SECONDARY STRUCTURE PREDICTION OF BACTERIAL PROTEIN (QUORUM-SENSING CONTROL REPRESSOR - *PSEUDOMONAS AERUGINOSA*) BASED ON NEURAL NETWORK

Saravanan K[1] Sivakumar S[2]

[1]*Department of Physics, AVS Engineering College, Salem, Tamil Nadu, INDIA*

[2]*Department of Physics, Government Arts College (Autonomous), Salem, Tamil Nadu, INDIA*

Corresponding Author: saroengg@gmail.com

ABSTRACT: *Quorum sensing, a cell-cell communication system, is broadly distributed among bacteria and is commonly used to regulate the production of shared products. An important consequence of quorum sensing is a delay in the production of certain products until the population density is high. The bacterium P. aeruginosa has a particularly complicated quorum sensing system involving multiple signals and receptors. One of these receptors, QscR, down-regulates gene expression, unlike the other receptors in P. aeruginosa. QscR does so by inducing the expression of a single operon whose function provides an element of resistance to a population reaching a quorum. This finding has importance for the design of quorum sensing inhibitory strategies and can also inform the design of synthetic biological circuits that use quorum sensing receptors to regulate gene expression. Precise prediction of protein secondary structures from the associated amino acid sequence is of great importance in bioinformatics and yet a challenging task for machine learning algorithms. In the present work, Multi-layered Neural Network (MLNN) was employed to predict the secondary structure of QscR protein (α-helix, β-sheet and C-coil). The results are compared with other prediction mechanisms. The obtained results are more accurate and better than the corresponding other mechanisms.*

Keywords: Protein structure prediction/QscR/Secondary structure/ Neural Network / Back-propagation/ Support vector machines.

1. INTRODUCTION

The prediction of protein secondary structure is an important step in the prediction of protein tertiary structure. Protein tertiary structure prediction is of great interest to biologists because proteins are able to perform their functions by coiling their amino acid sequences into specific three-dimensional shapes (tertiary structure). Therefore, this subject is of high importance in medicine (e.g. drug design) and biotechnology.

Three-dimensional structures for most proteins are determined by their one-dimensional sequences of amino acid residues. How to predict three-dimensional structures from one-dimensional sequences has been an unsolved problem for the last half-century. This problem is challenging because it demands an efficient technique to search in astronomically large conformational space and a highly accurate energy function to rank and guide the conformational search, both of which are not yet available. As a result, it is necessary to divide the structure prediction problem into many smaller problems with the hope that solving smaller problems will ultimately lead to the solution to the big problem. One of those smaller or sub-problems is the prediction of one-dimensional structural properties of proteins from their sequences. The most commonly predicted one-dimensional structural property of a protein is the secondary structure. Secondary structure describes each amino residue in a number of discrete states for which three state descriptions (helix, sheet and coil) are the most common.

Perception into the biological activity of a protein can be gained by identifying its large-scale functional movements. According to the well-established central dogma of biology, the amino acid sequence of the protein dictates its three-dimensional structure, which in turn determines and enables the molecule's function. This made to have quantitative connections with experimental quantities such as NMR order parameters [1,2] and TRP fluorescence spectra[3].

Hence, several studies have attempted to bridge the gap between feasible MD simulations [4]. In fact, the large-scale dynamical features encountered in MD trajectories can be conveniently interpreted at a first approximation, as a superposition of independent harmonic modes [5].This observation was complemented by Tirion [6] who pointed that, in a normal mode analysis of protein vibrations, the detailed classical force field could be replaced by suitable harmonic couplings with the same spring constants. These results stimulated a variety of studies where the elastic properties of proteins were described through coarse-grained models in which amino acids are replaced by effective centroids corresponding to the C(alpha) atoms and the energy function is reduced to harmonic couplings between pairs of spatially close centroids. These approaches, in particular, the Gaussian and anisotropic network models (GNM and ANM), have been found to be in accord with both experimental and MD results [7-8].

The secondary-structure prediction approaches in today can be categorized into three groups: neighbor-based, model-based, and Meta predictor-based [9]. The neighbor-based approaches predict the secondary structure by identifying a set of similar sequence fragments with known secondary structure; the model-based approaches employ sophisticated machine learning techniques to learn a predictive model trained on sequences of known structure,

whereas the Meta predictor -based approaches predict based on a combination of the results of various neighbor and/or model-based techniques.

Historically, the most successful model-based approaches, such as PSIPRED [10] were based on neural network (NN) learning techniques [11]. Protein secondary structures are traditionally characterized as 3 general states: helix (H), strand (E), and coil (C). From these general three states, the DSSP program [12] proposed a finer characterization of the secondary structures by extending the three states into eight states: 3_{10} helix (G), α-helix (H), π-helix (I), β-stand (E), bridge (B), turn (T), bend (S), and others (C). Prediction of the three states from protein sequences (i.e., the Q3 prediction problem) has been intensively investigated for decades using many machine learning methods, including the probability graph models [13, 14], support vector machines [15, 16], hidden Markov models [17, 18], artificial neural network [19-22], and bidirectional recurrent neural network(BRNN) [23].

Recently, the focus of secondary structure prediction has been on Q3 prediction to the prediction. Huge number of methods such as SC-GSN network [24], the bidirectional long short-term memory (BLSTM) method [25], the deep conditional neural field [26], DCRNN [27], the next-step conditioned deep convolutional neural network(CNN) [28] and Deep inception-inside-inception (Deep3I) network [29] have been widely explored. However, they have some drawbacks.

Hence, in this study, a multi-layered neural network (MLNN) is proposed to predict the structure of QscRQ3 secondary structure prediction. Firstly a local block comprising of original input combines local features and original sequence information. After local block filtering, the sequences are fed to an MLNN. This architecture can model the sequence structure and can capture long-range dependencies of the residues. The prediction results have demonstrated that the proposed ANN has better generalization performance in comparison with the other existing method.

2. MATERIAL AND METHODS

In this section, a few relevant concepts about cascaded NN and the models of protein models using elastic network models (ENMs) are discussed.

A crucial step in the construction of an MLNN based pipeline to analyze and process a given molecular structure is the identification of an appropriate molecular descriptor. From the b-GM framework, the Hamiltonian of the system depends only on the positions of the Ca atoms. Hence, it is concluded that the molecular descriptor considers the Cartesian coordinates of atoms as its input. However, for NN applications, it is not possible to characterize the biomolecule in

terms of Cartesian coordinates, since these are not invariant with respect to rotations and translations of the system.

A prominent example of a molecular descriptor is given by Behler and Parrinello's symmetry functions [30,31]. These functions describe the local environment of each atom in a molecular system based on a cut-off radius. The interatomic distances were larger than this value yield zero contribution to these descriptors. Symmetry functions have been mainly employed in order to provide accurate potential energy surfaces and to detect local atomic arrangements in liquids [26]. Although these descriptors proved to be extremely successful, the proposed task concerns the prediction of secondary structure using the chemical structure of the amino acids and a coding scheme is generated.

2.1. Cascaded Neural Network

- An Artificial Neural Network (ANN) shown in Figure 1 is a massively parallel distributed processor that has a natural propensity for storing experimental knowledge and making it available for use. They are made up of simple processing units called artificial neurons. It resembles the brain in two respects--
- Knowledge is acquired by the network from its environment through a learning process.
- Interneuron connection strengths known as synaptic weights are used to store the acquired knowledge.

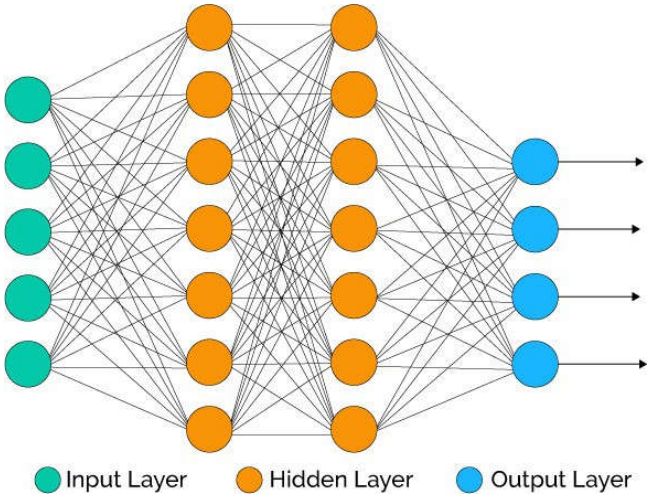

Figure 1.An Artificial Neural Network

2.2 Designing Neural Networks For Protein Secondary Structure Prediction

When designing NN for bioinformatics applications, there is a common designing issue that needs to be addressed. Figure 2 summarizes these issues.

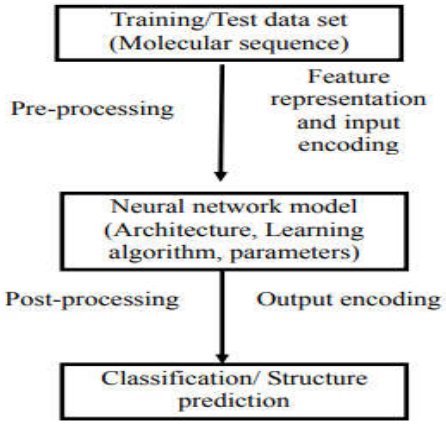

Figure 2. Design Neural Networks for Secondary Structure Prediction

Preprocessing of data involves feature presentation and input encoding. This is an important element that determines the performance and the information entered into the NN. In order to get the full benefit of the NN, the prior knowledge about sequence structure and functions is necessary. This allows the extraction of salient features in the given sequence.

Protein sequences in FASTA format were represented in 20 letters of amino acid, then each letter encoded to a unique number from 1 to 20. In this study, the sliding window scheme was used to cover protein-protein interactions. Then, each amino acid encoded by orthogonal encoding.

This work consists of several steps as described below:

Neural Network is one of the machine learning techniques that often used in amino acid prediction. This method aims to find input and output mapping of amino acids. The neural network is affected by stopping criteria, such as the number of iteration, MSE (Mean Square Error) and gradient descent. In this study, 1000 epoch interaction was used.

The system consists of two levels of the neural network. There is the first neural network (NN1) and the second neural network (NN2). Both NNI and NN2 were designed using feed forward architecture with three layers, they were trained separately with the back propagation method.

The input layer in NN1 contains L x 20 unit neurons, with L represents the length of windows schemes if the network trained with 17 windows schemes, so the input layer will have 17 x 20 = 340 unit neurons. The hidden layer has 3 unit neuron and the output layer has 3 unit neurons. The input layer in NN2 uses the output result from NN1. These classifiers receive identified Amino Acids for predicting alpha, beta and coils.

2.3. Training and Testing Sets

In this work, 100 proteins set for training and 5 protein set for testing were used.

2.4. Performance Measures

There are many ways to access the performance of the method for predicting secondary structures. The most commonly used measure is a simple success rate, Q3, which is the percentage of correctly predicted residues on three types of secondary structures:

Q3 denotes the overall accuracy.

Qobs = TP/(TP + FN)

Which shows the prediction sensitivity and

Qpred = TP/ (TP + FP) shows the prediction specificity.

Where TP: True Positive; FP: False Positive; FN: False Negative.

Prediction performance can be assessed using three following indicators, specificity, sensitivity and accuracy. These three indicators are measured from TN, TP, FN and FP; the formulas are as follows,

$$\text{Sensitivity} = TP(TP+FN)$$

$$\text{Specificity} = TN/ (TN+FP)$$

$$\text{Accuracy} = (TN+TP)/ TN+TP+FN+FP)$$

3. RESULTS

3.1. Neural Network

The ANN is configured to handle the protein structure. The training is carried out with (error) back propagation algorithm with Levenberg-Marquardt optimization. The structural assignment of the training data set, validation data set and testing data set are represented in Figure 3, and 4.Among the training data set, 47% were about the coil, 31% were strand and 21% were Helix. In the testing data set, it was about 48% of C, 31% of E and 21% of H.ANN gives a performance goal of around 0.296 which is attained after 74 epochs and is depicted in Figure 5.

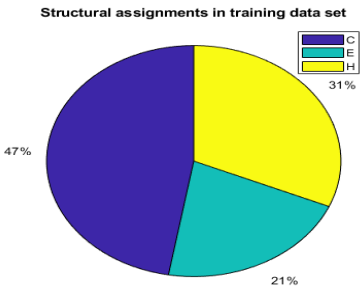

Figure3. Structural assignments in training data set

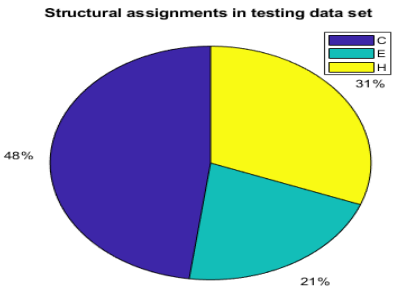

Figure 4. Structural assignments in testing data set

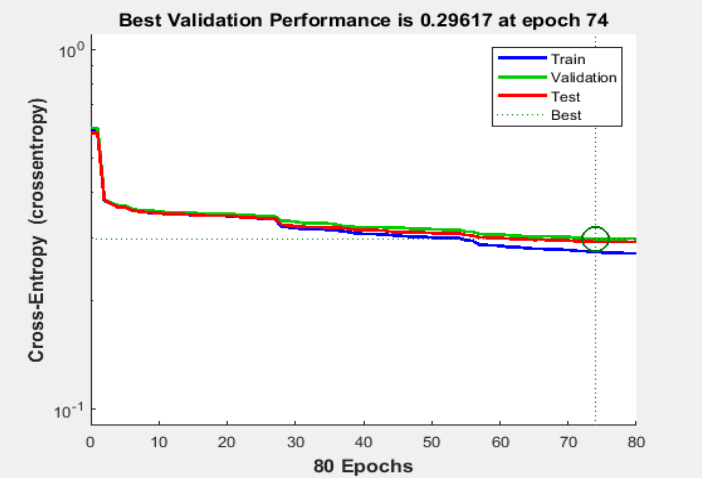

Figure 5.Best validation performance curve

3.2. Sensitivity and Specificity Analysis

The receiver operating characteristic (ROC) curve was implemented to evaluate the efficacy of the prediction. The ROC curve portrays the trade-off between the true positive rate (sensitivity) and the false positive rate (1—specificity) for different threshold values. The true positive rate is defined as the fraction of experimentally confirmed PPI pairs (all positives) that are correctly predicted. Likewise, the false positive rate is defined as the fraction of all potential interactions that are predicted and do not match with experimental pairs.

The area under the ROC curve (AUC) is a quantitative indicator for ranking the performance of prediction topology. At an AUC of 1, a perfect PPI prediction is obtained. The ROC curve is thus the sensitivity as a function of the fall-out. A perfect predictor would be described at 100% sensitive. The closer the ROC curve is to the upper left corner (100% sensitivity, 100% specificity), the higher the overall accuracy.Fig.6 represents the ROC curves for the proposed methodology.

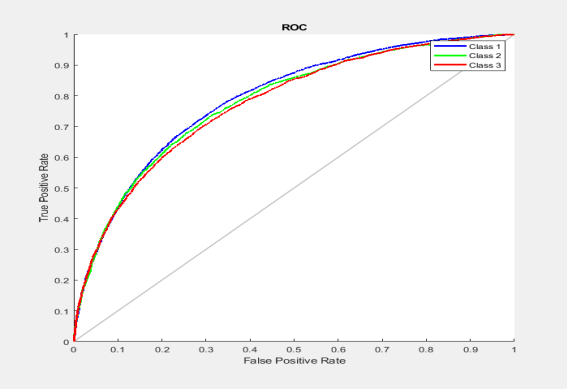

Figure 6. ROC curve

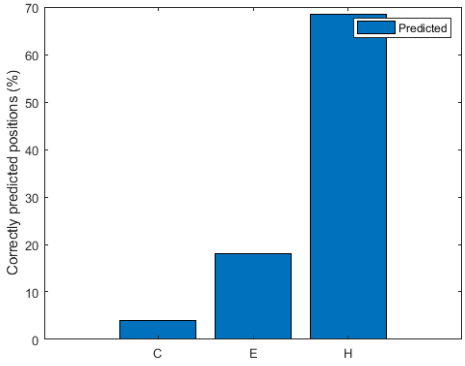

Fig. 7.Quality indices performance

Topologies predictions can be evaluated in more detail by calculation of assorted quality indices. Those indices demonstrate if the topology is accurately predicted and whether there was over-prediction or under-prediction. Fig.7 represents the correctly predicted positions in percentage. Bar (x,y) is the MATLAB function that was used to create the bar graphs and the same was tabulated in Table 1.

Table 1. Comparative analysis of the performance of the other methods in secondary structure prediction

S.No	Method	Alpha (%)	Beta sheet (%)	Coil
1.	**DSSP**	54	13	-
2.	**STRIDE**	54	11	-
3.	**Proposed NN**	68	15	4

From Figure 7 and Table 1, it is concluded that the proposed NN gives more and better prediction of secondary structure than the other topologies. Apart from this, by varying the size of the window as well as hidden layers, the accuracy can be varied and is depicted in Tables 2 and 3.

Table 2. Dependence of testing success on hidden units

S.No	Hidden Units	Q_3(%)
1.	**0**	61.39
2.	**25**	69.18
3.	**50**	65.64
4.	**75**	67.42

Table 2 shows that the peak performance on the testing set depends on the number of hidden units. Also, it is shown that having more hidden units is not always good because it can cause a high-variance problem.

Table 3. Dependence on the size of the window

S.No	Window size	Q_3(%)
1.	**1**	50.37
2.	**3**	54.04
3.	**5**	59.51
4.	**7**	63.17
5.	**17**	74.12

Table 3 shows the dependence of the testing accuracy rate on the size of the input window with 25 hidden units. The result shown in Table 3 indicates that when the size of the window was small the performance on the testing set was reduced, probably because information outside the window is not available for the prediction. When the size of the window was increased, the

performance reached a maximum at around 17 window size. For larger window sizes, the performance deteriorated, probably because of the effects of extra weights that could not contain any information about the secondary structure of the center. Thus, irrelevant weights can interfere with the performance of the network. Similarly; the accuracy of the proposed system at the end of the network is depicted in Table 4.

Table 4: Accuracy of the system

Q_3(%)	1st network	2nd network
	67.42	74.12

Thus, Predicted secondary for the sequence of chain A of QscR is tabulated below and its 3d structure is depicted in Figure 7.

Table 5: Predicted secondary structure odQscR under different topology

Methods		Secondary structure				
Sequence (1-50)		MHDEREGYLE	ILSRITTEEE	FFSLVLEICG	NYGFEFFSFG	ARAPFPLTAP
Structure	DSSP	******SHHH	HHHH** SHHH	HHHHHHHHHHH	HTT*SEEEEE	EE***STTS*
	Proposed MLNN	CHHHHHHHHH	HHHHCCCHHH	HHHHHHHHHH	HHCCCEEEEE	EECCCCCCCC
Sequence(51-100)		KYHFLSNYPG	EWKSRYISED	YTSIDPIVRH	GLLEYTPLIW	NGEDFQENRF
Structure	DSSP	*EEEEE*** H	HHHHHHHHTT	GGGT*HHHHH	HHHS*S* EEE	ETTT*SS*HH
	Proposed MLNN	CEEEECCCCH	HHHHHHHHHC	CHHHCHHHHH	HHHCCCCEEE	CCCCCHHHHH
Sequence(101-150)		FWEEALHHGI	RHGWSIPVRG	KYGLISMLSL	VRSSESIAAT	EILEKESFLL
Structure	DSSP	HHHHHHHTT*	*EEEEEEE*	GGG*EEEEEE	EESSS*** HH	HHHHHHHHHH
	Proposed MLNN	HHHHHHHHCC	CCEEEEEEEC	CCCCEEEEEE	ECCCCCCCHH	HHHHHHHHHH
Sequence(151-200)		WITSMLQATF	GDLLAPRIVP	ESNVRLTARE	TEMLKWTAVG	KTYGEIGLIL
Structure	DSSP	HHHHHHHHHH	HHHHHHHHSG	GGG**** HHH	HHHHHHHHTT	**HHHHHHHH
	Proposed MLNN	HHHHHHHHHH	HHHHCCCCCC	CCCCCCCHHH	HHHHHHHHHC	CCHHHHHHHH
Sequence(201-237)		SIDQRTVKFH	IVNAMRKLNS	SNKAEATMKA	YAIGLLN	--
Structure	DSSP	TS*HHHHHHH	HHHHHHHTT*	SSHHHHHHHH	HHTT***	--
	Proposed MLNN	CCCHHHHHHH	HHHHHHHHCC	CCHHHHHHHH	HHHCCCC	--

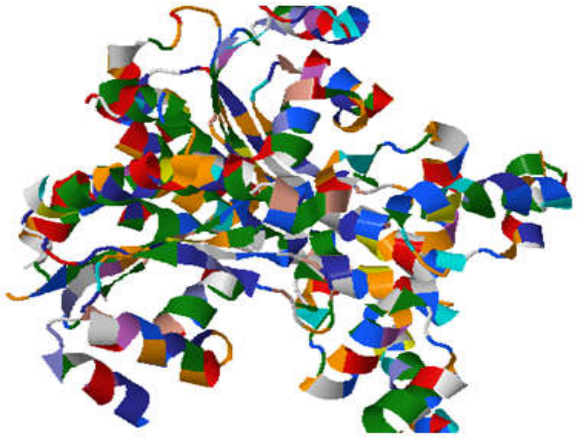

Figure 7. 3D structure of QscR

3.3. Ramachandran Plot

A graph between the dihedral angles of an amino acid in a protein. Due to steric hindrance from amino acid side chains, only certain angles are allowed in a folded protein. A plot between the dihedral angles of individual amino acids in a protein can serve to indicate how well the structure has been determined. Any deviations from the allowed values are called Outliers and usually indicate bad geometry. The Ramachandran plot for the proposed protein is displayed in Figure 8. Ideally, there should be no outliers in the Ramachandran plot, except for Glycine and Proline, which are "special" amino acids. However, there may be some rational explanation for outliers by the scientist depositing the structure. Expect to find more than 85-90% of residues to fall into the red regions.

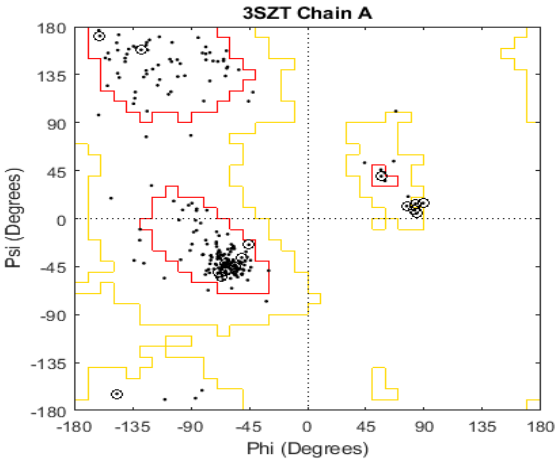

Figure 8. Ramachandran plot of QscR

From the above, it is observed that 92.4% of its residues lies in the favored region, 9% of residues are in the allowed region and the remaining 4% are in the not allowed region.

3.4. Heatmap of Protein

Proteins were prepared for docking followed by ligand removal from the original (downloaded from PDB) structures. The resulted geometric scores of the protein database are shown by color-coding heatmap (Fig. 9). In Fig. 9, x-axis and y-axis represent the protein counts and ligand counts which in turn indicate the corresponding ligand and protein leveling. The blue color indicates a lower binding score and the red signifies the top binding scores. This pattern shows potentially highly selective ligands which are good for potential drugs and drug delivery agents.

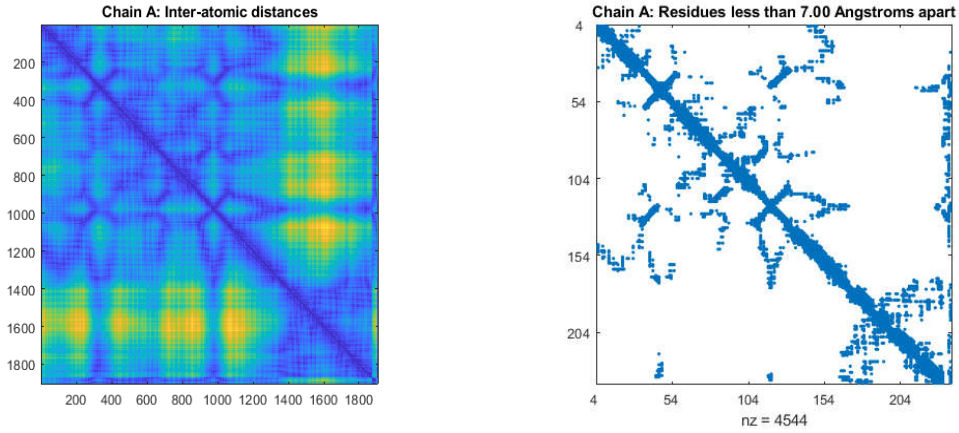

Figure 9. Heat map of QscR

4. CONCLUSION

In this work multilayered neural network (NN) was employed, to predict the protein secondary structure which aims to predict the secondary structure of QscR (α-helix, β-sheet and C-coil) using MATLAB Neural Network backpropagation algorithm. It also includes comparisons with the previous results obtained by other topologies. This MLNN describes the probabilities of the three secondary structural elements of an amino acid in sequence and includes information on the structural evolution. Third, in this pioneering work, a relationship is built between structural profiles and secondary structures; this is the simplest but best model now available. We believe that this methodology could be extended to other structural biological fields and that it will greatly improve prediction efficiencies.

REFERENCES

1. *G. Lipari and A Szabo, "Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules", Theory and range of validity. Journal of American Chemical Society V- 104 (1982),PP- 4546 – 4559.*
2. *G. Lipari, "Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules", Analysis of experimental results, Journal of American Chemical Society. 104, (1982),PP- 4559 – 4570.*
3. *CL Brooks, BM Pettitt. "Proteins: a theoretical perspective of dynamics, structure, and thermodynamics, (1988), Wiley publication.*
4. *F Pazos and M Chagoyen,"Practical protein bioinformatics, 1st edn. Berlin, Germany: Springer publication,(2016)*
5. *MG Saunders "Coarse-graining methods for computational biology", Annual Review Biophysics- vol.42, (2013),PP- 73–93.*
6. *MM Tirion. "Large amplitude elastic motions in proteins from a single-parameter, atomic analysis" Journal of physics Review and letter,vol.-77, (1996), PP-1905–1908.*
7. *BJ Alder and TE Wainwright, "Studies in molecular dynamics. I. General Method., Journal of Chemical Physics, vol.-. 31, (1959) PP-459–466.*
8. *M Karplus, "Molecular dynamics simulations of bimolecular", , ACC Journal of Chemical Research, vol.- 35, (2002) PP- 321–323.*
9. *Hae-Jin Hu, Robert W., "Current Methods for Protein Secondary-Structure Prediction Based on Support Vector Machine CH 1", Knowledge Discovery in Bioinformatics: Techniques, Methods, and Applications, (2007)*
10. *David T. Jones, "Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices", Journal of Material Biophysics. (1999)*
11. *S Wang , J Peng, "Protein secondary structure prediction using deep convolution neural fields", Journal of Science Report,vol.-6, (2016) PP-18962.*
12. *Kabsch, Wolfgang and Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features", Journal of Biopolymers vol-22,no. 12, (1983) ,PP-2577–2637.*
13. *SC Schmidler, JS Liu, "Bayesian segmentation of protein secondary structure", Journal of Computer Biology- vol.7,no.1, (2000),pp-233–48.*
14. *W Chu, Z Ghahramani, "Graphical model for protein secondary structure prediction, In: Proceedings 21st Annual International Conference on Machine Learning . New York: ACM:. (2004) pp- 161–168.*
15. *S Hua ,Z Sun, "A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach", Journal of Molecular Biology, vol.- 308,no.2, (2001), pp--397–407.*
16. *J Guo , H Chen , Z Sun , "A novel method for protein secondary structure prediction using dual-layer svm and profiles", Journal of Protein Structure Function Bioinformatics, vol.-54,no.4, (2004),pp -738–743.*
17. *K Asai , S Hayamizu , "Prediction of protein secondary structure by the hidden Markov model", Journal of Bioinformatics, vol. ,no.2,(1993),pp -141.*

18. *Z Aydin, Y Altunbasak, "Protein secondary structure prediction for a single-sequence using hidden semi-Markov models" Journal of Bioinformatics, vol.-7,no.1,2006,pp.178.*
19. *N Qian and TJ Sejnowski. "Predicting the secondary structure of globular proteins using neural network models" Journal of Molecular Biology, vol. - 202, no.4, (1988), pp -865–884.*
20. *DT Jones, "Protein secondary structure prediction based on position-specific scoring matrices", Journal of Molecular Biology, vol.- 292,no.2,(1999), pp-195.*
21. *DW Buchan, F Minneci, TC Nugent, "Scalable web services for the inspired protein analysis workbench", Journal of Nucleic Acids Research, vol. 41. (2013), pp.349–57.*
22. *E Faraggi,, "Improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles", , Journal of Computer science and Chemical, vol,33,no.3, (2012), pp- 259 – 67.*
23. *P Baldi, S Brunak, "Exploiting the past and the future in protein secondary structure prediction" Journal of Bioinformatics, vol. 15, no.11, (1999), pp-937–946.*
24. *J Zhou, OG Troyanskaya, "Deep supervised and convolution generative stochastic network for protein secondary structure prediction", In: Proceedings of the 31st International Conference on Machine Learning, Bejing, (2014), p- 745–753.*
25. *R Heffernan, Y Yang, "Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility", Journal of Bioinformatics, vol.- 33,no.8, (2017),pp.2842–2849.*
26. *A Busia, N Jaitly, "Next-step conditioned deep convolutional neural networks to improve protein secondary structure prediction", International Conference on Intelligent Systems for Molecular Biology & European Conference on Computational Biology, Leesburg: International Society of Computational Biology, (2017)*
27. *Z Li, Y Yu, "Protein secondary structure prediction using cascaded convolutional and recurrent neural networks", International Proceedings of the 25th International Joint Conference on Artificial Intelligence , California ,(2016),pp. 2560–2570.*
28. *S Wang, J Peng, "Protein secondary structure prediction using deep convolutional neural fields" ,Journal of Scientific Report,(2016),vol.-6.*
29. *C Fang , Y Shang , "New deep inception-inside-inception networks for protein secondary structure prediction", Journal of Protein Structure Function Bioinformatics,vol.86,,no.5, (2018),pp. 592–598.*
30. *J Behler and M Parrinello, "Generalized neural network representation of high-dimensional potential-energy surfaces", Journal of Physics Review Letter, vol. 98,(2007),pp. 146401.*
31. *J Behler,"Representing potential energy surfaces by high-dimensional neural network potentials", Journal of. Physics and Condensed Matter, vol.-26, (2014), pp- 183001.*