Statistical Assessment of Physicochemical Properties of Protein Tertiary Structure

Deok Hee Nam
Engineering and Computer Science, Wilberforce University, Wilberforce, Ohio, USA.

Abstract - *Proteins are large biomolecules, or macromolecules (heteropolymers) made up from C_α-amino acids, which referred to as residues, consisting of one or more long chains of amino acid residues. In addition, the residue-level structure-evolution relationship is sensitive to protein core size. For example, surface residues from large-core proteins evolve much faster than those from small-core proteins, while buried residues are equally constrained independent of protein core size. The proposed work shows how to predict the size of the residue through the root-mean-square-deviation (RMSD) values as one of the physicochemical properties of protein tertiary structure using various multivariate analyses with reduced dimensionalities. Neurofuzzy systems developed by the various reduced data are adapted to predict RMSD values through the physicochemical properties of protein tertiary structure data. Additionally, different paradigms of the reduced dimensionalities are compared to find the better performance between multivariate methods with five extracted features. Finally, various statistical categories are applied to determine the best performance among the applied techniques to evaluate the results through the neurofuzzy systems with the original and reduced data of physicochemical properties of protein tertiary structure.*

Keywords: data mining dimension reduction, feature extraction, multivariate analysis, neurofuzzy system, root-mean-square-deviation (RMSD).

1 Introduction

In many years, researches on the recognition of physicochemical properties of protein tertiary structure have been focused and received increasing attention. To control bioinformatic data sets to be implemented, system reduced techniques are frequently used for the effective implementation of the deployed data sets since it is also cost-effective and able to carry a relatively better performance. Nevertheless, the prediction of protein structure is still a difficult problem to improve the performance. Hence, the knowledge-based approach with the data mining methods, has been used for an alternative method. Through the acquired knowledge-based data sets, the characteristics of the protein structure can be used for predicting any known amino acid sequence (proteins) with known sequences but unknown structures [1]. The characterized protein data sets can be also categorized by the knowledge-based methods into two groups such as statistics-based methods and neural network based methods [2]. On the other hand, researches about the efficient data management are also getting more focused and important in these days. Simultaneously, data mining techniques to deal with the reduced data without losing any significant meaning, instead of using the original data, are developed more and more. Among data mining techniques, the most frequently used techniques are applying the various multivariate analysis techniques with the extracted features by reducing the dimensionalities. In this paper, the proposed dimension-reduction techniques are compared with the traditional techniques such as factor analysis, principal component analysis, and fuzzy c-means clustering analysis. Then, the results obtained by the proposed techniques through neurofuzzy systems with the reduced data are compared to the results from the tradition methods as described earlier. In this paper, the section 2 shows the review of the previous studies about the protein structures and briefly introduces the traditional multivariate techniques. In the section 3, the data sets of physicochemical properties of protein tertiary structure are briefly summarized. In the section 4, deployed neurofuzzy systems with applying the reduced data sets extracted by the proposed reduction techniques are presented. In the section 5, the results obtained by applying the proposed techniques are compared with the results applied by the traditional multivariate analyses techniques through the statistical evaluation methods. Finally, the section 6 shows the conclusion of the proposed study.

2 Literature Review

2.1 Previous studies of protein structures length

In general, to assess the quality of the protein structure, several characteristics of physicochemical properties of protein tertiary structure have been used such as Root Mean Square Deviation (RMSD), Template Modelling (TM-score) and Global Distance Test (GDT TS-score) [3]. As an alternative technique, using a supervised learning of recurrent neural networks can be also used for predicting the secondary structures of the protein from the underlying amino acid's sequence [4]. In Zhou and et al. [5], a model named severity

assessment code (SAC) method has been proposed based on the association rule classifier and its rule base structure are presenting the rules obtained by the Knowledge Discovery in the Databases (KDD) model with mining the secondary structure of the protein for the information of mutual impacts [5]. There is another technique based on neural networks to predict the protein secondary structures using a multi-modal back propagation neural network (MMBP) method [6]. In this model, three layers of the intelligent interface are composed by the integration of the multi-modal back propagation neural network (MMBP), mixed-modal SVM (MMS), modified Knowledge Discovery in the Databases (KDD) process and other techniques. In addition, a compact hybrid feature vector for an accurate secondary structure prediction as a method using the derivative feature vector was also proposed by Hassan and el at. [7]. Another protein structure recognition study was introduced by Franzosa and Xia [13] through the comparison study of the protein structure by measuring the independent effects of protein core size and expression on residue-level structure-evolution relationships with demonstrating yeast protein evolutionary rate at the level of individual amino acid residues scales linearly with degree of solvent accessibility. Aslam and el at. [14] shows the protein structural alignment as a study of the computational structural biology and compares a protein structure with know structures to classify into a new known group of the protein to determine the function of the protein, its evolution relationship with other protein molecules and embedded structures. In Zhang and Skolnick [15], a new algorithm for the template modeling (TM) alignment was introduced by comparing protein structures or models with specified equivalent structures between pairs of residues provided by sequence or threading algorithms. Zhou and el at. [20] proposed to measure the structural similarity between proteins by correlating the principle components of their secondary structure interaction matrix.

2.2 Factor Analysis (FA)

Factor analysis [8] is a method for explaining the structure of data by explaining the correlations between variables. Factor analysis summarizes data into a few dimensions by condensing a large number of variables into a smaller set of latent variables or factors without losing any significance of the given data. Since factor analysis is a statistical procedure to identify interrelationships that exist among a large number of variables, factor analysis identifies how suites of variables are related. Factor analysis can be used for exploratory or confirmatory purposes. As an exploratory procedure, factor analysis is used to search for a possible underlying structure in the variables. In confirmatory research, the researcher evaluates how similar the actual structure of the data, as indicated by factor analysis, is to the expected structure. The major difference between exploratory and confirmatory factor analysis is that researcher has formulated hypotheses about the underlying structure of the variables when using factor analysis for confirmatory purposes. As an exploratory tool, factor analysis doesn't have many statistical assumptions. The only real assumption is presence of relatedness between the variables as represented by the correlation coefficient. If there are no correlations, then there is no underlying structure. There are five basic factor analysis steps such as data collection and generation of the correlation matrix, partition of variance into common and unique components, extraction of initial factor solution, rotation and interpretation, and construction of scales or factor scores to use in further analyses. In addition, FA applies some rotational transformation based upon how each variable lies somewhere in the plane formed by the factors. The factor loadings, which represent the correlation between the factor and the variable, can also be thought of as the variable's coordinates on this plane. In un-rotated factor solution the Factor "axes" may not line up very well with the pattern of variables and the loadings may show no clear pattern. Factor axes can be rotated to more closely correspond to the variables and therefore become more meaningful. Relative relationships between variables are preserved. The rotation can be either orthogonal or oblique.

2.3 Principal Component Analysis (PCA)

Principal components analysis [9] is a procedure for identifying a smaller number of uncorrelated variables, called "principal components", from a large set of data. The goal of principal components analysis is to explain the maximum amount of variance with the fewest number of principal components without losing any significance of the given data. Principal Component Analysis (PCA) is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set. Hence, principal component analysis (PCA) is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The following characteristics explain the PCA. First, the first principal component accounts for as much of the variability in the data as possible, and each successive component accounts for as much of the remaining variability as possible. Second, PCA reduces attribute space from a larger number of variables to a smaller number of components and as such is a "non-dependent" procedure if it does not assume a dependent variable is specified. Third, PCA is a dimensionality reduction or data compression method. Since the goal is dimension reduction and there is no guarantee that the dimensions are interpretable, in order to select a subset of variables from a larger set based on the original variables, the highest correlations with the principal components need to be considered.

2.4 Fuzzy C-Means (FCM) Clustering [10]

For $2 \le c \le n$, consider a set of n vectors $X = \{x_1, x_2, ..., x_n\}$ to be clustered into c groups of data. Each of the groups, $x_i \in R^S$, is a feature vector consisting of s real-valued measurements describing the features of the object represented by x_i. The features could be length, width, color, etc. Fuzzy clustering of the objects can be represented by a fuzzy membership matrix called a fuzzy partition.

The set of all $c \times n$ non-degenerate constrained fuzzy partition matrices is denoted by M_{fcn} and is defined as

$$U = [u_{ij}]_{i=1,2,...,c,\, j=1,2,...,n} = M_{fcn} \quad (1)$$

where u_{ij} expresses the degree to which the element x_j belongs to the i^{th} cluster and is a numerical value in [0,1] such that the constraints in

$$\sum_{i=1}^{c} u_{ij} = 1 \quad \text{for all } j = 1, 2, \ldots\ldots, n \quad (2)$$

and

$$0 < \sum_{j=1}^{n} u_{ij} < n \quad \text{for all } i = 1, 2, \ldots, c. \quad (3)$$

For the fuzzy c-means algorithms, the objective is to find $U = [u_{ik}] \in M_{fcn}$ as the fuzzy c-partition matrix and $V = (v_1, \ldots, v_c)$ with $v_i \in R^s$ as the cluster center such that

$$J_m(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} (u_{ik})^m \left\| x_k - v_i \right\|^2 \quad (4)$$

is minimized, where u_{ik} is the value of the i^{th} membership function on the k^{th} data point x_k, n is the number of samples in X, and $m \in (1, \infty)$ is a weighting constant. The distances, $d_{ik} = \left\| x_k - v_i \right\|^2$, are weighted with the membership values u_{ik}^m, where $\left\| x_k - v_i \right\|^2$ is any inner product-induced norm on R^s, which is called the square of Euclidean distance for the i^{th} cluster center and the j^{th} data point. The Euclidean distance formula is

$$d_{ik} = \|x_k - v_i\| = (\{x_k - v_i\}^2)^{\frac{1}{2}} \quad (5)$$

where d_{ik} is the distance between the i^{th} cluster center and the k^{th} data point x_k.

The assumption is that the distance between their corresponding data vectors measures the similarity between objects. Then the necessary conditions to minimize the objective function, $J_m(U,V)$, can be

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{\frac{2}{m-1}}} \quad (6)$$

and

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m} \quad (7)$$

where $1 < i < c$, and $1 < k < n$. Therefore, the fuzzy c-means algorithm is an iterated procedure through these two necessary conditions to minimize the objective function, $J_m(U,V)$.

2.5 Varimax Rotation [11]

The varimax rotation procedure was first proposed by Kaiser in 1958. The procedure is to find an orthonormal rotation matrix T by multiplying the given number of points and the number of dimensions configuration A. Then, the sum of variances of the columns of B × B, is a maximum, where B = AT. Each factor has a small number of large loadings and a large number of Zero (or small) loadings. A direct solution for the optimal T is not available, except for the case when the number of dimensions equals two. Kaiser suggested an iterative algorithm based on planar rotations, i.e., alternate rotations of all pairs of columns of A. This simplifies that each original variable tends to be associated with one or a small number of factors, and each factor represents only a small number of variables. Then, the factors are interpreted from the opposition of few variables with positive loadings to few variables with negative loadings. The variance of the loadings is maximized by

$$\mathrm{V} = \sum (q_{j,l}^2 - \bar{q}_{j,l}^2)^2 \quad (8)$$

where $q_{j,l}^2$ is the squared loading of the j^{th} variable on the l factor and $\bar{q}_{j,l}^2$ is the mean of the squared loadings.

2.6 Root Mean Square Deviation (RMSD) for Protein Structures

The root-mean-square-deviation (RMSD) is the most commonly used metric in the detection of the protein structure in order to quantifies the deviation for the degree of the similarity in between two or more than two protein structures. RMSD values were calculated by

$$\mathrm{RMSD} = \sqrt{\frac{\sum d_i^2}{n}} \quad (9)$$

where n is the number of pairs of equivalent C_α atoms and d_i is the Euclidean distance between the two C_α atom of the i^{th} pair. By the standardization of RMSD values, $RMSD^{std}$ becomes

$$RMSD^{std} = \frac{RMSD}{1 + ln\sqrt{\frac{n}{100}}} \quad (10)$$

where n is the number of residues in the proteins that are compared and $RMSD^{std}$ values are the RMSD that would be measured if the structures that are compared contained 100 residues [19].

Basically, the RMSD quantifies how similar the three-dimensional structure of two proteins are [18]. In other words, for the protein structure comparison, the RMSD is the root mean square deviation between corresponding residues which is calculated after an optimal rotation of one structure to another. Since the RMSD weights the distances between all residue pairs equally, a small number of local structural deviations could result in a high RMSD even though the global

topologies of the compared structure are similar. Apart from the intrinsic uncertainty in structure determination, proteins are dynamic flexible entities that can undergo significant structural fluctuation. In addition, the average RMSD of randomly related proteins is based upon the length of compared structures [16]. RMSD for protein structure has often been used to measure the quality of reproduction of a know particular binding site cluster of a protein [17].

3 Physicochemical Properties of Protein Tertiary Structure Data [12]

The data set used in the presented paper is a data set of physicochemical properties of protein tertiary structure. The data set is taken from CASP 5-9. There are 45730 decoys and size varying from 0 to 21 Armstrong. Six variables of the data have been used to implement the prediction of the physicochemical properties of protein tertiary structure. The input variables are non-polar exposed data, fractional area of exposed non-polar part of residue, molecular mass weighted exposed data, Euclidian distance, secondary structure penalty, and special distribution constants (N.K. value),. The output is RMSD, which presents the size of the residue for the physicochemical property of protein tertiary structure.

4 Applied Neurofuzzy Systems

To predict the size of the residues for protein structures, the neurofuzzy systems are used for evaluating the predicted values applying the reduced data sets or original data set. Fig. 1 shows a neurofuzzy system using reduced five inputs and one output with the original data to evaluate the size of the residues for the protein structures through the root-mean-square-deviation (RMSD) values applying the physicochemical properties of protein tertiary structure.

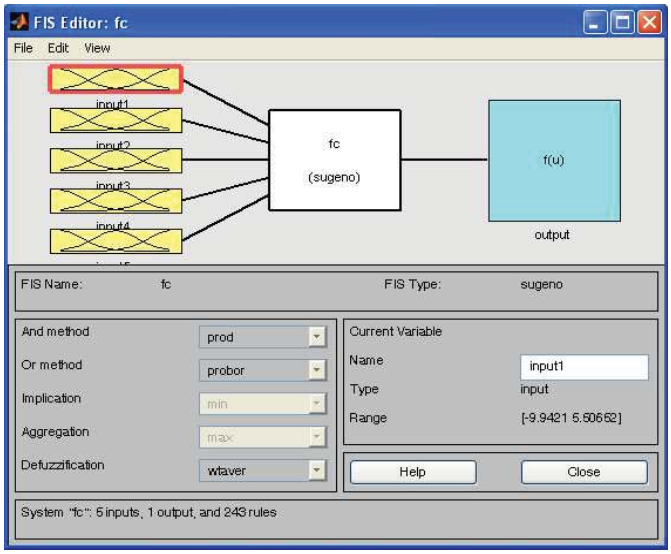

Fig. 1 Neurofuzzy Inference System of physicochemical properties of protein tertiary structure

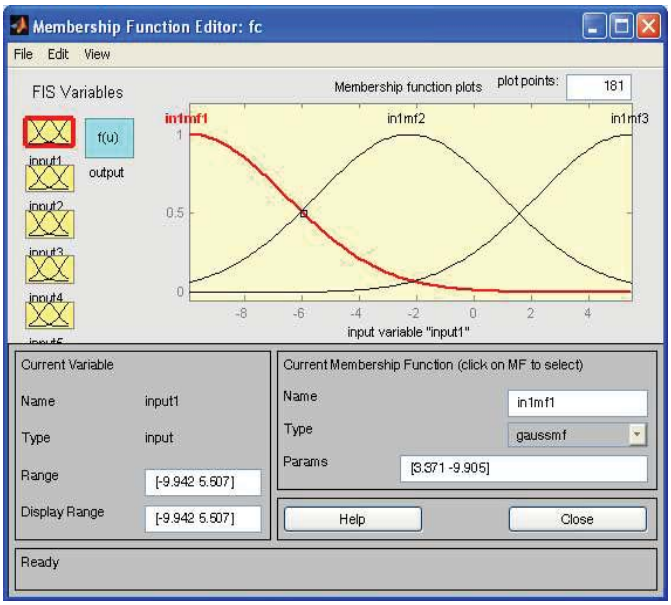

Fig. 2 Neurofuzzy Inference System with membership functions of Input Variables

Fig. 2 shows the membership functions for each input variables and an output variable for applied neurofuzzy systems. Fig. 3 shows the developed rulebase system for applied neurofuzzy systems with reduced five inputs and one output for the prediction of the size of the residues for the protein structures through root-mean-square-deviation (RMSD) for the physicochemical properties of protein tertiary structure.

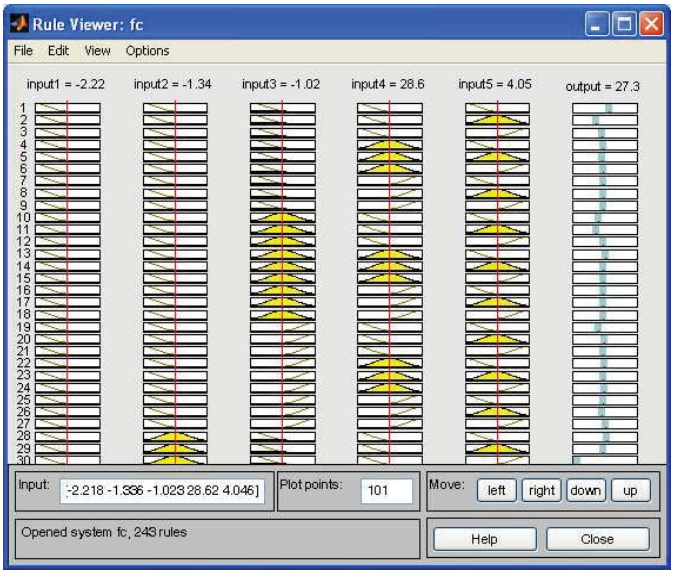

Fig. 3 Rulebase System for Reduced Components

5 Analyses and results

The proposed work has been analyzed by comparing the results with applying factor analysis, principal component analysis, and FCM clustering analysis with varimax rotation using the neurofuzzy systems. In Fig. 4, to determine the

reduced dimensionality through a proposed procedure, the number of the reduced components or factors is determined by the accumulation of the covariance and the significant eigenvalues of the system when the eigenvalues are plotted versus each factor or component extracted by the applied multivariate analyses. As an example of using the factor analysis among the other compared techniques, from Fig. 4, the first three or four newly extracted factors are relatively significant to implement the deployed data.

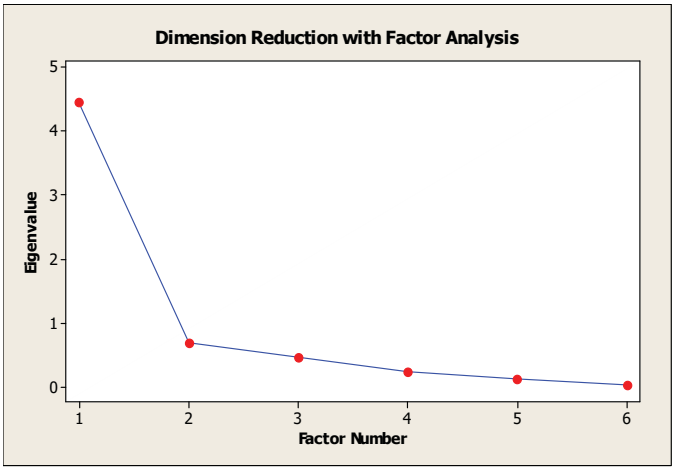

Fig. 4 Graph of Eigenvalues vs. Extracted Features

Hence, in the evaluation, three or four extracted factors by factor analysis may be enough to show its comparison, but five extracted factors are used for the predicted estimation of root-mean-square-deviation (RMSD) values applying the physicochemical properties of protein tertiary structure for the better accuracy with more accumulated data. Similarly, applying principal component analysis or other methods, five newly extracted components are selected and used to evaluate the prediction of the size of the residues for the protein structure through the root-mean-square-deviation (RMSD) values applying the physicochemical properties of protein tertiary structure.

TABLE 1 Multivariate Analyses including Five Reduced Dimension Analyses

	corr	trms	stdev	mad	ewi
pc	0.7397	1.239	2.178	1.203	4.8801
fc	0.9889	0.273	0.853	0.265	1.4024
pca	0.5629	4.068	2.968	4.069	11.542
fa	0.5731	4.033	2.938	4.033	11.431
clust	0.0803	2.983	5.514	2.896	12.313
org	0.4329	8.359	17.93	8.1927	35.055

In TABLE 1, the applied statistical categories are correlation (corr), total root mean square (trms), standard deviation (stdev), mean average distance (mad) and equally weighted index (ewi). The employed neurofuzzy systems are developed by the original data (org), the original data with applying FCM clustering analysis (clust), applying factor analysis (fa), applying factor analysis with FCM clustering analysis (fc), applying principal component analysis using correlation (pca), and applying principal component analysis with FCM clustering analysis (pc).

From TABLE 1, with applying the methodologies of dimensionality reduction and various multivariate analysis methods, the comparison between the applied methodologies using the original data and the reduced data of the physicochemical properties of protein tertiary structure with reduced dimension analyses is presented. In Fig. 5, the best performance among five reduced dimension techniques using the reduced dimension of the data set is the technique applying factor analysis with the FCM clustering analysis in order to predict the size of the residues for the protein structure through the root-mean-square-deviation (RMSD) values with applying the physicochemical properties of protein tertiary structure.

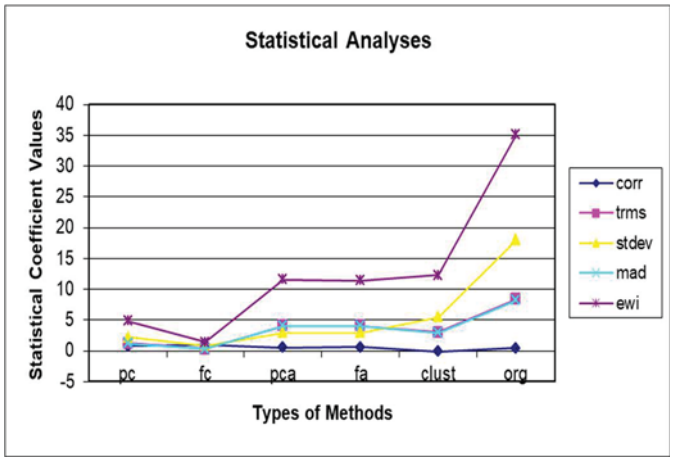

Fig. 5 Plot of types of methods vs. statistical coefficient values

The second-best performance to predict the size of the residues for the protein structures through the root-mean-square-deviation (RMSD) was the technique with applying FCM clustering analysis as a post operation with principle component analysis. In overall, among compared five reduced dimension analyses, the results predicted by applying the principle component analysis, factor analysis, and FCM clustering analysis showed the similar levels of the prediction of the size of the residues for the protein structures. The result predicted by applying the original data showed the relatively lower performance than other applied methods.

6 Conclusions

As a conclusion, the results of the size of the residue as one of the physicochemical properties of protein tertiary

structure through the root-mean-square-deviation (RMSD) with applying five reduced dimensions, are compared by the predicted values through the traditional multivariate analyses such as factor analysis and principle component analysis and the proposed techniques with applying the neurfuzzy systems implemented by the reduced data set as well as the original data set to estimate the predicted values. In overall, among five reduction analyses, the technique using factor analysis with the FCM clustering analysis as a post operation shows the best performance. Furthermore, the worst performance found the case with applying the original data. However, due to the data dependency, the results from the analyses using the reduced data or the original data, may cause the different tendancy of the performance even though the reduced cases show the better performance than the original data set case in this typical example. Therefore, for the future study, more various data may need to be explored to find out the more accurate prediction of the size of the residue for the protein structures through the root-mean-square-deviation (RMSD) values as one of the physicochemical properties of protein tertiary structure with the various dimensionality reduction methods.

ACKNOWLEGEMENT

Physicochemical properties of protein tertiary structure data set from UCI Machine Learning Repository are used and donated by Dua and Taniskidou [12].

7 References

[1] M. Iraji, and H. Ameri, "RMSD Protein Tertiary Structure Prediction with Soft Computing," International Journal of Mathematical Sciences and Computing, 2, pp. 24 – 33, 2016.

[2] S. Saha, "Protein Secondary Structure Prediction by Fuzzy Min Max Neural Network with Compensatory Neurons," Doctoral dissertation, Indian Institute of Technology, Kharagpur, India, 2008.

[3] P. Rana, H. Sharma, M. Bahattacharya, and A. Shukla, "Quality assessment of modeled protein structure using physicochemical properties," Journal of Bioinformatics and Computational Biology, November, pp. 1 – 14, 2014.

[4] S. Babaei, A. Geranmayeh, and S. Seyyedsalehi, "Protein secondary structure prediction using modular reciprocal bidirectional recurrent neural networks," Computer Methods and Programs in Biomedicine, Vol. 100, No. 3, pp. 237-247, 2010.

[5] Z. Zhou, B. Yang, and W. Hou, "Association classification algorithm based on structure sequence in protein secondary structure prediction," Expert Systems with Applications, Vol. 37, No. 9, pp. 6381-6389, 2010.

[6] W. Qu, H. Sui, B. Yang, and W. Qian, "Improving protein secondary structure prediction using a multi-modal BP method," Computers in Biology and Medicine, Vol. 41, No. 10, pp. 946-959, 2011.

[7] R. Hassan, R. Othman, P. Saad, and S. Kasim, "A compact hybrid feature vector for an accurate secondary structure prediction," Information Sciences, Vol. 181, No. 23, pp. 5267-5277, 2011.

[8] A. Wright, "The Current State and Future of Factor Analysis in Personality Disorder Research," Personality Disorders: Theory, Research, and Treatment, Vol. 8, No. 1, pp. 14 –25, 2017.

[9] H. Abdi, L. Williams2 and D. Valentin, "Multiple factor analysis: principal component analysis for multivariable and multiblock data sets," WIREs Computational Statistics, Vol. 5, pp. 149–179, 2013.

[10] M. Choudhry, and R. Procedia, "Performance Analysis of Fuzzy C-Means Clustering Methods for MRI Image Segmentation," The 12th International Multi-Conference on Information Processing-2016 (IMCIP-2016), Computer Science, Vol. 89, pp. 749 – 758, 2016. doi: 10.1016/j.procs.2016.06.052

[11] Henry F. Kaiser, "The varimax criterion for analytic rotation in factor analysis," Psychometrika, Vol. 23, Issue 3, September , pp. 187 – 200, 1958.

[12] UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, donated by Dua, D. and Karra Taniskidou, E. (2017), online at http://archive.ics.uci.edu/ml/datasets/Physicochemical+Properties+of+Protein+Tertiary+Structure

[13] E. Franzosa and Y. Xia, "Independent Effects of Protein Core Size and Expression on Residue-Level Structure-Evolution Relationships," Plos One, Volume 7, Issue 10, October 2012.

[14] N. Aslam, A. Nadeem, M. Babar, M. Pervez, M. Aslam, N. Nareed, T. Hussain, W. Shehzad, M. Wasim, Z. Bao, and M. Javed, "The accuracy of protein structure alignment servers," Electronic Journal of Biotechnology, Vol. 20, pp 9 – 13, 2016.

[15] Y. Zhang and J. Skolnick, "TM-align: a protein structure alignment algorithm based on the TM-score," Nucleic Acids Research, Vol. 33, No. 7, pp. 2302–2309, 2005.

[16] M. Betancourt and J. Skolnick, "Universal similarity measure for comparing protein structures," Biopolymers, Vol 59, pp 305 – 309, 2001.

[17] C. Cavasotto and R. Abagyan, "Protein Flexibility in Ligand Docking and Virtual Screening to Protein Kinases," Journal of Molecular Biology, Vol. 337, pp. 209–225, 2004.

[18] O. Carugo, "Statistical validation of the root mean square distance, a measure of protein structural proximity," Protein Engineering, Design and Selection, Vol 20, No. 1, pp. 33-38, 2007.

[19] O. Carugo and S. Ponger, "A normalized root-mean-square distance for comparing protein three-dimensional structures," Protein Science, Vol. 10 pp. 1470-1473, 2001.

[20] X. Zhou, J. Chou, and S. Wong, "Protein structure similarity from principle component correlation analysis," BMC Bioinformatics, Vol. 7, Issue 40, 2006.