SECONDARY STRUCTURE PREDICTION OF PROTEIN EGFR TYROSINE KINASES USING OPTIMIZED NEURAL TOPOLOGY

[1]Durairaj M, [2]Sivakumar S

[1]Department of Physics, Mahendra Engineering College, Mallasamudram West, Namakkal- 637503, Tamil Nadu, INDIA
[2]Department of Physics, Government Arts College (Autonomous), Salem- 636007 Tamil Nadu, INDIA

ABSTRACT

Receptor tyrosine kinases (RTKs) are essential components of signalling transduction that mediate cell-to-cell communication. The RTK family includes, epidermal growth factor receptor (EGFR), platelet-derived growth factor receptors (PDGFR), fibroblast growth factor receptors (FGFRs), vascular endothelial growth factor receptors (VEGFR), Ephs (ephrin receptors), and the insulin receptor, etc. Hence, it is essential to predict the secondary structure of Receptor tyrosine kinases (RTKs). However, Protein secondary structure prediction is one of the most important and challenging problems in bioinformatics. Machine learning techniques have been applied to solve the problem and have gained substantial success in this research area. Although, neural network based prediction becomes more popular, the training methodology involves more processing. Hence, in order to overcome this drawback, this work a new topology called PSO trained Neural Fields which can able to tune NN automatically is designed for protein SS prediction. The results are compared with other prediction mechanisms. The obtained results are more accurate and better than the corresponding other mechanisms.

Keywords: Receptor tyrosine kinases, epidermal growth factor receptor and neural network.

1. INTRODUCTION

Bioinformatics is the science of using computer and information technology in biology. One of the main challenges in structural bioinformatics is the prediction of protein tertiary structure which aims to find the 3D structure with the lowest free energy of a protein from its sequence. The significance of PSP is that scientists believe the function of a protein in a cell depends on its folding and structure. Findings of the native conformation can be used in designing drugs, recognizing the function of inheritance and contagious diseases and prevention of some diseases like Alzheimer, cystic fibrosis and mad cow. However, common computational methods are unable to solve PSP problem as their computational time and complexity are both expensive. At the other hand, experimental methods for PSP based on NMR and X-ray despite of significant development are still time consuming and expensive. The computational methods are divided into three groups: homology modeling, threading, and ab-initio. The first two are knowledge-based methods using the information of known protein structures, but the third is based on physicochemical parameters and solves the problem by inspiring what is happening in the nature; they predict the 3D structure of proteins given their primary sequences [1-6].The secondary-structure prediction approaches in today can be categorized into three groups: neighbor-based, model-based, and

metapredictor-based [7]. 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 metapredictor -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 [8] were based on neural network (NN) learning techniques [9]. 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 [10] 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 [11,12], support vector machines [13, 14], hidden Markov models [15, 16], artificial neural network [19], and bidirectional recurrent neural network(BRNN) [17].However, Artificial Neural Network (ANN) design is a complex task because its performance depends on the architecture, the selected transfer function, and the learning algorithm used to train the set of synaptic weights[20,21]. To overcome this drawback, a new methodology that automatically designs an ANN using particle swarm optimization algorithms (PSO) is implemented. Among various kinds of proteins, Protein receptor tyrosine kinases (RTKs) are the largest class of enzyme-linked receptors, both receptors and enzymes that bind to ligands and phosphorylate tyrosine residues of target proteins. All RPTKs are composed of three components: an extracellular domain containing a ligand binding site, a single transmembrane hydrophobic alpha helix region, and an intracellular domain containing tyrosine protein kinase (RTKs) activity.Receptor tyrosine kinase (RTKs) is a high affinity cell surface receptor for many polypeptide growth factors, cytokines and hormones. Of the 90 unique tyrosine kinase genes identified in the human genome, 58 encode receptor tyrosine kinase proteins. Receptor tyrosine kinases have not only been shown to be key regulators of normal cellular processes, but also play a key role in the development and progression of many types of cancer. In this work, in order to incorporate the prediction of structure of (RTKs) a new hybrid topology called PSO trained Neural Network is introduced.

2. PSO ALGORITHM

PSO Algorithm In comparison to the algorithms that are used to find the near optimal solutions, PSO produces acceptable results. PSO concept is close to evolutionary algorithms whilst requires simpler programming to be implemented. Every particle of the PSO is one solution of the search space. The aim of this algorithm is that the whole particle reaches the optimal global point. In this way, the speed of each particle changes dynamically, considering the last move and the neighbor particles. Therefore, position and velocity vectors are Xi = (Xi1,…,Xid) and Vi = (Vi1,…,Vid) for the ith particle in the D-dimensional search space, respectively. Velocity changes give each particle the ability to search around the best local or global point.

The velocity of the particle i is represented as shown in equation 1.

$$V_i^{(t+1)} = \omega \times V_i^{(t)} + c_1 \times r_1 \times (P_i^{best} - X_i^{(t)}) + c_2 \times r_2 \times (G^{best} - X_i^{(t)}) \quad (.1)$$

- $x_i^{(t)}$ - the present position of the particle i at iteration t
- t - Iteration pointer
- p_i^{best} - the best position of the particle i until iteration t
- G^{best} - the global best position of entire swarm until iteration t
- c_1, c_2 - the acceleration coefficients varies between 0 and 4
- $v_i^{(t)}$ - the velocity between the step size $x_i(t)$ and $x_i(t+1)$
- ω - the inertia weight/damping factor which decreases from 0.9 to 0.4 used to control the contact of new velocity with its previous velocity
- r_1,r_2 - Random variables with a range of [0, 2]

The inertia weight ω is calculated by the following equation 2.

$$\omega = \omega_{\max} - \frac{\omega_{\max} - \omega_{\min}}{iter_{\max}} \times iter \quad \text{.....(2)}$$

Where

ω_{max} - initial weight

ω_{max} - final weight

$iter_{max}$ - maximum iteration number

iter - current iteration number

A new velocity is calculated in the direction of p_i^{best} and G^{best} to execute a change in the current search point. Every particle attempts to migrate from its current position to the new position by using the modified velocity given below in equation 3.

$$X_i^{(t+1)} = X_i^{(t)} + V_i^{(t+1)} \quad \text{.....} \quad (3)$$

In PSO optimization, all the particles attempt to migrate for improved positions. The mutual effort of all the particles, the best position (optimal solution) is obtained. This iteration comes to an end after attaining the stopping condition. Common types of stopping conditions are number of iteration of the algorithm, number of iterations while the final update of the global best solution and a predefined fitness value. The PSO algorithm is completed after certain iterations after reaching the fitness value close enough to the desired output. There are surplus versions of PSO are available for discrete optimization, constrained optimization and for multi objective optimization.

2.1.ARTIFICIAL NEURAL NETWORK (ANN)

ANN is an approximation function mapping inputs to outputs. A typical network consists of three layers of neurons namely, input, hidden, and output; in which each neuron acting as an independent computational element. Input layer is defined as a layer of neurons receiving inputs directly from outside the network. Layer of a network that is not connected to the network output called hidden layer and layer whose output is passed to the world outside the network is output layer[18]. Weight functions apply weights to an input to get weighted inputs, as specified by a particular function. Different algorithms can apply to minimize the network error during the ANN training. This usually happens by finding the correct tune of the network in which is depended to weights, biases, number of neurons in the hidden layer and iteration number. Among these parameters, the training algorithm is important in a way it reaches to optimum weights and biases. Generally, the performance function of the ANN models during training process is assessed using sum of squares of the errors as follows:

$$E(W) = \frac{1}{2}\sum_{p=1}^{p}(d_p - y_p)^2 = \frac{1}{2}\sum_{p=1}^{p}(e_p)^2 \ , p = mT$$

Where

T is the total number of training samples,

m is the number of output layer neurons,

W represents the vector containing all the weights in the network,

yp is the actual network output, and

dp is the desired output.

Achieving optimum number of neurons in the hidden layer is often obtained through a trial and error procedure. Aside of the abovementioned parameters, correct selection of the input variables which affect the target variable is considered as one of the most important stages dealing with ANN models.

2.2. ANN TRAINING WITH PSO ALGORITHM

To train the ANN with PSO algorithm, the following procedures are taken under consideration. The algorithm is utilized to find the optimum weights and biases of the ANN model. Weights and biases' values form search space of the algorithm which is of n dimensions. The n is the total number of weights (and biases) that needs to be optimized. Each particle has a position vector and a velocity vector of n-dimensions. Here both weights and biases are shown by W. The optimal set of weights is obtained by flying the particles around the search space. At each iteration, the algorithm comes up with a set of weights that their fitness is assessed. It happens by assigning these weights to the nodes and predicting the target value. Afterwards, the accuracy of the prediction through assigned weights is evaluated

as the difference between actual and predicted values which should be minimized through optimization process. In this regard, the best fitness the particle has been achieved so far is considered as its personal best.

Similarly, the best fitness of the swarm is used as the global best. This process is repeated for a specific number of iteration until the optimized weights for the ANN are yielded. The steps for a PSO optimized ANN is given below. For a three-layered perceptron, W[1] and W[2] represent the connection weight matrix between the input layer and the hidden layer, and between the hidden layer and the output layer respectively. Applying PSO algorithm to train the multi-layer perceptron, the ith particle is denoted by:

$$W_i = \left\{W_i^{[1]}, W_i^{[2]}\right\} \quad \text{.....(3)}$$

$$P_i = \left\{P_i^{[1]}, P_i^{[2]}\right\} \quad \text{.....(4)}$$

$$P_b = \left\{P_b^{[1]}, P_b^{[2]}\right\} \quad \text{.....(5)}$$

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

where j = 1, 2; m = 1,. . . ,Mj ; n = 1,. . . ,Nj ;

Mj and Nj are the row and column sizes of the matrices

W, P, and V; r and s are positive constants;

a and b are random numbers in the range from 0 to 1;

t is the time step between observations and is often taken as unity;

V" and W" represent the new values.

Applying Equation,

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

the new velocity of the particle is computed by using its previous velocity and the distances of its current position from the best experiences both in its own and as a group. The second element on the right-hand side of Equation, represents the private thinking of the particle itself whilst the social part, i.e., the third element on the right hand side of Equation, denotes the collaboration among the particles as a group. The new position according to the new velocity can be determined by Equation

$$W_i^{"[j]} = W_i^{[j]} + V_i^{[j]}t$$

The fitness function f is the mean squared error and is defined as:

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

Where

F is the fitness value,

n is the number of data points.

2.3. STUDY AREA AND DATA

In this work, 75 proteins set for training and 5 protein set for testing were used. All these sets have a representative mix of the three secondary structure classes, α-helix, β-strand and coil. Each set was used as the validating set and as the testing set.

Table 1: The parameter configuration used in PSO

S.No	Parameters	Values
1.	Particles	100
2.	C1	1.2
3.	C2	1
4.	Max generation	500

3. RESULTS AND DISCUSSION

3.1. HYDROPHILICITY PLOT

A hydrophilicity plot is a quantitative analysis of the degree of hydrophobicity or hydrophilicity of amino acids of a protein. It is used to characterize or identify possible structure or domains of a protein.

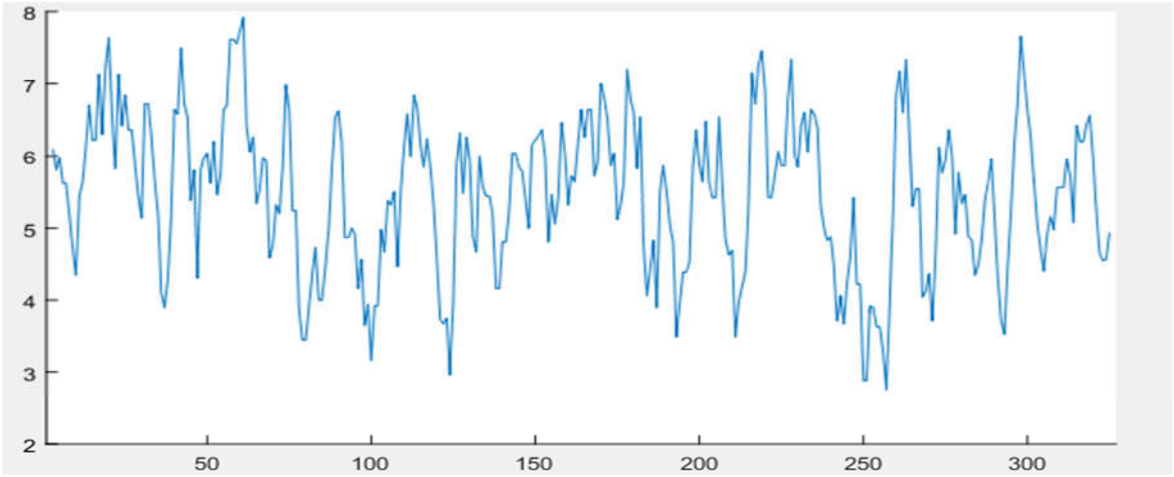

Figure 1. hydrophilicity plot of QscR

From the figure 1, it is concluded that amino acids shows positive for hydrophobicity, these amino acids may be part of alpha-helix spanning.

Table 2: Predicted secondary structure of QscR 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	CCCCHHHCHH	CCCHHHEEEE	EEECCCCCCE	EEEEEEECCC	CCCEEEEEEE
	Proposed PSONN	CCCHHHHCHH	CCCHHHEEEE	EEECCCHCCE	EEEEEEECHC	CCCEEEEEEE
Sequence(51-100)		ELREATSPKA	NKEILDEAYV	MASVDNPHVC	RLLGICLTST	VQLITQLMPF
Structure	DSSP	**SSTT*THH	HHHHHHHHHH	HTT***TTB*	*EEEEEESSS	EEEEEE**TT
	MLNN	EECCCCCHHH	HHHHHHHHHH	HHHCCCCCEE	EEEEEEECCC	CEEEEECCCC
	Proposed PSONN	EECCCCCHHH	HHHHHHHHHH	HHHCHCCCEE	EEEEEEECCC	CEEEEEHCCC
Sequence(101-150)		GCLLDYVREH	KDNIGSQYLL	NWCVQIAKGM	NYLEDRRLVH	RDLAARNVLV
Structure	DSSP	*BHHHHHHHHH	TTT**HHHHH	HHHHHHHHHH	HHHHHTTEE*	S***GGGEEE
	MLNN	CCHHHHHHHH	CCCCCHHHHH	HHHHHHHHHH	HHHHHHCCCH	HHHHHHHHHE
	Proposed PSONN	CCHHHHHHHH	CCCCCHHHHH	HHHHHHHHHH	HHHHHHCCCH	HHHHHHHHHE
Sequence(151-200)		KTPQHVKITD	FGLAKLLGAE	EKEYHAEGGK	VPIKWMALES	ILHRIYTHQS
Structure	DSSP	EETTEEEE**	*TT*EESS**	*********	**GGG**HHH	HHS****HHH
	MLNN	CCCCEEEEEC	CCCCEECCCC	CCEEECCCCC	EEEECCCHHH	HHHCCCCCHH
	Proposed PSONN	CCCCEEEEEC	CCCCEECCCC	CCEEECCCCC	EEEECCCHHH	HHHCCCCCHH
Sequence(201-250)		DVWSYGVTVW	ELMTFGSKPY	DGIPASEISS	ILEKGERLPQ	PPICTIDVYM
Structure	DSSP	HHHHHHHHHH	HHHTTS**TT	SSS*GGGHHH	HHHHT*****	*TTB*HHHHH
	MLNN	HHHHHHHHHH	HHHHCCCCCC	CCCCHHHHHH	HHHHCCCCCC	CCCCCHHHHH
	Proposed PSONN	HHHHHHHHHH	HHHHCCCCHC	CCCCHHHHHH	HHHHCCCCCC	CCCCCHHHHH
Sequence(251-300)		IMVKCWMIDA	DSRPKFRELI	IEFSKMARDP	QRYLVIQGDE	RMHLPSPTDS
Structure	DSSP	HHHHHT*SSG	GGS**HHHHH	HHHHHHTTSH	HHHB**TT*S	S*********
	MLNN	HHHHHHHCCH	HHCCCHHHHH	HHHHHHHHCC	HHHCCCCCCC	CCCCCCCCCC
	Proposed PSONN	HHHHHHHCCH	HHCCCHHHHH	HHHHHHHHCC	HHHCCCCCCC	CCCCHCCCCC
Sequence(300-327)		NFYRALMDEE	DMDDVVDADE	YLIPQQG	---	---
Structure	DSSP	********SS	**TTB**TTT	*******	---	---
	MLNN	CCCCCCCCCC	CCCCCCCCCC	CCCCCCC	---	----
	Proposed PSONN	CCCCCCCCCC	CCCCCCCCCC	CCCCCCC	----	---

3.2.ASSESSMENT OF PREDICTION ACCURACY

Four routinely used assessment criteria were adopted here, that is, sensitivity (SN), specificity (SP), accuracy (ACC), and AUC (area under Receiver Operating Characteristic curve): where TP, TN, FP, and FN were the abbreviations of true positives, true negatives, false positives, and false negatives. The experimental results were given in Table 3.

Table 3: The prediction performance of algorithms

S.No	Methodology	SN (%)	SP (%)	ACC (%)
1.	MLNN	90.55	93.17	93.28
2.	PSO-NN	91.73	94.59	94.04

The performances of our proposed method PSONN in comparison with MLNN is shown in Figure2. Figure 2 show that PSONN achieves the best performance than MLNN.

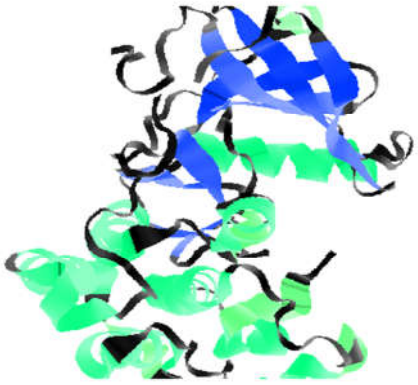

Figure 2. 3D structure of QscR

3.2. COMPARISON WITH OTHER METHODS

Listed in Table 4 were the experimental results performed by state-of-the-art methods.

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

S.NO	Method	Alpha (%)	Beta sheet (%)	Coil
1.	DSSP	33	15	-
2.	STRIDE	35	16	-
3.	MLNN	54.13	16	40
4.	PSO-NN	55.09	17	40

From the figure 7 and table 1, it is concluded that the proposed NN gives more and better prediction of secondary structure than the other topologies.

4. CONCLUSION

In this work, a novel method PSO based ANN is implemented to identify secondary structure of a protein. The proposed predictor achieved a promising results and outperformed many other state-of-the-art predictors. This scheme automatically tunes the neural network using optimization topology called PSO. It achieved the accuracy about 97% on the independent dataset. The experimental performance indicated that the proposed method could

be useful in assisting the discovery of important protein modifications and would be powerful in protein structure prediction research domains.

REFERENCES

1. Rylance, G. 2004. Applications of genetic algorithms in protein folding studies. First year report, School of Chemistry, University of Birmingham, England.
2. Sikder, A.R., Zomaya, A.Y. 2005. An Overview of protein-folding techniques: issues and perspectives. Int J Bioinf Res Appl 1, 121–143.
3. B. Rost, C. Sander, Prediction of protein secondary structure at better than 70% accuracy, J. Mol. Biol. 232 (1993) 584.
4. P. Baldi, S. Brunak, P. Frasconi, G. Soda, G. Pollastri, Ploiting the past and the future in secondary structure prediction, Bioinformatics 15 (11) (1999) 937.
5. G. Pollastri, A. McLysaght, Porter: a new, accurate server for protein secondary structure prediction, Bioinformatics 21 (2005) 1719.
6. P. Baldi, S. Brunak, Bioinformatics – The Machine Learning Approach, MIT Press, Cambridge, MA, 2001.
7. Hae-Jin Hu, Robert W. Harrison, Phang C, Tai, and Yi Pan, "Current Methods for Protein Secondary-Structure Prediction Based on Support Vector Machines", Ch.1, Knowledge Discovery in Bioinformatics: Techniques, Methods, and Applications, 2007.
8. David T. Jones, "Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices", JMB, 1999.
9. Wang S, Peng J, Ma J, Xu J. 2016 Protein secondary structure prediction using deep convolutional neural fields. Sci. Rep. 6, 18962. (doi:10.1038/srep18962)
10. Kabsch, Wolfgang and Sander, Christian. Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers, 22(12):2577–2637, 1983. ISSN 1097-0282.
11. Schmidler SC, Liu JS, Brutlag LD. Bayesian segmentation of protein secondary structure. J Comput Biol A J Comput Mol Cell Biol. 2000; 7(1-2):233–48.
12. Chu W, Ghahramani Z, Wild DL. A graphical model for protein secondary structure prediction. In: Proceedings 21st Annual International Conference on Machine Learning(ICML). New York: ACM: 2004. p. 161–8.
13. Hua S, Sun Z. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J Mol Biol. 2001; 308(2):397–407.
14. Guo J, Chen H, Sun Z, Lin Y. A novel method for protein secondary structure prediction using dual-layer svm and profiles. Protein Struct Funct Bioinform. 2004; 54(4):738–43.
15. Asai K, Hayamizu S, Handa K. Prediction of protein secondary structure by the hidden markov model. Bioinformatics. 1993; 9(2):141.
16. Aydin Z, Altunbasak Y, Borodovsky M. Protein secondary structure prediction for a single-sequence using hidden semi-markov models. BMC Bioinformatics. 2006; 7(1):178.

17. Qian N, Sejnowski TJ. Predicting the secondary structure of globular proteins using neural network models. J Mol Biol. 1988; 202(4):865–84
18. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 1999; 292(2):195.
19. Buchan DW, Minneci F, Nugent TC, Bryson K, Jones DT. Scalable web services for the psipred protein analysis workbench. Nucleic Acids Res. 2013; 41(Web Server issue):349–57.
20. Faraggi E, Al E. Spine x: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. J Comput Chem. 2012; 33(3):259–67.
21. Baldi P, Brunak S, frasconi P, Soda G, Pollastri G. Exploiting the past and the future in protein secondary structure prediction. Bioinformatics. 1999; 15(11):937–46.