International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

Deep Learning Mechanism Augmented with 16-Hybrid Cellular Automata For Secondary Structure Prediction

Pokkuluri Kiran Sree

Abstract*: A protein plays various role in our human body like cellular development, reproduction, endurance and regulation of human body. Based on the structure of the genes we can extract lots of information regarding the human body. It is very easy to extract lots of information from a structure than a sequence. Identifying the protein structure helps in drug design. The secondary structure, to some extent tells about the effect of amino acid changes and explains the reason for the disease of an individual. A doctor can suggest medicines without any side effects to a patient based on the protein structure acquired from DNA.We have developed a classifier DL-16-MACA which can predict the secondary structure of an amino acid sequence of different lengths. In this prediction we have considered three classes Helix (H), Strands (E), Coiled(C). For Helix class the sensitivity, percentage accuracy is 0.923 and 90.6% respectively. For Strands class the sensitivity, percentage accuracy is 0.852 and 85.55%respectively. For Coiled class the sensitivity, percentage accuracy is 0.789 and 77.1% respectively. The percentage accuracy when tested with PDB datasets is 85.4% which substantially comparable with existing literature.

Keywords: **Cellular Automata, Deep Learning, Secondary Structure.**

I. INTRODUCTION

Secondary structure prediction is termed as a very important problem in bioinformatics. Based on the structure of the genes we can extract lots of information regarding the human body. As per the existing research, it is very easy to extract lots of information from a structure than a sequence. Identifying the protein structure helps in drug design.

Cellular Automata is set of cell on a grid, where sixteen cells are considered as a state and the rules are applied to make the transitions. A cellular automaton is a versatile classifier which performs even better when augmented with innate classifier with deep learning. The basic CNN classifiers are modified considerably to address the problem of secondary structure prediction.

We have considered modified CNN technique augmented with 16- Hybrid cellular automata to predict the secondary structure of a protein. In the section II summary of literature survey was provided, section III consists of the design of the classifier, section IV consists of implementation details with comparisons with the existing literature.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
Dr P.Kiran Sree*, Professor, Dept of CSE, Shri Vishnu Engineering College for Women, Bhimavaram, drkiransree@gmail.com

II. LITERATURE SURVEY

Many authors have proposed various models for predicting the secondary structure of the human protein. Garnier et al. has used amino acid frequency [1] to predict the secondary structure, with an accuracy of 60%. Robust et al. have proposed a method PHD[2] which depends on sequence family alignment with Neural Networks for secondary structure prediction with an accuracy of 78%. Muguffin et al. have proposed a method PSIPRED [3], which depends on PSI-Blast profiles with Neural Networks for secondary structure prediction with an accuracy of 78%. These two methods considered long range interactions to build a classifier based on NN. Perrakis et al. has proposed a method based on sequence family alignments with iterative structure refinement, which is also promising. Methods reported in [4], [5], [6], [7] and [8] for secondary structure predictions are also reviewed.

After reviewing various papers from the literature, we strongly indentified a room for a new method to address this problem with more preciseness working on alpha helix and the beta sheet. We conclude representing these sheets with 16-CA and process these with the corresponding rules(Complemented & Non Complemented)

III. DESIGN OF DL-16-MACA FOR PROTEIN SECONDARY STRUCTURE PREDICTION

For addressing the problem of secondary structure prediction of protein, we use 3-neighborhood, p-state cellular automata named as AIS-PSMACA. The rules reported in deep learning will model every amino acid of the protein sequence. These parameters DL-16-MACA tree, depth (d), the number of transitions (t), repeated transitions (rt) and unique transitions (ut) are used to predict the secondary structures. For the analysis of structural data available, the datasets are retrieved from PDB[168] (Protein Data Bank). The physical parameters in the data sets are mapped to DL-16-MACA model. Indirect mapping between the physical parameters and CA parameters will occur. In the context of pattern classification the input to DL-16-MACA will be an amino acid sequence and the output will be any one of the secondary structures of the protein i.e. Helix (H) or Strands(S) or Coiled(C).

Datasets The datasets are retrieved from PDB [8] (Protein Data Bank). We have chosen 11660 protein sequences randomly from PDB, each of length 1141.

Among them, 3000 sequences are used for constructing the tree, 5500 are used for mapping the CA constraints to the physical structure of protein, 2000 sequences are used for checking the accuracy of the constructed structure of DL-16-MACA and the rest of 600 sequences are used for testing the proposed classifier.

Mapping of Physical Parameters with DL-16-MACA model

1. Depth(d) is the number of steps taken by the encoded amino acid to reach a recognized basin when processed with a rule vector.
2. The number of transitions (t) is the number of times the same intermediate node is visited when processed with a rule vector.
3. Repeated transitions (rt) is the number of wasted transitions between two intermediate nodes.
4. Unique transitions (ut) are the number of unique transitions between two intermediate nodes, which leads to the prediction of Helix (H), Strands (E) and Coiled(C).

A. ALGORITHM (TRAINING)

This algorithm deals with training of amino acid sequences extracted from PDB to predict the secondary structure of human protein. This algorithm is carefully articulated to build the tree and also the important mapping tables. The rules which are reported in CA are used and extra four parameters are added. The maximum depth of the tree is fixed as 10(threshold). The output of the algorithm is sets of attractor basins and the mapping tables.

Algorithm-DL-16-MACA training

Input: Amino acid sequences from PDB

Output: Mapping tables (Maps structural equivalents to the parameters of CA) and attractor basins.

Step 1: Read the amino acid sequence in the multiples of three.
Step 2: Encode the sequence.
Step 3: Take a rule vector (Highest Fitness Rule) and apply this to get mapping tables.
Step 4: Apply the rule vector again to form an DL-16-MACA tree with 3cells, 6 attractors and 3 classes.
Step 5: For every transition note d (depth), t (transitions), ut(unique transitions), rt (repeated transitions).
Step 6: For all 20 amino acids, map the structural and CA parameters.
Step 7: When all the 20 amino acids are processed and 6 attractor basins are formed, output the mapping tables and stop.

B. ALGORITHM (TESTING)

The main aim of this testing algorithm is to direct the amino acid sequences into a set of basins formed in 8.3.3.1 based on the mapping table and information base developed. The input to this algorithm will be an amino acid sequence and output will be one of the possible three classes (Helix (H), Strands (E), Coiled(C)).

Algorithm-DL-16-MACA testing

Input: Amino Acid Sequence

Output: Class of the sequence (H/E/C)

Step 1: Read the amino acid sequence in the multiples of three.

Step 2: Encode the sequence.

Step 3: Take a rule vector and based on the mapping function formulated in algorithm , distribute the sequences into any one of the basins.

Step 4: For every transition, note the values for d (depth), t (transitions), ut (unique transitions), rt (repeated transitions). When d>threshold go to step 6.

Step 5: For all the amino acid entities mark the classes based on mapping of physical parameters and CA parameters (information base).

Step 6: Stop.

IV. EXPERIMENTAL RESULTS OF DL-16-MACA

A. OUTPUT

The datasets used to process the classifier and evaluate our works are taken from Protein Data Bank (PDB). We have also considered data sets from Protein Data Sets to verify the accuracy when tested with different dataset.

In the output-1, we can see the input was given as an amino acid sequence of length 141 and the possible three classes are predicted for each input.

Example : Consider the subsequence of the input say MNIFEMLRID

The subsequence is of length 10.

The output is CCHHHHHHHH.

To display this output, the probabilities of each alphabet represent any of these three classes in amino acid is calculated as shown in table 8.

The data we use to adjust the network weights is called the TRAINING SET. To make sure we have not over-fitted our network to our training set, we should test the network on a completely separate TESTING SET. This splitting of training and testing data is called CROSSVALIDATION and is an important concept in statistics and machine learning as it allows us to predict how well a method is likely to work on entirely new data.

OUTPUT -1:

Input: Sequence_human_Kiran_Protein_2nitt
MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNSLDAAKSELDKAIGRNTNGVITKDEAEKLFNQDVDAAV RGILRNAKLKPVYDSLDAVRRAALINMVFQMGETGVAGFTNSLRMLQQKRWDEAAVNLAKSRW
Sequence Sequence_human_Kiran_Protein_2jntuh length 141
Output: (C for Coiled, H for Helix, E for Strands)
CCCCCHHHHHHCEEEEEECCCCCEEEEEECEEECCCCCCHHHHHHHHHHHHHHHCCCCCCHHHHHHHHHHHHH HHHHHHHHHHHHCCCCCCCCCHHHHHHHHHHHHHHHHCCCCCCCCHHHHHHHHHHHHHHHHHHCCCC

Table 1: Computation of AIS-PSMACA

SNO	Amino Acid Sequence	Helix(H)	Sheet(E)	Coiled(C)	Prediction
1.	M	0.062	0.123	0.815	C
2.	N	0.070	0.169	0.761	C
3.	I	0.345	0.323	0.332	H
4.	F	0.626	0.261	0.113	H
5.	E	0.702	0.199	0.099	H
6.	M	0.617	0.270	0.113	H
7.	L	0.600	0.235	0.165	H
8.	R	0.536	0.294	0.170	H
9.	I	0.655	0.186	0.158	H
10.	D	0.628	0.158	0.214	H
11.					

B. PERFORMANCE OF DL-16-MACA

The results shown in this section are calculated for 4600 sequences extracted from PDB, which are of length 141. We have shown the frequencies of helix, strands and coiled categories in the dataset. The sensitivity and precision are calculated as per the equations in 6.1. The highest sensitivity (0.923) and precision (0.889) are reported as helix class as shown in table 8.4 and figure 8.3

Table 2: Performance of AIS-PSMACA

Class	Sensitivity	Precision	Frequency	Accuracy
H(Helix)	0.92	0.889	0.489	0.90
E(Strand)	0.85	0.859	0.295	0.85
C(Coiled)	0.78	0.754	0.216	0.77

The performance of DL-16-MACA is compared with GOR[9], PSIPRED[10], PHD[11], JPred[12] and R5NCA[13] as shown in table2 and figure .GOR reports very less prediction accuracy of 65%, as it depends only on amino acid frequency

.

Table 3: Comparison of DL-16-MACA with standard Techniques

Prediction Method	Prediction Accuracy
GOR[159]	65%
PSIPRED[161]	71%
PHD[160]	70%
JPred[167]	78%
R5NCA[169]	88%
DL-16-MACA	85.4%

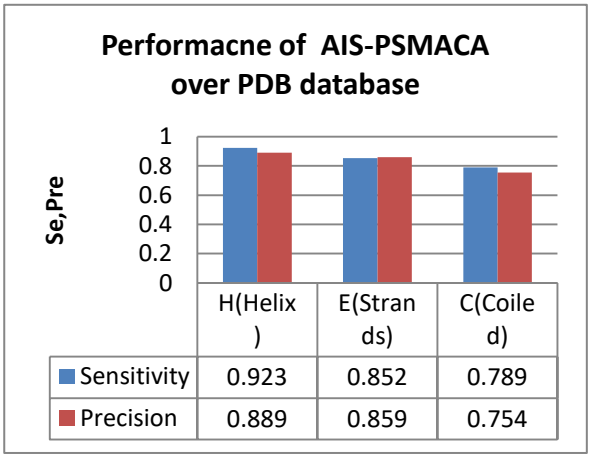

Figure 1 Sensitivity, Precision calculation for the three classes

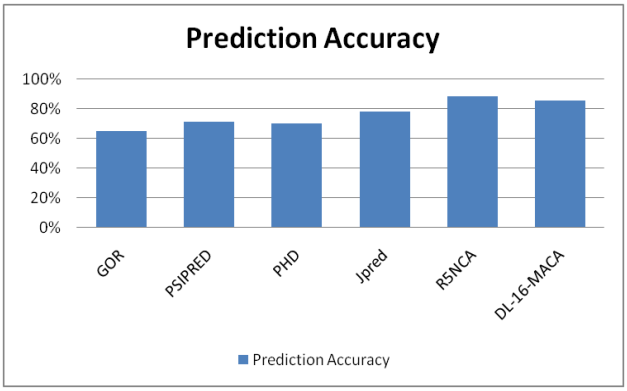

Figure 2: Comparison of DL-16-MACA with Standard Methods

V. CONCLUSION

The learning mechanism employed in DL-16-MACA predicts Helix class with high accuracy, but is the not the case in Stands and Coiled classes The lesser accuracy in prediction is due to the poor mapping of physical parameters with the CA parameters . DL-16-MACA prediction for Stands and Coiled classes is to be explored for correction. We have successfully developed a versatile classifier which can identify the secondary structure with an accuracy of 87%. In future we are striving to use this framework to predict the complete protein structure.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to our beloved Principal **Dr. G. Srinivasa Rao** , Vice-Principal **Dr. P. Srinivasa Raju,** HOD-CSE **Dr V. Purushothama Raju,** of Shri Vishnu Engineering College for Women (A), Bhimavaram for their constant motivation & support in providing facilities to pursue the research.

REFERENCES

1. Cuff, J.A., Clamp, M.E., Siddiqui, A.S., Finlay, M. and Barton, G.J., 1998. JPred: a consensus secondary structure prediction server. Bioinformatics (Oxford, England), 14(10), pp.892-893.
2. Rost, Burkhard. "Protein secondary structure prediction continues to rise." Journal of structural biology 134, no. 2-3 (2001): 204-218.
3. Hirokawa, Takatsug, Seah Boon-Chieng, and Shigeki Mitaku. "SOSUI: classification and secondary structure prediction system for membrane proteins." Bioinformatics (Oxford, England) 14, no. 4 (1998): 378-379.
4. Cole, Christian, Jonathan D. Barber, and Geoffrey J. Barton. "The Jpred 3 secondary structure prediction server." Nucleic acids research 36, no. suppl_2 (2008): W197-W201.
5. Rost, B., Sander, C. and Schneider, R., 1994. Redefining the goals of protein secondary structure prediction. Journal of molecular biology, 235(1), pp.13-26.
6. McGuffin, Liam J., Kevin Bryson, and David T. Jones. "The PSIPRED protein structure prediction server." Bioinformatics 16.4 (2000): 404-405.
7. Sree, P. Kiran, and Inampudi Ramesh Babu. "Identification of Protein Coding Regions in Genomic DNA Using Unsupervised FMACA Based Pattern Classifier." IJCSNS 8.1 (2008): 305.
8. Sree PK, Babu IR. Improving quality of clustering using cellular automata for information retrieval. arXiv preprint arXiv:1401.2684. 2014 Jan 13.
9. Sree, Pokkuluri Kiran. "AIS-INMACA: A Novel Integrated MACA Based Clonal Classifier for Protein Coding and Promoter Region Prediction." J Bioinfo Comp Genom 1 (2014): 1-7.
10. Sree, Pokkuluri Kiran, Inampudi Ramesh Babu, and Smt SSSN Usha Devi Nedunuri. "HMACA: Towards Proposing a Cellular Automata Based Tool for Protein Coding, Promoter Region Identification and Protein Structure Prediction."
11. Sree, P. Kiran, I. Ramesh Babu, and NSSSN Usha Devi. "Investigating an Artificial Immune System to strengthen protein structure prediction and protein coding region identification using the Cellular Automata classifier." International journal of bioinformatics research and applications 5.6 (2009): 647-662.
12. Sree, Pokkuluri Kiran, Inampudi Ramesh Babu, and SSSN Usha Devi Nedunuri. "PRMACA: A promoter region identification using multiple attractor cellular automata (MACA)." In ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India-Vol I, pp. 393-399. Springer, Cham, 2014.
13. Sree, Pokkuluri Kiran, and Nedunuri Usha Devi. "Achieving Efficient File Compression with Linear Cellular Automata Pattern Classifier." International Journal of Hybrid Information Technology 6.2 (2013): 15-26.

AUTHORS PROFILE

Dr Pokkuluri Kiran Sree obtained his B.Tech degree in Computer Science & Engineering with distinction from JNTU Hyderabad and M.E in Computer Science & Engineering with distinction from Anna University. He has obtained his Ph.D in Artificial Intelligence from Jawaharlal Nehru Technological University-Hyderabad. His areas of interests include Cellular Automata, Parallel Algorithms, Artificial Intelligence, and cloud computing. He was the reviewer for some reputed International Journals and IEEE Society Conferences on Artificial Intelligence, Image Processing and Bioinformatics. He has published 86 articles in various international journals and conferences. He has authored six text books on Artificial Intelligence. He is working as Professor in the department of CSE at Shri Vishnu Engineering College for Women, Bhimavaram.