Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8629 - 8643
Received 16 February 2021; Accepted 08 March 2021.

Efficient Secondary Structure Prediction of Qscr -Protein Using Ann Coupled Pso Algorithm

[1]Saravanan K, [2]Sivakumar S, [3]Sridevi K P

[1]Department of Physics, AVS Engineering College, Salem, Tamilnadu, INDIA
[2]Department of Physics, Government Arts College (Autonomous), Salem, Tamilnadu, INDIA
[3]Department of Physics, Sri Kailash Women's College, Thalaivasal, Salem,Tamilnadu, INDIA

ABSTRACT

Identifying Secondary structure of protein becomes more important in designing drugs. Presently, machine learning algorithms like Artificial Neural Network (ANN) and Support Vector Machine (SVM) has been utilized to identify the Protein Secondary Structure (PSS). However, in order to improve more accuracy of PSS prediction, this work proposed PSO to search the best parameters of theANN predictor developed for QscR protein. From the results, it is found that the proposed topologyexhibits high accuracy than the other existing techniques.

Keywords: Artificial Neural Network,Protein Secondary Structure, Particle Swarm Optimization, QscR.

1. INTRODUCTION

Enormous progress in genome sequencing and the consequent attainability of more protein sequences find its usefulness in the field of computational biology and hence, it is made necessary to predict the PSS. Developments in predicting the SS is used in the field of protein engineering and in designing drugs. To determine the 3D structure of protein in huge scale, the present crystallization techniques are too costly and consumes more time [1].Prediction of SS can be used as intermediate to increase the predicting rate of protein structure. It also helps in recognition of genes, categorization of structures, functional patterns and finding the defective structures which are the sources for human diseases.For determining the secondary structure, various computational technologies have been utilized in a successful way [2,3]. It has been proved that machine learning and empirical methods are considered as the most successful method. The prime method GOR relies on information theory [4]. Metamorphic information is utilized for enhanced structure prediction. Neural networks which relies on multiple sequence arrangement is used by the protein predicting server. The PSIBLAST and NNs are used by PSIPRED algorithm.Based on the Jnet algorithm, the Jpred prediction server predicts PSS. By making use of the

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8629 - 8643
Received 16 February 2021; Accepted 08 March 2021.

currently available structural information and computational methods, SS determination techniques have been developed extensively. To attain better categorization efficiency, it is advised to consider long range interactions as a key factor.Hence, by implementing ANN, an innovative approach for predicting the PSS has been designed. It is a tedious task to frame an ANN, as the fabrication relies on the construction, the chosen transfer function and the learning algorithm which is utilised to train the synaptic weights. Hence, it is advised to use the PSO to enhance the efficiency of ANN classifier by adjusting its specifications like hidden neurons, bias and input weights which are utilized to train the ANN [5,6].

2. DATA GENERATION

By deriving the details from relative similarity groups of QscR, which includes both short and long range interactions existing among the AA of proteins, the database is created. The dictionary of database of PSS (DSSP),has 8 categories of protein structures. In these categories, a shortened set of 3 SS termed as α- helix (H), β-Strand (E) and Coil (C) are considered in this work. By using 100 non-homologous protein sequences, a profile matrix is designed.

2.1 Methods and optimization

In this, for categorizationANN is used. The variables of the ANN are tuned using the PSO.

Encoding of ANN

The input of an ANN has prearranged patterns (Residues). Each pattern is said to have 27 features with values lying between 0 and 1. There are three units in the output and these units correlate any one of the three secondary structure elements. The correlation is defined as 1 for a class of interest and a – 1 for the remaining two classes. At the hidden layer, the given input along with a bias and weights generates an activation function. The hidden layer's output associated with a different set of weights earns three outputs [7]. The predicted class which has the maximum value is treated as an output with lowest MSE.To construct a model, the ANN makes use of a set of training samples. At the time of training phase, the variables like weights, bias and deigned hidden neurons of ANN is enhanced using the PSO [8-12]. By doing so, the accuracy of categorization can be improved.The variables are stored and used at the time of the testing stage. Initially, input weights are selected randomly and later on they are adjusted by the PSO. The output weights obtained from the hidden layer are empirically determined by utilizing the pseudo inverse. In the hidden layer, sigmoidal activation function is utilized and regarding the output neurons, linear activation function is implemented [6].

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8629 - 8643
Received 16 February 2021; Accepted 08 March 2021.

In ANN algorithm, the elementary steps involved are as follows

- Choose the relevant activation function (G) and number of hidden neurons from the given class labels and training samples.

- Choose bias (b) and input weights (V) randomly and the output weights (W).

- The class label can be calculated from the determined weights (W ; V; b) and the inaccuracy between the observed and predicted values should be reduced. The best performances are later tuned by the PSO.

2.2 PSO

One of the stochastic optimization methods is PSO. This technique imitator of intelligent social act of colony of birds (or) schools of fish, identified as a particles in a community. In a minimum possible time, these particles by working together give a transparent and excellent solution to a problem. A random set of values called particles initialise the formation of PSO algorithm. These particles contribute collectively to achieve the desired solution. Several parameters defined by these values will increase the performance of the system. To attain the finest possible solution, this method constantly explore a multi- dimensional space set on by a fitness basis, The best values for input weights, hidden layer neurons and bias values are found using a PSO [13,14].

2.3 PSO trained ANN

The following strategies are taken into discussion to accommodate ANN with PSO algorithm [15-17]. In the ANN models by using the PSO algorithm, the optimum weights and the bias are realized. The search space of the algorithm with ‘n’ dimensions is formed by the weights and biases. Here ‘n’ represents the total number weights and biases that are to be developed. There are n-dimensions of position vector and velocity for every single particle. The letter ‘w’ denotes both the weights and biases. By flying the particles on all sides of the search space, the exclusive set of weights are acquired. On every repetition, a set of weights with their fitness accessed comes up along with the algorithm. This occurs by applying these weights to the nodes and by determining the value to be achieved. Subsequently, the correctness of the forecasting using the allocated weights is calculated by the variation between the original and forecasted values, the variation must be reduced using the optimization techniques[24][25][26]. Using this view, the particle with best fitness has been attained and until now it is treated as the individual best.Correspondingly, the swarm with the best fitness is treated as the best one globally. The present procedure is reproduced for the definite number of repetitions until the correct weight for the ANN is earned. The procedures for a ANN optimized by PSO is stated below. For a perceptron with three layers, W[1] denotes the relation

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8629 - 8643
Received 16 February 2021; Accepted 08 March 2021.

among the input layer and the hidden layer whereas W[2] denotes the relation among the hidden layer and the output layer commonly. Multi-layer perceptron is trained by using the PSO method. The i^{th} particle can be represented by

$$W_i = \left\{W_i^{[1]}, W_i^{[2]}\right\} \quad (1)$$

$$P_i = \left\{P_i^{[1]}, P_i^{[2]}\right\} (2)$$

$$P_b = \left\{P_b^{[1]}, P_b^{[2]}\right\} (3)$$

$$V_i = \left\{V_i^{[1]}, V_i^{[2]}\right\} (4)$$

where

j = 1, 2;

m = 1,. . . ,M_j ;

n = 1,. . . ,N_j;

M_j&N_j - row and column size

W, P, V, r , s - Constants;

a& b - Random numbers ranges within 0 to 1;

t - Time step between observations. Usually it is unity;

V" & W" - New values.

Applying Equation,

$$V_i^{"[j]}(m,n) = V_i^{[j]}(m,n) + \left\{r\alpha\left[P_i^{[j]}(m,n) - W_i^{[j]}(m,n)\right] + s\beta\left[P_b^{[j]}(m,n) - W_i^{[j]}(m,n)\right]\right\}/t \quad (5)$$

The current velocity of a particle is calculated by applying its earlier velocity and the distance of its recent location based on Pbest and Gbest value. On the right hand side of the equation, the second element serves as the exclusive thinking of the particle by its own and at the same time the third element, represents the combination in the midst of the particle as a group. The current location based on the recent velocity can be resolved by the equation as follows

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8629 - 8643
Received 16 February 2021; Accepted 08 March 2021.

$$W_i^{"[j]} = W_i^{[j]} + V_i^{[j]}t \qquad (6)$$

The MSE is the fitness function and is given by

$$f = \frac{1}{n}\sum_{i=1}^{n}(K_{oi} - K_{pi})^2 \text{n} \qquad (7)$$

Where

F - Fitness value,

n – No. of data points.

3. Results and Discussion

For training 100 protein set and to test 5 protein set are used. All the 3 SS (α-helix, β-strand and coil) exists in the form of a mixture in every sets mentioned above.

Table 1. The parameter configuration used in PSO

Parameters	Values
No. of Particles	100
C1	1
C2	2
Max. Iteration	1000

Among the training data set, 47% were about 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.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8629 - 8643
Received 16 February 2021; Accepted 08 March 2021.

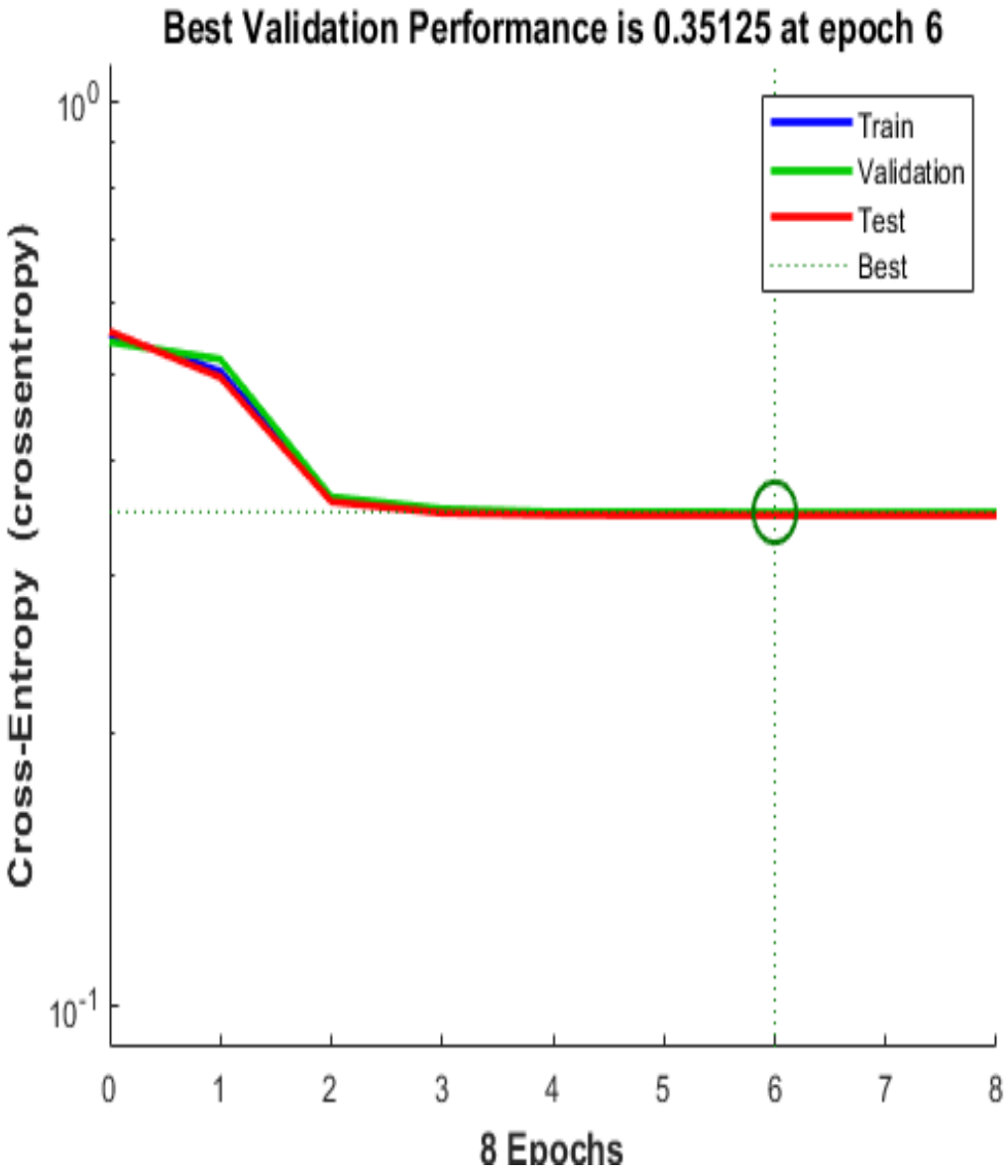

Figure 1. Best validation performance curve

For training, validation and test steps the figure (1) depicts the MSE of ANN model. With reference to the graph at the epoch 6, the least MSE occurs in the validation step and it has best validation performance equivalent to 0.35. Unless the network error on the validation vectors gets decreased, it is important to mention that the training model keeps going. Thus, for the sequence of chain A of QscR, the predicted SS is tabulated below

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8629 - 8643
Received 16 February 2021; Accepted 08 March 2021.

Table 2. Predicted secondary structure of QscR under different topology

Methods		Secondary structure				
Sequence (1-50)		MHDEREGYLE	ILSRITTEEE	**FFSLVLEICG**	**NYGFEFFSFG**	ARAPFPLTAP
Structure	**DSSP**	******SHHH	HHHH** SHHH	HHHHHHHHHH	HTT*SEEEEE	EE***STTS*
	MLNN	CHHHHHHHHH	HHHHCCCHHH	HHHHHHHHHH	HHCCCEEEEE	EECCCCCCCC
	Proposed PSONN	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
	MLNN	CEEEECCCCH	HHHHHHHHHC	CHHHCHHHHH	HHHCCCCEEE	CCCCCHHHHH
	Proposed PSONN	HEEEECCCCH	HHHHHHHHHC	CHHHHHHHHH	HHHCCCHEEE	CCCCHHHHHH
Sequence(101-150)		FWEEALHHGI	RHGWSIPVRG	KYGLISMLSL	VRSSESIAAT	EILEKESFLL
Structure	**DSSP**	HHHHHHHTT*	*EEEEEEEE*	GGG*EEEEEE	EESSS*** HH	HHHHHHHHHH
	MLNN	HHHHHHHHCC	CCEEEEEEEC	CCCCEEEEEE	ECCCCCCCHH	HHHHHHHHHH
	Proposed PSONN	HHHHHHHHCC	CCEEEEEEEC	CCCCEEEEEE	ECCCCCCCHH	HHHHHHHHHH
Sequence(151-200)		WITSMLQATF	GDLLAPRIVP	ESNVRLTARE	TEMLKWTAVG	KTYGEIGLIL
Structure	**DSSP**	HHHHHHHHHH	HHHHHHHHSG	GGG**** HHH	HHHHHHHHTT	**HHHHHHHH
	MLNN	HHHHHHHHHH	HHHHCCCCCC	CCCCCCCHHH	HHHHHHHHCC	CCHHHHHHHH
	Proposed PSONN	HHHHHHHHHH	HHHHCCCCCH	CCCCCCCHHH	HHHHHHHHHE	HCHHHHHHHH
Sequence(201-237)		SIDQRTVKFH	IVNAMRKLNS	SNKAEATMKA	YAIGLLN	
Structure	**DSSP**	TS*HHHHHHH	HHHHHHHTT*	SSHHHHHHHH	HHTT***	
	MLNN	CCCHHHHHHH	HHHHHHHHCC	CCHHHHHHHH	HHHCCCC	
	Proposed PSONN	CCCHHHHHHH	HHHHHHHHCC	CCHHHHHHHH	HHHCCCC	

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8629 - 8643
Received 16 February 2021; Accepted 08 March 2021.

To conclude, with reference to the above table it is observed that the suggested PSOANN shows better accuracy in predicting PSS than that of the other conventional methods. Apart from the detection of SS of QscR, the domination of the configuration of the amino acids is also investigated for determining the structures.The results obtained from many SS prediction methods are expressed in terms of commonly used three SS. At the AA level, it is strange to notice an investigation of the results with few deviations. To recognize the designs of prediction accuracies in connection to various AAs, the results of categorization from many SS prediction servers are observed and reanalysed [18-21]. Thus, the influence of the composition and physicochemical properties of AAs is investigated.In SS analysis, the AAs are not present in uniform quantities[22-23]. Very few amino acids such as Glutamine, Leucine and Isoleucine possess the highest number of helix residues while Cysteine as well as Proline possesses the exact number helix residues in them. The Strand, Serine, Leucine as well as Phenylalanine are having the highest content. On the other hand, most of the acids like Asparagine, Aspartic acid do not exist strand. Similarly, Cysteine and Glutamine does not exhibit coil property. Serine and Proline are the acids which bear more number of residues in the coil structures. Alanine and Asparagine seem to have the same content in the coil and the same is depicted in figure 2.

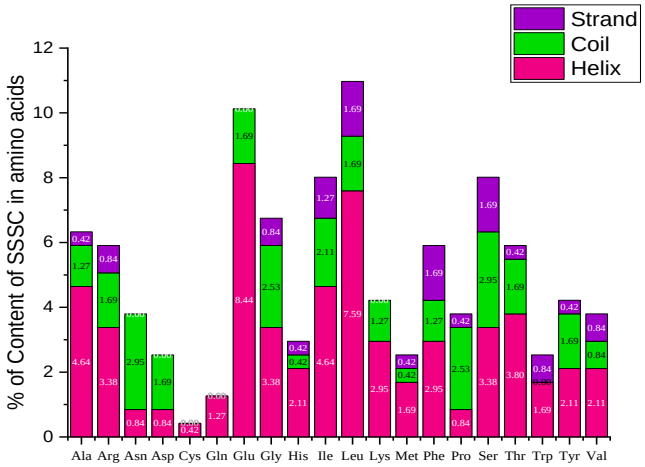

Figure 2a. Content of AAs in SS of QscR

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8629 - 8643
Received 16 February 2021; Accepted 08 March 2021.

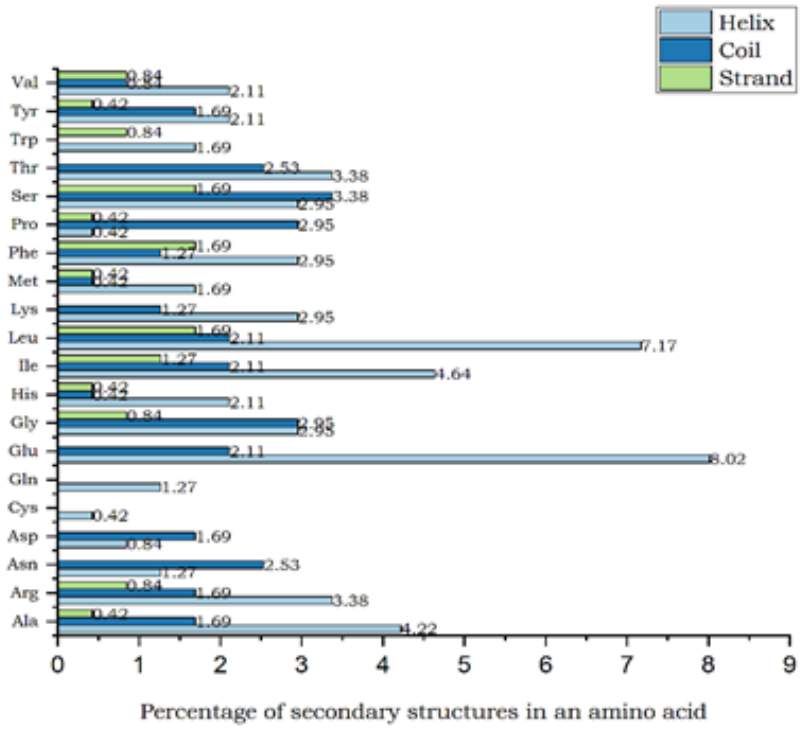

Figure 2b.Percentage of secondary structures in an amino acid

From the above graph, it is concluded that cys and gln are helix in nature and amino acids such as asn, glu and thr have not exhibited strand nature. From the analysis, it is concluded that α helix plays a vital role in QscR protein.Various steps to calculate the forecasting efficiency have been used in this approach. In the three states of secondary structures, the percentage of residues which are correctly predicted is illustrated by Q3 accuracy. By adding together the identified and recognized segments and by calculating their overlap, the precision is furnished by segment overlap measure (SOV) and it is depicted in the table (3).

Table 3. Performance analysis of Q3 and SOV of the proposed topology

	Qoverall (%)	**Q_H (%)**	**Q_E (%)**	**Q_C (%)**
Q3	78.9	84.6	58.3	85.0

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8629 - 8643
Received 16 February 2021; Accepted 08 March 2021.

SOV	77.6	85.4	72.9	52.7

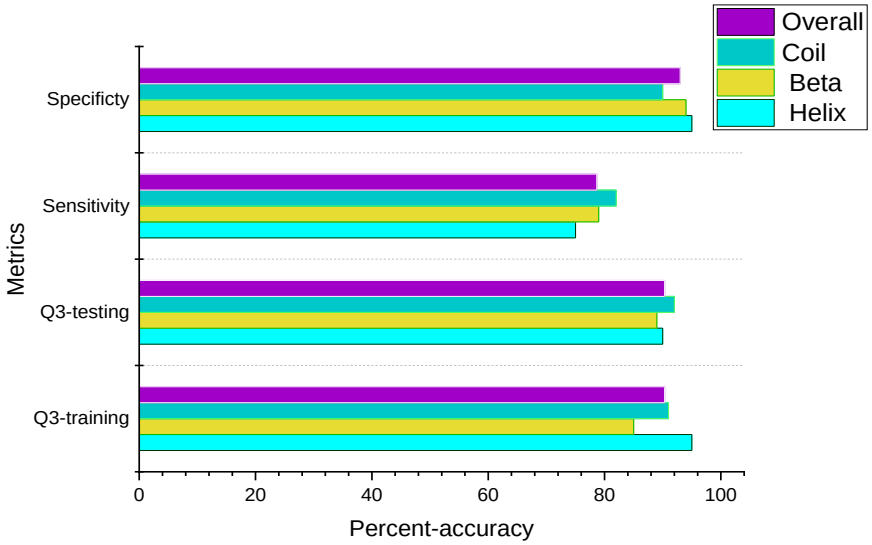

Figure 3. Accuracy of the proposed topology

Figure 3 shows the parameters like Q3-training, Q3-testing, Sensitivity, Specificity, for a given set of proteins which is tested. By using the terms of TP correct and FPerror, the achievement of allotment can be checked out. Regarding TN and FN, they are influenced by identical explanation. The forecasted class conforming to a pre-set threshold is determined by the calculated probabilities supplied by the output of a classification. The ROC is graphed by taking the TP rate and FP rate as the co-ordinate pairs. The region below the ROC assists to combine the achievement of all the tasks which are tested.

5. Comparative Analysis

Finally, the results obtained using proposed methodology is compared with the performance of other networks which depicted the secondary structure of QscR and is depicted in table 4 and 5 and in figure 4 to 6.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8629 - 8643
Received 16 February 2021; Accepted 08 March 2021.

Table 4. Comparative analysis of performance of the other methods in secondary structure prediction

Methods	Alpha (%)	Beta sheet(%)	Coil (%)
DSSP	54	13	-
STRIDE	54	11	-
MLNN	56.96	11.39	31.64
PSO-NN	59.07	12.23	28.69

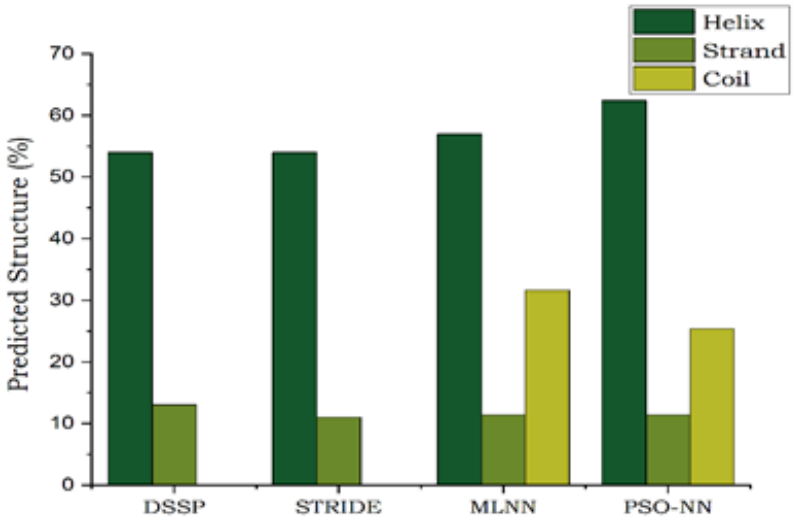

Figure 4. Comparative analysis of performance of the other methods in secondary structure prediction

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8629 - 8643
Received 16 February 2021; Accepted 08 March 2021.

Table 5. Performance comparison of Q3 and SOV of the proposed topology

	Qoverall (%)	Q_H (%)	Q_E (%)	Q_C (%)	Methodology
Q3	72.6	83.9	58.3	52.5	MLANN
SOV	75.6	85	72.9	42.6	
Q3	78.9	84.6	58.3	85.0	PSOANN
SOV	77.6	85.4	72.9	52.7	

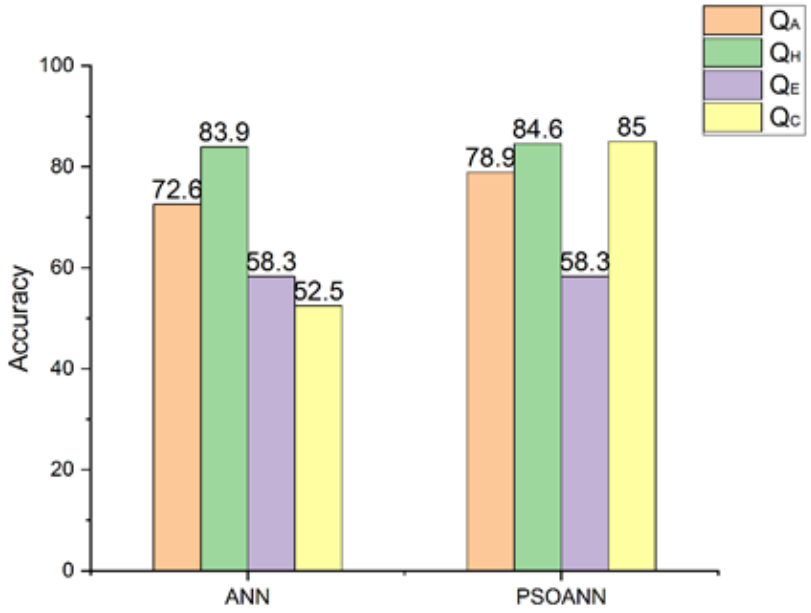

Figure 5. Overall Q3 comparison with other topologies

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8629 - 8643
Received 16 February 2021; Accepted 08 March 2021.

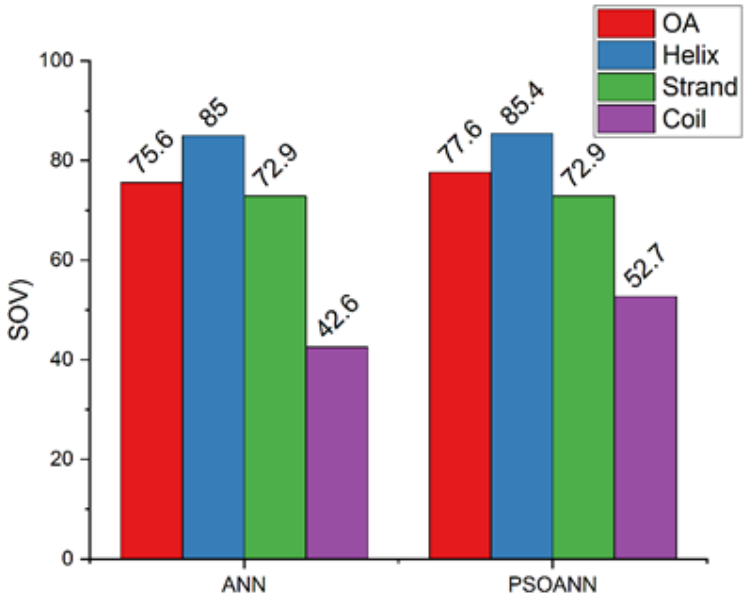

Figure 6. Overall SOV comparison with other topologies

From the overall comparison study, it is concluded that the proposed PSO-NN gives more and better prediction of secondary structure than the other topologies. Similarly, it exhibits greater accuracy and high per-residue accuracy than MLNN topology.

6. CONCLUSION

An innovative method, ANN which depends on PSO, is implemented to identify the secondary structure of a QscRprotein. The recommended predictor has achieved encouraging results and has surpassed a lot of other advanced predictors. In an individual dataset, an accuracy of 95% is achieved. The empirical performance determined by the suggested technique put up helpful hands for the detection of major protein modifications and this approach will be dynamic in the research domains where structure of proteins are predicted

Reference

1. Garnier, J., Osguthorpe, D. J., and Robson, B., “Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins”, Journal of Molecular Biology, 1978, 1, 97–120.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8629 - 8643
Received 16 February 2021; Accepted 08 March 2021.

2. Montgomerie, S., Sundaraj, S., Gallin, W., and Wishart, D., "Improving the accuracy of protein secondary structure prediction using structural alignment", BMC Bioinformatics, 2006, 301:301.
3. Kim, H. and Park, H., "Protein secondary structure prediction based on an improved support vector machines approach", Protein Engineering, 2003, 16, 553–560.
4. Garnier, J., Gibrat, J. F., and Robson, B., "GOR secondary structure prediction method version IV", Methods in Enzymology, 1996, 226:540–553.
5. E. Alba and R. Martí, "Metaheuristic Procedures for Training Neural Networks, Operations Research/Computer Science Interfaces Series", Springer, New York, NY, USA, 2006.
6. J. Yu, L. Xi, and S. Wang, "An improved particle swarm optimization for evolving feedforward artificial neural networks", Neural Processing Letters, 2007, vol. 26, no (3), pp. 217–231.
7. M. Conforth and Y. Meng, "Toward evolving neural networks using bio-inspired algorithms", in *IC-AI,* H. R. Arabnia and Y. Mun, Eds., CSREA Press, 2008, pp. 413–419.
8. Y. Da and G. Xiurun, "An improved PSO-based ANN with simulated annealing technique", Neurocomputing, 2005, vol. 63, pp. 527–533.
9. K. K. Kuok, S. Harun, and S. M. Shamsuddin, "Particle swarm optimization feedforward neural network for modeling runoff", International Journal of Environmental Science and Technology, 2010, vol. 7, no(1), pp. 67–78.
10. B. A. Garro, H. Sossa, and R. A. Vazquez, "Design of artificial neural networks using a modified particle swarm optimization algorithm", in Proceedings of the International Joint Conference on Neural Networks (IJCNN '09), IEEE, Atlanta, Ga, USA, 2009. pp. 938–945.
11. Huang C. L., & Dun J. F., "A distributed pso—svm hybrid system with feature selection and parameter optimization", Applied Soft Computing Journal, 2008, 8(4), 1381–1391.
12. ling Chen, H., Yang, B., jing Wang, S., Wang, G., zhong Li, H. and bin Liu, W., "Towards an optimal support vector machine classifier using a parallel particle swarm optimization strategy", Applied Mathematics and Computation, 2014, 239, pp.180-197.
13. C. Liu, W.-B. Du and W.-X. Wang, "Particle swarm optimization with scale-free interactions", PLoS ONE, 2014, 9.
14. H. X. Long, S. L. Wu and Y. Lv, "Protein structure prediction based on profile HMM and QPSO", Advanced Materials Research, 2014, 1004-1005:853–856,
15. J. Sun, V. Palade, Y. Cai, W. Fang and A. X. Wu, "Biochemical systems identification by a random drift particle swarm optimization approach", BMC Bioinformatics, 2014, 15.
16. A.S. Mohais, R. Mohais, C. Ward, and C. Posthoff, "Earthquake classifying neural networks trained with random dynamic neighborhood PSOs", in Proceedings of the 9th Annual Genetic and Evolutionary Computation Conference (GECCO '07), ACM, New York, NY, USA, 2007, pp. 110–117.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 8629 - 8643
Received 16 February 2021; Accepted 08 March 2021.

17. Chang Y, Yu G., "Multi-Sub-Swarm PSO Classifier Design and Rule Extraction", Int. Work. Cloud Computing Information Security, 2013, 104–107.
18. L. G. P. Hernández, K. R. Vázquez and R. G. Juárez, "Estimation of 3D protein structure by means of parallel particle swarm optimization", IEEE Congress on Evolutionary Computation (CEC), 2010, pp. 1–8.
19. M. H. Scalabrin, R.S. Parpinelli, C.M. Benitez and H.S. Lopes, "Population-based harmony search using GPU applied to protein structure prediction", Int. J. of Computational Science and Engineering, 2014, 9, 106–118.
20. Venter G., &Sobieszczanskisobieski J., "Particle swarm optimization", Aiaa Journal, 2013, 41(8), 129–132.
21. Ratnaweera A., Halgamuge S. K., & Watson H. C., "Self-organizing hierarchical particle swarm optimizer with time-varying acceleration coefficients", IEEE Press, 2004.
22. Amin Salih Mohammed, Saravana Balaji B, Saleem Basha M S, Asha P N, Venkatachalam K(2020),FCO — Fuzzy constraints applied Cluster Optimization technique for Wireless AdHoc Networks,Computer Communications, Volume 154,Pages 501-508.
23. Ponmagal, R.S., Karthick, S., Dhiyanesh, B. et al. Optimized virtual network function provisioning technique for mobile edge cloud computing. J Ambient Intell Human Comput (2020).
24. Ramamoorthy, S., Ravikumar, G., Saravana Balaji, B. et al. MCAMO: multi constraint aware multi-objective resource scheduling optimization technique for cloud infrastructure services. J Ambient Intell Human Comput (2020).
25. Basha, A.J., Balaji, B.S., Poornima, S. et al. Support vector machine and simple recurrent network based automatic sleep stage classification of fuzzy kernel. J Ambient Intell Human Comput (2020)
26. Balaji, B.S., Balakrishnan, S., Venkatachalam, K. et al. Automated query classification-based web service similarity technique using machine learning. J Ambient Intell Human Comput (2020)