ISSN No:-2456-2165

Comparative Analysis of Protein Secondary Structure Prediction Methods

[1]Priyanka. B. V, [2]Rachitha. K. T, [3] Sanchitha. N, [4]Srinidhi. H. S, [5] Pavan Kumar S P, [6] Shashank N
[5,6]Assistant Professor
Vidyavardhaka College of Engineering, Mysuru, India

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 predictions which are in great demand to the continuous discovery 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.

Keywords:- *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 structure is the three dimensional form of local segments of proteins. The two most common secondary structure elements 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.

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 direct 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]. 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 accuracy for 25PDB is 86.38% ,CB513 is 84.11%, CASP10 is 83.07%, CASP11 is 81.98%, CASP12 is 82.35%, CASP9 is 83.92%.The main drawback is that long range interactions of the protein are not captured.

Auto encoder classifier incorporates radical group encoding and position specific scoring matrix (PSSM) composing a new encoding method for predicting the protein secondary structure. Bayes classifier, single layer auto encoder and stacked auto encoder with two hidden layers are used. The protein features extraction is done using auto encoder [4].The single layer auto encoder extracts 1500 features .The stacked auto encoder extracts features in two layers. 1500 features are extracted in the first layer and 800 features in the second layer. The radical group encoding method is used to encode amino acids sequence based on the presence of radical groups considering 42 features.Blosum62 matrix is a variant of position specific scoring matrix. First 20 columns of Blosum 62 matrix are combined with radical group encoding to form a new encoding method .Database of secondary structure assignments for all protein entries in the protein data bank(DSSP) is used for structure simplification [5] .Auto encoder is used for the purpose of protein features extraction [6-7] and prediction is done using Bayes classifier [8-9] . CB513 is the dataset used. The result of the accuracy of various classifiers for the best sliding window length is shown as in the below table.

ISSN No:-2456-2165

Method name	Sliding window length	Accuracy
Bayes classifier	21	70.98%
Single-layer auto encoder	13	71.95%
Stacked auto encoder	13	72.96%

Table 1:- Result of Various Classifiers

Comparing the accuracies of various classifiers, the accuracy of single-layer auto encoder and the Bayes Classifier differs by 1.3%, which means single-layer auto encoder has higher accuracy. Similarly the accuracy of stacked auto encoder and the single-layer auto encoder differs by 1.39%, stacked auto encoder having the higher accuracy. 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 HHBlits[13].

The neighbor residual uni[NRU]t, 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 are used. Every NRU consists of convolutions with a window size of three. The features with repect 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. These predictions help in providing Q3 and Q8 accuracies. Q3 and Q8 accuracies measure the percentage of residues that are correctly predicted among the 3-stage and 8-stage respectively as shown the below table. The Q3 and the Q8 accuracies obtained from DeepNRN with respect to other state-of-art tools are 83.7% and 71.1%, respectively.

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.

	Q3	Q8
CB513	83.7	71.1
CASP10	85.6	75.33
CASP11	83.6	72.9
CASP12	81.6	70.8

Table 2:- Q3 and Q8 Accuracy Results (%)

Margin infused relaxed algorithm (MIRA) is used to prevent the proteins from increasing the variety of its structure by improving the accuracy limit [14]. Secondary structure prediction of a protein has a major importance for the tertiary model. Identification and prediction of random coils which are in the folded state in the secondary structure of a protein has a major importance in the biological analysis. Binomial distribution is used to improve the tetrapeptide structure of a protein. CullPDB and the CB513 are the basic datasets used. By using the 10-cross validation method, overall accuracy rate is increased by 90.76% but the predicted accuracy rate is 88.45%. MIRA algorithm[17] is compared with well-known 9 existing approaches but the result which is obtained from the MIRA algorithm is more efficient than any other result set.

Forecasting the 3D structure of protein will directly help in the prediction of the protein function. Secondary protein structure will always acts as an intermediate stage between the primary and the tertiary structure of a protein. The accurate forecasting of 2D structure of a protein will give rise to more accurate and high resolution of the tertiary structure of a protein. Secondary structure of a protein is outlined by three-forms, alpha-helix, beta-strand and random coil which are extracted from the neighbor protein folds. Alpha-helix and the beta-strand are the dominant secondary structure of a protein and they are grouped as a standard secondary structure of a protein. Some of the tools for estimating secondary structure of proteins are PSIPRED [15], JPRED [16], SPIDER2, S2D, RaptorX-SS8, PSSpred, Frag1D and many more.

Secondary structure of a proteins are predicted by using the following procedures: In the first step, the dataset for training and testing is selected; secondly, the peptide or protein models that can really reproduce their fundamental connection with the characteristics to be projected are formulated; thirdly, an algorithmic method is established to function the projection; finally, the results are evaluated using cross-validation to assess the estimated precision of the predictor.

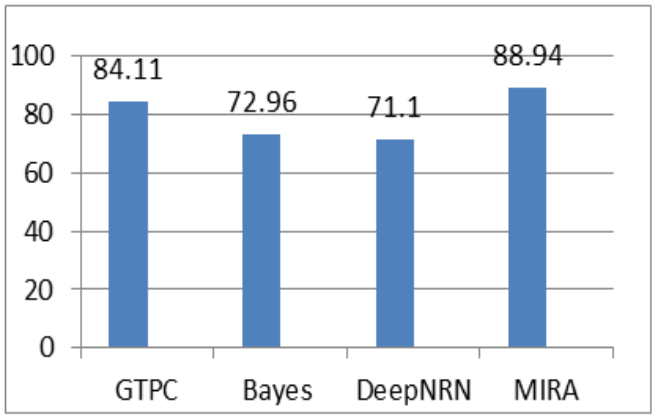
Algorithm	Q3		QR	
	CullPDB	CB513	CullPDB	CB513
MIRA	90.15	88.94	88.45	89.52
IDQD	87.8	86.85	86.1	87.57
DCRNN	87.9	85.3	77.9	76.9
ANN	76.9	75.4	71.2	68.4

Table 3:- Accuracy Values of Q3 and QR (%) for CullPDB and CB513 Datasets

ISSN No:-2456-2165

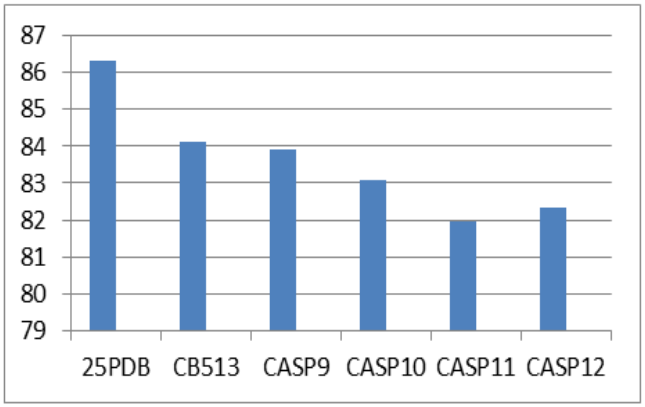
III. COMPARITIVE ANALYSIS

The comparative accuracies between the different classifiers for the same input dataset is shown in the below graph 1. Margin infused relaxed algorithm(MIRA) gives the highest accuracy for the input dataset CB513 which is 88.94.



Graph 1:- Comparison of Accuracy of Classifiers for Input Dataset CB513.

When the accuracies between the different datasets used in the Group Template Pattern Classifiers(GTPCs) are compared, 25PDB dataset has the highest accuracy of 86.31 as shown in the graph 2.

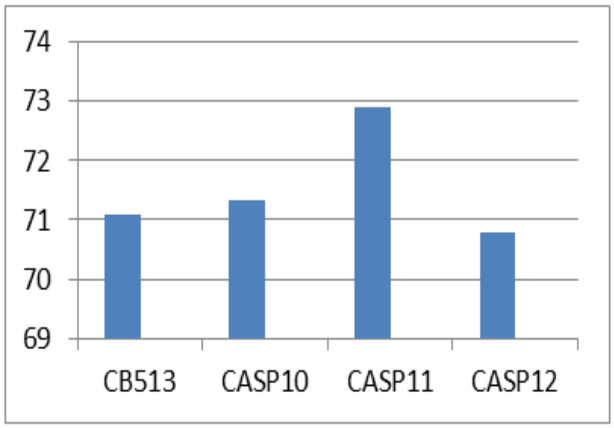


Graph 2:- Comparison of Accuracy of GTPC for Different Input Datasets.

IV. CONCLUSION

From the above methods, it can be inferred that there is difficulty to capture long range interactions produced by the whole protein, there is less accuracy and protein sequences are limited. Some protein properties such as solvent accessibility, protein order and disorder, region cannot be predicted and even the datasets used is limited. Henceforth, in order to overcome the above drawbacks, a novel method has to be designed and implemented.

When the accuracies between the different datasets used in the Deep Neural Residual Network(DeepNRN) are compared, CASP11 dataset has the highest accuracy of 72.9 as shown in the graph 3.



Graph 3:- Comparison of Accuracy of DeepNRN for Different Input Datasets

REFERENCES

[1]. 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.
[2]. D. T. Jones, “Protein secondary structure prediction based on position-specific scoring matrices,” J. Mol. Biol., vol. 292, 1999, pp.195–202.
[3]. V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1992.
[4]. YihuiLiu ,Yuming 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.
[5]. W. Kabsch□C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features Biopolymers,1983, 22(12): pp: 2577□2637.
[6]. K. X. Zhang, Unsupervised learning of Chinese lexical features based on automatic encoder. Chinese Journal of information science,2013, 27(5): pp: 1-7.
[7]. D. E. Rumelhart, G. E. Hinton, Eilliams R J. Learning representations by back-propagating errors. Nature, 1986, 323: pp: 533-536.
[8]. S. Theodoridis, K. Koutroumbas, Pattern Recognition, Third Edition,Academic Press, 2006.
[9]. Raymer M L, Doom T E, Kuhn L A, et al. Knowledge discovery in medical and biological datasets using a hybrid Bayes classifier/evolutionary algorithm. IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics, 2003, 33(5):802-813.
[10]. 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).
[11]. Rost, Burkhard, and Chris Sander. "Prediction of protein secondary structure at better than 70% accuracy." *Journal of molecular biology* 232, no. 2 (1993): 584-599.

[12]. Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic acids research* 25, no. 17 (1997): 3389-3402.
[13]. Remmert, Michael, Andreas Biegert, Andreas Hauser, and Johannes Söding. "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment." *Nature methods* 9, no. 2 (2012): 173-175.
[14]. 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.
[15]. Kevin Bryson, Liam J. McGuffin, Russell L. Marsden, Jonathan J. Ward, Jaspreet S. Sodhi and David T. Jones, “Protein structure prediction servers at University College London”, Nucleic Acids Research, Vol. 33, Web Server issue, W36–W38, 2005.
[16]. Alexey Drozdetskiy, Christian Cole, James Procter and Geoffrey J. Barton, “JPred4: a protein secondary structure prediction server”, Nucleic Acids Research, Vol. 43, Web Server issue W389–W394, 2015.
[17]. Crammer Koby and Singer Yoram, "Ultraconservative Online Algorithms for Multiclass Problems". Jnal of Mache Learn Res. 3: 951–991, 2003.