Protein Secondary Structure Prediction

Priyanka B V, Rachitha K T, Sanchitha N, Srinidhi H S,

Abstract: ***Proteins are made up of basic units called amino acids which are held together by bonds namely hydrogen and ionic bond. The way in which the amino acids are sequenced has been categorized into two dimensional and three dimensional structures. The main advantage of predicting secondary structure is to produce tertiary structure likelihoods that are in great demand for continuous detection of proteins. This paper reviews the different methods adopted for predicting the protein secondary structure and provides a comparative analysis of accuracies obtained from various input datasets [1].***

Index Terms: ***Protein secondary structure, auto encoder, Bayes classifier, Margin Infused Relaxed Algorithm(MIRA), Deep Neural Residual Network (DeepNRN), PSI-BLAST, CullPDB, support vector machines, Position Specific Scoring Matrix(PSSM).***

I. INTRODUCTION

Protein secondary assembly is the three dimensional procedure of local segments of proteins. The two common secondary structure components are alpha helices and beta helices. It is very important to define a meaningful secondary structure of protein as it helps in providing a successful study of the relation between the protein structure and the amino acid sequence. Every protein secondary structure differs in their hydrogen bonding patterns, repeating turns, bridges and ladders [1].

Secondary structure is being used to understand how proteins interact with other molecules such as small molecules or ligands that can become a drug candidate. Secondary structure of proteins directs to the identification of a protein function. It is also helpful in the production of drugs, monitoring the functionalities of bacteria, to make a study on restricted enzymes. It is even used in predicting three dimensional protein structure. Site specific mutation experiments are also conducted using the secondary structure of proteins. Hence, Secondary structure plays a very important role. This Paper reviews various methods used to predict the secondary structure of protein.

II. LITERATURE SURVEY

Group Template Pattern classifiers is a method which is used to search patterns where the protein lengths are similar. It divides the provided training data set into many categories based on length which helps in building the prediction model [1].

Revised Manuscript Received on July 05, 2019.

Priyanka B V, Vidyavardhaka College of Engineering, Mysuru, India
Rachitha K T, Vidyavardhaka College of Engineering, Mysuru, India
Sanchitha N, Vidyavardhaka College of Engineering, Mysuru, India
Srinidhi H S, Vidyavardhaka College of Engineering, Mysuru, India
Pavankumar S P, Vidyavardhaka College of Engineering, Mysuru, India
Shashank N, Vidyavardhaka College of Engineering, Mysuru, India

The main datasets used are ASTRAL, CullPDB, which, together consists of 15696 proteins. The other data sets used are 25PDB, CB513, CASP9, CASP10, CASP11, CASP12. The pattern representation of the secondary structure of proteins from the above datasets is stored in a matrix called Position Specific Scoring Matrix (PSSM) and the respective software used is PSI-BLAST [2]. This software devices PSSM, and finds the region of similarities between the input data and the data which is already stored in the database.

The Support Vector Machines (SVMs) are an algorithm which helps in separating different classes of patterns and Vapnik developed this machine [3]. The main drawback is that long range interactions of the protein are not captured [2].

Auto encoder classifier consolidates radical gathering encoding and position-explicit scoring network (PSSM) making another encoding strategy for foreseeing the protein optional structure. Bayes classifier, single layer auto encoder and stacked auto encoder with two concealed layers are utilized. The protein highlights extraction is finished utilizing auto encoder [4].

The single layer auto-encoder separates 1500 highlights. The stacked auto encoder concentrates include in two layers. 1500 highlights are separated in the primary layer and 800 highlights in the second layer. The extreme gathering encoding strategy is utilized to encode amino acids arrangement dependent on the nearness of radical gatherings considering 42 features. Blosum62 framework is a variation of the position-explicit scoring grid.

Initial 20 segments of the Blosum 62 network are joined with radical gathering encoding to frame another encoding strategy. Database of auxiliary structure assignments for all protein passages in the protein information bank(DSSP) is utilized for structure rearrangements [5].

The auto-encoder is utilized with the end goal of protein highlights extraction [6-7] and forecast is finished utilizing Bayes classifier [8-9]. CB513 is the dataset utilized. The consequence of the precision of different classifiers for the best sliding window length is appeared in the beneath table 2.1.

Method name	Sliding window length	Accuracy
Bayes classifier	21	66.98%
Single-layer auto encoder	13	63.95%
Stacked auto encoder	13	67.96%

Table 2.1: result of various classifiers

The main drawbacks are stacked auto encoder has a less predicting accuracy.

The dataset is with less protein sequence

The main objective of the new deep neighbor residual network (DeepNRN) is to predict secondary structures of the proteins [10]. The DeepNRN architecture uses window size of 3. The neighbor residual unit is the main part of this network. This unit is connected in a short cut manner with two types which are more detailed than the previous units. There is a different block called struct2struct network [11], which helps in refining the output obtained from the DeepNRN network to make it look like a real protein. There are mainly three types of inputs used which are, sequence of the protein features, features of the profiles obtained from PSI-BLAST [12] and also from Habits [13].

The neighbor residual unit[NRU], which is the basic block of DeepNRN consists of convolutions and concatenation sequences which have two short-cuts. To reduce the cost, a hierarchical layer of convolutions is used. Every NRU consists of convolutions with a window size of three. The features with respect to the input, sent to the DeepNRN comprises of profile with sequences of protein drawn by PSI-BLAST and HHBlits. The datasets used are CullPDB which mainly helps in training the deep networks, CB513, CASP10, CASP11 and CASP12 which are used in testing and comparison. This method overcomes a machine learning problem called as vanishing-gradient problem. But the main drawback is that there are no interactions between the different residues that are in bound to the 3D structural space

Margin infused relaxed algorithm (MIRA) (MIRA) is utilized to keep the proteins from expanding the assortment of its structure by improving as far as possible [14]. Optional structure expectation of a protein has real significance for the tertiary model. Recognizable proof and forecast of irregular loops which are in the collapsed state in the optional structure of a protein have real significance in the organic investigation. The binomial appropriation is utilized to improve the tetra-peptide structure of a protein. CullPDB and the CB513 are the essential datasets utilized. By utilizing the 10-cross-approval strategy, MIRA calculation [17] is contrasted and understood 9 existing methodologies yet the outcome which is gotten from the MIRA calculation is more proficient than some other outcome set. Estimating the 3D structure of the protein will legitimately help in the expectation of the protein work. Optional protein structure will dependably go about as a middle of the road organize between the essential and the tertiary structure of a protein. The precise gauging of the 2D structure of a protein will offer ascent to the more exact and high goals of the tertiary structure of a protein. Optional structure of a protein is a plot by three-shapes, alpha-helix, beta-strand and arbitrary loop which are extricated from the neighbor protein folds.

Alpha-helix and the beta-strand are the prevailing auxiliary structure of a protein and they are gathered as a standard optional structure of a protein. A portion of the devices for assessing the optional structure of proteins are PSIPRED [15], JPRED [16], SPIDER2, S2D, RaptorX-SS8, PSSpred, Frag1D and some more.

Optional structure of a proteins are anticipated by utilizing the accompanying methodology: In the initial step, the dataset for preparing and testing is chosen; furthermore, the peptide or protein models that can truly replicate their central association with the qualities to be anticipated are planned; thirdly, an algorithmic technique is set up to capacity the projection; at long last, the outcomes are assessed utilizing cross-approval to survey the evaluated accuracy of the indicator.

III. PROPOSED METHOD

Neural network models are like brain where every neuron are connected to every other neuron. If the one neuron stimulate the stimulation is passed to another neuron and this continue to perform a particular task.

The Roast Sander dataset consists of wide range of proteins which consists of primary structure, secondary structure composition. Here we have taken a subset of protein dataset in its wide range.

We build a neural network to train the model so that it can identify the secondary structure that is whether it is helix, coil or beta plated as shown in figure 3.1. Due to the random nature of the neural network every time the neural network is processes we get the slightly different results.

A. Defining the Network Design

For this issue, we characterize a neural system with one information layer, one shrouded layer, and one yield layer. The info layer comprises of a sliding window for each contribution of the amino corrosive succession. The yield is anticipated dependent on the focal buildup of the window. A window size of 17 is utilized. Every window position is encoded utilizing a paired exhibit of size 20, having one component for every amino corrosive sort. As the amino acids comprise 20 distinct sorts of amino acids, the components that concur with the specific amino acids are set to 1. while the other amino acids are set to 0. thus the info layer is encouraged with 17x20 units. which means 17 distinct gatherings comprising of 20 units for each situation.

Next, we decide all the potential subsequences of protein arrangements relating to the sliding window measure by making the Hankel grid. where ith section speaks to the protein subsequences beginning from the ith position. what's more, the jth component to 1if the given position has a numeric portrayal which is equivalent to j.

The yield layer of the neural system comprises of a twofold plan of three units each speaking to one of the auxiliary states. By acquiring the auxiliary assignments of all subsequence with identified with the sliding window by thinking about the focal position in every window. the double estimations of curl are 1 0, sheet is 0 1 0, helix is 0 1.

When we characterize the information and target lattices for each grouping, we make an info framework, P, and target grid, T, speaking to the encoding for every one of the successions bolstered into the system.

B. Creation of Neural Network

This auxiliary structure expectation can be thought as an example acknowledgment where the system is now prepared so that if the given info succession coordinates the prepared arrangements dependent on the focal buildup and considering the sliding window then the outcome can be gotten. Henceforth it is comparable to design acknowledgment.

C. Training the Neural Network

For training, we use the Scaled Conjugate Gradient algorithm as The pattern recognition network. even though we have other

methods such as Deep Learning Toolbox documentation). for every cycle of training, we provide the training sequences using the sliding window which sends one residue at one time. we use the logsig transfer function to transform the signals from the inputs and this job is done by the hidden layer. This produces an output in the form of ones or zeros. The weights are adjusted accordingly to increase the accuracy and to reduce the error by looking into the target matrix. The training tool of the network displays the updates. Details such as the algorithm, error types and performance criteria are displayed.

Data overfitting is the main problem encountered during the training of neural network. overfitting means it does not learn how to generalize to new situations, but it directly tends to recollect the training examples. Early stopping is the default method used for this problem and the current training set is divided into three subsets which are training set, validating set and test set. Training set is used to compute the gradient and progress with weights and biases of the network. Validating set is to increase the over fitted data and the test set is used to look into the quality of the data set's division [16].

To divide the data into the three sets, we use the function train which divides into 60% of train set,20% each of validating and test set. We can divide the way we want using the function net.divideFunc (default dividerand).

D. Analyzing the Network Response

To analyze the network response, we examine the confusion matrix by considering the outputs of the trained network and comparing them to the expected results (targets).

E. Improvements in Neural Network for More Accurate Results

1. By increasing the window size the number of protein subsequences obtained are also more and helps in obtaining the more accurate results.

2. By increasing the number of hidden layers the network can be more sophisticated and helps in overcoming the over-fitting of data disadvantage.

3. Even though the number of datasets fed into the network is fairly large, the number of datasets can be enhanced which can train the network in a better way. This will eventually lead to better accuracy score.

IV. RESULTS

Input:

```
prompt =

    'enter the dataset value'

enter the dataset value6

A =

     6


id =

    '1CRN-1PLANTSEEDPROTEIN30-AP'

seq =

    'TTCCPSIVARSNFNVCRLPGTPEAICATYTGCIIIPGATCPGDYAN'
```

Figure 4.1 Input protein its primary sequence and secondary structure

Output:

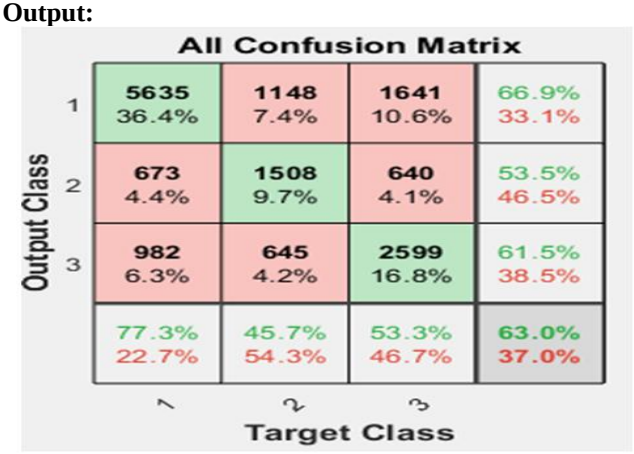

Figure 4.2 Confusion matrix

Utilizing an important arrangement of factors and information occurrences as appeared in figure 4.1, a neural system has been made. The system is then being utilized to foresee other information examples and the precision rate. At the point when a genuine positive information point is sure, that is a right expectation, called a genuine positive(TP). So also, when a genuine negative information point is named negative, that is valid negative(TN) In addition to it, when the true encouraging data point is categorized by network as undesirable, which is an improper calculation, it is known as false negative(FN). Likewise, when a factual negative data point is categorized as positive, it is classified as false positive(FP). It is denoted in the confusion matrix as shown in diagram 4.2. The amount of excess positions which were properly categorized for each organizational class are shown along diagonal cells. The number of residue positions that were misclassified are shown along the off-diagonal cells (example helical positions are projected as coiled locations). The diagonal cells are correctly classified that are corresponding to observations.

In each cell, both the number of observations and the percentage of the total number of observations are shown.

The percentages of all the examples predicted which belongs to each class that is correctly and incorrectly classified that are plot along the column on the far right of the graph. The above-mentioned metrics are often called the precision matrix (or positive predictive value) and false discovery rate, respectively. The percentages of all the examples belonging to each class that is correctly and incorrectly classified are shown along the row at the bottom of the plotted graph. These metrics are often called the recall (or true positive rate) and false negative rate, respectively. The overall accuracy will be shown in the cell in the bottom right of the plot.

V. CONCLUSION

The technique exhibited here predicts the basic condition of a given protein buildup dependent on the auxiliary condition of its neighbors.

Notwithstanding, there are further limitations when anticipating the substance of auxiliary components in a protein, for example, the base length of each basic component. In particular, a helix is relegated to any gathering of at least four touching deposits, and a sheet is appointed to any gathering of at least two adjacent buildups. To fuse this sort of data, an extra system can be made with the goal that the primary system predicts the basic state from the amino corrosive grouping, and the second system predicts the auxiliary component from the basic state. Using this novel method, we are able to increase the accuracy compared to the previous methods. This method is faster at predicting the structure when compared to previous methods. The number of datasets fed into the feed forward network is much larger than fed into the previous networks. We are able to output the different structures of the protein such as coil, helix and sheets which are represented as C, H and E respectively.

REFERENCES

1. Rost, B., and Sander, C., "Prediction of protein secondary structure at better than 70% accuracy", Journal of Molecular Biology, 232(2):584-99, 1993.
2. Holley, L.H. and Karplus, M., "Protein secondary structure prediction with a neural network", PNAS, 86(1):152-6, 1989.
3. [3] Kabsch, W., and Sander, C., "How good are predictions of protein secondary structure?", FEBS Letters, 155(2):179-82, 1983.
4. M. Spencer, J. Eickholt, and J. Cheng, "A deep learning networkapproach to ab initio protein secondary structure",IEEE/ACM Transactions on Computational Biology & Bioinformatics, vol. 12, 2015, pp. 103-112.
5. YihuiLiu ,Fuming Ma andJinyong Cheng,"A Novel Group Template Pattern Classifiers (GTPCs) Method in Protein Secondary Structure Prediction", 2017 3rd IEEE International Conference on Computer and Communications.
6. Zhang Shuai-yan, Liu Yi-hui and Cheng Jin-yong3, " The prediction of protein secondary structure based on auto encoder", 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD 2017).
7. Chao Fang, Yi Shang, and Dong Xu, " A New Deep Neighbor Residual Network For Protein Secondary Structure Prediction", 2017 International Conference on Tools with Artificial Intelligence.
8. Dr.D.Ramyachitra, R.Ranjani Rani, V.Kamalakkannan predicting secondary structure of protein using MIRA algorithms based on Tetra peptide structural word Proceedings of the Second International Conference on Inventive Systems and Control (ICISC 2018).
9. Akosua Busia, Jasmine Collins, Navdeep Jaitly, Protein Secondary Structure Prediction Using Deep Multi-scale Convolutional Neural Networks and Next-Step Conditioning. 4 Nov 2016.
10. Yuming Ma, Yihui Liu andJinyong Cheng Protein Secondary Structure Prediction Based on Data Partition and Semi-Random Subspace Method.
11. Yanchun Wang , Jinyong Cheng, Yihui Liu , Yehong Chen Prediction of protein secondary structure using support vector machine with PSSM profiles.
12. Wei Chu Zoubin Ghahramani Computational Neuroscience Unit, University College London, London, WC1N 3AR, UK David L. Wild, A Graphical Model for Protein Secondary Structure Prediction.
13. Sj¨olander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I. S., & Haussler, D. (1996). Dirichlet mixtures: A method for improving detection of weak but significant protein sequence homology. Computing Applications in the Biosciences, 12, 327–345.
14. Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensistivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research, 22, 4673–4680.
15. Yel, R. F., Lim, L. P., & Burge, C. B. (2001). Computational inference of homologous gene structures in the human genome. Genome Res., 11, 803–816.
16. Zhang, L., Pavlovic, V., Cantor, C. R., & Kasif, S. (2003). Human-mouse gene identification by comparative evidence integration and evolutionary analysis. Genome Res., 13, 1190–1202.

AUTHORS PROFILE

Ms. Priyanka B V is a student of eighth semester from computer science and engineering branch. She is pursuing her bachelor of engineering from Vidyavardhaka College of Engineering, Mysuru, Karnataka.

Ms. Rachitha K T is a student of eighth semester from computer science and engineering branch. She is pursuing her bachelor of engineering from Vidyavardhaka College of Engineering, Mysuru, Karnataka.

Ms. Sanchitha N is a student of eighth semester from computer science and engineering branch. She is pursuing her bachelor of engineering from Vidyavardhaka College of Engineering, Mysuru, Karnataka.

Ms. Srinidhi H S is a student of eighth semester from computer science and engineering branch. She is pursuing her bachelor of engineering from Vidyavardhaka College of Engineering, Mysuru, Karnataka.

Mr. Pavankumar S P is currently working as an assistant professor at Vidyavardhaka College of Engineering. He is having a teaching experience of 5 years. His areas of interest are Internet of Things, Bioinformatics, Digital Image Processing.

Mr. Shashank N is currently working as an assistant professor at Vidyavardhaka College of Engineering. He is having a teaching experience of 1 year. His areas of interest are Digital Image Processing.