Prediction of Secondary Structure of Protein using Support Vector Machine

Shivani Agarwal
Assistant Professor
IMS Engineering College, Ghaziabad.

Pankaj Agarwal
Professor & Head
IMS Engineering College, Ghaziabad.

Deepali Mendiratta
Student
IMS Engineering College, Ghaziabad.

ABSTRACT

The tertiary structure of protein is difficult to predict accurately directly from a protein sequence. The intermediate step is required to predict the structure which project the one dimensional structure into the three dimensional structure. This intermediate step is called secondary structure of protein. The secondary structure of protein plays a key role in the designing of drugs. There are many machine learning algorithms such as HMM (hidden markov model), SVM (Support vector machine), NN (neural network), Fuzzy Logic. A technique which we used to predict the secondary structure of protein is Support Vector Machine (SVM) with Hidden markov transition encoding matrix. Support vector machine is a supervised machine learning method and is based on the principle of the structural risk minimization. The concept of SVM is based on the construction of hyper plane in the high dimensional space to classify the data into the categories. The main objective is to increase the accuracy and decrease the error of prediction.

Keywords

Data Set, Kernel function, Markov Transition Encoding Scheme, Secondary structure, Support Vector Machine.

1. INTRODUCTION

Secondary structure of protein prediction is a big problem in molecular biology. Secondary structure of protein means the intermediate step between the primary and tertiary protein structure. If we find out the protein secondary structure so we easily discover the drugs of a particular disease that is a big achievement in the field of molecular biology. These problems can be solve by using the neural networks .Neural network divided in to two parts biological neural network and artificial neural network. This problem is not easy to solve by the use of biological neural network. So these problems can be solved by the computational biology that known as Artificial Neural Network. Artificial Neural Network is a Technique of Soft Computing. Soft computing is different from the hard computing. Use of Soft computing methods to solve the real word problem which do not produce exact results which totally believe on the partial truth or can say to solve the NP- hard problems. In this paper we use the technique of soft computing that is SVM and HMM for solve the problem of bioinformatics.

1.1 Structure of protein

Proteins are the basic components in the organism. Protein macromolecules are made up from the linear sequences of 20 amino acids, the structural unit of proteins, joined together by peptide bonds. This is known as primary structure of protein. There are number of protein structure formation with the combination of 20 amino acids. Each amino acid is made up of a combination of a carboxyl atom and a hydrogen atom. The structures of proteins are determined by the use of spectroscopy or crystallography. The next level of protein structure depends on this sequence of amino acids. The secondary structure is the result of folding and twisting of the primary structure sequence

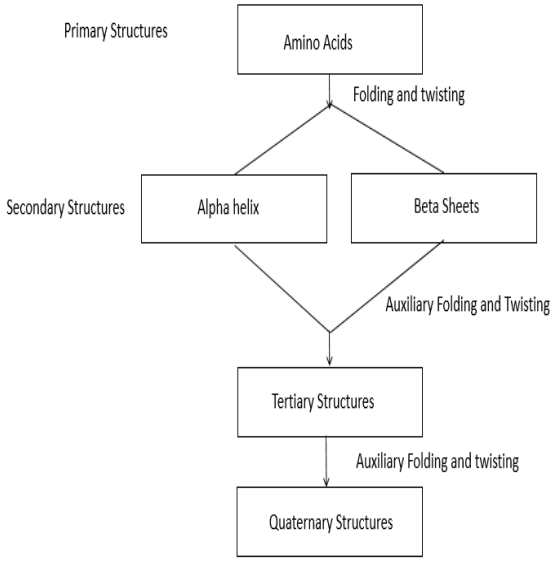

Fig 1: Structure of Protein [1]

. The tertiary structure is the three dimensional structure of the polypeptide chain. The last structure formed of protein is the quaternary dimensional structure formed by more folding and the twisting of polypeptide chain. For the prediction of tertiary structure of protein, an intermediate structure of the sequence of proteins is created which is called secondary structure of protein. The three classes which are used to classify the secondary structure are, alpha helix (H), beta sheets (E) and coil (C).

1.2 Need of secondary structure of protein

Prediction of secondary structure provides an environment for better drug designing. Tertiary structure is a very complex structure that is not the better way for drug in a short duration

and in a better way.so we discover the better technique for drug designing that is the secondary structure prediction that is a intermediate step between the primary and the secondary structure that is in the form of H,E and C.

1.3 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning method developed for the purpose of pattern recognition and regression. This method is proposed by **Vapnik and Corinna Cortes.**

SVM transforms the input sample data into a high dimensional Hilbert space. It tries to search a hyper plane which optimally separates the two different classes by maximizing the distance between the hyper plane and the input sample data. This hyper plane is known as Optimal Separating Hyper plane (OSH). Using the equation, $w^Tx+b=q$, we decide the line with largest margin.

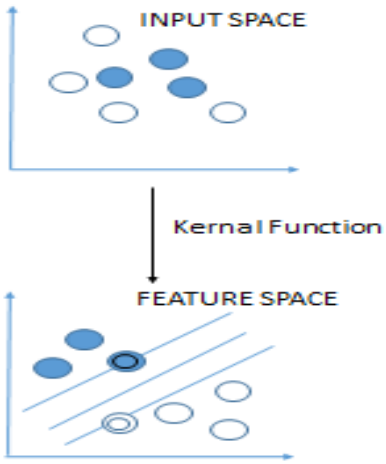

Fig 2: Input space is mapped into Feature space using the kernel function

We find the parameters w and b and maximize the distance $2/\|w\|$.

For the prediction of the secondary structure of protein, a sequence uses a Markov Transition Matrix encoding scheme, whose output is provided to a single layer of Support Vector Machine. We use three binary classifiers for the prediction. In the first layer, we use three one-versus-rest (H/~H, E/~E and C/~C) binary classifiers which are combined together to handle multi class case which compares the given probabilities with the defined range of the classes. If it comes zero, the residue is classified into it calculated class.

There are various advantages of using the SVM method for the prediction of secondary structure of proteins. First, reducing the risk of over-fitting problem, as SVM is based on the principle of structural risk minimization that guarantees the lowest error. Second, the training of machine can certainly converge to a global optimal. Third, there is no loss of information of useful data.

1.4 Kernel Method

The kernel function denoted by k (x, x'), increase the computational power of learning of the linear learning machines by projecting the data into the high dimensional feature space. The choice of kernel affects the performance of the machine. By using the kernel function like Linear Function, Polynomial Kernel function, sigmoidal function, Radial Basis Function, we can make the SVM, a non-linear classifier. Most effective kernel function used by various scientists is Radial Basis Function (RBF) which gives a largest performance as compared to other kernel functions. The RBF is given by the equation,

$$k(x,x') = \exp(-\gamma(x-x')^2)$$

Where, γ is the gamma parameter

1.5 Markov Transition Encoding Scheme

Hidden markov model are the naturally represented probabilistic models which generates a sequence in which the probability of occurrence of one state depend on the probability of occurrence of previous state. For the representation we derive three Markov models for the prediction of classes of secondary structure, H (Helix), E(Strand) and C(Coil).

Transition probability is given as,

$$\alpha_{st}^{i} = \frac{C_{st}^{i}}{\sum_i C_{st}^{i}}$$

C_{st}^{i} is the frequency of amino acid T followed S for sequences in class i belongs to {H,E,C}.

The general model for hidden markov model is given as,

$$P(x_i|x_{i-1}, x_{i-2} \ldots . x_1) = P(x_i|x_{i-1}, x_{i-2} \ldots . x_{i-n})$$

Using this general form, the three transition probability becomes,

$$\alpha_{qrst} = P(x_i|x_{i-1} = r, x_{i-2} = s, x_{i-3} = t)$$

Where,α(qrst) is residue of amino acid at state *i* and *s*, *r* and *q* are residues of amino acid at the Previous state, *i*-1, *i*-2 and *i*-3, respectively.

The pre-processing is a major step in feature extraction in which primary sequences are encoded. It transforms the original variables into new inputs which are then used for classification. Various procedures can be used for the encoding. One of which is a third-order Markov Model in which we use a Markov transition matrix encoding. This encoding extracts the essential feature of protein sequences and reduces the dimensional space. By this, the performances of the predicting model increases and it improves the speed and memory learning model.

1.6 Data Set

The selection of the data set is taken from Compilation and creation of datasets from Protein Data Bank (PDB) source which is datasets generated from the DSSP and PDB_Select [2]. The dataset which predicts 3-states are used to train the machine.

1.7 Performance measures

The performance of the secondary structure prediction can be measured by various methods. One of the commonly used method for the prediction of secondary structure of protein is Q3 (Q helix, Q sheet and Q coil).

$$Q_3 = (N_{match} / N_{total}) \text{ X } 100$$

Where, N_{match} is the number of correct predictions

N_{total} is the total number of predictions

2. Methodology

The obtained input vector from the encoding scheme is used as an input to the layer of support vector machine. Using the SVM, a hyper plane is created which separates the two classes as H/H~, C/C~ and E/E~.

Step 1: Finding the input vector

The algorithm for the creation of Markov transition matrix and the input sequence is given as:

1) The patterns of sequence are taken from the database and transition matrix is prepared using the sliding window.
Example, if the sliding window is 7 of a residue, the dimension becomes 7*20=140 dimensions per pattern.

2) The transition probability can be calculated using the frequency of the residues at each sequence. Example, considering the input sequence and the obtained output sequence:-

Input: F K V P E K V A A K W

Output: E E E C E E H C E E E

Transition probability can be calculated as

For input sequence:

	P_1	P_2	P_3	P_4	P_5	P_6	P_7
Seq1	0.143	0.286	0.286	0.143	0.143	0.286	0.286
Seq2	0.571	0.571	0.286	0.143	0.571	0.571	0.286

Step 2: Calculating the output sequence

For output sequence:

	P_1	P_2	P_3	P_4	P_5	P_6	P_7
Seq1	0.429	0.429	0.429	0.429	0.429	0.429	0.143
Seq2	0.571	0.286	0.571	0.143	0.571	0.286	0.571

Similarly we can find out the probabilities of occurrence from other residues. Using this calculated transition matrix, we can obtain the input vector for our input data and the actual output data.

Step3: Training the machine using the input data

The input data is trained using the Support vector machine layer which uses a case of multi class classifier as H/H~, E/E~ and C/C~. The residues are classified by subtracting the range of the classes defined initially. If we get the output as the value zero, we come to halt and a class is defined for the calculated value.

Example, if we have already defined range of class H as 0.15, C as 0.16, E as 0.12. For first value, we get output as 0.143-0.15= -0.007, 0.143-0.16= -0.017, 0.143-0.12= 0.023. Ignoring the negative values, as classifying using positive values, we classify the given input to class E which is the actual obtained class. Similarly we obtain the classes for other values.

Step 4: Testing of the machine using Test data

The test data is provided to the machine and we calculate the class for that sequence of using the given machine.

Step 5: Calculation of Accuracy

Accuracy is calculated using the performance measure, Q3. The obtained accuracy for the sequence given is 75.14%.

3. Results and Conclusion

SVM is a classifier that can be used for prediction of secondary structure of prediction that is a better learning approach for prediction with the use of hidden markov model transition matrix as a input and output encoding scheme. Support Vector Machine have outperformed in various application of bioinformatics achieving a higher accuracy in prediction which was not given by other methods like artificial neural network etc.

Support vector machine technique has a various advantages of using as it reduces over-fitting problem, no information of data set is lost and training occurs at least possible error. In this paper we use the concept of one versus one binary classifier with the use of sliding window concept.

Using the SVM, the accuracy obtained till now for the secondary structure prediction of protein is 78.20%. This accuracy can be improved by the removal of error in prediction and by adopting various other schemes of encoding.

4. REFERENCES

[1] Shivani Agarwal, Arushi Baboota and Deepali Mendiratta 2013. Design and Implementation of an Algorithm to

predict Secondary Structure of Proteins using Artificial Neural Network. International Journal of Emerging Research in Management and Technology.

[2]http://crdd.osdd.net/raghava/ccpdb/help.html#regsn

[3] Hae-Jin Hu, Yi Pan. 2004. Improved Protein Secondary Structure Prediction Using Support Vector Machine With a New Encoding Scheme and an Advanced Tertiary Classifier.

[4] Jung-Ying Wang. 2002Application of Support Vector Machines in Bioinformatics.

[5] Jian Guo, Hu Chen, Zhirong Sun, Yuanlie Lin. 2004. A Novel Method for Protein Secondary Structure Prediction Using Dual-Layer SVM and Profiles.

[6] Kasemsant Kuphanumat and Chidchanok Lursinsap. Highly Accurate Protein Secondary Structure Prediction by Combination of nth- order Markov Transition Matrix and Support Vector Machine

[7] Anjum B Reyaz-Ahmed. 2007. Protein Secondary Structure Prediction Using Support Vector Machines, Neural Networks and Genetic Algorithms.

[8] Sujun Hua and Zhirong Sun. 2001. A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach.

[9] Lipontseng Cecilia Tsilo. 2008. Protein Secondary Structure Prediction Using Neural Networks and Support Vector Machines.