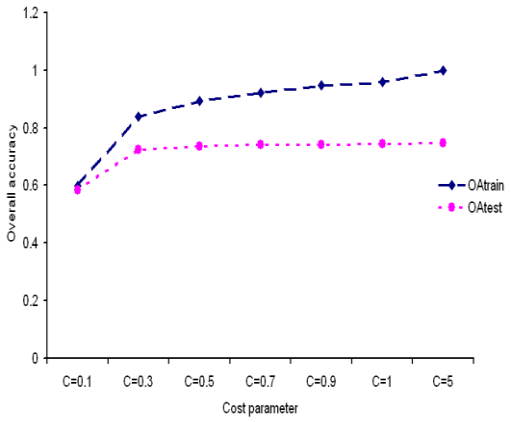


Fig.5. Training time vs. cost parameters in Support Vector Machines

CONCLUSION

Protein structure prediction is an important step towards predicting the tertiary structure of proteins. The reason is that knowing the tertiary structure of proteins can help to determine their functions. The main aim of this paper was to compare performance of Support Vector Machines in predicting the secondary structure of proteins from their amino acid sequences. The following conclusions were derived:

1. This approach created six binary classifiers. The results are obtained from the same window length of 13

2. The experimental shows that increasing the size of the training data set
improves the performance significantly

3. Choosing window size affected in the results . Choosing an appropriate window length also helps to improve the performance.

REFERENCES

[1] Wikipedia Online. Http://en.wikipedia.org/wiki

[2] Lipo Wang, “Support Vector Machines: Theory and Applications,” Springer.

[3] Pongsametrey Sok and Nguonly Taing “Support Vector Machine (SVM) Based Classifier For Khmer Printed Character-set Recognition” APSIPA 2014

[4] https://www.khanacademy.org/science/biology/macromolecules/proteins-and-amino-acids/a/orders-of-protein-structure.

[5] http://antheprot-pbil.ibcp.fr.