I.J. Mathematical Sciences and Computing, **2016, 2, 24-33**
Published Online April 2016 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijmsc.2016.02.03

***Available online at** http://www.mecs-press.net/ijmsc*

RMSD Protein Tertiary Structure Prediction with Soft Computing

Mohammad Saber Iraji [a], Hakimeh Ameri [b]

[a] *Faculty Member of Department of Computer Engineering and Information Technology, Payame Noor University, I.R. of Iran*
[b] *Teacher of Department of Computer Engineering and Information Technology, Payame Noor University, I.R. of Iran*

Abstract

Root-mean-square-deviation (RMSD) is an indicator in protein-structure-prediction-algorithms (PSPAs). Goal of PSP algorithms is to obtain 0 Å RMSD from native protein structures. Protein structure and RMSD prediction is very essential. In 2013, the estimated RMSD proteins based on nine features were obtained best results using D2N (Distance to the native). We presented in This paper proposed approach to reduce predicted RMSD Error Than the actual amount for RMSD and calculate mean absolute error (MAE), through feed forward neural network, adaptive neuro fuzzy method. ANFIS is achieved better and more accurate results.

Index Terms: Root-mean-square-deviation (RMSD), protein, native structure, neural network, fuzzy.

1. Introduction

Finding an accurate algorithm to predict the protein structure is proved as an extremely difficult problem. There are two fundamentally different approaches: one is ab initio prediction that did not take any help of previously known protein structures, which employ computer-based applications to minimize very large functions corresponding to the free energy in the molecules. The other one is known as a knowledge-based approach which used amino acid sequences as a source of knowledge. This knowledge then extracts and stores to be used to predict any knew amino acid sequence (proteins) with known sequence but unknown structure. We can categorize knowledge-based methods into two groups: statistics based, and neural network based [1].

During the past four decades, several proteins secondary structural class prediction algorithms based on the sequence of protein amino acids have been studied and developed. Techniques such as neural networks, Fuzzy neural network, Support Vector Machines (SVM), component-coupled algorithm, and nearest neighbor classifier with a complexity-based distance measure (NN-CDM) are some of them. Among all these works the

* Corresponding author.
E-mail address: iraji.ms@gmail.com, ha.amery@gmail.com

neural network based methods showed the most efficiency. Neural networks are an artificial intelligence technique and a family of statistical learning algorithms that usually involves a large number of processors operating in parallel, each with its own small knowledge to create a secondary structure model with optimal parameters. Neurons in a neural network can be processed to encode the amino acid sequences into a usable numerical form. The classification performance of a neural network prediction algorithm can be considerably differing depending on the derivation of a sustainable parametric model from the training data set [2]. A fuzzy neural network or neuro-fuzzy system is a learning machine that finds the parameters of a fuzzy system by exploiting approximation techniques from neural networks. This technique used to calculate the degree of confidence for each prediction result and overcome the uncertainties prediction problem [3]. Another technique that can be used to predict the secondary structural protein based on the extracted feathers of predicted burial information from amino acid residues is SVM.

2. Related Work

Sepideh Babaei and et al. (2010) studied a supervised learning of recurrent neural networks, which is well-suited for prediction of protein secondary structures from the underlying amino acid's sequence. They devised a modular prediction system based on the interactive integration of the MRR-NN and the MBR-NN structures. To model the strong correlations between adjacent secondary structure elements a Modular mutual recurrent neural network (MRR-NN) has been proposed. And, MBR-NN is a multilayer bidirectional recurrent neural network which can capture the long-range intra-molecular interactions between amino acids in formation of the secondary structure. The proposed model used to arbitrarily engage the neigh boring effects of the secondary structure types concurrent with memorizing the sequential dependencies of amino acids along the protein chain. They used their model on PSIPRED data set into three-fold cross-validation. The reported accuracy was 79.36% and boosts the segment overlap (SOV) was up to 70.09% [4]. Zhun Zhou and et al. (2010) proposed a model named SAC method based on association rule classifier, and rules are the rules are obtained by the KDD* model mining secondary structure for information of mutual impact. The KDD* Model focused on high confidence and low support rules, which is called ‘knowledge in shortage’. The proposed model was based on CMAR (Classification based on Multiple Association rules) algorithm. The accuracies of the proposed model tested on RS126 and CB513 data sets are reported as 83.6% and 80.49%, respectively [5].Wu Qu and et al. (2011) proposed a multi-modal back propagation neural network (MMBP) method to predict the protein secondary structures. The created model is a compound pyramid model (CPM), which is composed of three layers of the intelligent interface that integrate multi-modal back propagation neural network (MMBP), mixed-modal SVM (MMS), modified Knowledge Discovery in the Databases (KDD) process and so on. The claimed accuracy of the proposed model on a non-redundant test data set of 256 proteins from RCASP256 was 86.13% [6]. Rohayan tihasan and et al. (2011) recommended a derivative feature vector, DPS that utilizes the optimal length of the local protein structure to predict the secondary structures of proteins. The DPS denotes the unification of amino acid propensity score and dihedral angle score. Secondary structure assignment method (SSAM) and secondary structure prediction method (SSPM) generate class labels for DPS feature vectors and the Support Vector Machine (SVM) used for prediction. The accuracy of the proposed model on the RS126 sequences was 84.4% [7]. Sepideh Babaei and et al. (2012) used a multilayer perceptron of recurrent neural network (RNN) pruned for optimizing size of hidden layers to enhance the secondary structure prediction. A type of reciprocal recurrent neural networks (MRR-NN) and the long-range interactions between amino acids in formation of the secondary structure (bidirectional RNN) and a multilayer bidirectional recurrent neural network (MBR-NN) consecutively applied to capture the predominant long-term dependencies. Finally, a modular prediction system (the MRR-NN and MBR-NN) used on the trusted PSIPRED data set and report the percentage accuracy (Q3) up to 76.91% and augment the segment overlap (SOV) up to 68.13% [8].Mohammad Hossein Zangooei and et al. (2012) proposed a method based on Support Vector Regression (SVR) classification. They used non-dominated Sorting Genetic Algorithm II (NSGAII) to find mapping points (MPs) for mapping a real-value to an integer one. We applied non-dominated Sorting Genetic Algorithm II (NSGAII)

to find mapping points (MPs) to round a real-value to an integer one. In order to enhance the performance of the proposed model, they used the NSGAII to find and tune the SVR kernel parameters optimally. To improve the prediction result, the Dynamic Weighted Kernel Fusion (DWKF) method for fusing of three SVR kernels was used. The obtained classification accuracy of the proposed model on RS126 and CB513 data sets reported as 85.79% and 84.94% respectively [9].

Maqsood Hayat and et al. (2014) proposed a model employing hybrid descriptor space along with the optimized evidence-theoretic K-nearest neighbor algorithms. The hybrid space is a composition of two descriptor spaces, including Multi-profile Bays and bi-gram probability. The high discriminative descriptors from the hybrid space have been selected by use of a well-known evolutionary feature selection technique named particle swarm optimization. These features are extracted to enhance the generalization power of the classifier. They used the jack knife test on three low similarity benchmark databases, including PDB, 1189, and 640 to evaluate the performance of their proposed model. The success rates of their proposed model are 87.0%, 86.6%, and 88.4%, respectively on the three benchmark data sets [10]. Maulika S. Patel and et al. (2014)provided a hybrid novel algorithm, KB-PROSSP-NN, which is a combination of knowledge base method (KB-PROSSP) and modelling of the exceptions in the knowledge base using neural networks (NN) for protein secondary structure prediction (PSSP). These two methods are used in cascade to optimize the results. They used two popular data sets RS126 and CB396 to evaluate the accuracy of the proposed model. The Q3 accuracy of 90.16% and 82.28% achieved on the RS126 and CB396 test sets respectively [11]. Yong Tat Tan and et al. (2015) claimed that nearest neighbor – complexity distance measure (NN-CDM) algorithm using Lempel–Ziv (LZ) complexity-based distance measure had a problem. NN-CDM algorithm is slow and ineffective in handling uncertainties. To solve this problem, they proposed fuzzy NN-CDM (FKNN-CDM) algorithm that combines the confidence level of prediction results and enhances the prediction process by designing hardware architecture. The high average prediction accuracies for Z277 and PDB data sets using the hybrid proposed algorithm are 84.12% and 47.81% respectively [3].

Avinash Mishra and et al. (2014) used random forest machine learning methods to predict the protein structure. The two types of data sets have been chosen are: (i) Public decoys from Decoys 'R' us (http://dd.compbio.washington.edu/) and, (ii) Server predictions of CASP experiments CASP5 to CASP9 (http:// predictioncenter.org/download area/). This contains 409 systems with their decoys covering a Root-mean-square-deviation (RMSD) range of 0–30 Å There are 278 systems belonging to CASP decoys while 131 systems are from public decoy dataset. Three training models based on RMSDs are designed. The first model named "Model-I" trained on the complete training set consist of 64,827 structures that covered the whole range of RMSDs (0–30 Å) The second one named "Model-II" trained on 13,793 structures which covered RMSD range of 0–10 Á and the last model named "Model-III" trained on 13,793 structures which covered an RMSD range of 0–5 Á They combined these tree models together to set their final prediction structure. These models are combined as three different layers and used to predict the most accurate distance of any given structure. A 10-fold cross validation performed and Correlation, R2 and accuracy used for evaluation. The best reported results are as follow: that native can be predicted as less than 1.5 Å RMSD structure in~89.3% of the cases. High quality structures (0–3 Å) can be predicted to within ±1.5 Å error in ~98.5% of the cases. It means a 2 Å structure may be reported as either 1 Å or 3 Å structure or anywhere in between with ~ 98% accuracy [4].

Yuehui Chen and et al. (2012) proposed a novel method based on an integrated RBF neural network to predict the protein interaction sites. Features named sequence profiles, entropy, relative entropy, conservation weight, accessible surface area and sequence variability are extracted in the first step. They made six sliding windows on these features that contained 1, 3, 5, 7, 9 and 11 amino acid residues respectively, and used them as input layer of six radial basis functional (RBF) neural networks. These RBF neural networks were trained to predict the protein–protein interaction sites. Decision fusion (DF) and Genetic Algorithm based Selective Ensembles (GASEN) are used to obtain the final results. The recall and accuracy of their proposed methods are and 0.8032 respectively [12].

We proposed a model for prediction the distance of an estimate structure from native based on Physicochemical Properties of Protein Tertiary Structure In this paper. The paper is organized in five sections.

After the introduction Section 1 which introduces the related works of protein structure estimate, Section 2 continues with Mathematical Model in section 3. Section 4 and 5 presents the results, conclusions of the research. The paper ends with a list of references.

3. Mathematical Model

Soft computing are included, different types of neural networks, fuzzy systems, genetic algorithms, etc. that used in information retrieval applications. Fuzzy theory was developed by Zadeh, is a new intelligent method, stated to solve unlike problems more efficient than the old calculations.

3.1. Neural Network

Developing a neural net solution means teaching the net a desired behavior. This is called the learning phase. Either sample data sets or a "teacher" can be used in this step. A teacher is either a numerical function or a person that rates the quality of the neural net performance. Since neural nets are mostly used for complex applications where no adequate mathematical models exist and rating the performance of a neural net is difficult in most applications, most are trained with sample data (figure 1).

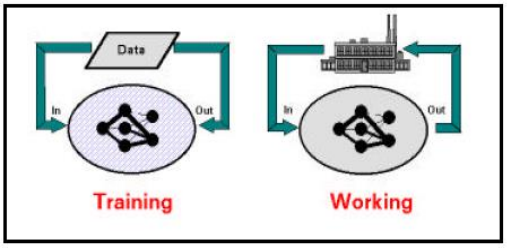

Fig.1. Training and Working Phase for Supervised Learning

3.1.1. Neuron Model

An elementary neuron with R inputs is shown in figure 2. Each input is weighted with an appropriate w. The sum of the weighted inputs and the bias, forms the input to the transfer function f. Neurons may use any differentiable transfer function f to generate their output.

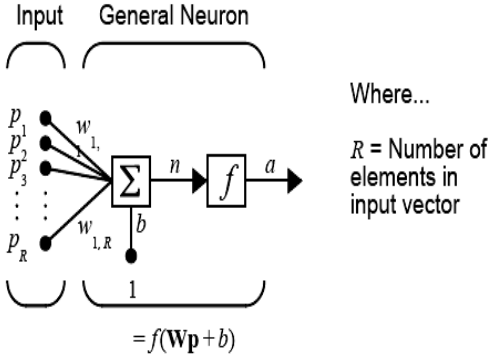

Fig.2. Structure a Neuron

Feed forward networks often have one or more hidden layers of sigmoid neurons followed by an output layer of linear neurons. Multiple layers of neurons with nonlinear transfer functions allow the network to learn nonlinear and linear relationships between input and output vectors(figure 3). The linear output layer lets the network produce values outside the range –1 to +1. On the other hand, if you want to constrain the outputs of a network (such as between 0 and 1), then the output layer should use a sigmoid transfer function (such as logsig). This network can be used as a general function approaching. It can approximate any function with a finite number of discontinuities, arbitrarily well, given sufficient neurons in the hidden layer [14].

3.1.2. Neural Network Step for Estimating Rsmd

Back-propagation feed forward neural network approach is used to predict the RMSD from native structure. A neural network is constructed with six inputs and one output RMSD.

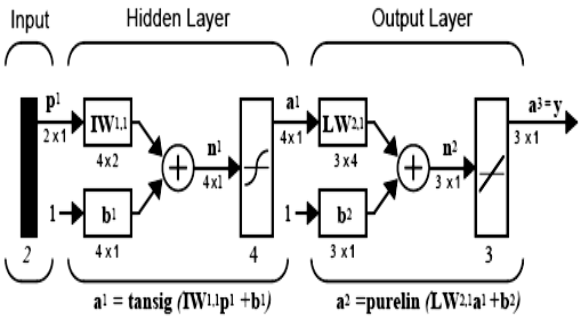

Fig.3. Feed Forward Neural Network with Two Layers

The input nodes represent the distinguishing parameters of proteins approach, and the output nodes represent the RMSD. The network is constructed by using MATLAB. In our experiment, the training function we have considered is a train GDA, the adaptation learning function considered learns GDM, and the performance function used was Mean Square Error (MSE). The projects considered for these researches are taken from the UCI. This is a data set of Physicochemical Properties of Protein Tertiary Structure. The data set is taken from CASP 5-9. There are 45730 decoys and size varying from 0 to 21 Armstrong [13]. After determining affect following features was extracted between the data set. The input nodes represent the physic-chemical properties of proteins.

F1-Total surface area.
F2- Non polar exposed area.
F3- Fractional area of exposed non polar residue.
F4- Fractional area of exposed non polar part of residue.
F5- Average deviation from standard exposed area of residue.
F6- Special Distribution constraints (N, K Value).

3.2. ANFIS (Adaptive Neuro Fuzzy Inference System)

ANFIS (Adaptive Neuro Fuzzy Inference System) is based sugeno (Jang, Sun & Mizutani, 1997; Jang & Sun, 1995) A generic rule in a Sugeno fuzzy pattern has the form If Input 1 = x and Input 2 = y, then output is z = ax + by + c. Figure 4 explains the ANFIS neural network [14]. In Figure 4 first layer are the degree of membership

of linguistic variables. The second layer is 'rules layer'. After the linear composition of rules at third layer then specify the degree of belonging to a special class by Sigmund's function in layer 4. ANFIS is a type of fuzzy neural network with a learning algorithm based on a set of training data for tuning an available rule base that permits the rule base to reconcile the training data [14].

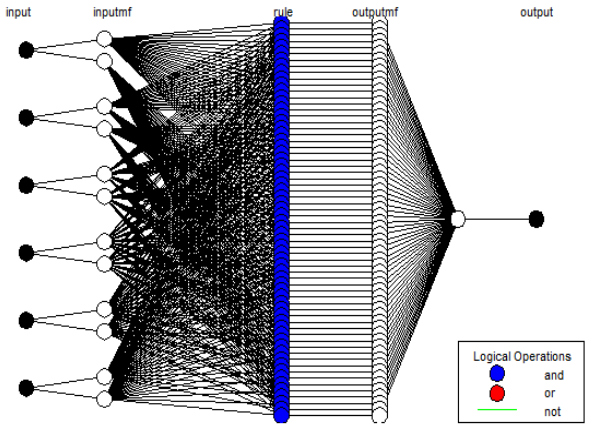

Fig.4. Adaptive Neuro fuzzy Network (anfis) for rmsd

We have applied features f1...f6 to ANFIS the given training data, the related rules is set, and obtain more accurate output RMSD (Figure 4).

4. Experimental Results

We implement our proposed system in MATLAB version 7.12 on Laptop, 1.7 GHZ CPU. In Anfis proposed system was considered a database of CASP 5-9 with 45730 records, In order to train and test the fuzzy neural network. After calculate six Features: Total surface area, Non polar exposed area, Fractional area of exposed non polar residue, Fractional area of exposed non polar part of residue, Average deviation from standard exposed area of residue, Special Distribution constraint (N, K Value) described above for 38110 records was considered for train, and 7620 records were allocated to Test system. Neural network set six input, seven

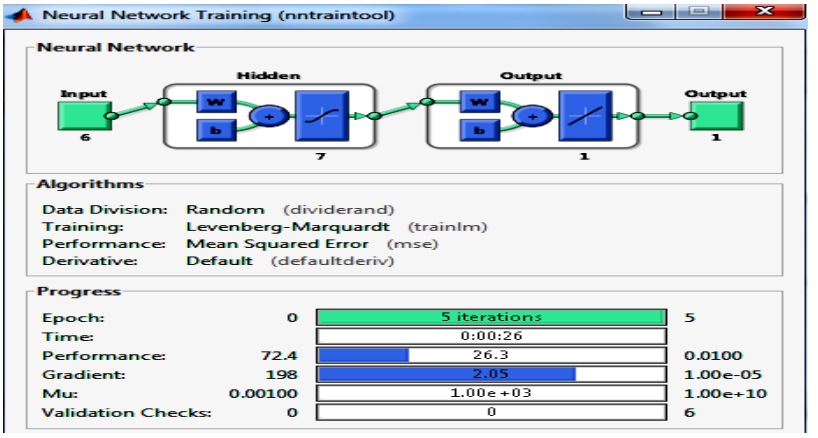

Fig.5. Adaptive Neuro fuzzy Network (ANFIS) for RMSD

hidden neuron, one output RMSD(figure 5) and number of epochs 5 were considered. After train neural network, Rmse(Root mean squared errors) was obtained for train data 5. 1044 and for test data 5.1008 (figure 6, 7).also Rmse of test data for six hidden neurons and 12 hidden neuron 5.1347 and 5.1226 was obtained. However, in Adaptive Fuzzy Neural Network After setting network parameters to generate fis =grid partition, optim. method= hybrid, linear, train fis epochs=5 Rmse for training data Obtained 4.7386 (Figure 8).

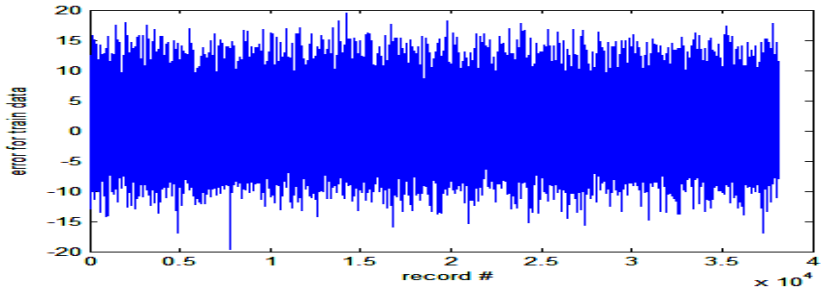

Fig.6. Sqrt mse=5.1044 for Train Data

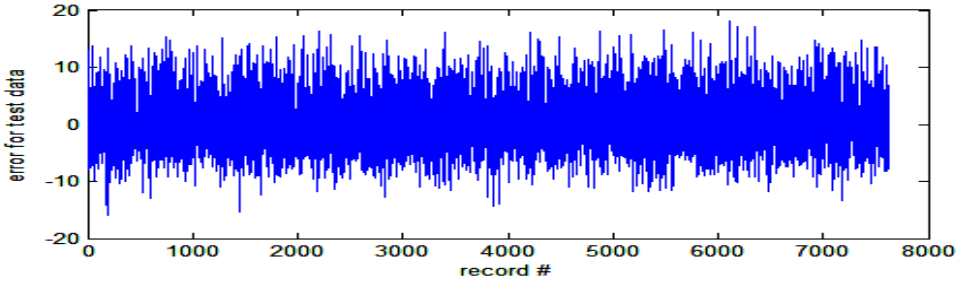

Fig.7. Error for Test Data Sqrt(mse)= 5.1008

After completing the process the training adaptive fuzzy neural network, fuzzy input variables were calibrated (Figure 9) and the number 7620 record was to test.

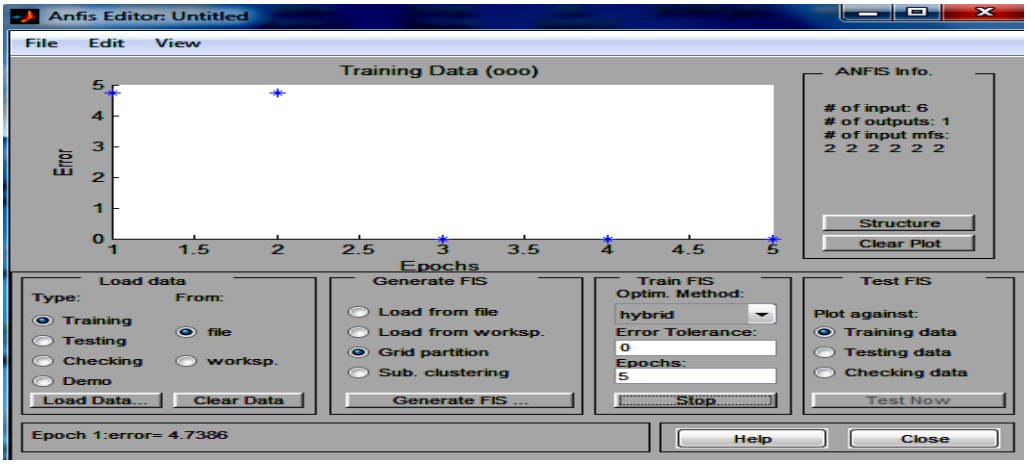

Fig.8. CalculateRMSE for Train Data with Proposed ANFIS System

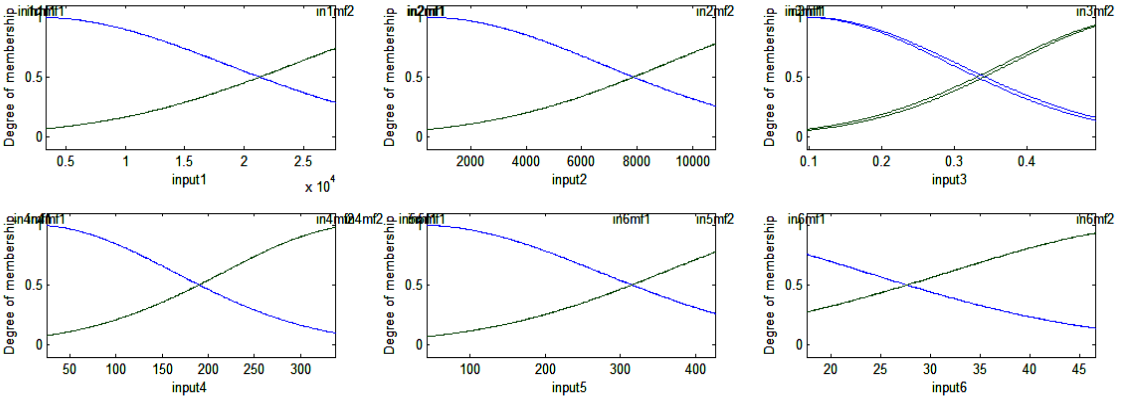

Fig.9. Input Variable Fuzzy Membership For Proposed ANFIS System Before and After Calibrate

Table 1. Comparison Actual RMSD and Predicted RMSD for Test Data

#	F1	F2	F3	F4	F5	F6	Actual RMSD	Neural network predicted RMSD	Adaptive fuzzy Neural network predicted RMSD
1	5915.26	1855.78	0.31372	55.4459	84.3491	39.4094	7.878	7.535813	7.715751
2	7511.48	1975.72	0.26302	85.9864	114.437	37.2338	2.862	4.605313	3.250838
3	9489.2	2632.9	0.27746	76.1758	123.313	37.0363	15.358	9.76639	10.46009
4	8946.25	2600.19	0.29064	75.657	112.681	34.5376	2.527	9.848676	9.867808
5	13179	3998.54	0.3034	149.945	218.111	29.1008	4.43	5.89325	4.935356
6	6493.69	1705.78	0.26268	52.9662	76.9265	38.6855	2.28	8.218051	8.726524
7	10204.9	3381.03	0.33131	108.207	159.552	36.0665	9.815	7.800493	7.154989
8	8569.51	2095.52	0.24453	85.6867	118.469	34.4689	5.233	6.27075	3.639084
9	9880.43	2824.41	0.28585	107.046	136.504	35.878	9.165	6.301614	5.01185

Several methods exist to compare cost estimation models. Each method has its advantages and disadvantages. In this work, mean absolute error (MAE) of RMSD will be used. AE for each observation i can be obtained as:

$$\text{AEi} = |\text{Actualusability}_i - \text{Predicatedusability}_i| \quad (1)$$

MAE can be achieved through the summation of MAE over N observations:

$$\text{MAE} = \frac{1}{n}\sum_1^N \text{AE}_i \quad (2)$$

After training adaptive fuzzy neural network by uci data, we have applied our proposed neuro fuzzy system on 7620 records. In the proposed system predicted for row, 2 RMSD obtain 3.250838 When F1=7511.48, F2=1975.7, F3=0.26302, F4=85.9864, F5=114.437, F6=37.2338.

5. Conclusions

Our purpose in this research is to predict the exact quantity of the Root-mean-square-deviation (RMSD) protein-structure from native protein structure's Protein and creating an adaptive neuro fuzzy model for this purpose, as yet. The Total surface area, Non polar exposed area, Fractional area of exposed non polar residue, Fractional area of exposed non polar part of residue, Average deviation from standard exposed area of residue, Special Distribution constraints (N, K Value) features were considered In order to The RMSD protein-structure estimation. Using adaptive neuro fuzzy combinatorial proposed system offered a2lgorithm, the amount of mean absolute error (MAE) indicator for ANFIS is3.845204 and neural network is 4.202547. ANFIS model is better, performed calculations in experimental results in table 1, prove this claim. In order to perform future works, the proposed model for distance from native protein structure can be developed with Raising The data relating to projects, and also other's neural methods can be used in order to determine the exact amount of RMSD in industrial environments and other data sets. Other features of protein can also be considered for prediction protein structure. Probably better results can be achieved by changing the number of linguistic variables or the type of membership function.

References

[1] Saha, S. (2008). *Protein Secondary Structure Prediction by Fuzzy Min Max Neural Network with Compensatory Neurons* (Doctoral dissertation, Indian Institute of Technology, Kharagpur).

[2] E. Sakk, A. Alexander, On the variability of neural network classfication measures in the protein secondary structure prediction problem, Appl. Comput. Intell. Soft Comput. (2013) 1–9. Available online at 4/16/2015.(http://www.hindawi.com/journals/acisc/2013/794350/).

[3] Tan, Y. T., &Rosdi, B. A. (2015). FPGA-based hardware accelerator for the prediction of protein secondary class via fuzzy K-nearest neighbors with Lempel–Ziv complexity based distance measure. *Neurocomputing*, *148*, 409-419.

[4] Babaei, S., Geranmayeh, A., &Seyyedsalehi, S. A. (2010). Protein secondary structure prediction using modular reciprocal bidirectional recurrent neural networks. *Computer methods and programs in biomedicine*, *100*(3), 237-247.

[5] Zhou, Z., Yang, B., &Hou, W. (2010). Association classification algorithm based on structure sequence in protein secondary structure prediction. *Expert Systems with Applications*, *37*(9), 6381-6389.

[6] Qu, W., Sui, H., Yang, B., &Qian, W. (2011). Improving protein secondary structure prediction using a multi-modal BP method. *Computers in biology and Medicine*, *41*(10), 946-959.

[7] Hassan, R., Othman, R. M., Saad, P., &Kasim, S. (2011). A compact hybrid feature vector for an accurate secondary structure prediction. *Information Sciences*, *181*(23), 5267-5277.

[8] Babaei, S., Geranmayeh, A., &Seyyedsalehi, S. A. (2012). Towards designing modular recurrent neural networks in learning protein secondary structures.*Expert Systems with Applications*, *39*(6), 6263-6274.

[9] Zangooei, M. H., &Jalili, S. (2012). Protein secondary structure prediction using DWKF based on SVR-NSGAII. *Neurocomputing*, *94*, 87-101.

[10] Hayat, M., Tahir, M., & Khan, S. A. (2014). Prediction of protein structure classes using hybrid space of multi-profile Bayes and bi-gram probability feature spaces. *Journal of theoretical biology*, *346*, 8-15.

[11] Patel, M. S., &Mazumdar, H. S. (2014). Knowledge base and neural network approach for protein secondary structure prediction. *Journal of theoretical biology*, *361*, 182-189.

[12] Chen, Y., Xu, J., Yang, B., Zhao, Y., & He, W. (2012). A novel method for prediction of protein interaction sites based on integrated RBF neural networks.*Computers in biology and medicine*, *42*(4), 402-407.

[13] Mishra, A., Rana, P. S., Mittal, A., &Jayaram, B. (2014). D2N: Distance to the native. *Biochimica et BiophysicaActa (BBA)-Proteins and Proteomics*, *1844*(10), 1798-1807.
[14] Iraji, M. S., &Motameni, H. (2012). Object Oriented Software Effort Estimate with Adaptive Neuro Fuzzy use Case Size Point (ANFUSP). *International Journal of Intelligent Systems and Applications (IJISA)*, *4*(6), 14.

Authors' Profiles

Mohammad Saber Iraji received B.Sc in Computer Software engineering from Shomal university, Iran, Amol; M.Sc1in industrial engineering (system management and productivity) from Iran, Tehran and M.Sc2 in Computer Science. Currently, he is engaged in research and teaching on Computer Graphics, Image Processing, Fuzzy and Artificial Intelligent, Data Mining, Software engineering and he is Faculty Member of Department of Computer Engineering and Information Technology, Payame Noor University, I.R. of Iran.

Hakimeh Ameri is graduated in M.S at K.N.Toosi Univerity of science and technology in information technology. Her main research interest is on bioinformatics, Data analysis and big data. She has 7 published papers in this filed. She now teaching Artificial Intelligence, data mining, Information technology, programming languages and data structure in University.

How to cite this paper: Mohammad Saber Iraji, Hakimeh Ameri,"RMSD Protein Tertiary Structure Prediction with Soft Computing", International Journal of Mathematical Sciences and Computing(IJMSC), Vol.2, No.2, pp.24-33, 2016.DOI: 10.5815/ijmsc.2016.02.03