Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5257 - 5266
Received 15 May 2021; Accepted 20 May 2021.

Protein Secondary Structure Prediction Using FFA Optimized ANN

[1]Durairaj M, [2]Sivakumar S, [3]Sangeetha B, [4]Saravannan K, [5]Saravanakumar K

[1]Department of Physics, Mahendra Engineering College, Mallasamudram West, Namakkal, Tamil Nadu, INDIA
[2]Department of Physics, Government Arts College (Autonomous), Salem, Tamilnadu, INDIA
[3]Department of EEE, AVS Engineering College, Salem, Tamilnadu, INDIA
[3]Department of Physics, AVS Engineering College, Salem, Tamilnadu, INDIA
[5]Department of Physics,Mahendra Institute of Technology, Namakkal,Tamilnadu, INDIA

ABSTRACT

The epidermal growth factor (EGF) family of RTKs plays a vital role in regulation of cell proliferation, differentiation, survival and migration. It carries both restricted and redundant function during developed stage of mammalian and maintains tissues at the adult stage. While the regulation carried by these receptors gets decreased, it will lead to many diseases like cancer among humans. Hence, the understanding about the function and regulation of RTK is essential for the development of drugs for human diseases. As the SS of a protein is responsible for the interactions among proteins, it is difficult for the scientists to understand their mutual relationships and functions. Hence the prediction of PSS is considered as a difficult task. Even though PSO based optimization topology exhibits higher accuracy in PSS prediction there may be some drawbacks, when it is subjected to high-dimensional space. Its convergence rate under high-dimensional space is very low. Hence to overcome this drawback, this work utilized a nature inspired firefly algorithm to tune ANN. It will results in improved convergence rate and also consumes less time with high accuracy.

Keywords

Protein Secondary Structure (PSS);Firefly; ANN; RTK

Introduction

Growth factors are essential for the growth, development and homeostasis of multicellular organisms. These growth factors are essential for cell to cell communications underlying embryonic tissue induction, cell survival, fate determination, tissue specialization, apoptosis and cell migration. These receptors transduce extracellular signals through the activation of intracellular messengers or through receptor translocation to the nucleus. The epidermal growth factor (EGF) family of RTKs is also referred as ErbB or HER receptors [1]. Hence, it is most widely studied for its role in development and physiology.

BPNN can be utilized in many applications. It relies on bias, weight and learning algorithm adopted in a NN design [2-8]. In this, steepest descent algorithm is normally implemented in BPNN topology. It is also called gradient method. In this topology, convergence time is key factor. Conjugate BP and LM BPNN is normally utilized to reduce the convergence time. As conjugate utilizes second order derivative, it converges with a less number of iterations[9,10].

LM BP is the standard topology, which is based on the nonlinear least square algorithm[11-13]. However, performance of these algorithms is high it needs a memory for optimization. Hence, numerous topologies is adopted to tune ANN.

The recent development in optimization topologies is extremely helpful to solve complex problems. They can also be implemented in the form of hybrid nature and are adopted as standard algorithm.

Among various types of algorithms GA plays a vital role. Thus, BPNN associated with GA will provide best solution and hence, it is named as GA BPNN. It performs well in any dynamic nature of environment. However, it exhibits slow convergence because of large search space [14-18].

In all the optimization procedures, weights, bias are updated periodically. Hence, this work proposed a novel nature based topology to optimize the function of ANN. Thus, the proposed FFA will exhibit convergence with short duration of time.

It is also observed that this proposed topology is never confined to a local minimum, and it is completed quickly if the convergence condition of a data set is already predefined.

FFA

In this topology, flies utilize the flash signals to fascinate other flies for mating. The movement of firefly would be random if there are no other insects in its proximity and its movement. This is because of light intensity of the firefly. The intensity (L) of light decreases as the distance (r) increases and thus most flies can communicate only up to several hundred meters. Thus, the formulated fitness function is always associated with the brightness of flash[19-22].

Accordingly, the intensity of the flashing light is calculated as

$$L_i = L_o e^{\left(-nd^2\right)} \qquad (1)$$

The distance between two flies can be calculated as

$$d_{ij} = \|f_i - f_j\| = \sqrt{\sum \left(f_i - f_j\right)^2} \qquad (2)$$

Thus by modifying the flash light absorption parameter, a quick convergence can be obtained. But in most of the cases it is fixed. Initially, any one of the fly is considered as a brightest one and rest of the flies are moved towards to that.

During this process, the distance and attractiveness of every firefly with respect to the brighter one is estimated. Finally, the flies are arranged on the basis of their performance.

Thus, the procedure adopted by FFA ANN topology is depicted in Figure 1.

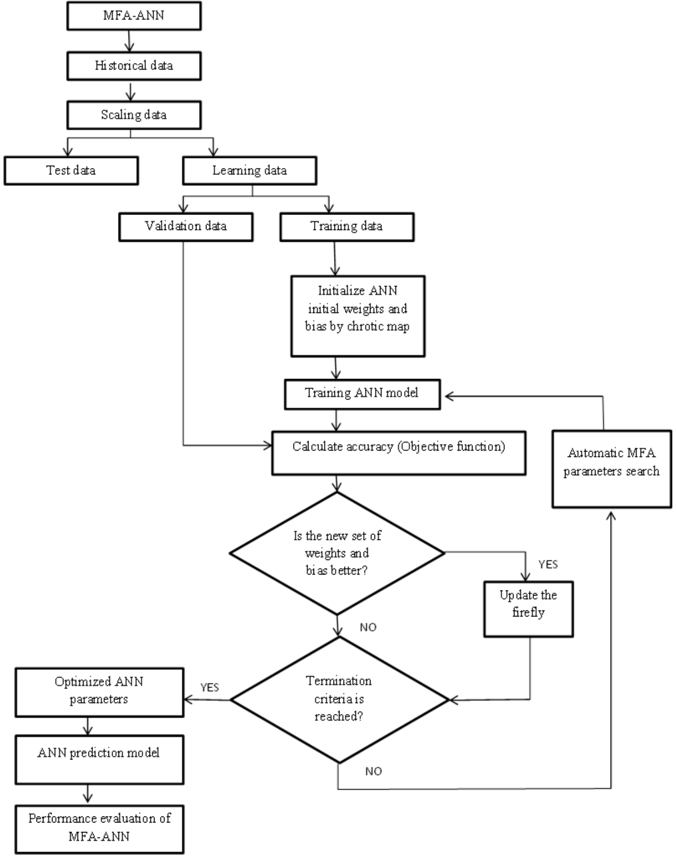
MFA-ANN
Historical data
Scaling data
Test data
Learning data
Validation data
Training data
Initialize ANN initial weights and bias by chrotic map
Training ANN model
Calculate accuracy (Objective function)
Automatic MFA parameters search
Is the new set of weights and bias better?
YES
Update the firefly
NO
Termination criteria is reached?
YES
Optimized ANN parameters
ANN prediction model
Performance evaluation of MFA-ANN
NO

Figure 1.FFA ANN topology

The proposed FF based BPNN converges with less number of iteration. The number of iteration and the rate of correct classification can be enhanced by increasing the population.

Study Area and Data

In this work, 100 proteins set for training and 5 protein set for testing were used. All these sets have 3 secondary structure classes (α-helix/ β-strand/coil). Each set was utilized as the validating and testing set. The parameter configuration used for FFA is depicted in Table 1.

Table 1. Parameter configuration of FFA

Parameters	Values
No of parameter	2
Range of the chosen parameter	-10 to 10
No. of Fireflies	50
No.of Iterations	200
α	0.81
γ	1.00
Δ	0.98

Results and Discussion

Among the training data set 47% were about coil, 31% were helix and 21% were strand. In the testing data set, it was about 48% of C, 31% of H and 21% of E.

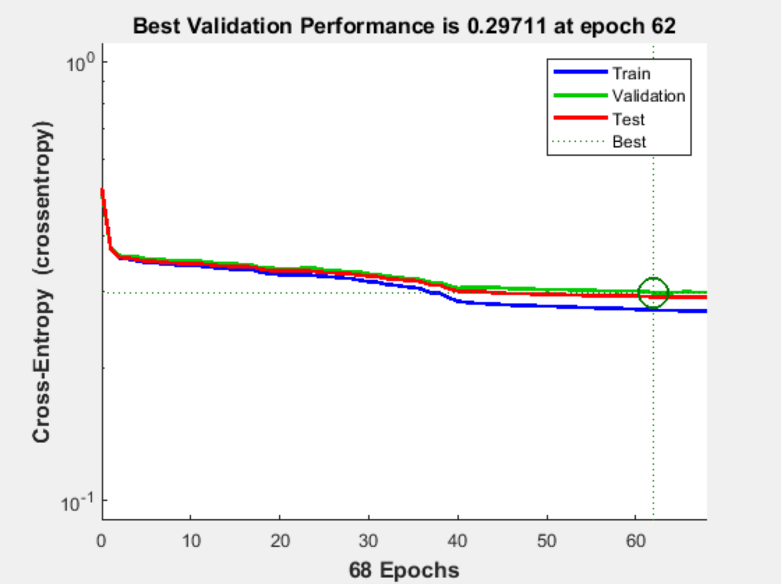

Figure 2. Best validation performance curve

The MSE obtained for an ANN model is depicted in Figure2. Thus from the above figure it is found at 62 epoch, the best validation is about 0.29. Thus, it is concluded that as the error of the network gets reduced, the efficiency is high.

Thus, the predicted secondary for the sequence of chain A of RTKs using FFA tuned NN is tabulated below.

Received 15 May 2021; Accepted 20 May 2021.

Table 2. Predicted secondary structure of RTKs under different topology

Methods		**Secondary Structure**				
Sequence (1-50)		GEAPNQALLR	ILKETEFKKI	KVLGSGAFGT	VYKGLWIPEG	EKVKIPVAIK
Structure	**DSSP**	**S*******	B**GGG***E	EEEEE*SS*E	EEEEEE**TT	S**BEEEEEE
	MLNN	CCCCCHHHHH	HHHHHHHCCC	CCCCCCCCCE	EEEEEECCCC	CCCCEEEEEE
	PSONN	CCCHHHHHHH	HHHHHHHCCC	EEEEECCCCC	EECCEEECCC	CEEEEEEEHH
	Proposed FFANN	CCCCHHHCHH	CCCHHHEEEE	EEECCCCCCE	EEEEEEECCC	CCCEEEEEEE
Sequence(51-100)		ELREATSPKA	NKEILDEAYV	MASVDNPHVC	RLLGICLTST	VQLITQLMPF
Structure	**DSSP**	**SSTT*THH	HHHHHHHHHH	HTT***TTB*	*EEEEEESSS	EEEEEE**TT
	MLNN	HHHCCCCCCC	HHHHHHHHHH	HHHCCCCCHE	EEEEEECCCC	HHHHHHHHHH
	PSONN	HHCCCCCHHH	HHHHHHHHHH	CCCCCCCCEE	EEEEEEECCC	EEEEEHHHCC
	Proposed FFANN	EECCCCCHHH	HHHHHHHHHH	HHHCCCCCEE	EEEEEEECCC	CEEEEECCCC
Sequence(101-150)		GCLLDYVREH	KDNIGSQYLL	NWCVQIAKGM	NYLEDRRLVH	RDLAARNVLV
Structure	**DSSP**	*BHHHHHHHH	TTT**HHHHH	HHHHHHHHHH	HHHHHTTEE*	S***GGGEEE
	MLNN	HHHHHHHHHC	CCCCCHHHHH	HHHHHHHHHH	HHHHHHHHHH	HHHHHHHEEE
	PSONN	CCHHHHHHHC	CCCEEEEHHH	HHHHHHHHHH	HHHHHHCHHH	HHHHHHHHHH
	Proposed FFANN	CCHHHHHHHH	CCCCCHHHHH	HHHHHHHHHH	HHHHHHCCCH	HHHHHHHHHE
Sequence(151-200)		KTPQHVKITD	FGLAKLLGAE	EKEYHAEGGK	VPIKWMALES	ILHRIYTHQS
Structure	**DSSP**	EETTEEEE**	*TT*EESS**	**********	**GGG**HHH	HHS****HHH
	MLNN	ECCCCEEEEH	HHHHHHHCCC	CCHEEHCCCC	CCEEEECHHH	HHHHCCCCCC
	PSONN	CCCCEEEEEC	CCCEEEECCC	CCCCCCCCCC	EEEECCCHHH	HHHHHCCCCC
	Proposed FFANN	CCCCEEEEEC	CCCCEECCCC	CCEEECCCCC	EEEECCCHHH	HHHCCCCCHH
Sequence(201-250)		DVWSYGVTVW	ELMTFGSKPY	DGIPASEISS	ILEKGERLPQ	PPICTIDVYM
Structure	**DSSP**	HHHHHHHHHH	HHHTTS**TT	SSS*GGGHHH	HHHHT*****	*TTB*HHHHH
	MLNN	CHEEEEEEHH	HEEECCCCCC	CCCCCHHHHH	HHHCCCCCCC	CCCCCHHHHH
	PSONN	CCEEEEEEEE	ECCCCCCCCC	CCCCCHHHHH	HHHHCCCCCC	CCCCHHHHHH
	Proposed FFANN	HHHHHHHHHH	HHHHCCCCCC	CCCCHHHHHH	HHHHCCCCCC	CCCCCHHHHH
Sequence(251-300)		IMVKCWMIDA	DSRPKFRELI	IEFSKMARDP	QRYLVIQGDE	RMHLPSPTDS
Structure	**DSSP**	HHHHHT*SSG	GGS**HHHHH	HHHHHHTTSH	HHHB**TT*S	S*********
	MLNN	HHHHHHCCCC	CCCCCHHHHH	HHHHHHCCCC	CCEEEEECCC	CCCCCCCCCC
	PSONN	HHHHHHHHCC	CCCCHHHHHH	HHHHHHHCCC	HHHHHCCCCC	CCCCCCCCCC
	Proposed FFANN	HHHHHHHCCH	HHCCCHHHHH	HHHHHHHHCC	HHHCCCCCCC	CCCCCCCCCC
Sequence(300-327)		NFYRALMDEE	DMDDVVDADE	YLIPQQG		

Structure	**DSSP**	********SS	**TTB**TTT	*******		
	MLNN	HHHHHHHHHH	HHHHHHHHHH	HHHCCCC		
	PSONN	CCCCCCCCHH	HHHHHHHHCC	CCCCCCC		
	Proposed FFANN	CCCCCCCCCC	CCCCCCCCCC	CCCCCCC		

From the above analysis, the predicted H is about 40.36%and E is about 17.43%, hence the proposed RTKs is of mixed class rather than all helix or beta.

The impact of AA's composition in PSS is investigated. This is displayed in terms of 3 type SS. Similarly, the effect of physiochemical property of AA over SS prediction is also investigated here.

It can be seen that AAs are not present in uniform quantities in SS. There are very few AAs such as Glutamine andLeucine and have the highest number of helix residues while Alanine and Serine have the exact number of helix residues in it. In the case of strand, Lysine, and Leucine seem to have the highest content while most of the acids like Asparagine, Aspartic acid does not exist strand. Similarly, large numbers of residues of coil are in Glutamic acid, Leucine and while Alanine, Asparagine seem to have the same content in the coil and the same is depicted in Figure 3.

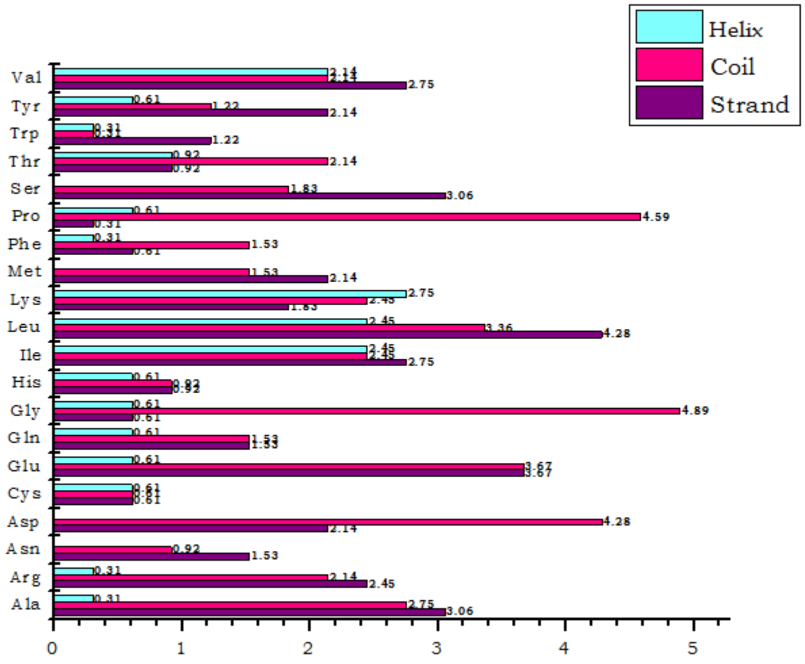

Figure 3. Content of AAs in SS

Received 15 May 2021; Accepted 20 May 2021.

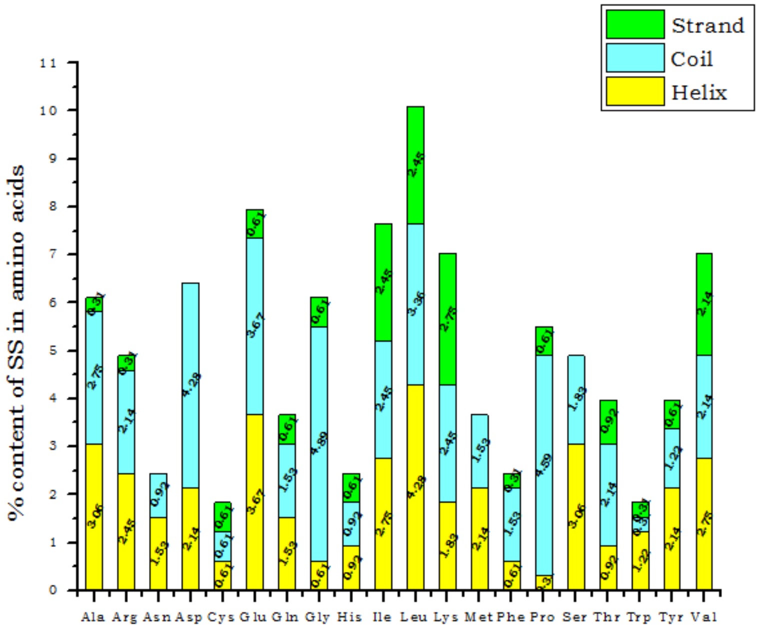

Figure 4. Percentage of SS in an AA

From the above Figure 4, it is concluded that all AAs exhibit helix nature and AAs such as Asparagine, Aspartic acid have not exhibit strand nature. Thus, it is concluded that α-helix plays a vital role in 2J5F protein.

A lot of measurements are adopted to examine the accuracy of prediction topology. Q3 accuracy represents the correctly predicted state in percentage. Thus, the Q3 accuracy is represented as Q_H, Q_E and Q_C[23].

SOVprovides the overlap of predicted 3 SS by observing the predicted and measured segments [24,25] and is shown in Table 3.

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

	Q overall (%)	**Q_H (%)**	**Q_E (%)**	**Q_C (%)**
Q3	67.9	95.4	32.7	77.1
SOV	53.8	95.6	31.7	42.0

Comparison with Other Methods

Finally, the results obtained using this proposed methodology is compared with the performance of other networks which depicted the SS of RTKs.

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

Method	**Alpha (%)**	**Beta sheet(%)**	**Coil (%)**
DSSP	33	15	-
STRIDE	35	16	-

Received 15 May 2021; Accepted 20 May 2021.

MLNN	38.53	16.20	45.25
PSO-NN	38.83	19.26	31.19
FFA-NN	40.36	17.43	42.21

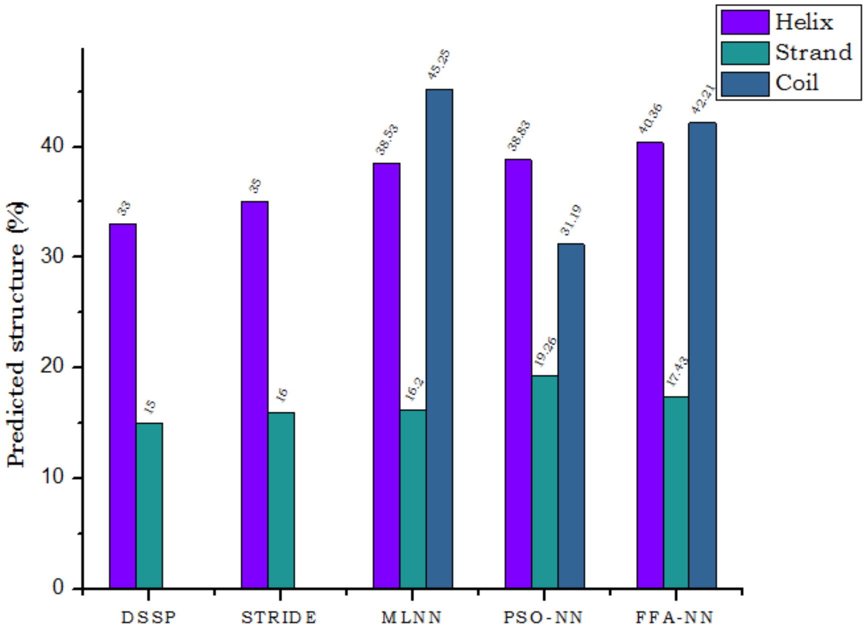

Figure 5. Predicted secondary structure

From the Table 4 and Figure 5, it is concluded that the proposed FFA-NN gives more and better prediction of SS than other topologies.

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

	Qoverall (%)	**QH (%)**	**QE (%)**	**QC (%)**	**Methodology**
Q3	51.7	75.2	24.8	56.2	**MLNN**
SOV	44.7	62.8	27.2	43.9	
Q3	55	74.3	29.2	62.9	**PSONN**
SOV	52.3	74.8	32.0	54.9	
Q3	67.9	95.4	32.7	77.1	**FFANN**
SOV	53.8	95.6	31.7	42.0	

Received 15 May 2021; Accepted 20 May 2021.

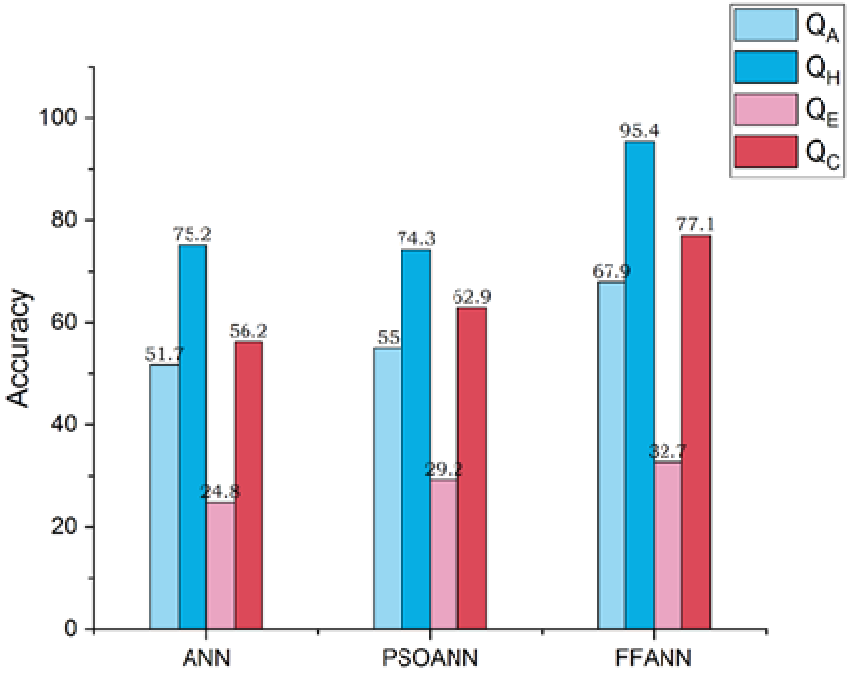

Figure 6. Accuracy of the prediction methodology

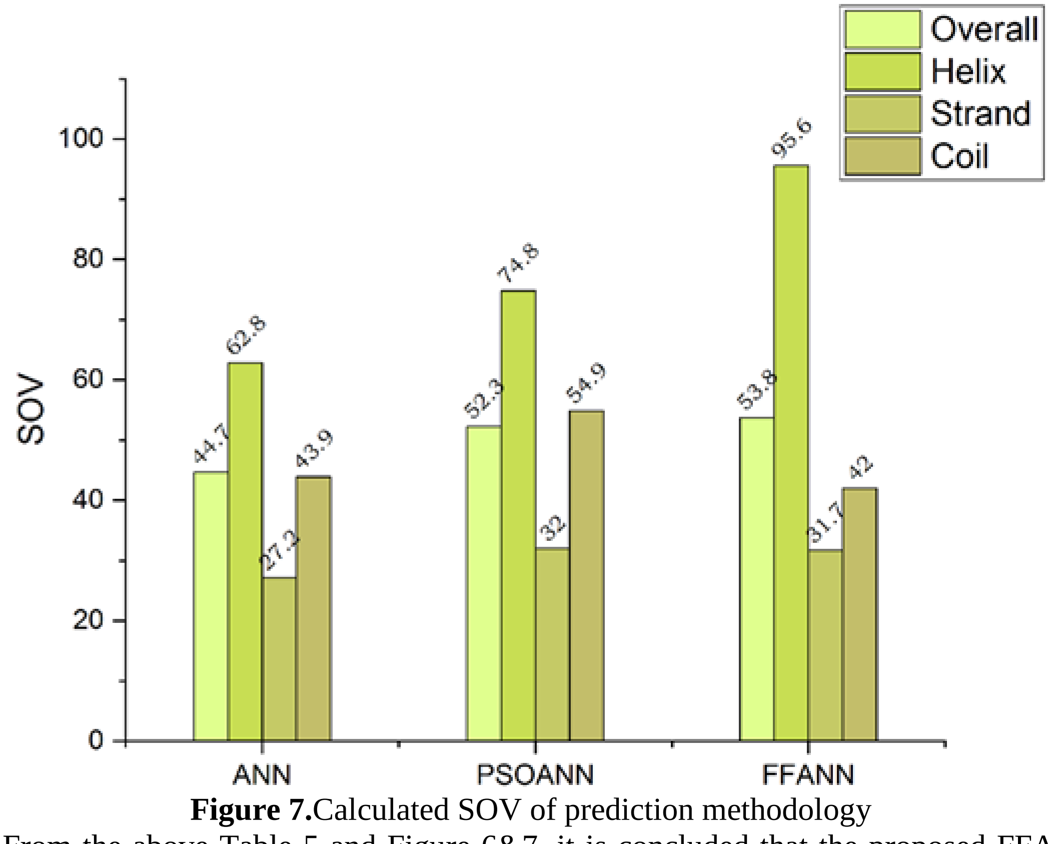

Figure 7. Calculated SOV of prediction methodology

From the above Table 5 and Figure 6&7, it is concluded that the proposed FFANN has improved prediction accuracy than the other conventional methodology.

Conclusion

A new topology called firefly trained neural fields which can able to tune NN automatically is designed for PSS prediction. In this work, a novel method firefly based ANN is implemented to identify SS of a protein. This proposed topology obtained is a promising accuracy over SS prediction than other topologies illustrated previously. This scheme automatically tunes the NN using optimization topology called firefly. The results obtained from the experimental setup have proven that this proposed topology can be utilized for predicting SS of any protein and hence, it would be more powerful in SS prediction research domain.

References

[1] Karpov, O. A., Fearnley, G. W., Smith, G. A., Kankanala, J., McPherson, M. J., Tomlinson, D. C., ... &Ponnambalam, S. (2015). Receptor tyrosine kinase structure and function in health and disease. *AIMS Biophysics*, *2*(4), 476-502.2.

[2] Baker, D., &Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, *294*(5540), 93-96.

[3] Quan, L., Lv, Q., & Zhang, Y. (2016). STRUM: structure-based prediction of protein stability changes upon single-point mutation. *Bioinformatics*, *32*(19), 2936-2946.

[4] Wang, S., Peng, J., Ma, J., &Xu, J. (2016). Protein secondary structure prediction using deep convolutional neural fields. *Scientific reports*, *6*(1), 1-11.

[5] Whisstock, J. C., &Lesk, A. M. (2003). Prediction of protein function from protein sequence and structure. *Quarterly reviews of biophysics*, *36*(3), 307.

[6] Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., ...& Friedberg, I. (2013). A large-scale evaluation of computational protein function prediction. *Nature methods*, *10*(3), 221-227.

[7] Eisenberg, D. (2003). The discovery of the α-helix and β-sheet, the principal structural features of proteins. *Proceedings of the National Academy of Sciences*, *100*(20), 11207-11210.

[8] Yoo, P. D., Zhou, B. B., &Zomaya, A. Y. (2008). Machine learning techniques for protein secondary structure prediction: an overview and evaluation. *Current Bioinformatics*, *3*(2), 74-86.

[9] Qian, N., &Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *Journal of molecular biology*, *202*(4), 865-884.

[10] Holley, L. H., &Karplus, M. (1989). Protein secondary structure prediction with a neural network. *Proceedings of the National Academy of Sciences*, *86*(1), 152-156.

[11] Rost, B., & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *Journal of molecular biology*, *232*(2), 584-599.

[12] Zvelebil, M. J., Barton, G. J., Taylor, W. R., & Sternberg, M. J. (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *Journal of molecular biology*, *195*(4), 957-961.

[13] Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology*, *292*(2), 195-202.

[14] Busia, A., &Jaitly, N. (2017). Next-step conditioned deep convolutional neural networks improve protein secondary structure prediction. *arXiv preprint arXiv:1702.03865*.

[15] Buchan, D. W., Minneci, F., Nugent, T. C., Bryson, K., & Jones, D. T. (2013). Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic acids research*, *41*(W1), W349-W357.

[16] Mirabello, C., &Pollastri, G. (2013). Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility. *Bioinformatics*, *29*(16), 2056-2058.

[17] Sun, T., Zhou, B., Lai, L., & Pei, J. (2017). Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC bioinformatics*, *18*(1), 1-8.

[18] Gomes, J., Ramsundar, B., Feinberg, E. N., &Pande, V. S. (2017). Atomic convolutional networks for predicting protein-ligand binding affinity. *arXiv preprint arXiv:1703.10603*.

[19] Yang, X. S. (2009, October). Firefly algorithms for multimodal optimization. In *International symposium on stochastic algorithms* (pp. 169-178). Springer, Berlin, Heidelberg.

[20] Gandomi, A. H., Yang, X. S., Talatahari, S., &Alavi, A. H. (2013). Firefly algorithm with chaos. *Communications in Nonlinear Science and Numerical Simulation*, *18*(1), 89-98.

[21] Yang, X. S. (2010). *Nature-inspired metaheuristic algorithms*. Luniver press.

[22] Yang, X. S. (2010). Firefly algorithm, stochastic test functions and design optimisation. *International journal of bio-inspired computation*, *2*(2), 78-84.

[23] Kryshtafovych, A., Barbato, A., Fidelis, K., Monastyrskyy, B., Schwede, T., &Tramontano, A. (2014). Assessment of the assessment: evaluation of the model quality estimates in CASP10. *Proteins: Structure, Function, and Bioinformatics*, *82*, 112-126.

[24] Hua, S., & Sun, Z. (2001). A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *Journal of molecular biology*, *308*(2), 397-407.

[25] Liu, T., & Wang, Z. (2018). SOV_refine: A further refined definition of segment overlap score and its significance for protein structure similarity. *Source code for biology and medicine*, *13*(1), 1-10.