International Journal of Scientific Research and Management (IJSRM)
||Volume||5||Issue||07||Pages||6019-6033||2017||
Website: www.ijsrm.in ISSN (e): 2321-3418
Index Copernicus value (2015): 57.47 DOI: 10.18535/ijsrm/v5i7.27

Quality Assessment of modelled protein structure using Back-propagation and Radial Basis Function algorithm

Er. Amanpreet Kaur[1], Dr. Baljit Singh Khehra[2]

[1] Department of Computer Science and Engineering
Baba Banda Singh Bahadur Engineering College
Fatehgarh Sahib, India
[2] Department of Computer Science and Engineering
Baba Banda Singh Bahadur Engineering College
Fatehgarh Sahib, India

Abstract: Protein structure prediction (PSP) is the most important and challenging problem in bioinformatics today. This is due to the fact that the biological function of the protein is determined by its structure. While there is a gap between the number of known protein structures and the number of known protein sequences, protein structure prediction aims at reducing this structure –sequence gap. Protein structure can be experimentally determined using either X-ray crystallography or Nuclear Magnetic Resonance (NMR). However, these empirical techniques are very time consuming. So, various machine learning approaches have been developed for protein structure prediction like HMM, SVM and NN. In this paper, general introductory background to the area is discussed and two approaches of neural network i.e back-propagation and radial basis function are used for the prediction of protein tertiary structure. The aim of the study is to observe performance and applicability of these two neural network approaches on the same problem. More specifically, feed-forward artificial neural networks are trained with backpropagation neural network and radial basis function neural networks. These algorithms are used for the classification of protein data set, trained with the same input parameters and output data so that they can be compared. The advantages and disadvantages, in terms of the quality of the results, computational cost and time are identified. An algorithm for the selection of the spread constant is applied and tests are performed for the determination of the neural network with the best performance. These approaches depends on the chemical and physical properties of the constituent amino acids. Not all neural network algorithms have the same performance, so we represent the general success keys for any such algorithm. The data set used in the study is available as supplement at *http://bit.ly/RF-PCP-DataSets*.

Keywords: ***Data mining, Protein Structure Prediction (PSP), Neural Networks, back-propagation, radial basis function (RBF).***

1. Introduction

Data mining involves the use of practical data analysis tools to design previously exotic, valid patterns and relationships in huge data set. There are several applications for Machine Learning (ML), the most powerful of which is data mining. People are often prone to making mistakes during analysis or, possibly, when trying to establish relationships between multiple features. This makes it difficult for them to find solutions to specific problems. Machine learning can often be strongly applied to these problems, improving the designs of machines and the efficiency of the systems. Classification is the most important technique to identify a specific character or group of them.

Different classification algorithms have been proposed by various researchers for classification of protein sequences. The Protein sequence consists of twenty different amino acids which are aligned in some specific sequences. Popular protein sequence classification techniques involve extraction of particular features from the sequences. These features depend on the structural and functional properties of amino acids. These features can be compared with their predefined values.

1.1 Proteins

Proteins are main building blocks of our life. Proteins form the basis of structures such as skin, hair, and tendon and they are responsible for catalyzing and regulating biochemical reactions, transporting molecules. The shape of protein is given by its amino acid sequence. There are 20 distinct types of amino acid and each amino acid is known by its side chain that determines the properties of amino acid [16].

Currently, protein plays a vital role in the research human body. The structural class, which is one of the important attribute of a protein plays an important role in both theoretical and experimental studies in protein science. In generally speaking, protein is the chief executor of important movement. On the one hand, the data of Protein sequence database has been growing very fast. On the other hand, the structure of the protein is comparably less identified. Protein tertiary structures have vital influence on the behaviour of the protein from long-term study. According to the definition by Levitt and Chothia, proteins are classified into the following four structural classes: (1) all-a class, which are essentially formed by helices and only includes small amount of strands, (2) all-b class, which are essentially formed by strands and only includes small amount of helices, (3) a/ b class, which includes both helices and mostly parallel strands, and (4)a+ b class, which includes both helices and mostly anti parallel strands. Prediction of tertiary structure, however, still remains as an unsolved problem and various solution methods are urgently needed.

In its native environment, the chain of amino acids (or residues) of a protein folds into local secondary structures including alpha helices, beta strands, and non regular coils. The secondary structure is specified by a sequence classifying each amino acid into the corresponding secondary structure element (e.g., alpha, beta, or gamma). The secondary structure elements are further packed to form a tertiary structure depending on hydrophobic forces and side chain interactions, such as hydrogen bonding, between amino acids. The tertiary structure is described and coordinates of all the atoms of a protein or, in a more coarse description, by the coordinates of the backbone atoms. Finally, several related protein chains can interact or assemble together to form protein complexes. These protein complexes said to be the protein quaternary structure. The quaternary structure is described by the coordinates of all the atoms, or all the backbone atoms in a coarse version, associated with all the chains participating in the quaternary organization, given in the same frame of reference[4].

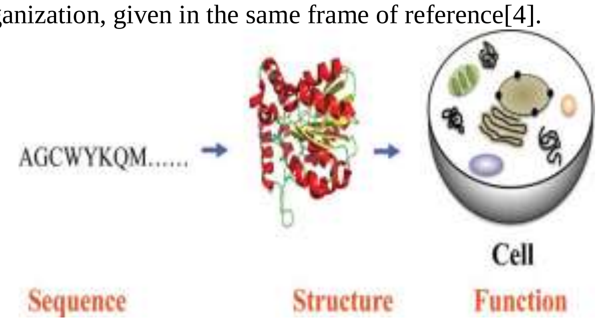


Figure 1: Protein sequence-structure-function relationship. A protein is a linear polypeptide chain composed of 20 different kinds of amino acids represented by a sequence of letters (left). It folds into a tertiary (3-D) structure (middle) composed of three kinds of local secondary structure elements (helix – red; betastrand– yellow; loop – green). The protein with its native 3-D structure can carry out several biological functions in the cell (right). [4]

1.2 Protein Structure

Proteins are large, organic molecules and are among the most vital components in the cells of living organisms. They are more diverse in structure and function than any other kind of molecule. It can act as Enzymes, antibodies, hormones, transport molecules, hair, skin, muscle, tendons, cartilage, claws, nails, horns, hooves, and feathers are all made of proteins. Protein structure has a basically four levels of category: Primary Structure, Secondary structure, Tertiary structure and Quaternary structure. Fig. 2 shows different levels of protein structure [16].

The structure of protein can be determined by experimental methods such as NMR and X-ray crystallography. These experimental methods, however, are waste of time and they are not feasible to everyone. Currently, with the development of machine learning, a great number of researchers are fond of taking advantage of machine learning methods to solve the problem. A variety of intelligent machine learning tools, such as Support Vector Machine (SVM),Hidden Markov Model(HMM), Neural Networks (NN) are widely used in protein structure prediction.

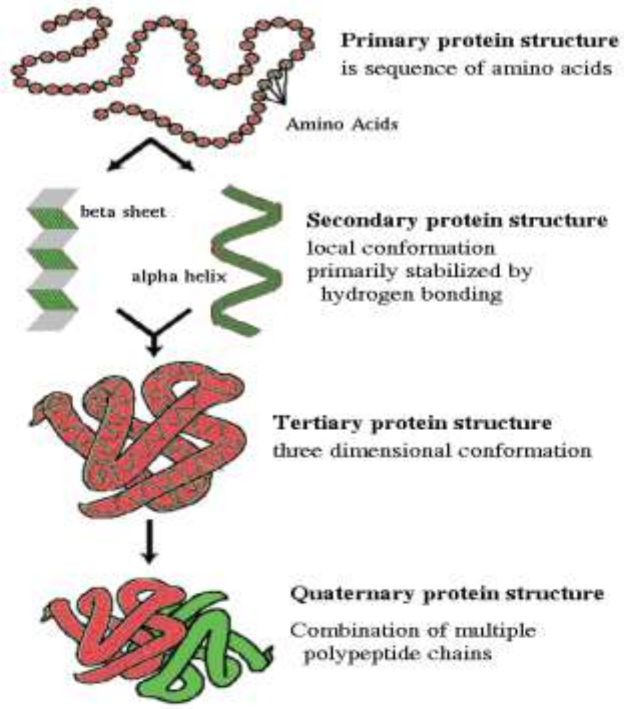


Figure 2: Levels of protein structure [13]

Primary Structure:

Chain of amino acid sequence is referring as primary structure. Every α-amino acid has of a backbone part that is present altogether in amino acid varieties, and a side chain that is distinctive to every variety of residue. Proline is an exception from this rule. The primary structure is held together by peptide bonds, that are created during the process of protein biosynthesis or translation. The primary structure of a protein is determined by the gene equivalent to the protein. A specific sequence of nucleotides in DNA is transcribed into mRNA, that is read by the ribosome in a process called translation. The sequence of a protein is exclusive to that protein, and defines the structure and function of the protein. The sequence of a protein may be determined by strategies such as Edman degradation or tandem mass spectrometry [13].

Secondary Structure:

The secondary structure consists of native folding regularities maintained by hydrogen bonds and is historically subdivided into 3 classes: alpha-helices (H), beta-sheets (E), and coil(C).Secondary structure contained localized and recurring fold of peptide chain, wherever 2 main regular structures are the α-helix and β-sheet. Hydrogen bond is answerable for secondary structure-helix may be considered the default state for secondary structure. It is most significant for higher understanding tertiary structure. It is extremely necessary because knowledge of secondary structure helps in the prediction of tertiary structure when structure discovery without sequence similarity within the datasets [13].

Tertiary Structure:

Tertiary structure refers to 3-dimensional structure of a one protein molecule. It involves localized spatial interaction among primary structure parts, i.e. the amino acids. The alpha-helices and beta-sheets are folded into a compact ball. The folding is driven by the non-specific hydrophobic interactions, however the structure is stable only if when the elements of a protein domain are locked into place by specific tertiary interactions, like salt bridges, hydrogen bonds, and the tight packing of side chains and disulfide bonds. The disulfide bonds are extremely rare in cytoplasm proteins, since the cytoplasm is generally a reducing environment [13].

Quaternary Structure:

Quaternary structure is that the arrangement of multiple folded protein or coiling protein molecules in a multi-subunit complex. Several proteins are literally assemblies of more than one polypeptide chain, that within the context of the larger assemblage are called as protein subunits. Additionally to the tertiary structure of the subunits, multiple-subunit proteins possess a quaternary structure, that is the arrangement into which the subunits assemble .Enzymes composed of subunits with various functions are typically known as holoenzymes, in which some elements could also be called regulatory subunits and the functional core is called the catalytic subunit. Examples of proteins with quaternary structure include hemoglobin, DNA polymerase, and ion channels. Different assemblies referred to instead as multi-protein complexes also possess quaternary structure [13].

1.3 Protein Tertiary Structure Prediction

Protein structure prediction is that the prediction of the 3-dimensional structure of a protein from its amino acid sequence therefore all activities of protein area unit depends upon its three dimensional structure. Structure prediction is essentially different from the inverse drawback of protein design. The 3-dimensional structure of a protein is decided by the network of covalent and non-covalent interactions. Though protein is built by the chemical process of 20 different amino acids into linear chains, proteins form an incredible array of various tasks. A protein chain folds into a novel shape that is stabilized by non covalent interactions between regions within the linear sequence of amino acids. This spatial organization of a protein its shape in three dimensions could be a key to understanding its function. Only when a protein is in its correct three-dimensional structure, or conformation**,** is it ready to perform efficiently. A key idea in understanding how proteins work is that function is derived from 3-dimensional structure, and 3-dimensional structure is specified by amino acid sequences [16].

Protein structure prediction is that the prediction of the 3-dimensional structure of a protein from its amino acid sequence, i.e., the prediction of its secondary, tertiary, and quaternary structure from its primary structure. Structure prediction is basically different from the inverse problem of protein design. In bioinformatics several prediction methods are available such as:

- Ab-initio, theoretical modelling, and conformation space search
- Homology modelling and threading

Primary and Secondary structure prediction:

Primary structure could be chain of 20 amino acid sequence, which is described as:

Protein Sequence: Input 1D

GRPRAINKHEQEQISRLLEKGHPRQQLAIIF

↓

HCCCCCCCHECECCCCCCECHHCCCCCCCCC

Protein Structure: Output 1D

Protein output 1D structure is getting using dictionary of secondary structure prediction (DSSP) methodology. Secondary structure prediction is that the classification of primary 1D structure in to three classes: Helix (H), Strand (E) and Coil(C).

Techniques used for Protein structure prediction are:

- Soft Computing Techniques like Artificial Neural networks.
- Probabilistic techniques like Hidden Markov Model.
- Evolutionary Computation like Genetic Algorithm.
- Statical techniques like SVM.
- Clustering algorithms etc. [14]

Bioinformatics techniques to protein secondary structure prediction largely depend upon the information out there in amino acid sequence. Evolutionary algorithms are like simple genetic algorithms (GA), messy GA, fast messy GA have addressed this problem. Support Vector Machine (SVM) represents a replacement approach to supervised pattern classification that has been with success applied to a large range of pattern recognition issues, as well as object recognition, speaker identification, gene function prediction with micro array expression profile, etc. In these cases, the performance of SVM either matches or is considerably higher than that of ancient machine learning approaches, as well as neural networks. However still SVMs are blackbox models. ANN is a good technique of protein structure prediction that relies on the sound theory of Back Propagation Algorithm. Protein secondary structure prediction has been satisfactorily performed by machine learning techniques like Artificial Neural Network and Support vector machines. Most secondary structure prediction programs target alpha helix and beta sheet structures and summarize all different structures within the random coil pseudo category. For the classification, ANN is employed as a binary classifier.

1.4 Necessities of PSSP(Protein Secondary Structure Prediction)

PSSP is receiving significance in the recent area of research due to the following:

- Being the difficulty of structural bioinformatics, protein secondary structure prediction can give prediction and analysis of macromolecules which are the basis of an organism.
- Protein secondary structure prediction(PSSP) give structure function relationship. That is which particular protein structure is responsible for which particular function would be known by PSSP. So by changing the protein's structure or by synthesizing new proteins, functions could be added or removed or required functions could be attained.
- Structure of the viral proteins can be examined by PSSP and this examination of the structures of the viral proteins provides the way to design drugs for specific viruses.
- PSSP lowers the sequence structure gap. The sequence structure gap can be best defined by giving the example of large scale sequencing projects like Human Genome Project. In these type of projects, protein sequences are produced at a very rapid speed which results in a huge gap between the number of known protein structures (>150,000) and the no. of known protein sequences (>4,000).This gap is called sequence structure gap and PSSP can successfully minimize this gap. Experimental approaches are not capable of structure determination of few proteins like

DOI: 10.18535/ijsrm/v5i7.27

membrane proteins. So the prediction of protein structure using computational tool is of good interest.[2]

Sequence-Structure Gap and the Need for Structure Prediction

With the advent of recombinant DNA technology it has become possible to signify the amino acid sequences of proteins quite fast. However, signifying the 3- dimensional structure of proteins is a time taking task and hence there exists a large gap between the number of proteins of known amino acid sequence and that of known structures. This is known as the sequence-structure gap. As the knowledge of the 3-D structure of a protein is very important to understand its function, it is essential to develop techniques to predict the structure of a protein from its amino acid sequence.

1.5 Fold Recognition

Proteins fold due to hydrophobic effect, electrostatic forces, Vander Waals interaction and Hydrogen bonding. Protein threading, also called fold recognition, is a methodology of protein modelling (i.e. computational protein structure prediction) which is used to model those proteins those have the same fold as proteins of known structures, but do not have similar proteins with known structure. Protein folding is the method by which a protein assumes its 3D structure. All protein molecules are endowed with a primary structure having the polypeptide chain. Fold recognition need a criterion to identify the best template for single target sequence. The protein fold-recognition method to structure prediction aims to identify the known structural framework that accommodates the target protein sequence in the best way. Typically, a fold-recognition program comprises four components:

(1) The representation of the template structures (usually corresponding to proteins from the Protein Data Bank database),

(2) The evaluation of the compatibility between the target sequence and a template fold,

(3) The algorithm to compute the optimal alignment between the target sequence and the template structure, and the method the ranking is computed and the statistical significance is estimated [16]

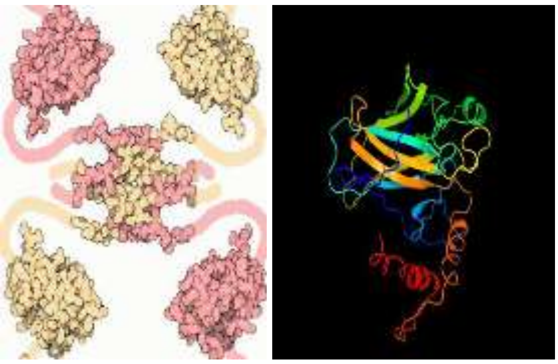


Figure 3: 3.1 Structure of protein P53, 3.2 Folded structure of protein P53 [16]

1.6 Approaches for Protein Tertiary Structure Prediction

Presently 3 approaches are followed for the estimating of the tertiary structure of proteins. These are 1) Homology modelling, 2) Threading and 3) Ab initio structure prediction.

1.6.1 Homology Modelling

This is the simplest and very reliable approach. The observation that proteins with same sequences tend to fold into same structures forms the basis for this method. It can be noticed that even proteins with 25% sequence identity fold into same structures. This approach does not work for remote homologs (< 25% pair wise identity). The method for homology modelling may be briefly defined as: given a query sequence Q, and a sequence database of known protein structures, find a protein P such that P has same sequence as to Q and return P’s structure as an approximation to Q’ structure. The following are the main steps in homology modelling:

1) Finding known structures linked to the query sequence whose structure has to be modeled

2) Aligning the query sequence to the templates

3) Constructing variable side-chains and main-chains and

4) Model refinement, assessing the model built and choosing the most native conformations.

1.6.2 Threading

Threading is a approach for fold recognition. This is employed for sequences having sequence identity ≤ 30%. In this approach, given a sequence and the set of folds available in the Protein Data Bank (PDB) the target is to see if the sequence can accept one of the folds of known structure. This approach takes merit of the knowledge of existing structures and the principles by which they are stabilized. Fold assignment and alignment are attained by threading the sequence via every structure in a library of all known folds.

1.6.3 Ab initio (de novo) structure prediction

While homology modelling and threading needs knowledge of known structures, ab initio structure prediction has no limitations like above approaches. It starts with the estimation that the real structure of a protein is at the global free energy minimal. In previous years a particularly successful approach called Rosetta has been developed by Baker and colleagues (Simons et al, 1997). This approach has assimilated information got from known structures and is depends upon a picture of protein folding in which small segments of the protein chain flicker between distinct local structures consistent with their local sequence, and folding to the real state occurs when these local segments are oriented so that low free energy interactions are made throughout the protein.[22]

Applications and limitations of methods for structure prediction

Each of the above methods described provide structural description to different extent. While homology modelling can give atomic level details of the target protein, threading can help only to know the fold of the protein. Baker and Sali (2001) have describes the accuracy and application of protein structure models with examples. Large and medium level homology models with sequence identity > 30% are convenient in refining functional prediction like ligand binding. Low accuracy models of several of the ribosomal proteins were favorable in building the molecular model for whole yeast ribosome. The correctness and applicability of models produced by ab initio methods are in general of lesser accuracy compared to models obtained from either homology modelling or threading. These are convenient in predicting functional relationships from structural similarity and for identification of patches of conserved surface residues.

2. Related Work

W. Dianhui, L. K. Wung *et al*.[34] presents a modular neural classifier for protein sequences with improved classification criteria. The intelligent classification techniques described in this paper aims to enhance the performance of single neural classifiers based on a centralized information structure in terms of recognition rate, generalization and reliability. F. Mhamdi *et al*. [15] presents the classification of proteins by basing on its primary structures. The sequence of proteins can be collected in a file. The application of text mining technique is proposed for extracting the features. An algorithm is also developed which extracts all the n-grams existing in the file of data and produced a learning file. D. H. Shing, Y. Y. Chi [11] implements a genetic algorithm to cluster the training set before a prediction model is built. Using position specific scoring matrix (PSSM) as part of the input, the hybrid method achieves good performances on sets of 513 non redundant protein sequences and 294 partially redundant sequences. The results also show that clustering achieves the goal of data preprocessing differently on redundant and non-redundant sets, and it seems almost preferable to cluster the data before prediction is preformed. C. Jianlin, N. T.Allison *et al*. [8] reviews the development and application of hidden Markov models, neural networks, support vector machines, Bayesian methods, and clustering methods in 1-D, 2-D, 3-D, and 4-D protein structure predictions. P. Sun and J. Zhang [28] presents a prediction method of protein contact on the basis of information granules and RBF neural network have been brought forward. This method improved the encoding approach of protein structure data and classifier performance to enhance the predicting accuracy of protein contact. V. Swati, S. K. Bithin *et al*. [33] discusses three types of neural networks such as feed forward neural network, probabilistic neural network and radial basis function neural network. The main objective of the paper is to build up an efficient classifier using neural networks. The measures used to estimate the performance of the classifier are Precision, Sensitivity and Specificity. N. Mathuriya *et al*.[24] observed that the K-means clustering algorithm is not very much suitable for the problem and the back propagation neural network has the high performance. The artificial neural network (ANN) is the technique of data mining that is different from traditional techniques. It is the nonlinear auto-fit dynamic system made of various cells with simulating the construction of biology neural systems. S. Saha *et al*.[31] presents a review with three different classification models such as neural network model, fuzzy ARTMAP model and Rough set classifier model. This is followed by a new technique for classifying protein sequences. The proposed model is typically implemented with an own designed tool and tries to reduce the computational overheads encountered by earlier approaches and increase the accuracy of classification.

K. Yasaman, F. Mahmood *et al*.[20] discuss the most important algorithms like evolutionary algorithms, particle swarm or ant colony optimization and computational methods introduced in this field and their challenges. B. Wenzheng, C. Yiming et al.[7] purposed forward novel approach for predicting the tertiary structure of protein and construct an Error Correcting Output Codes(ECOC) classification model on the basis of Particle swarm optimization(PSO) and neural network(NN).Three feature extraction methods, which are Amino Acid Composition, Amino Acid Frequency and Hydrophobic Amino Acid Combination, respectively, are employed to extract the features of protein sequences. To evaluate the efficiency of the proposed method we choose a benchmark protein sequence dataset (640 dataset) as the test data set. The final results show that our method is efficient for protein structure prediction. R. N. Chandrayani, K. Manali [29] presents the results of protein p53. P. Mayuri, S. Hitesh [25] uses model based (i.e., supervised learning) approach for protein secondary structure prediction and our objective is to enhance the prediction of 2D protein structure problem using advance machine learning techniques like, linear and non-linear support vector machine with different kernel functions. M. Vidyasagar [22] reviews some of the many challenging problems in computational biology that are amenable to treatment using a systems approach. Specific problems discussed include string alignment and protein structure prediction. M. Sonal, P. Yadunath *et al*.[21] explored the machine learning classification models with six physical and chemical properties to classify the root mean square deviation (RMSD) of the protein structure in absence of its true native state and each protein structure lies between 0A˚ to 6A˚ RMSD space. Physical and chemical properties used in this paper are total surface area, Euclidean distance, total empirical energy, secondary structure penalty, residue length, and pair number. Artificial bee colony algorithm is used to determine the feature importance. To measure the robustness of the best classification model, K-fold cross validation is used. R. S. Prashant, S. Harish *et al*.[30] explore nine machine learning methods with six physicochemical properties to predict the RMSD (Root Mean Square Deviation), TM-score (Template Modelling) and GDT TS-score (Global Dis- tance Test) of modelled protein structure in the absence of its true native state. Physicochemical properties namely total surface area, euclidean distance, total empirical energy, secondary structure penalty, sequence length and pair numbers are used. The K-fold cross validation is used to measure the robustness of the best predictive method.

C. Nandini *et al*.[9] explains several techniques used by different researches for the classification of proteins and also provides an overview of different protein sequence classification methods. From the vast data we have to derive the hidden knowledge so that it is used in wide range of areas to design drug, to indentify diseases, and in classification of protein sequence etc. I. S. Mohammad, A. Hakimeh [16] 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. W. Bo,L. Yongkui *et al*.[35] summarise some of the recent studies adopting this SVM learning machine for prediction structure prediction are the one which used frequent profiles with evolutionary information. W. Jian and L. Jian-Ping [39] discussed about neural network, an improvement scheme that iterative matrix replace secondary derivative has been developed by introduced quasi-Newton algorithm. Profile code based on probability has been used and comparison of window width and learning training has been completed. The experiment results indicate that the prediction for secondary structures of protein obtain a very good effect based on neural network and quasi-Newton algorithm. W. L. George, P. Marius *et al*.[40] describe a large scale application of a back-propagation neural network to the analysis, classification and prediction of protein secondary and tertiary structure from the sequence information alone. W. J. Barry [37] presented the tutorial. this tutorial begins with a short history of neural network research, and a review of chemical applications. The bulk, however, is devoted to providing a clear and detailed introduction to the theory behind backpropagation neural networks, along with a discussion of practical issues facing developers.

Z. Zhen, J. Nan [42] proposed a new technique based on radial basis function neural networks for prediction of protein secondary structure. To make the technique comparable to other secondary structure prediction methods, they used the benchmark evaluation data set of 126 protein chains in this paper. They also analyzed how to use evolutionary information to increase the prediction accuracy. The paper discussed the influence of data selection and structure design on the performance of the networks. The results show that this method is feasible and effective. W. Leyi and Z. Quan [41] reviews some machine learning methods. They conduct a detail survey of recent computational methods, especially machine learning-based methods, for protein fold recognition. This review is anticipated to assist researchers in their pursuit to systematically understand the computational recognition of protein folds. P. Rojalina, D. Nilamadhab *et al.* [27] investigates the protein secondary structure prediction problem by ancient learning techniques such as Artificial Neural Network where Back propagation algorithm is used for learning. It measures the efficiency and accuracy of the machine learning methods through Mean Square Error. B. Hemashree, S.K. Kandarpa [5] discuss about the ANN approach for protein structure prediction. The Artificial Neural Network (ANN) technique for prediction of protein secondary structure is the most successful one among all the techniques used. In this method, ANNs are trained to make them capable of performing recognition of amino acid patterns in known secondary structure units and these patterns are used to differentiate between the different types of secondary structures. This work is related to the prediction of secondary structure of proteins employing artificial neural network though it is restricted initially to three structures only. A. Shivani, B. Arushi *et al.*[2] discuss multilayer feed forward artificial neural network. A tool used for the secondary structure prediction of proteins from the amino acid sequence is multilayer feed forward artificial neural network with back propagation. This approach is a machine learning methodology in which the network is trained using the recognized data sets for which the widely used benchmark is Protein Data Bank (PDB), maintained by Research Collaboratory for Structural Bioinformatics (RCSB). The algorithm used for the classification is Define Secondary Structure Prediction (DSSP) that classifies the sequences in the 3-level subclasses: Helix (H), Sheet (E) and Coil (C). The objective is to get the maximum predictive accuracy with the minimalized error.

3. Materials and Methods

3.1 Approach

The approach is described in Figure 5. Firstly, load the dataset of the protein sequence in the MATLAB. In second phase, categorize all the input parameters and output results to classify the protein in appropriate group. Then data normalization has performed which will let the parameters in a particular range to get the desired and accurate results. In third phase, apply the Back Propagation Neural Network algorithm on the protein data set. In forth phase, apply the Radial basis function algorithm on the same protein data set. In the fifth phase, evaluate the performance after applying both algorithms on the protein data set. Finally, compare the results of both techniques and check that which technique gives better results.

The key issue resolved through this approach is that many models have similar structures for the same target or same target may have different modelled structures. All the physicochemical properties of such structures may be the same for few cases and different for most of the cases, but removal of such duplicate entries ensures the uniqueness in the dataset.

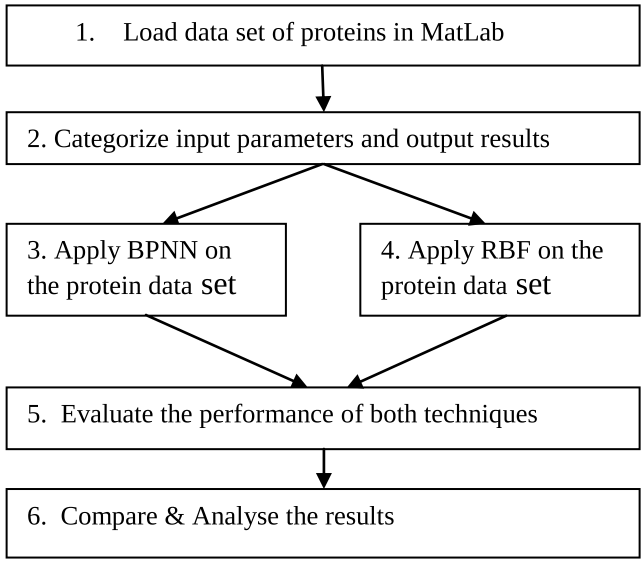


Figure 4: Methodology used

3.1.1 *Neural Network*

Neural network is used in several fields of study and a high degree of attention, made some boost progress. The behaviour of the neural network depends mainly on two aspects: one is the topology of the network, a network learning rules. Neural network comprises of many nodes mesh structure, each node in the network structure has some assigned values. Neural network generally includes three levels i.e input layer, hidden layer and output layer. The organization form of internal nodes based on neural network, neural network can be divided into distinct types of structures. Multilayer neural network consists of input layer, hidden layer and output layer; this builds a multi-layer neural network. Analysis shows that by comparing with the single layer network, multilayer neural network has better capability to process information, especially for complicated information processing ability.

This technique is based on the operation of synaptic connections in neurons of the brain, where input is processed in various levels and generalized to a final output. The neural network is often “trained” to generalize definite input signals to a desired output. In the secondary structure prediction of neural networks, input is a sliding window with 13-17 residue sequence. Information from the central amino acid of each input window is adjusted by a weighting factor, calculates and sent to a next level, termed the hidden layer, where the signal is transferred into a number near to either 1 or zero, and then transported to three output units representing each of the possible secondary structures. The output units each weigh the signal again and sum them, and convert them into either a 1 or a 0. An output signal near to 1 for a secondary structure unit demonstrates that the secondary structure of that unit is predicted and an output signal near to 0 indicated that it is not predicted. Neural network are trained by adjusting the values of the weights that modify that signals using a training set of sequences with known structure.

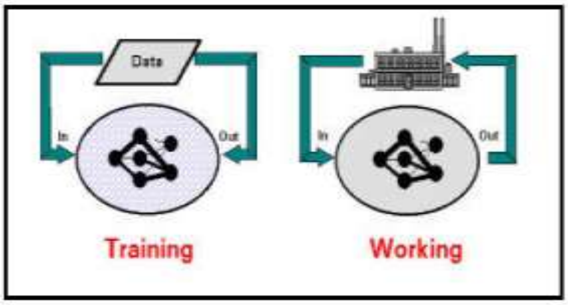


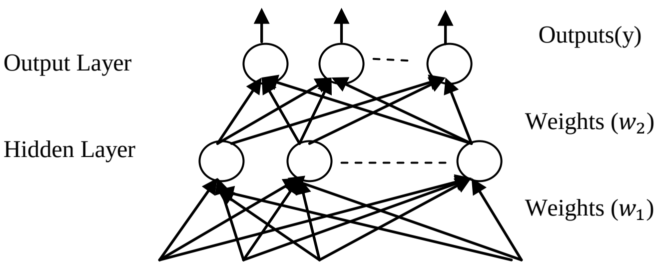
Figure 5: Training and Working Phase for Supervised Learning [8]

Artificial Neural Network - Supervised Training Algorithms.

There are many several neural network models and algorithms. Supervised learning infers the function from the supervised training data, and every example in the training data consists of an input and a target output. The supervised learning approach analyses the link between the inputs and target outputs within the training set, and produces an inferred function. The function is then employed by the network to predict the output for a given input. The prediction is predicated on the learned relationship between known amino acid sequences and 3D shapes, which is a method of supervised learning. A commonly employed algorithm for the multi-layer feed-forward neural network is the backpropagation algorithm. Its learning process can be divided into two sub-processes. First, the data feeds forward from the input layer to the output layer. Second, the error is back-propagated from the output layer to the input layer to raise the difference in between the actual output and the target output. The structure of multilayer feed-forward neural networks and the feed-forward process is presented in the next section. [28]

- **A Multilayer feed-forward Neural Network**

A Multilayer feed-forward neural network comprises of various simple, highly interconnected computing units, called neurons, which are coordinated in layers. Recall that every neuron acts a simple task of information processing, i.e. converting received inputs into processed outputs. The neurons are connected via linking edges. The knowledge learned between network inputs and outputs is acheived and stored as edges weights and biases, according to the strength of the links between different nodes. Although the information is shared at each node, overall the neural network acts well and efficiently. The architecture of a 3-layer feed-forward neural network is shown in Figure 6. The neurons in this network are coordinated into three layers (i.e. input layer, hidden layer and output layer), and the neurons in every layer are fully connected to neurons in the adjacent layer. In the feed-forward network, the data flows in one direction, from the input layer to the output layer, without feedback from the output layer. [28]



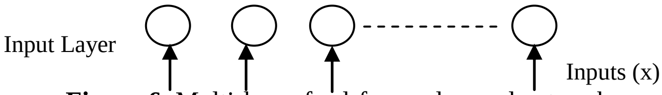
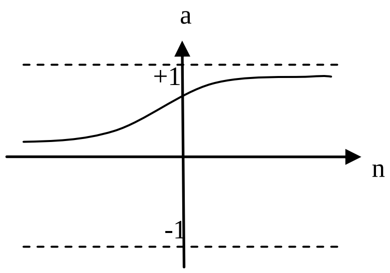


Figure 6: Multi-layer feed-forward neural networks

Neurons in the input layer are static, since they simply get data and pass it on to the neurons in the hidden layer, without any data conversion. Neurons in the hidden layer are fully connected with neurons in both the input and output layers, and are critical for neural networks to learn the mapping relationships between inputs and outputs. After getting information from neurons in the input layer, the hidden neurons process the information using the transfer function, and then propagate the processed information to the output layer for further processing to generate the outputs. Although neural networks may have more than one hidden layer, most applications only use one. Thus, the multi-layer feed-forward neural network architecture is represented by the number of layers, the number of nodes in every layer, and the transfer function employed in every layer, since the nodes in the similar layer use the similar transfer function. However, there is no widely accepted procedure to resolute the architecture of an MLP, like the number of hidden layers and the number of neurons in every layer. Therefore, it is a 22 complex process to construct a neural network. In general, the number of hidden layers and number of neurons in every layer depends on the categories of transfer functions and learning algorithms and the problems that required to be solved, and is usually resolved empirically. Due to the complication of establishing a neural network, the cost of overly huge neural networks may be more, particularly when the model structure is huge and the input has a large number of dimensions. In order to know how a neural network works, we first need to understand how the neurons in the hidden layers and output layers process data; this mechanism is shown in Figure 5. Information is prepared in two steps at each neuron in the hidden and output layers. In the first step, inputs are multiplied by the weights, related to every corresponding edge, and then gather to form a weighted sum. In the second step, the neuron uses the transfer function to transfer the sum into the output. In several cases there is an additional step between the formation of the weighted sum and the transformation, as few networks may add a bias onto the weighted sum before it is converted by the transfer function. The bias is a constant, which helps to increase the flexibility of the network. [28]

There are various options for the transfer function, but only a few are commonly used in practice. In general, the transfer function is bounded and non-decreasing. The logistic function is usually used transfer function, particularly in the hidden layers, because it is simple, non linear, bounded and monotonically increasing.



$$a = logsig(n)$$

Log-Sigmoid Transfer Function

Figure 7: The Sigmoid (logistic) Function

If the logistic function is employed in the hidden layer and the linear function is employed in the output layer, the neural network structure for a feed-forward neural network may be written as:

$$y = b_2 + \sum_{j=1}^{q} w_{1j} f\left(\sum_{i=1}^{p} w_{ij} x_i + b_i\right) \quad (11)$$

where y is the predicted output and $\{x_{i,} i = 1,2, \ldots, p\}$ are the input tuples to the network, p is the number of nodes in the input layer, q is the number of nodes in the hidden layer, $\{w_{ij,} i = 0,1,2, \ldots, p$; j=1,2,….,q} show the weights located between the input and hidden layers, $\{w_{ij,} i = 0,1,2, \ldots, p$; j=1,2,….,q} show the weights between the hidden and output layers, b_1 *and* b_2 are biases in the hidden and output layers, respectively, and f is the logistic function described above. After determining the neural network topology, we train the neural network with the training set, which comprises of the input tuples and the known output tuples. Then the input values are weighted and added at the hidden layer, and the weighted sum is converted through a specific transfer function to form the inputs to the output layer or the inputs to another hidden layer. The similar operation is conducted on the next layer until the network developed its outputs. The actual outputs give by the network can be compared with the desired (target) outputs to actuate how closely they match. This is measured by a score function, the mean squared errors (MSE), which is denoted by E. The target of training a neural network is to develop the score function, namely the mean squared error (MSE), to attain its global minimal. During the developing processes, the set of weights and biases keeps restoring. Thus, the network training process is actually a nonlinear optimization problem. [28]

3.1.2 Back-propagation network

A BP network learns by example, that is, we must provide a learning set that consists of some input examples and the known-correct output for each case. So, we use these input-output examples to show the network what type of behavior is expected, and the BP algorithm allows the network to adapt.

The BP learning process works in small iterative steps: one of the example cases is applied to the network, and the network produces some output based on the current state of it's synaptic weights (initially, the output will be random). This output is compared to the known-good output, and a mean-squared error signal is calculated. The error value is then propagated backwards through the network, and small changes are made to the weights in each layer. The weight changes are calculated to reduce the error signal for the case in question. The whole process is repeated for each of the example cases, then back to the first case again, and so on. The cycle is repeated until the overall error value drops below some pre-determined threshold. At this point we say that the network has learned the problem "well enough" - the network will never exactly learn the ideal function, but rather it will asymptotically approach the ideal function.

Generation of BP neural network

BP neural network is the most representative and usually applied learning algorithm in artificial neural network which is the acyclic multi-level network training algorithm. 3-layered network input layer, hidden layer, and output layer can be selected considered network topology because correctness of network and expression ability can't always be enhanced when hidden layer and its neurons can be enhanced in BP neural network. Generally, for an amino acid, its amino acid residues have statistical correlation which could affect the secondary structure of amino acid. Thus, the input window of neural network can be designed. If 9 amino acids VRKKRWACD can be inputted, the number of neurons is 9x21 in input layer which is coding bit rate of amino acid. The output layer consisted of 3 neurons corresponding to three secondary structures of protein alpha helix, beta sheet, and gamma crimp. Comparing 3 results of output layer, alpha helix coding is 100, beta sheet coding is 010, and gamma crimp is 001. Adaptive adjustment strategy can be employed in this algorithm [24]. BP network can be defined with Mat lab as follows

Net=newff(TEMP,[30,3],
{ 'tansing','purelin' },'traingd');

Protein Structure Prediction using Back Propagation Neural Network

The algorithm which is employed to train the ANN having 3-layer is as follows:
• Initialize the random weights in the network (often randomly)
• Do
For each example e in the training set
O = neural-net-output(network, i) ; forward pass
T = teacher output for i
Calculate error (T - O) at the output units
Compute delta_wh for all weights from hidden layer to output layer ; backward pass
• Compute delta_wi for all weights from input layer to hidden layer ; backward pass continued
• Update the weights in the network until all examples classified correctly or stopping criterion satisfied
• Return the network [14]

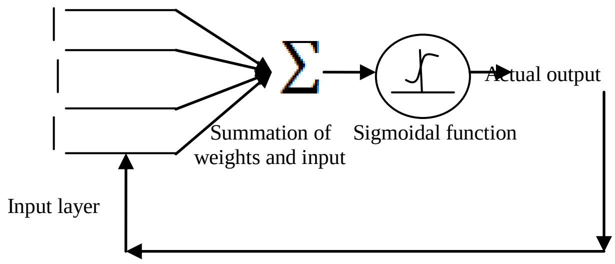


Figure 8: Artificial neural network model with back propagation

Training Parameters

The performance parameters are learning algorithm, transfer function and the number of cycles the network consumes, to be converged.
• Epochs: This decides for how many cycles we would like to train the network. For example, if the number of epochs is 100 then the whole training data will be presented 100 times.
• Learning Algorithm: the learning algorithm is LEARNGDM (Learning by Gradient Descent Method).
• Min_grad: This decides the acceptable error that we would choose. If this MSE (mean squared error) is reached the network has converged and training will stop. Generally this value is 1e-5.
• Transfer Function: The transfer function selected is Log Sigmoidal function (LOGSIG)
• Gradient descent (GDM) was used to lower the mean squared error between network output and the actual error rate. After initializing all the parameters, the network is ready to be trained. Repeated experiments were performed to get the

neural network converged. Weights were initialized to random values and networks were run until at least one of the following termination conditions was satisfied

1. Maximum Epoch
2. Minimum Gradient
3. Performance Goal

If the MSE reaches the set value the network has converged and training stops. If the network does not converge and the number of epochs has reached the set value, then the network has not converged. We will have to retrain the network, by changing some of the parameters or the training algorithm or the network architecture. Once, the network has been trained will be simulated with a set of inputs that the network has not seen. We have classified the data sets into 2 parts. i.e training set and testing set which is not used in the training process, and is used to test and then we have simulated our results with these datasets. Almost 2/3rd of the total dataset can be taken as training set and 1/3rd of the rest can be taken as test set. This is done through the analysis of the correctness attained via testing against these set. [14]

3.1.3 RBF Neural Networks Model

Radial basis functions were introduced in 1985 by Powell and Broomhead. Lowe was the first to exploit the use of Radial basis functions in neural networks. Radial basis function neural networks comprises of 3 layers, as shown in figure

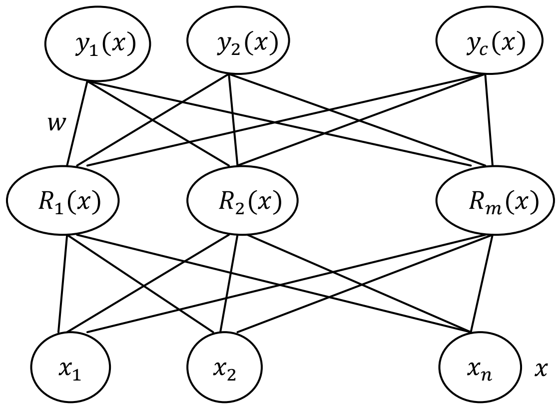


Figure 9: Radial basis function neural networks

Radial basis function neural networks are feed-forward networks. RBF networks stand for a class of neural network models, in which the hidden units are activated according to the distance between the input units. RBF networks join two different types of learning: supervised and unsupervised. At this time we used supervised learning. A special class of functions is used in RBF networks to perform a nonlinear transformation on a given network input. These functions give the RBF network its name. Radial basis functions are defined by the fact that their response varies (decreases or increases) monotonically with distance from a mid point. A typical radial function is the Gaussian function that is given by

$$R_t(x) = \exp(-\frac{\|X-Ci\|^2}{2\sigma_i^2}) \quad , i = 1,2,3, \ldots\ldots, m \qquad (12)$$

The weight correction can be calculated by back-propagation. The hidden neurons serve as computing units that functions a nonlinear transformation on the input vector by means of radial basis functions that is responsible as a basis for this transformation. Referring to every neuron comprises two parameters: The neuron specific center and the radius of the RBF, which is similar to all hidden neurons. The nonlinear transformation of the input vector is carried out in every hidden neuron by counting the response of the radial function. The Euclidean distance between the input vector and the center vector should be counted [27].

The Prediction Method Based on RBFNN

Basically there are twenty kinds of amino acids. In the first step of the analysis each amino acid might be converted into binary code of 21 units (each unit representing 1 of the amino acids) before it is delivered to neural networks. In the second step twenty units are valued to enter the percentage of every amino acid in the multiple sequence alignment at that position.

The training of the cluster layer executes first. Once finished, the output layer is subjected to supervised learning as employed in the feed-forward network with back-propagation. To train the network, various parameters required to be specified. The maximum learning rate parameter (L_{max}) ranging between 0 and 1, the minimum learning rate parameter (L_{min}) ranging between 0 and 1 and the maximum number of epochs used for training (*N*). They are used to construct the learning gradient. This gradient is required to update the weights of a cluster node whose input weight vector has the least squared Euclidean distance to the input vector pattern. The maximal learning rate also serves as the initial learning rate. In updating the weight of every dimension, the following formula is applied:

$$W_{new} = W_{old} + [L \times (I_{value} - W_{old})] \qquad (13)$$

W_{old} means the weight before updating. W_{new} means the updated weight. L is learning rate. I_{value} is input pattern value. After all input patterns have been run via the network, the learning rate itself is updated via:

$$L_{new} = G \times L_{old} \qquad (14)$$

L_{old} means the learning rate before updating. L_{new} means the updated learning rate. G is learning gradient. The gradient is calculated by the following formula:

$$G = L_{max} - E_{complete} \times (L_{max} - L_{min}) \qquad (15)$$

$E_{complete}$ means epochs completed. The weight correction w_{ij} of neuron i and j is calculated by back-propagation:

$$w_{ij} = L_{constant} \times G_{local} \times I_j \qquad (16)$$

The learning rate constant $L_{constant}$ and the non-normalized minimum average squared error of BP algorithm must be specified. G_{local} means local Gradient. *I*j means input signal of neuron j [27].

3.2 Dataset and its features

A key to comprehend the function of organic macromolecules, e.g., proteins, is the determination or expectation of its structure. Vast scale gene-sequencing projects accumulate a countless protein sequences [19]. However, information around three-dimensional structures is available for just little fractions of known proteins. In this manner experimental structure prediction has improved. This makes a requirement for seperating structural information from sequence databases. To encourage the need different protein databases are available online.

We used the database as alluded by the authors R. S. Prashant, S. Harish, B. Mahua and S. Anupam. Authors have used training, testing and validation dataset. There are a total of 95091 modelled structures of 4896 native targets. The modelled structures are taken from protein structure prediction center (CASP-5 to CASP-10 tests), public decoys database and native structure from protein data bank (RCSB). Table 1 portrays the physicochemical properties used in this study. [30]

3.3 Qualitative assessment

The Root Mean Square Deviation (RMSD), Template Modelling (TM-score) and Global Distance Test (GDT TS-score) survey the nature of the protein structure with respect to the experimental structure.

Feature	Information
Area	Total Surface Area
ED	Euclidean distance
Energy	Total empirical energy
SS	Secondary structure penalty
SL	Sequence length
PN	Pair number

Table 1: Description of the features

3.3.1 Root Mean Square Deviation (RMSD)

The RMSD is computed utilizing the superposition between matched pairs of Cα between two protein sequences. This superposition is processed using the Kabsch rotation matrix. The RMSD is computed as:

$$RMSD = \sqrt{\sum_{i}^{N}(d_i * d_i)/N} \qquad (17)$$

where, d_i is the distance between matched pair i, N is the quantity of matched pairs. RMSD is computed using the freely available program at 13.

3.3.2 Template Modelling (TM) score

TM-score is an algorithm to compute the structural comparability of two protein models. It is utilized to quantitatively evaluate the accuracy of protein structure predictions with respect to the experimental structure. TM-score weights the close atom pairs stronger than the far off matches, it is more sensitive to the topology fold than the regularly utilized RMSD since a local variation can result in a high RMSD value. TM-score has the value in (0,1] and autonomous on the length of the proteins. In light of measurements, the expected TM-score value for a random pair of proteins is ≤ 0.17 and for accurately aligned proteins ≥ 0.5. The TM-Score is computed as:

$$\text{TM score} = \frac{1}{\text{L}}\sum_{\text{i}=1}^{\text{N}}(1/(1+\text{d}_\text{i}^2/\text{d}^2)) \qquad (18)$$

where, d_i is the distance between identical residues i, d is the distance threshold, N is the quantity of residue pairs and L is the quantity of residues in the experimental structure.TM-score is calculated using the freely available program at 14.

3.3.3 Global Distance Test (GDT TS) score

GDT TS-score [3] [4] is another measure that quantitatively survey the accuracy of protein structure predictions in respect to the experimental structure. It is a measure of similarity between two protein structures with same amino acid sequences but distinct tertiary structures. GDT TS-score has the value in (0,1]. Similar to TM-score, it is also free in the length of the proteins. The GDT TS-score is computed as:

$$\text{GDT TS score} = (C1 + C2 + C3 + C4)/4N \qquad (19)$$

where, C1 is the quantity of residues superposed below (threshold/4), C2 is the quantity of residues superposed below (threshold/2), C3 is the quantity of residues superposed below (threshold), C4 is the quantity of residues superposed below (2*threshold), N is total quantity of residues and threshold used for the GDT TS-score is 4 °A. GDT TS-score is utilizing the freely available program at 14.

3.4 Feature Measurement

Here, six physicochemical properties named as total surface area (Area), euclidean distance (ED), total empirical energy (Energy), secondary structure penalty (SS), sequence length (SL) and pair number (PN) are selected. A brief discussion on the selected properties is given below:

3.4.1 Total surface area (Area)

Protein folding is led by different driving forces, which looks for towards minimization of its total surface area. Level of these external forces relies upon the surface of protein exposed to the solvent, which convey the strong dependency of free energy on solvent accessible surface area (SASA). SASA has been broadly utilized as one of the vital properties to evaluate the nature of protein structures. Hydrophobic collapse is considered as a major consideration in protein folding and this can be estimated as a loss of SASA of non-polar residues. Each amino acid shows a different affinity to be found on the surface of the protein based on the functional groups present in its side chain [17]. Some questions arise with regard to the usage of SASA: (i) should it be the total area or is it the area of the non-polar residues, (ii) what is the standard fixed value of SASA for a native structure and (iii) is the rule of minimum area applicable to non-globular proteins. Here, total SASA have been calculated using Lee & Richards [17] method as absolute value.

3.4.2 Euclidean distance (ED)

Spatial positioning of Cα or Cβ atoms are a decisive factor in providing the 3D conformation of a protein structure. Recently, neighborhood profiles of Cα atoms for each pair of residues have been characterized and observed to be invariant in 3618 native proteins suggesting certain universal geometrical constraints in their positioning [1]. Here four aliphatic non polar residues are considered i.e. Alanine (ALA), Valine (VAL), Leucine (LEU) and Isoleucine (ILE); collectively they formed 6 unique pairs among each other. Cumulative inter-atomic euclidean distance of their respective Cβ atoms for aliphatic nonpolar residues were calculated for each residue pair. Euclidean distance is given as:

$$\text{E}_\text{d} = \sum_{\text{i}=1}^{\text{n}}\sum_{\text{j}=1}^{\text{n}}\text{e}_{\text{ij}} \qquad (20)$$

where, n is the total number of aliphatic nonpolar residues; i and j are individual aliphatic nonpolar residues. e is the euclidean distance between i and j.

3.4.3 Total empirical energy (Energy)

The total empirical energy is the absolute sum of electrostatic force, van der Waals force and hydrophobic force [23][26]. Molecular dynamics simulation package AMBER1226 is used to compute total empirical energy. It is computed as given below:

$$E^{ij}_{elec} = \frac{332*qi*qj}{r_{ij}} \quad (21)$$

$$E^{ij}_{vdW} = \frac{C^{ij}_{12}}{r^{12}_{ij}} - \frac{C^{ij}_{6}}{r^{6}_{ij}} \quad (22)$$

$$E^{ij}_{hyd} = \frac{M^{ij}_{12}}{r^{12}_{ij}} - \frac{M^{ij}_{6}}{r^{6}_{ij}} \quad (23)$$

where, r_{ij} is the distance between pair of atoms i and j, $C^{ij}_{12} = \in \sigma^{12}$, $C^{ij}_{6} = 2 \in \sigma^{6}$, σ is the vander Waals radii, ∈ is the well depth, $M^{ij}_{12} =\in R^{12}, M^{ij}_{6} =\in R^{6}$, R is the distance variable and ∈ is set to 1. Finally total empirical energy is given as:

$$E_{total} = \sum_{i}^{n-1} \sum_{j=i+1}^{n} \left(E^{ij}_{elec} + E^{ij}_{vdW} + E^{ij}_{hyd}\right) \quad (24)$$

3.4.4 Secondary Structure penalty (SS)

Secondary structure prediction has come to 82% accuracy 6, 32, 18 in the course of the last few years. In this manner deviation from ideal predicted secondary structures can be utilized as a measure to evaluate the quality of a structure. Secondary structure penalty is measured from the secondary structure sequence. It is processed as the missmatches in the helix, sheet and coil of the STRIDE [10] and the PSIPRED [12] prediction. STRIDE get the real number of helix, sheet and coil exhibit in the secondary structure sequence where as PSIPRED utilize neural network to predict the likelihood for the similar secondary structure classes.

$$SS = \sum_{i=1}^{n} q_i \quad (25)$$

$$\begin{cases} 1 & \text{if } S_{stride}(P_i) = \text{if } S_{psipred}(P_i) \\ 0 & \text{otherwise} \end{cases} \quad (26)$$

where, P is the protein secondary structure sequence; S_{stride} (P) and $S_{psipred}$ (P) are the quantity of helix, sheet and coil returned by STRIDE and PSIPRED individually for every amino acid P_i . SS is figured by counting the total number of miss-matches found. It is found that SS has bring down a lower value for native and higher for non-native structure.

3.4.5 Sequence Length (SL)

Sequence length is the total number of amino acid present in the protein structure. It is computed from actual sequence.

3.4.6 Pair Number (PN)

Pair number is the aggregate number of aliphatic hydrophobic residue pairs in the protein structure and it is figured by counting the total number of pairs between the Cβ carbons in the protein structure.

4. Model Evaluation

There are various ways to measure the performance of the prediction, where some are more suitable than others depending on the application considered. A brief discussion on the performance measures is explained below. Here, we created three different models to predict the three output variables (i.e. RMSD, TM-score and GDT TS-score) by using same number of input variables (i.e. features). The formula used for all the machine learning models is given by

RMSD ~ f(Area, ED, Energy, SS, SL, PN) (27)
TM ~ f(Area, ED, Energy, SS, SL, PN) (28)
GDT TS ~ f(Area, ED, Energy, SS, SL, PN) (29)

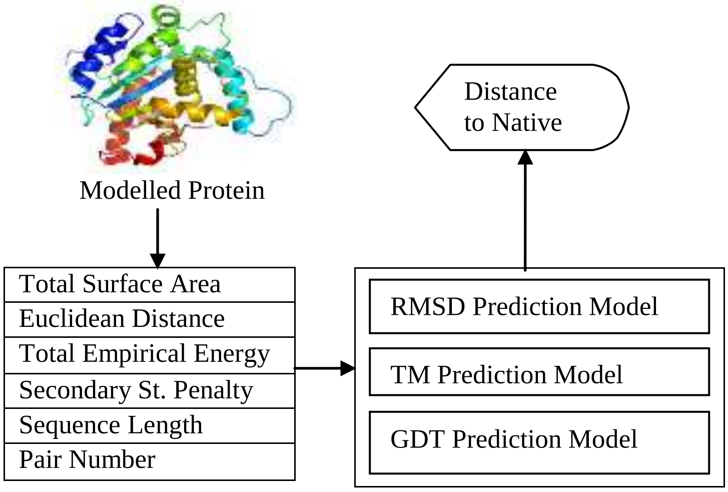


Figure 10: Prediction Method

4.1 Root Mean Squared Error

RMSE is a prominent equation to measure the error rate of a regression model. However, it must be compared between models whose errors are measured in the similar units. It is computed as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}} \quad (30)$$

where, a is actual target, p is predicted target and n is the total quantity of instances.

4.2 Correlation (r)

Correlation depicts the statistical relationships between actual and predicted values. It is described as follows:

$$Corr = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (31)$$

where, x is the actual value, y is the predicted value, $\bar{x}$ is the mean of the all actual values, $\bar{y}$ is the mean of the all predicted values and n is the quantity of instances. Correlation lies in [0,1] and thought to be good if its value tends towards 1.

4.3 Coefficient of Determination (R^2)

DOI: 10.18535/ijsrm/v5i7.27

The coefficient of determination (R^2) summarizes the illustrative power of the regression model. R^2 portrays the extent of change of the dependent variable clarified by the regression model. If the regression model is perfect then R^2 is 1 and if the regression model is an aggregate failure then R^2 is zero i.e. no variance is clarified by regression. The Coefficient of Determination is calculated by taking the square the r (i.e. Correlation). It is defined as follows:

$$R^2 = r * r \tag{32}$$

4.4 Accuracy

The accuracy is computed as percentage deviation of predicted target with actual target with acceptable error.

$$Accuracy = \frac{100}{n} \sum_{i=1}^{n} q_i \tag{33}$$

$$q_i = \begin{cases} 1 & if\ S_{stride}(P_i) = if\ S_{psipred}(P_i) \\ 0 & otherwise \end{cases} \tag{34}$$

where, a is actual target, p is predicted target, err is the acceptable error and n is the total quantity of instances.

5. Results

In this section, the prediction results of two neural network algorithms on the training-testing dataset are evaluated. For training testing experiment, dataset comprises of protein structures from CASP-5 to CASP-9 experiments, public decoys database and native structure from protein data bank (RCSB). All the models are evaluated on RMSE, correlation, R2 and accuracy.

5.1 Training-Testing Experiment

The distribution of data in training-testing experiment are set to 70% and 30% respectively for all the methods. Table 2 shows the comparative performance of two neural network algorithms that is BPNN and RBF based on RMSE, correlation, R2 and accuracy. The performance result shows that Backpropagation algorithm (BPNN) outperforms over Radial Basis Function algorithm in the prediction. The RMSE is used to measure the difference between actual and predicted values. The RMSE is calculated using equation 30 and Table 2 shows the RMSE of these two methods. The Radial basis function (RBF) outperforms in case of RMSE because RBF has the lowest RMSE of 0.166, 0.333, 0.280 and 0.500 with learning rate 0.15, 0.3, 0.2 and 0.45 respectively.

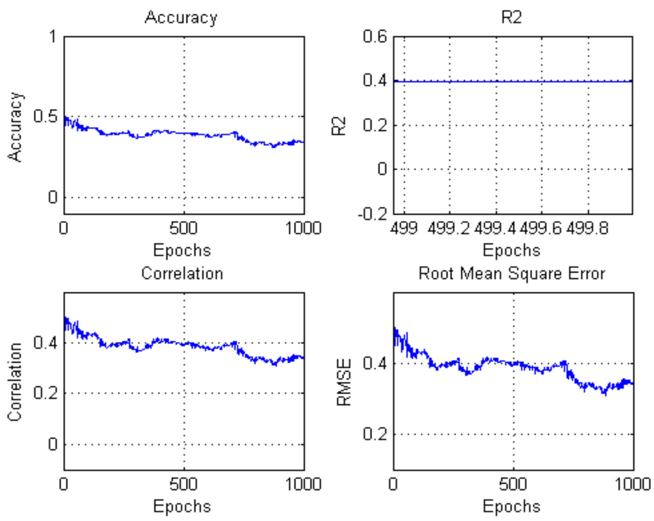
The correlation describes the statistical relationship between actual and predicted values. The correlation is calculated using equation 31 and Table 2 shows the correlation of BPNN and RBF. The BPNN has the highest correlation of 0.355, 0.412, 0.365 and 0.319 with learning rate 0.15, 0.3, 0.2 and 0.45 respectively.

The R^2 is the sum of squares regression between actual and predicted values. The R^2 is calculated using equation 32 and Table 2 shows the R^2 of all the methods. The BPNN has the highest R^2 of 0.477, 0.513, 0.483 and 0.452 with learning rate 0.15, 0.3, 0.2 and 0.45 respectively.

Accuracy is the degree of consistency of measured quantity to its true actual. The accuracy is calculated using equation 33 with some acceptable error and Table 2 shows the accuracy of all the methods. The BPNN has the highest accuracy of 0.775, 0.834, 0.786 and 0.734 with learning rate 0.15, 0.3, 0.2 and 0.45 respectively.

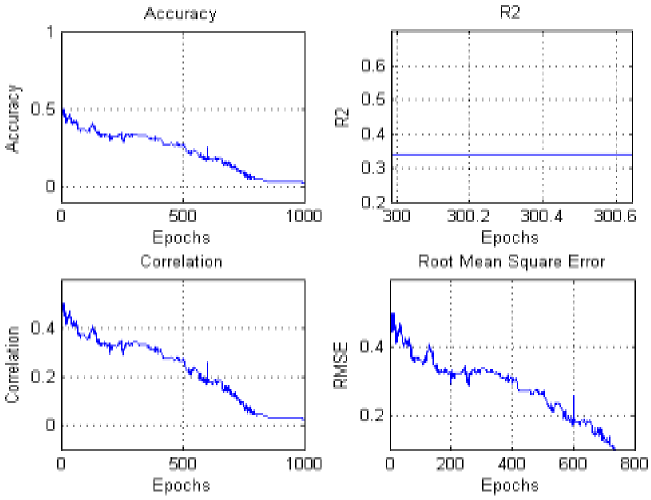
Table 2: Performance comparison of BPNN and RBF

Learning Rate	Accuracy		R^2		Correlation		RMSE	
	BPNN	RBF	BPNN	RBF	BPNN	RBF	BPNN	RBF
0.15	**0.775**	0.228	**0.477**	0.140	**0.355**	0.027	0.596	**0.166**
0.3	**0.834**	0.457	**0.513**	0.281	**0.412**	0.111	0.642	**0.333**
0.2	**0.786**	0.384	**0.483**	0.236	**0.365**	0.078	0.604	**0.280**
0.45	**0.734**	0.685	**0.452**	0.422	**0.319**	0.250	0.565	**0.500**



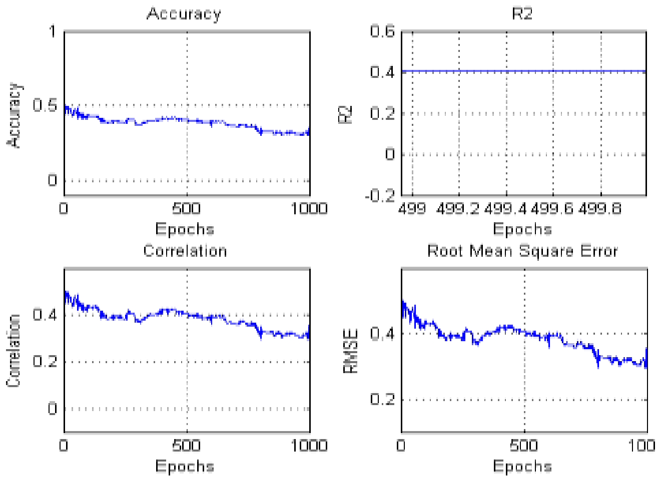
(a)

DOI: 10.18535/ijsrm/v5i7.27

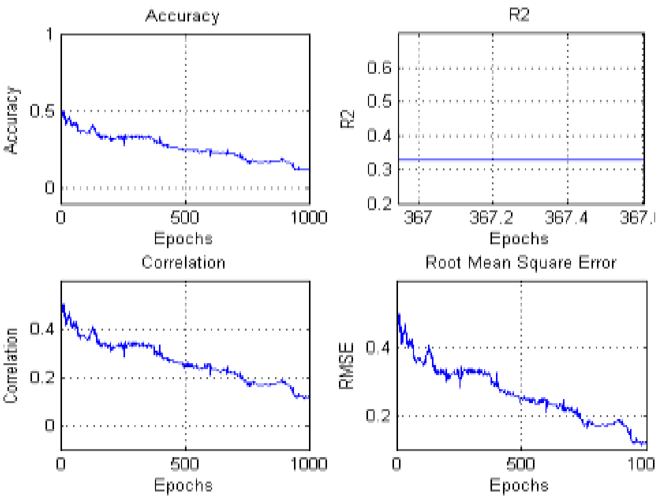


(b)

Figure 11: Performance of (a) BPNN & (b) RBF with learning rate 0.15

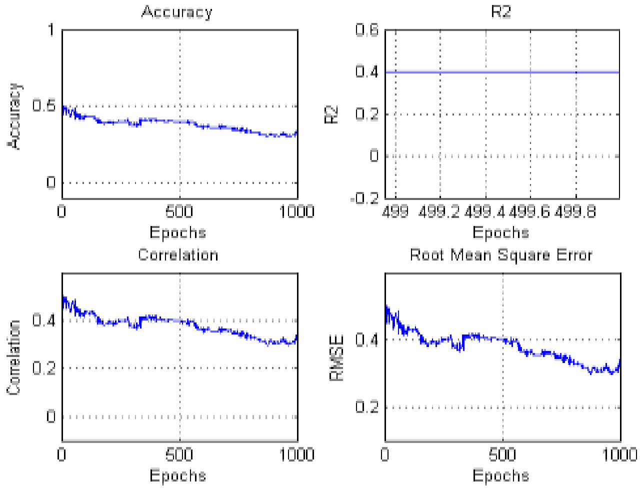


(a)

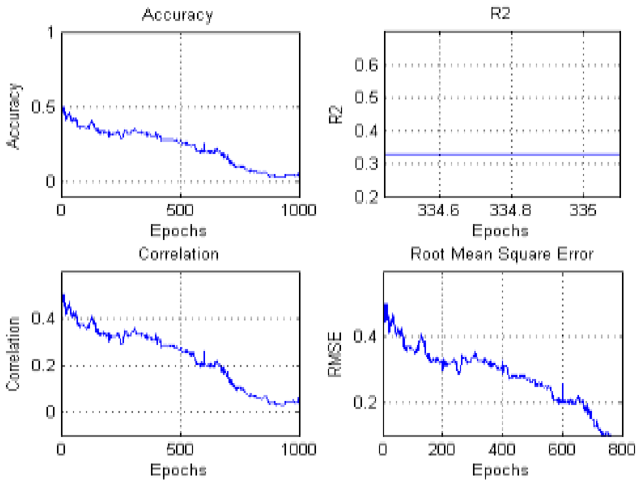


(b)

Figure 12: Performance of (a) BPNN & (b) RBF with learning rate 0.3

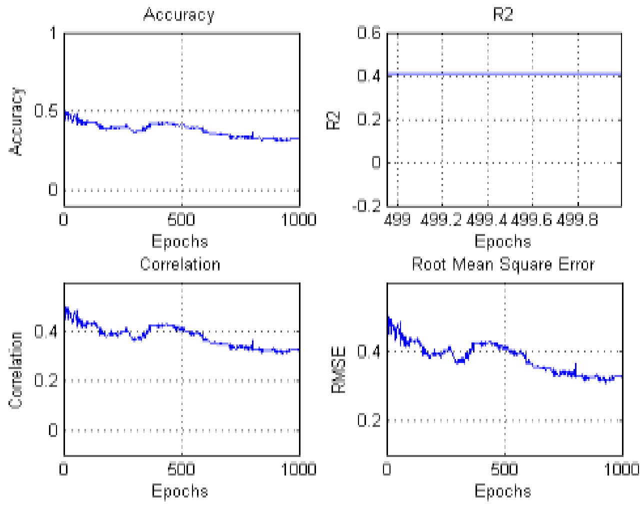


(a)

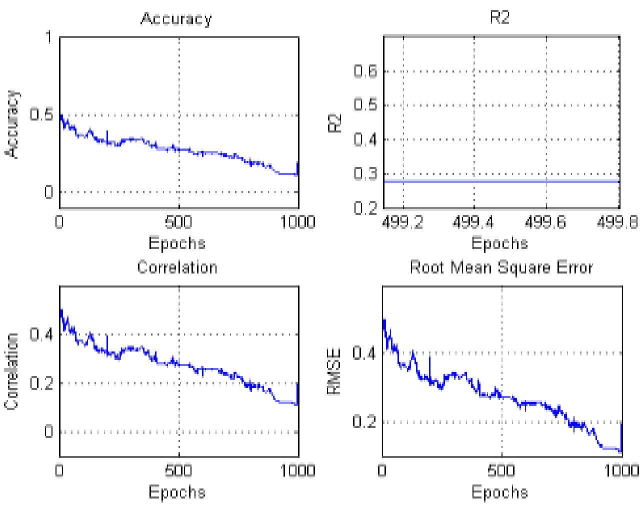


(b)

Figure 13: Performance of (a) BPNN & (b) RBF with learning rate 0.2



(a)



(b)

Figure14: Performance of (a) BPNN & (b) RBF with learning rate 0.45

6. Conclusion

In this work, two neural network techniques that is BPNN and RBF are explored having six physicochemical properties for estimating the absolute quality of a modelled protein structure in the absence of its true native state. The absolute quality of a model is expressed in terms of how well the model score agrees with the expected values from a representative set of high resolution experimental structures. In this paper, four parameters are discussed that is accuracy, R^2, correlation and RMSE with the help of which we measure the quality of these two neural network techniques. Here, neural network techniques does not include any additional information from other models or alternative template structures. The dataset used in this study is low in features and very high in observation values. All the models are evaluated on RMSE, correlation, R^2 and accuracy. Through the intensive experiments, it is found that Back-propagation neural network algorithm outperforms in most of the parameters over the Radial basis function algorithm.

7. Acknowledgement

I would like to thank my guide Dr. BALJIT SINGH KHEHRA, Professor and Head of Computer Science and Engineering department, for providing necessary infrastructure for the research work and helping me to write this paper.

References

[1] A. Mittal and B Jayaram, "Backbones of folded proteins reveal novel invariant amino acid neighborhoods", JBSD, Vol. 28, No. 4, pp. 443–454, 2011.

[2] A. Shivani, B. Arushi, M. Deepali , "Design and Implementation of an Algorithm to Predict Secondary Structure of Proteins using Artificial Neural Network", International Journal of Emerging Research in Management &Technology ISSN: 2278-9359, Vol. 2, No. 12, pp. 34-40, 2013.

[3] A Zemla, C Venclovas, J Moult, and K Fidelis, "Processing and analysis of CASP3 protein structure Predictions", Proteins, Vol. 37, pp. 22–29, 1999.

[4] A Zemla, "LGA: a method for finding 3D similarities in protein structures", NAR, Vol. 31, No. 13, pp. 3370–3374, 2003.

[5] B. Hemashree, S.K. Kandarpa, "Protein Structure Prediction using Artificial Neural Network", Special Issue of International Journal of Computer Applications (0975 – 8887) on Electronics, Information and Communication Engineering - ICEICE , No.3, pp. 22-24, 2011.

[6] B. Marco, B. Stefan, W. Andrew and A. Konstantin, "SWISS-MODEL: modeling protein tertiary and quaternary structure using evolutionary information", Nucleic acids research, page gku340, 2014.

[7] B. Wenzheng, C. Yiming, C. Yuehui, "Multiple Feature Fusion Protein Tertiary Structure Prediction", International Conference on Information Science and Cloud Computing Companion, pp. 751-756, 2013.

[8] C. Jianlin, N. T. Allison, and B. Pierre, "Machine Learning Methods for Protein Structure Prediction", IEEE reviews in biomedical engineering , Vol. 1, pp. 41-49, 2008.

[9] C. Nandini, "A Survey on Protein Sequence Classification with Data Mining Techniques", International Journal Of Scientific & Engineering Research, Vol. 7, No. 7, pp. 1442-1449, July 2016.

[10] D Frishman and P Argos, "Knowledge based protein secondary structure assignment", Proteins, Vol. 23, No. 4, pp. 566–579, 1995.

[11] D. H. Shing, Y. Y. Chi, "Secondary Structure Prediction Using SVM and Clustering", Proceedings of the Fourth International Conference on Hybrid Intelligent Systems, 2004.

[12] D. T. Jones, "Protein secondary structure prediction based on position specific scoring matrices", Journal of Molecular Biology, Vol. 292, No. 2, pp. 195–202, 1999.

[13] http://zhanglab.ccmb.med.umich.edu/TM-score/RMSD.f.

[14] http://zhanglab.ccmb.med.umich.edu/TM-score/TMscore.f.

[15] F. Mhamdi, "Text mining, feature selection and data mining for proteins classification", International conference on information and communication technologies, pp. 457-458, April 2004.

[16] I. S. Mohammad, A. Hakimeh, "RMSD Protein Tertiary Structure Prediction with Soft Computing", International Journal of Mathematical Sciences and Computing, Vol. 2, pp. 24-33, 2016.

[17] J Janin, "Surface and inside volumes in globular proteins", 1979.

[18] K. Andriy, M. John, B. Patrick, and J Fernando Bazan, "Challenging the state of the art in protein structure prediction: Highlights of experiment

CASP10 structures", Proteins: Structure, Function, and Bioinformatics, Vol. 82, No. S2, pp. 26–42, 2014.

[19] K. T. Simons, C. Kooperberg, E. Huang, and D. Baker. "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions", Journal of Molecular Biology, Vol.268, No. 1, pp. 209–225, 1997.

[20] K. Yasaman, F. Mahmood, K. Hamed, S. Hossein, "Protein Structure Prediction using Bio-Inspired Algorithm: a review", The 16th CSI International Symposium on Artificial Intelligence and Signal Processing, pp. 201-206, 2012.

[21] M. Sonal, P. Yadunath and A. Anamika, "Classification of Protein Structure (RMSD <= 6A°) using Physicochemical Properties", International Journal of Bio-Science and Bio-Technology, Vol.7, No.6, pp.141-150, 2015.

[22] M. Vidyasagar, "Some Challenges in Computational Biology", European control conference, pp. 3364-3369, 2013.

[23] N Arora and B Jayaram, "Strength of hydrogen bonds in a helices", Journal of Computational Chemistry, Vol. 18, No. 8 pp. 1245–1252, 1997.

[24] N. Mathuriya, "Comparison of K-means and Backpropagation Data Mining Algorithms", International Journal of Computer Technology and Electronics Engineering, Vol. 2, No. 2, pp. 151-155, April 2012.

[25] P. Mayuri, S. Hitesh, "Protein Secondary Structure Prediction Using Support Vector Machines (SVMs)", International Conference on Machine Intelligence Research and Advancement, pp. 594-598, 2013.

[26] P Narang, K Bhushan, S Bose, and B Jayaram, "Protein structure evaluation using an all-atom energy based empirical scoring function", JBSD, Vol. 23, No. 4, pp. 385–406, 2006.

[27] P. Rojalina, D. Nilamadhab, R. Samita, "A Novel Approach for Protein Structure Prediction using Back Propagation Neural Network", International Journal of Computer Science And Technology, Vol. 3, No. 2, pp. 600-603, 2012.

[28] P. Sun and J. Zhang, "Improved Prediction Method of Protein Contact Based on RBF Neural Network", 3rd International conference on bioinformatics and biomedical engineering, pp. 1-4, 2009.

[29] R. N. Chandrayani, K. Manali, "Bioinformatics: Protein Structure Prediction", 4th IEEECNT, 2013.

[30] R. S. Prashant, S. Harish, B. Mahua and S. Anupam, "Quality assessment of modelled protein structure using physicochemical properties", Journal of Bioinformatics and Computational Biology, Vol. 13, No. 02, pp. 1-16, 2015.

[31] S. Saha, "Application of Data Mining In Protein Sequence Classification", International Journal of Database Management Systems, Vol. 4, No. 5, pp. 103-118, October 2012.

[32] T. Z. Sen, R.L. Jernigan, J. Garnier and A Kloczkowski, "GOR V server for protein secondary structure prediction", Bioinformatics, Vol. 21, No. 11, pp. 2787–2788, 2005.

[33] V. Swati, S. K. Bithin, R. K. Santanu, "An Efficient Technique for Protein Classification Using Feature Extraction by Artificial Neural Networks", Annual IEEE India conference (INDICON), 2010.

[34] W. Dianhui, L. K. Nung, D. S. Tharam, "Data Mining for Building Neural Protein Sequence Classification Systems with Improved Performance", IEEE Publications, pp. 1746-1751, 2003.

[35] W. Bo,L. Yongkui , Y. Jian , L. Shuang, "Application Research of Protein Structure Prediction Based Support Vector Machine", International Symposium on Knowledge Acquisition and Modeling, pp. 581-584, 2008.

[36] https://www.researchgate.net/publication/228894892_Approaches_to_Protein_Structure_Prediction_and_Their_Applications

[37] W. J. Barry, "Backpropagation Neural Network: A tutorial", Chemometrics and Intelligent Laboratory System Elsevier Science Publishers B.V., Amsterdam, Vol. 18, pp. 115-155, 1993.

[38] W. Jian, L. Jian-Ping, "Protein Secondary Structure Prediction Based on BP Neural Network and Quasi- Newton Algorithm", International conference on apperceiving computing and intelligent analysis (ICACIA) IEEE, pp. 128-131,2008.

[39] W. L. George, P. Marius, L. N. Michael, Neural Network Analysis Of protein tertiary structure, Tetrahedron Computer Methodology, Vol. 3, No. ¾, pp. 191-211, 1990.

[40] W. Leyi and Z. Quan, "Recent Progress in Machine Learning-Based Methods for Protein Fold Recognition", International Journal of molecular sciences, pp. 1-13, 2016.

[41] Z. Zhen, J. Nan, "RADIAL BASIS FUNCTION METHOD FOR PREDICTION OF PROTEIN SECONDAR STRUCTURE", Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, Vol. 3, pp. 1379-1383, 2008.

[42] *http://bit.ly/RF-PCP-DataSets*.

[43]https://www.ruor.uottawa.ca/bitstream/10393/23636/3/Zhao_Jing_2013_thesis.pdf

[44] www.scfbio-iitd.res.in/software/pcsm/dataset/PublicDecoys.