Prediction of Tertiary Protein Structure using NNE and SVM Approach

Amanpreet Singh Garcha[1], Gurjot Kaur Walia[2],
[1]Guru Nanak Dev Engineering College, Ludhiana.
[2] Guru Nanak Dev Engineering College, Ludhiana.
(E-mail : apsinghgarcha@gmail.com)

***Abstract* —** Protein structure prediction is the method to predict the 3-dimensional geometry of protein structure by using machine learning methods in bio-informatics. Proteins in accordance to their structural folds are categorized into primary, secondary and tertiary level structures. In this approach there is an application of neural network ensemble techniques and combination with supporting vector machines for the prediction of protein tertiary structures. In result section, the second NNE algorithm value is closer to the actual value that comes 3.2095*. Well, ensemble techniques performed better than previous solo type neural network methods. Here, along with combination of Support Vector Machine each predicted ensemble result categorized in a much better way on the basis of Angstrom Unit. SVM technique is quite useful to decide us whether the predicted value is useful or rejected value. The results are on the basis of three columns headed under a) NN Ensemble b) Angstrom classification and c) Error difference between actual and predicted values. In the end, these three columns based result gives more clarity and satisfaction with simple and systematic analysis of tertiary protein structures. The GUI using Matlab is quite easy, fast and user friendly at larger manufacturing scales.

Keywords— *Classification, Ensemble, Homology, Neural Network, Protein tertiary structure, SVM.*

I. INTRODUCTION

Understanding 3-d protein [10] structure is the key to understand the protein functions and is very helpful in drug designing. Centuries ago X-ray or NMR Crystallography techniques were followed to produce the 3-dimensional [8] structures of proteins structures as an image on monitor screens. Those methods were based on calculations of electron map densities. Once the adequate amount of protein crystal is available, it's easy to calculate protein end co-ordinate. In this method of geometry such as bond length, bond angles etc were calculated. But there was a problem, this method require some specific and important laboratory conditions also very time consuming and not very feasible. The problem in this approach was that it works on actual biological protein crystals only. Now-a-days this type of protein structures can be studied or predicted without biological protein crystals which are based on their physico-chemical properties. It is quite fast and easy to solve this urgency with this successful [7] approach via various neural network methods [2]. These techniques even further useful in studying protein functions [6], [9] and the same can be in drug designing also. In structural and geometrical values the data is easy to store in computers for a long time than biological crystals. The stored calculations require very less amount of memory while actual crystals were getting spoiled too soon comparatively

II. LITERATURE REVIEW

Avinash Mishra, P. Rana, Amittal, B. Jayaram had applied random forest machine learning supervised methods for prediction of protein structures. Three working out models on the basis of RMSD were developed. The primary layer [1] named model 1 trained on the whole training set consists of 64,827 structures that covered the whole range of RMSD from 0-30 Angstroms. The secondary layer under model 2and trained only 13,793 protein structures had covered RMSD 0-10 Angstroms. Finally model 3 trained using 13,793 protein structures ranges from 0-5Angstroms. On basis of these layers results were predicted with higher accuracies.

Mohammad Saber Iraji and Hakimeh Ameri the systematic system used in his paper is to get slightly less predicted RMSD Error [11] than the real amount of RMSD and the mean Absolute error (MAE) is calculated in feed forward network neural network, adaptive neuro-fuzzy method. ANFIS produced precise and improved result than feed forward network.

Mathuriya The artificial neural network (ANN) is a method of data mining unlike from traditional techniques [13]. It is a non-linear auto-fit dynamic system made from range of cells with simulating the construction of biological neural systems.

Sonal Mishra , Yadunath Pathak and Anamika Ahirwar had explored the nine machine learning categorization of models with six physical and chemical properties to categorize the RMSD of the protein structure in the non-existence of its true native state and each protein lies among 0A to 6 A RMSD space. Together both properties [12] used six physical features. ABC algorithm is used to identify the protein structures.

III. ENSEMBLE APPROACH

Neural network ensemble is a kind of learning network where many neural networks are used together to work out the problem. In this approach, [4] neural network mechanism improves the output. Ensemble approach basically designed for making of linearity and non-linearity in data and to produce desirable practical outputs. In this ensemble neural networks plays major role as it combines various neural networks to form a single output model. In this, feed-forward and back propagation neural network [5] types of algorithms can be used frequently.

IV. PERFORMANCE OPERATIONS

In this work, Mean square error (MSE) is used as performance function. From 'F1 to F6' are the input nodes and data set is used to relate RMSD value which is taken from data set named CASP 5-9 of Physico-chemical properties of protein TERTIARY Structures.. There are 45730 [11] decoys and size varying from 0 to 21 angstrom with physicochemical features. In this 'Neural Network Ensemble' training has advantage that it needs very less formal training and can handle very large number of calculations as dataset. However, the bigger is the dataset can predict very closer to the actual calculated values needed for our work.

V. NNE AND SVM APPROACH

Neural network ensemble [3] is composed of multiple models and their average result is manipulated according to the requirement of result. Further Support Vector Machine is [14] a discriminative classifier. It performs [6], [12] classification by finding the hyper plane that maximizes the margin between classes and final classification unit is Angstrom based (nomenclature sec. VIII). In this work range of multi-dimensional hyper planes are set according to the range of values of Root means square deviations as discussed in next section VI.

VI. CLASSIFICATION ON THE BASIS OF SVM

TABLE 1.0 CLASSIFICATIONIS BASED ON ANGSTROM VALUES

Class [12]	**Classifications according to discreet values**	**Angstrom values**
If	1.0< =RMSD<=2.0	$1A^o$
If	2.0< RMSD<=3.0	$2A^o$
If	3.0< RMSD<=4.0	$3A^o$
If	4.0< RMSD<=5.0	$4A^o$
If	5.0< RMSD<=6.0	$5A^o$
If	6.0< RMSD<=7.0	$6A^o$

$1A^o$ (angstrom) = 10^{-10} m (S.I. unit)

VII. LIST OF PHYSICO-CHEMICAL PROPERTIES

As input side of neural ensemble networks receives the following inputs:

1 Feature (F1) = Total surface area.
2 Feature (F2) = Non-Polar exposed Area.
3 Feature (F3) = Fractional area of exposed non-polar residue.
4 Feature (F4) = Fractional area of exposed non-polar part of residue.
5 Feature (F5) = Average deviation from standard exposed area of residue.
6 Feature (F6) = Special distribution constraints (N, K Value).

VIII. NOMENCLATURE, EQUATION AND UNIT

A. RMSD is the actual numerical form assigned to each of protein tertiary structure and all of these values can be downloaded once from already calculated web datasets in excel sheet format document. The predicted RMSD value shows us the divergence [12] between actual and predicted protein structures. RMSD stands for root means square deviation. Mathematically the equation is given below:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} D_i^2}$$

Di=Distance between the matched pairs, N=Number of matched pairs available.

B. Angstrom is S.I. unit used for measurement of length dimension ie. $1A^o= 10^{-10}$ m

IX. FLOWCHART

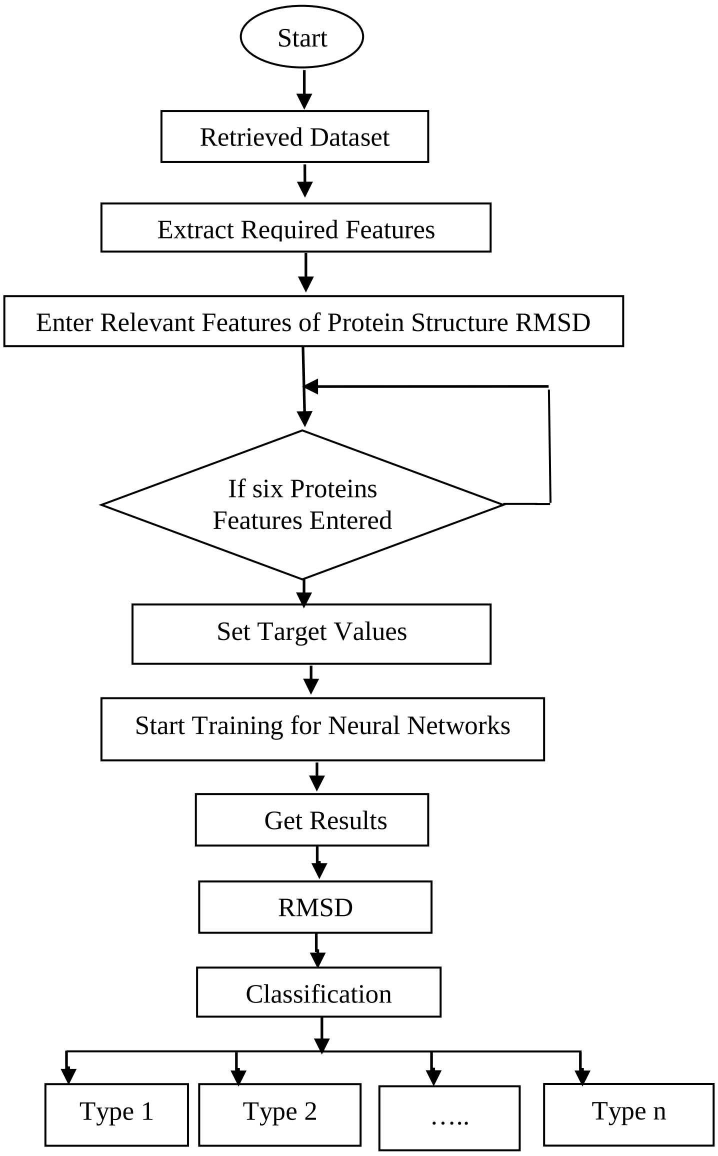
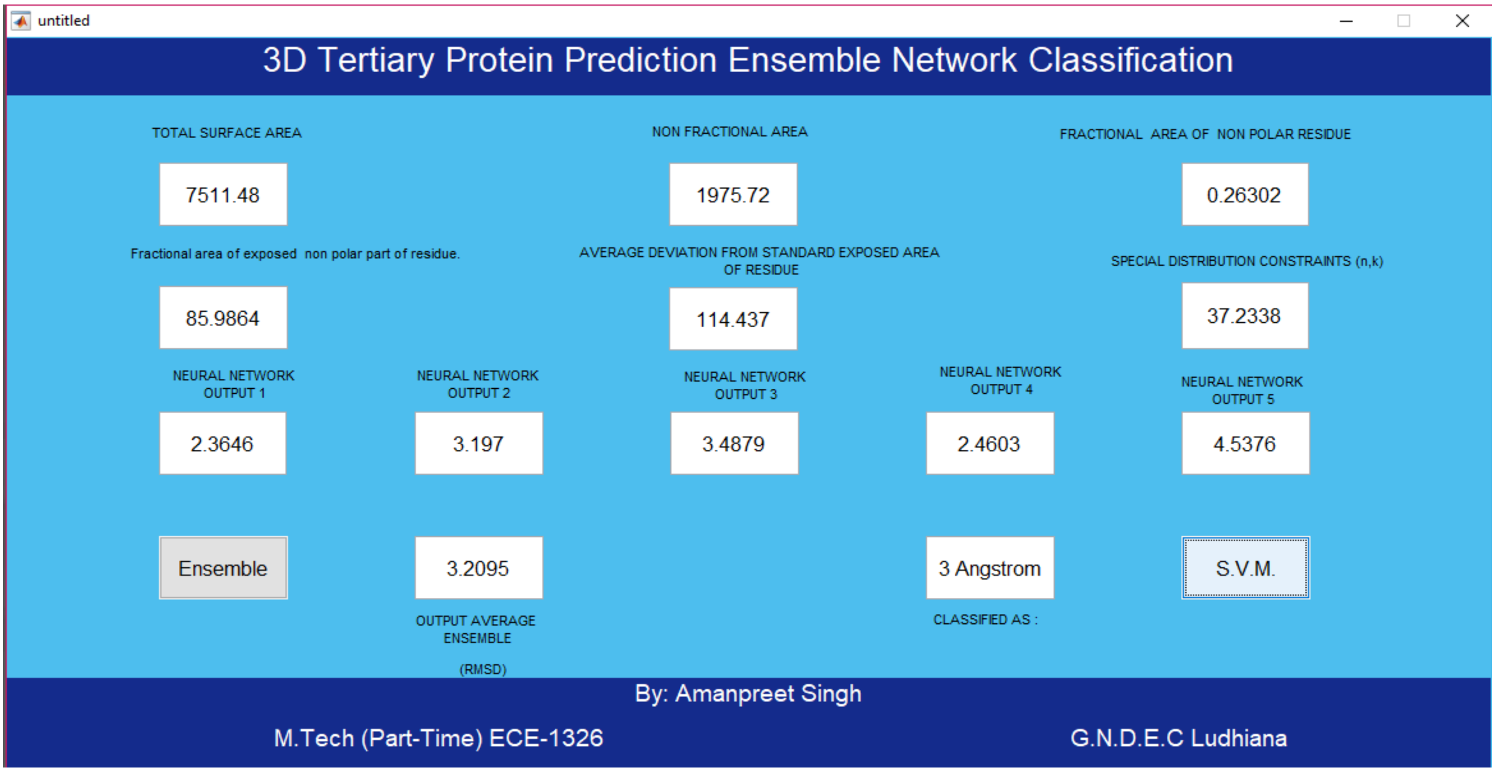

Start
Retrieved Dataset
Extract Required Features
Enter Relevant Features of Protein Structure RMSD
If six Proteins Features Entered
Set Target Values
Start Training for Neural Networks
Get Results
RMSD
Classification
Type 1
Type 2
…..
Type n

FIGURE 1.0 ALGORITHM DEVELOPMENT AND RESULT CLASSIFICATIONS

X. RESULT S



untitled

3D Tertiary Protein Prediction Ensemble Network Classification

TOTAL SURFACE AREA: 7511.48

NON FRACTIONAL AREA: 1975.72

FRACTIONAL AREA OF NON POLAR RESIDUE: 0.26302

Fractional area of exposed non polar part of residue.: 85.9864

AVERAGE DEVIATION FROM STANDARD EXPOSED AREA OF RESIDUE: 114.437

SPECIAL DISTRIBUTION CONSTRAINTS (n,k): 37.2338

NEURAL NETWORK OUTPUT 1: 2.3646

NEURAL NETWORK OUTPUT 2: 3.197

NEURAL NETWORK OUTPUT 3: 3.4879

NEURAL NETWORK OUTPUT 4: 2.4603

NEURAL NETWORK OUTPUT 5: 4.5376

Ensemble

OUTPUT AVERAGE ENSEMBLE (RMSD): 3.2095

CLASSIFIED AS : 3 Angstrom

S.V.M.

By: Amanpreet Singh

M.Tech (Part-Time) ECE-1326   G.N.D.E.C Ludhiana

Figure.1.1.Screenshot of second algorithm of NNE to get RMSD value and Angstrom classification using Matlab GUI

TABLE 1.1 COMPARISON OF TECHNIQUES WITH ACTUAL RMSD VALUES ALONG WITH PHYSIO-CHEMICAL FEATURES

Sr. no.	F1	F2	F3	F4	F5	F6	Actual RMSD |y|	Neural Network [11]	Proposed Method |x|	Ensemble classified as:	Error Difference = |x-y|
1	5915.26	1855.78	0.31372	55.4459	84.3491	39.4094	7.878	7.535813	7.3661	7 A°	0.5119
2	7511.48	1975.72	0.26302	85.9864	114.437	37.2338	2.862	4.605313	3.2095*	3 A°	0.3475
3	9489.2	2632.9	0.27746	76.1758	123.313	37.0363	15.358	9.76639	9.9033	9 A°	5.4547
4	8946.25	2600.19	0.29064	75.657	112.681	34.5376	2.527	9.848676	10.4675	10 A°	7.9405
5	13179	3998.54	0.3034	149.945	218.111	29.1008	4.43	5.89325	4.8608	4 A°	0.4308
6	6493.69	1705.78	0.26268	52.9662	76.9265	38.6855	2.28	8.218051	8.5908	8 A°	6.3108
7	10204	3381.03	0.33131	108.207	159.552	36.0665	9.815	7.800493	6.4068	6 A°	3.4082

*Closely predicted value.

XII. GRAPH FOR COMPARISON OF TECHNIQUES

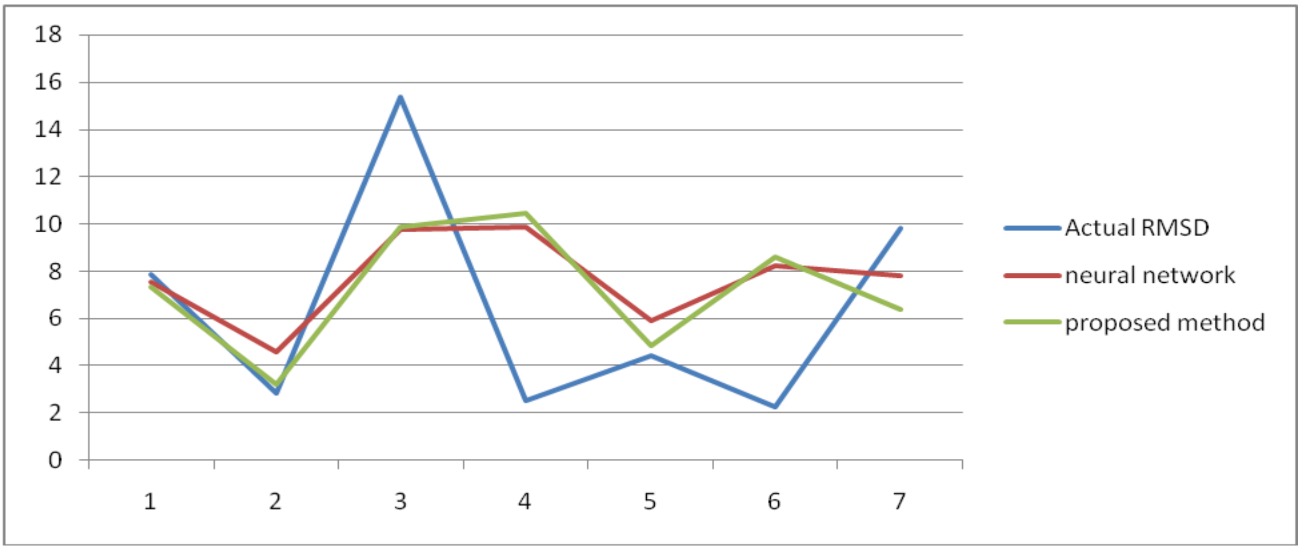


Fig. 1.2 Comparing proposed method with actual values and neural network

XIII. IMPROVEMENT

Here is an improvement 'in reduction' of mean square error in "second ensemble network algorithm" in proposed work value, before this in neural network it was 4.605313 and was diverted at large extent from actual real value of Root Mean Square Deviation. However, in this purposed work it comes more close to actual RMSD i.e. 3.2095*(means less getting diverting from actual Root Mean Square Deviation). Hence, it shows less error from its actual Root Mean Square Deviation and visibly good work for protein tertiary structure prediction. Further, when we use this resultant value for Support Vector Machine (SVM) classification it fits into 'near Angstrom' (native structure). SVM is frequently used technique for protein structure classifications as it fits as perfect Classification' technique by using 'combination' of NNE and SVM. Dataset surely can be trained every time, adjustment of weights done accordingly to predict more close finally. Algorithms and accuracies become better and enhanced each time.

XIV. CONCLUSION

In result (as listed in the table 1.1) RMSD and further classifications according to Angstrom values.

- During result of second algorithm features F1 to F6 (corresponding values 7511.48, 1975.72, 0.26302, 85.9864, 114.437, and 37.2338. The three satisfactory postulates are given as:

a) Smallest error difference i.e. **0.3475**.

b) The predicted RMSD after NNE = **3.2095 better than** previous technique(s).

c) It corresponds to prediction of smallest angstrom classified value which is 'lowest predictability' as i.e. **3 Unit Angstroms.**

- WHILE the "actual" small Root Mean square deviation equals to **2.862** valued.

XV. FUTURE SCOPE

In this work, algorithm is very customizable for the requirements of predictable work as needed. The purposed method can be further urbanized in future work with assorted types of ensemble neural networks. In future, to determine much similar prediction of protein structures even at industrialized level on larger scales. Lastly, the quantity of these Physical-Protein features can be used in a way to predict the structures from lesser features or different set of physical features. It will be very helpful for protein study or data science fields.

REFERENCES

[1]. Mishra, P. S. Rana, A. Mittal, and B. Jayaram, "D2N: Distance to the native," Biochimica et Biophysica Acta - Proteins and Proteomics, vol. 1844, no. 10, pp. 1798–1807, 2014.

[2]. Tramontano and K. Büssow, "Protein structure prediction: Concepts and applications," Analytical and Bioanalytical Chemistry, vol. 386, no. 6, pp. 1579–1580, 2006.

[3]. L. N. Ã and A. Lumini. "MppS : An Ensemble of Support Vector Machine Based on Multiple Physicochemical Properties of Amino Acids", Italy, vol. 69, pp.1688–1690, 2006

[4]. Parekh, "Bond Market Prediction using an Ensemble of Neural Networks," vol. 82, no. November, pp. 21–27, 2013.

[5]. E. A. Kaur and B. S. Khehra, "Quality Assessment of modelled protein structure using Back-propagation and Radial BasisFunction algorithm," vol. 5, no. 07, 2017.

[6]. J. Cheng, A N. Tegge, and P. Baldi, "Machine Learning Methods for Protein Structure Prediction," IEEE Reviews in Biomedical Engineering, vol. 1, pp. 41–49, 2008.

[7]. J. V. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," Journal of Clinical Epidemiology, vol. 49, no. 11, pp. 1225–1231, 1996.

[8]. K. Mizuguchi and N. Go, "Seeking significance in three-dimensional protein structure comparisons Kenji Mizuguchi and Nobuhiro Go," 1995.

[9]. K. Taha and P. Yoo, "An information extraction system for protein function prediction," Biology (CIBCB), 2015 IEEE Conference, 2015.

[10]. M. Dorn, M. B. E Silva, L. S. Buriol, and L. C. Lamb, "Three-dimensional protein structure prediction: Methods and computational strategies," Computational Biology and Chemistry, vol. 53, no. PB, pp. 251–276, 2014.

[11]. M. Saber Iraji and H. Ameri, "RMSD Protein Tertiary Structure Prediction with Soft Computing," International Journal of Mathematical Sciences and Computing, vol. 2, no. 2, pp. 24–33, 2016.

[12]. Y. Pathak, M. Saraswat, P. S. Rana, and P. K. Singh, "Classification of Protein Structure (RMSD Properties 6Å)using Physicochemical," 2014, vol. 7, no. 6, pp. 141–15