International Journal of Scientific Engineering and Technology
Volume No. 12, Special Issue No. 1

ISSN : 2277-1581
29 Dec. 2023

Predictive Analysis of Protein Secondary Structure Employing Neural Network Approach

Roli Shrivastava, Shatendra Kumar Dubey
Department of Information Technology,
NRI Institute of Information Science and Technology, Bhopal (M.P), India
roli.rlrl@gmail.com,shatendradubey@gmail.com

Abstract: There are many protein bio-structures available. It is an challenging problem to predict the suitable protein structure for the drug analysis and the human bio analysis. network (NN) based modeling and simulation approach is proposed in this paper to address the structure prediction approach. The proposed method uses the log sigmoidal function to train the NN. The Rost-SanderDataset is used for the NN based prediction methodology. The design parameters and NN layer parameters are varied for improved performance. The performance is evaluated based on the ROC curves and the confusion metrics for the three independent classes of data.

Key words Protein structure Prediction, Neural Network, ROC, Training, Validation, Confusion matrices

1. Introduction

Proteins are extremely adaptable bio-molecules [1]. According to a functional standpoint, precise knowledge about protein structure is critical. Protein structure prediction models a protein's 3D structure from its amino acids sequencing employing knowledge-driven techniques. In the recent times many methods have been proposed to investigate the thermal stability [1-9] and the structural analysis [10-15] of the various protein compositions.

. This study is a fundamental instance meant to demonstrate the methods for configuring a NN to predict the SS of proteins. Their configuration as well as training procedures are not intended to offer the optimal answer for the given situation. NN based models are widely utilized in a range of applications, including automatic pattern recognition, which seek to imitate the information processing which takes place in the brainThe PSSP is widely being used for the bio modeling the most frequent applications of protein structures are shown in the Figure 1.

The main Goal of this paper is to design the improved NN based prediction model for solving the protein secondary structure prediction (PSSP) problem. These application uses PSSP since methods are short and fast in predicting this process requires significant fewer budgets comparative to state of art bio and chemical processes. And also the prediction accuracy is also far better

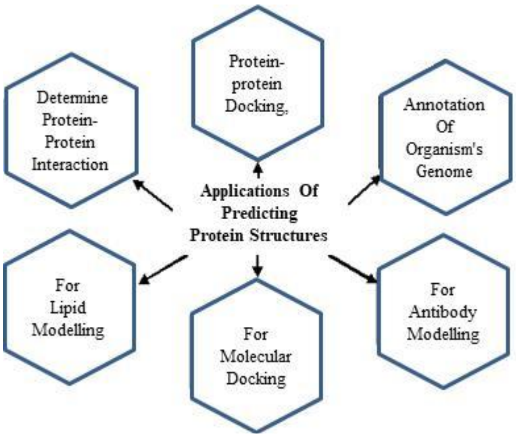

Determine Protein-Protein Interaction
Protein-protein Docking,
Annotation Of Organism's Genome
Applications Of Predicting Protein Structures
For Lipid Modelling
For Molecular Docking
For Antibody Modelling

Figure 1 wide scope of Image based Watermarking

Contribution of Work

This paper demonstrates a secondary structure prediction approach which makes utilization of a feed-forward neural network as well as the deep learning toolbox. In the proposed approach the window size is increase to around 25% and the mask size of the feed forward NN layer architecture is increased to 5 and it makes the significant improvement in ROC curve efficiency. This alloys the more time for training and validation. The mean square error is estimate d for different epochs. Paper evaluated the performance for the three different set of structural classes based on the confusion matrices.

2. Review of Related Works

Therehas been lot of research carried out for protein thermal stability measure and the structural prediction. Our prime concern is to review both the case in this section. Although the case study is based on specific study of protein secondary structures predictions. The broad classification of the various protein structure prediction methodologies are shown in the Figure 2. The prime concern of the paper is on the learning based prediction. The contributions in recent deep learning based prediction approaches are reviewed in the section.

A. Thermal Stability Review

Minoru Saito et al [1]had designed by enhancing techniques of simulation of molecular dynamics (MD) and free energy disturbance computations, melting-temperature changes of mutation proteins are effectively and highly accurately estimated. In order to run mathematical simulations, far-reaching Electrostatic connections were first explicitly determined via the Particle-Particle and Particle-Cell

(PPPC) approach, which avoided trimming the relationships as would have happened with the traditional cutoff approach.

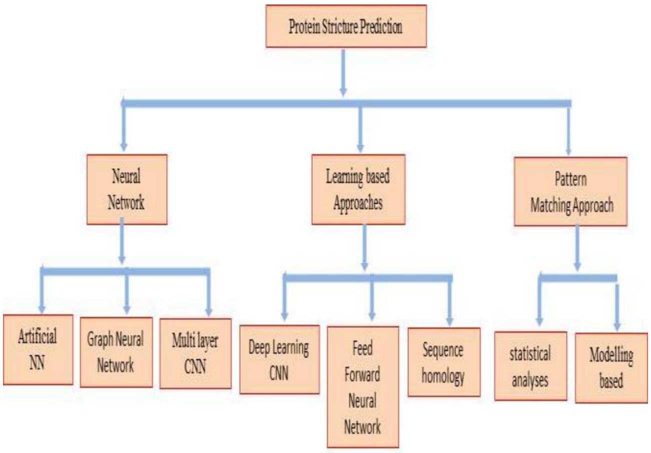


Figure 2 Classification of the Protein Structure prediction approaches

Secondly, the Accepting Ratio technique (ARM) was utilized to assess the intrinsic energy differential within the recombinant and the normal form proteins, rather than the standard free energy perturbation technique (FEPM).

Eric A. Franzosa et al [2]preseted mechanisms associated with protein bending, operation, and development are significantly impacted by the long-term viability of a molecule's native form. This study examines the connections between a protein's structure and sequence-based characteristics and its experimentally assessed security, as evaluated by temperature-dependent melting studies.

ChulYeh et al [3] have preseted the relationship between the macroeconomic and nano-scale characteristics of protein and its disorder-order phase change of folding has been largely attributed to their heat capacity. The intricate and variable structure of structure in proteins, complex solvent surroundings, and configurationally averages make it difficult to calculate using atomistic modeling techniques. They have presented a comparison of simulation frameworks that vary in their explicit solvent descriptions and field force precision in order to further clarify these variables on estimating protein heat conductivity.A good descriptive report of the protein thermal stability is given in the [4]. In this sequence Anthony P. Russell et al [5] have derived the equation in the shape of a diagram is provided to calculate the nucleotide percentage of melting DNA. The above equation was used to determine the content of patterns that melt at various locations on the boiling curves of mammalian DNA, Clostridium coil DNA, and Candida DNA. The percentage of the two amino acids (AT) in the melting strands drops exponentially with warmth as the DNA melts. The base pair ratio in the DNA determines the typical composition of strands that melt within a particular region of their melting curve.

Yang Yang et al [6]stated that a crucial characteristic of peptides that has numerous physiological ramifications is their stability. Understanding the stability of proteins is crucial for a variety of purposes, including applications in biotechnology, cell equilibrium, protein cleansing, and structural identification. The identification of temperature stability by experimentation has been laborious, and there is a scarcity of data. The development of mass spectroscopy and restricted proteolysis techniques has made it easier to produce more comprehensive data on protein expression in cells retention. Wayne J. Becktel et al [7] proposed plotting the unrestricted power of unfolded as an indicator of the ambient temperature is known as a protein's instability curve. The majority of proteins undergo a significant but roughly constant shift in thermal resistance during denaturation, also known as unfolding. The majority of the noteworthy characteristics of stable protein profiles can be deduced from the fact that folding constitutes a two-state operation.

MattiaMiotto et al [8]without of any prior knowledgehave introduced a noval graph-theoretical approach to evaluate thermal endurance based just on architecture. Our method uses populations of contact networks for assessing protein molecules, which we define as energy-weighted diagrams.

Shashi Kumar et al [9] in their research have tried to learn more about SazCA's thermodynamic durability and how it affects the structure and unwinding of proteins. We provide computational models of water-solvated SazCA at 293–393 K to examine the connection between suppleness and thermal stability. According to our structural research, the protein remains most structurally stable around 353 K, and above that point, the chains of protein become extremely flexible. a hydrogen bond analysis revealed larger liquid media accessibility of slippery surface residue for configurations over 353 K. Peishan Huang et al [10]preseted a method that is frequently used to create catalysts with industrial relevance is the modification of proteins to increase their thermal endurance. Work is still ongoing in the field of creating new test data and mathematical instruments to direct these technical endeavors.

Review of PSSP

L. Howard Holley et al[11] have presented the neural network-based approach for predicting the extracellular structures of protein is given. The neural network was trained to identify the relationship among the amino acid sequence and secondary structural features using a collection of samples of 48 proteins with known structures.ShengWang et al [12] concluded that predicting the secondary structure (SS) of proteins is crucial for understanding how they operate. The most effective predictors can currently achieve ~80% Q3 reliability when using simply the order of (profile) info as the starting point feature; this has not improved over the last ten years. For the purpose of predicting protein SS, we now use DeepCNF (Deep Convex Neural Field).

Since Proteins Secondary Protein Structure Predictions (PSSP) is one of the most difficult tasks in biology, numerous approaches have been put out in an attempt by Ali Abdulhafidh Ibrahim et al [13] to address it by attempting to produce predictions that are more accurate. This research aims to create and deploy a smart system that uses five different neural network (NN) models to forecast the protein's second-order structure based on its main amino acid sequencing.

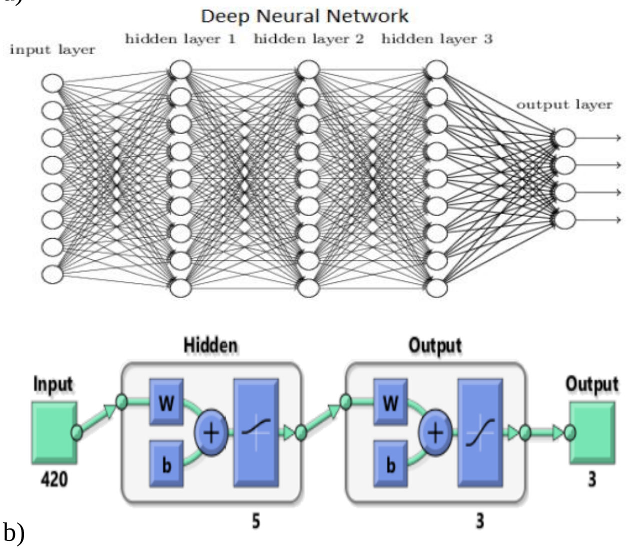
WafaaWardah et al [14]have addressed the research for in silico secondary structure of proteins prediction is contained in the scientific literature. Artificial neural network-based techniques make up a sizable portion of this library; this branch of computer learning and AI is becoming more and more well-liked across a range of use areas. Romana Rahman Ema et al [15] designed an approach to enhance forecasting efficiency, Hybrid Recurrent Neural Networks (HRNN) were suggested in this study for forecasting of secondary protein structures. The goal of the effort is to forecast the second-order structure of peptides and derive a highly precise answer that is amenable to mathematical modeling.

Mt. AkhiKhatun et al [16] compared the 3 machine learning-based prediction techniques have been presented in this research to predict the primary structure of peptides. Convolutional neural network neural networks' (CNN) design has enhanced these forecasting techniques. The function of activation that has been employed is the rectangular linear unit (ReLU). Since existing approaches are either computationally complex or too complex for implementation thus PSSP is still a open challenge.

3. Proposed NN Hidden layer Arrangements

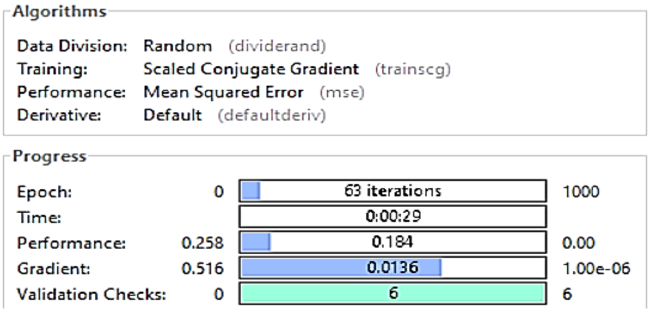
Paperhas defined a NN with one input layer, two hidden layer, and one output layer for the current challenge. Every input sequence of amino acid has a sliding window encoded by the input layer, while a prediction is produced regarding the structural condition of the core residue inside the window. The NN architecture is shown in Figure 3.

a)



b)

Figure 3 representation of the Hidden and output NN arrangements


Algorithms

Data Division: Random (dividerand)
Training: Scaled Conjugate Gradient (trainscg)
Performance: Mean Squared Error (mse)
Derivative: Default (defaultderiv)

Progress

Epoch:	0	63 iterations	1000
Time:		0:00:29	
Performance:	0.258	0.184	0.00
Gradient:	0.516	0.0136	1.00e-06
Validation Checks:	0	6	6

Figure 4 training performance and timings

The respective training performance and timings utilization are presented in the Figure 4. In accordance with the statistical association among the secondary structure (SS) of a certain residue location with the eight residues on each side of the prediction point, we select a window of size 21. A binary array of size 20 is used to encode each window position, with one element representing each type of amino acid. The element corresponding to the kind of amino acid at the specified position is set to 1 in each batch of 20 input. Therefore, input layer has R = 21x20 as the input units

4. Proposed System Architecture

The proposed design block diagram is shown in the Figure 5 representing sequential modeling steps.

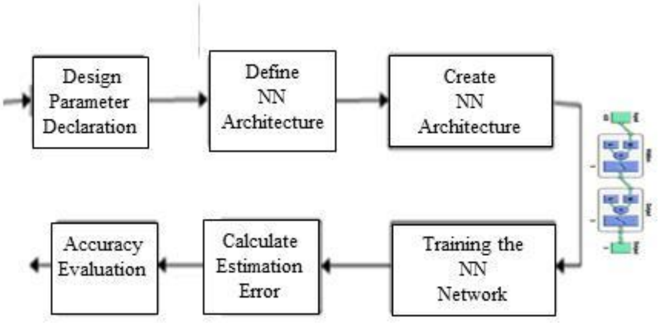


Figure 5 block diagram of Watermark embedding

5. Results of the PSSP

The proposed method uses the log sigmoidal function to train the NN. The Rost-Sander Dataset is used for the NN based prediction methodogoloy. The design parameters and NN layer parameters are varied for improved performance. Proteins with structures spanning a comparatively broad variety of domain types, structure, and length comprise up the Rost-Sander data set.The results of the production stage are preseted in the three pass. In first pass the arrangement of data a=used for the class vise production Training, and Validation are preseted. In second pas the mean square error (MSE) is estimated to determine the best possible solution for epoch. Finally the ROC curves and confusion matrix are preseted for the production.

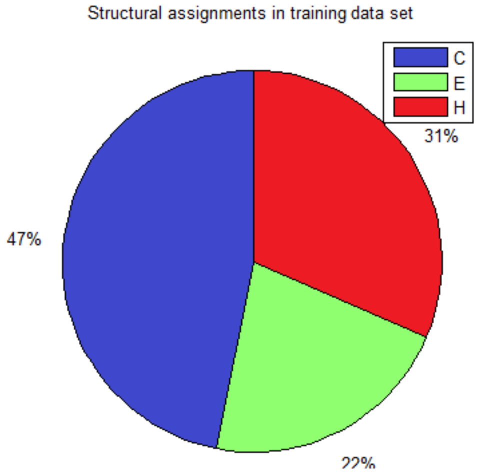


Figure 6 the ratio of the structure arrangements used for training of data.

The ratio of the structure arrangements used for training of data used for the pridction is shown in the Figure 6 the similar percentage of data breackup s used for the vaidation phase of the NN. The maximum data paret is usedfor the class1 and this can be clearly reflected in the pricition accuarcy and beter resulst.

Parametric evaluation

In order to predict the best optimal possible results the mean square error (MSE) is estimated and plotted against the number of epochs as shown in the Figure 7. The 19 epochs are used for the predictions.

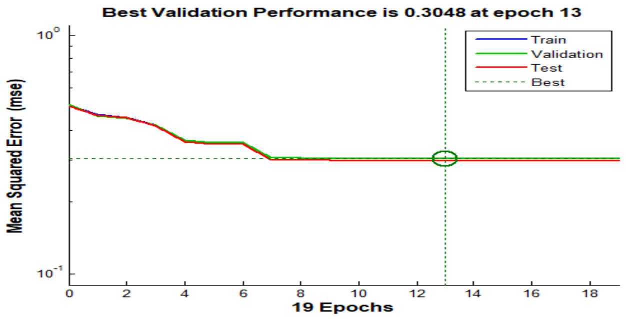


Figure 7 Estimation MSE errors vs. epochs for optimal solutions.

It can be concluded from the Figure 7 that the testing results offers the reduced MSE values comparatively. And the best possible optimal results is found corresponding to the 13th epoch with 0.3048 MSE value. It can also be observed that initially huger MSE is observed. The MSE start converging nearly after 8th epoch.

In order to further justify the results accuracythe results of the ROC curves are plotted for the three classes of data as in Figure 8. It can be clearly accuracy of the ROC curves is better for the Class 1 and Class 3 since the data sample size used correspondingly are 47% and 31 % respectively. The improvement in the ROC is offered since the optimal increased window size and the layer sixe is used to tune the network in this work.

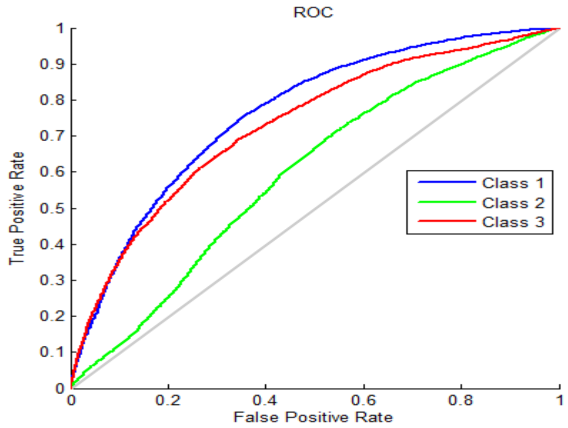


Figure 8 ROC curves comparison for three classes of data

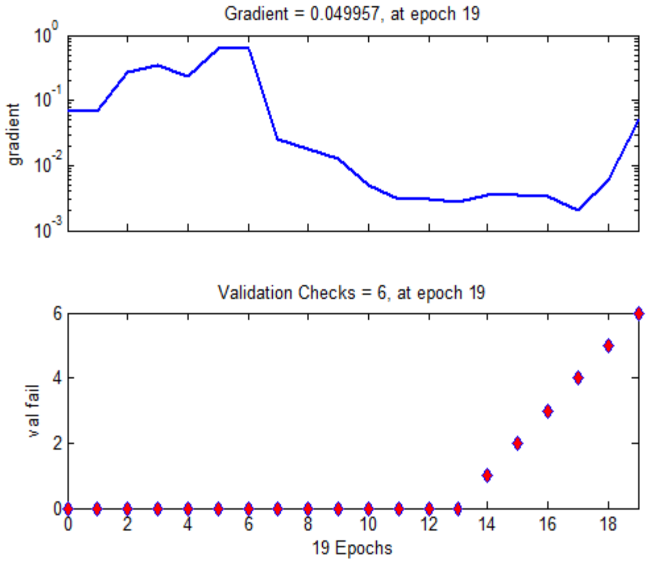


Figure 9 Result of validation check and gradients vs. epoch

The result of validation check and gradients are plotted in the Figure 9 for the respective epochs. The maximum values are reached at the epoch 19 respectively.

Results of Network Analysis

By comparing the trained network's outputs to the anticipated outcomes (targets), paper may analyses the network response. In order to analyses the network response the confusion matrix are plotted for each class of data as shown in the Figure 10. By looking at the confusion matrix it is concluded that there is significant chance of improvement in the prediction accuracy is required. in this context the optimal window size may be selected for improvement. Optimally selecting the

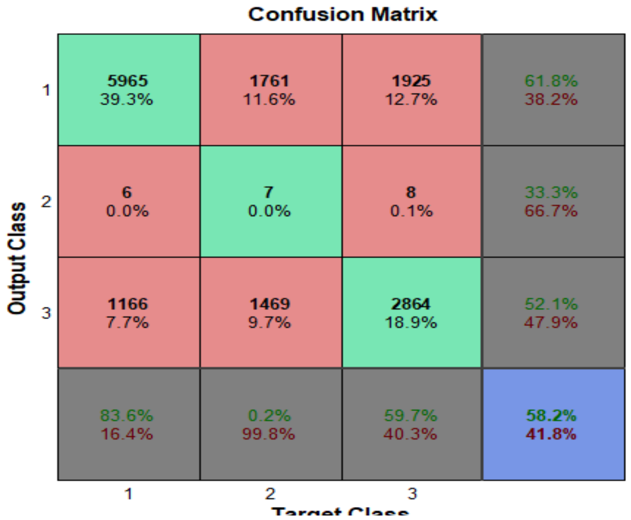


Figure 10 results of the confusion matrix class vise for prediction accuracy description.

7. Conclusion and Future Work

This research addresses the structural prediction technique by proposing a modelling and simulation approach based on neural networks (NNs). The suggested technique trains the NN using the log sigmoidal function. The NN based prediction methodogoloy uses the Rost-Sander Dataset. To achieve better performance, the NN layer and design parameters are changed. The three distinct classes of data's ROC curves and confusion metrics are used to assess the performance.

It can be clearly accuracy of the ROC curves is better for the Class 1 and Class 3 since the data sample size used correspondingly are 47% and 31 % respectively. The improvement in the ROC is offered. Paper has demonstrated the good use of NN based prediction .

There is significant chance of minimizing the MSE and the improvement in confusion matrix in near future by optimally tuning the NN parameters and layer architecture.

References

- *Minoru Saito "7 Accurate Calculations of Relative Melting Temperatures of Mutant Proteins by Molecular Dynamics/Free Energy Perturbation Methods" Y. Taniguchi et al. (eds.), Biological Systems Under Extreme Conditions © Springer-Verlag Berlin Heidelberg 2002*
- *Eric A. Franzosa "Structural Correlates of Protein Melting Temperature" Experimental Standard Conditions of Enzyme Characterizations, September 13th – 16th, 2009, Ru¨desheim/Rhein, Germany.*
- *In-ChulYeh "Calculation of Protein Heat Capacity from Replica-Exchange Molecular Dynamics Simulations with Different Implicit Solvent Models" J. Phys. Chem. B, 2008, 112 (47), 15064-15073 • Publication Date (Web): 30 October 2008 Downloaded from http://pubs.acs.org on November 24, 2008 10.1021/jp802469g This article not subject to U.S. Copyright. Published 2008 by the American Chemical Society Published on Web 10/30/2008*
- *A case report onProtein Thermal Shift technology by applied system thermofisher scientific.*
- *Anthony P. Russell "Determination Of Melting Sequences in DNA And Dna-Protein Complexes By Difference Spectra" From the Veterans Adninistration Hospital, Bedford, Massachusetts 01730 and the Department of Biochemistry, Boston University School of Medicine, Boston, Massachusetts 02118*
- *Yang Yang "ProTstab2 for Prediction of Protein Thermal Stabilities" Yang, Y.; Zhao, J.; Zeng, L.; Vihinen, M. ProTstab2 for Prediction of Protein Thermal Stabilities. Int. J. Mol. Sci. 2022, 23, 10798. https:// doi.org/10.3390/ijms231810798*
- *WAYNE J. BECKTEL" Protein Stability Curves" Biopolymers, Vol. 26, 1859-1877 (1987) 0 1987 John Wiley & Sons, Inc. CCC 0006-3525/87/111859-19$04.00"*
- *MattiaMiotto" Insights on protein thermal stability: a graph representation of molecular interactions" n. bioRxiv preprint doi: https://doi.org/10.1101/354266; this version posted August 1, 2018. The copyright holder for this preprint (which was notcertified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.*
- *Kumar S, Deshpande PA (2021) Structural and thermodynamic analysis of factors governing the stability and thermal folding/ unfolding of SazCA. PLoS ONE 16(4): e0249866. https://doi.org/10.1371/journal.pone.0249866*
- *Downloaded via 171.48.59.67 on November 15, 2023 at 05:09:12 (UTC). See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.*
- *L. HOWARD HOLLEY" Protein secondary structure prediction with a neural network" Proc. Nati. Acad. Sci. USA Vol. 86, pp. 152-156, January 1989 Biophysics*
- *ShengWang" Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields" Scientific Reports | 6:18962 | DOI: 10.1038/srep18962*
- *Ali Abdulhafidh Ibrahim" Using Neural Networks to Predict Secondary Structure for Protein Folding" Journal of Computer and Communications, 2017, 5, 1-8 http://www.scirp.org/journal/jcc ISSN Online: 2327-5227 ISSN Print: 2327-5219*
- *WafaaWardah" Protein secondary structure prediction using neural networks and deep learning:areview" https://doi.org/doi:10.1016/j.compbiolchem.2019.107093*
- *Romana Rahman Ema" Protein Secondary Structure Prediction using Hybrid Recurrent Neural Networks" Romana Rahman Ema et al. / Journal of Computer Science 2022, 18 (7): 599.611 DOI: 10.3844/jcssp.2022.599.611*
- *Romana Rahman Ema" Protein Secondary Structure Prediction based on CNN and Machine Learning Algorithms" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 13, No. 11, 2022*