A Two Stage Model on Prediction of Protein Stability Changes in Case of Uncertainty using Fuzzy K-Means Clustering and Fuzzy Artificial Neural Networks

Juliet Rozario, B.Radha

ABSTRACT--- In both industrial applications and basic research the manipulation of protein stability is essential for knowing the principles which govern protein thermostability. This leads to hotspot in data mining based protein engineering and stability prediction. There are so many works related to the prediction of protein stability but they all lack in data preprocessing, presence of duplicates in the dataset and ability to handle uncertainty present in them. The main aim of this paper is to enhance the quality of the protein stability dataset and to increase the accuracy rate of prediction system. For deduplication process fuzzy K-means (FKM)based clustering is applied to cluster and match the duplicate records and eradicate them. To handle the uncertainty Fuzzy Artificial Neural Network (FANN) is used to perform prediction on protein stability. Simulation results proved the efficiency of FKM-FANN which yields excellent results comparing the existing methods.

Keywords— Protein stability, fuzzy k means, fuzzy artificial neural network, prediction and deduplication.

INTRODUCTION

In current era, there is an exponential growth in the field of genomics and proteomics which generated a huge volume of biological data. Bioinformatics which is also referred as computational biology is the interdisciplinary art of understanding biological data using information processing system. Analyzing such voluminous biological data necessitates sufficient knowledge to infer the structure of these kinds of data. The research starts focusing on a significant field which efficiently handles large quantities of biological data, is data mining. With the advent of mining especially to analysis protein structure prediction, classification of gene, using microarray data for classification cancer, gene expression data clustering, protein-protein interaction based modeling are greatly improved. Thus it proves the significant increase in interaction among bioinformatics and data mining.

Mutant protein prediction based on its stability variation is one of the challenging and vital tasks among protein engineering. It is essential to discover about the mechanisms which are responsible for stability of protein along with developing sensitive protein mutants under temperature [1]. In order to learn about the specific importance of amino acid in the functionality of protein and the biological stability the protein mutants are frequently used [2]. Single amino acid mutation may suggestively change the stability of structure of protein. Therefore, protein designers and biologist requires perfect prediction of in what way single amino acid mutations will affect the stability of a protein structure [3].Conversely, instead of using linear combination of energy terms based statistical methods, machine learning approaches can study complex nonlinear functions of input protein sequence, mutation and information of protein structure. This greatly helps in discovering complex interactions which affect the stability of the proteins.

The input data may be often incomplete and unavailable in real dataset [4]. This work aims at developing an optimal method to produceconsistent prediction from limited input data and examining the protein stability deviations upon single mutation. This work hascreated a non-redundant dataset of protein mutants using fuzzy K-means clustering. Furthermore, for classification fuzzy artificial neural network is used.

RELATED WORK

There are many works done by the researches in the field of protein stability prediction depending in the mutation information. This subsection discusses about such existing models in protein stability prediction.

In the work [5], the authors aimed to introduce mining approaches to discover the hidden knowledge about the protein stability variation upon double point mutation. They built a unique dataset of mutant’s protein with different features to perform systematic investigation on the dataset. They used decision tree and table for comparison.

In [6] the authors presented a thermodynamic stability of protein using neural networks which involves in discovering whether the provided mutant will increase or decrease the stability of protein.

The work in [7], determines the change in stability of protein using double mutants. They used rule induction model to develop a knowledge pattern using the generated rule. To identify the significant rule pattern for a given input, the authors used a fuzzy query-based model.

Bordner and Abagyan in their work [8] used sequence

Revised Manuscript Received on July 10, 2019.

Juliet Rozario, Assistant Professor in PG Department of Research and Computer Science Nehru Arts and Science College Coimbatore. Tamil Nadu, India.

B.Radha, Assistant Professor in Department of Information Technology Krishna Arts and Science College Coimbatore. Tamil Nadu, India.

information to forecast the energy change using empirical study. But structural information produces lessaccuracy comparing this approach.

Frenz [9]in their work discovered dissimilarity scores using sequence information about position of nutation. The protein stability in Staphylococcal is predicted in this process with nuclease at 20 different residue positions.

Casadio et al [10] developed radial bases function to forecast variation of energy due to mutation. They stated that local interactions among amino acid formation is considered as an important factor for energy prediction.

In paper [11] described about a meta searching technique which predicts the protein stability based on mutation of amino acids. It uses 8 different existing classification tools to perform a consistent desision making. It also proved the overall improvement in prediction process.

In [12] authorsproposed a new gaussian method which predicts protein stability variation with single and double mutations. This approach is simulated with high volume of molecular data. In this work Bayesian network which produces prediction process with combined data.

In [13] a mutagenesis with computational intelligence is developed based on statistical and knowledge-oriented systems. If a single amino acid is replaced during a mutation in protein, the approach generates an optimal regularized situation during every remainder position. The protein structure and its functionalities can be well defined using accurate model for prediction.

Partiality in forecasting free energy changes depending on point mutation is focused in the work [14]. They performed dataset construction with equal proportion of stabilizing as well as destabilizing mutation with the corresponding structure of wild protein and mutant protein. They used fifteen different measures of ΔΔG predictors are used.

PROBLEM DEFINITION & RESULTS

- In S1615 dataset there is the existence of the redundant records
- Data preprocessing on the existing data set is lacked
- Presence of redundant data may disable the ability of precise development of perdition of mutant stability
- Accurate classification of the mutant stability in case of uncertainty is highly leads to false alarm

Methodology for predicting Protein Stability using Fuzzy K-means based Deduplication and fuzzy neural network based Protein stability prediction

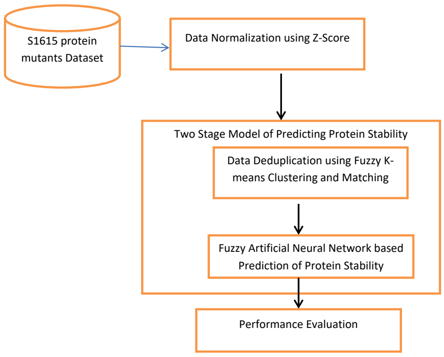
S1615 protein mutants Dataset
Data Normalization using Z-Score
Two Stage Model of Predicting Protein Stability
Data Deduplication using Fuzzy K-means Clustering and Matching
Fuzzy Artificial Neural Network based Prediction of Protein Stability
Performance Evaluation

Figure 1:Framework of the proposed Method

The proposed method handles two factors in case of the predicting the mutation stability they are applying DeDuplicaiton to eliminate the redundant records and developing the prediction model to determine the mutation stability.It performs data cleaning process by detecting and removing duplicate entries in the dataset. The data deduplication process is applied to clean the dataset for identifying and removing the duplicated records. Presence of duplicate datasets affects the memory space of the data warehouse and increases the complexity in extraction of hidden knowledge using mining approaches. In this work deduplication is done chiefly discover and eliminate the duplicate records in database. Thus, it improves the quality and the performance of the database. The proposed deduplication model determines both partial and fully duplicated instances in the protein database. The steps involved in deduplication are transformation, clustering of similar instance's and determining matching among them. Fuzzy artificial neural network is used for determining the stability change in advance. Figure 1 illustrates the overall framework of the proposed methodology.

Dataset Description of S1615 Dataset

- This research work used the dataset of Capriotti et al. [15] whose name is S1615 dataset
- S1615 is collected from ProTherm [1] repository of mutant and proteins.
- S1615 dataset comprised of 1615 single site mutations received from 42 various proteins.
- In this dataset, each mutation is composed of six attributes they are PDB code, mutation, solvent accessibility, pH value, temperature, and energy change (ΔΔG).
- If the value of ΔΔG energy is changed positive then there is a increase in the stability by the protein and that instance is classified as positive
- If the value of ΔΔG energy is negative, then mutation is destabilizing and is classified as a negative instance.

Datacleaning

This process involves in discovering and cleaning the inappropriate records from a set of instances or S1615 database. Cleaning process involves in modification, deletion or replacing the values of data with relevant information of S1615 database. In this paper, the cleaning process removes the duplicate records from the database, so that memory space can be utilized in an optimal way. Duplication occurs when storing a specific record more than one time which cannot contribute more in protein stability prediction. Because the redundancy of records may lead to incorrect results which results in wrong conclusion and thus the completed process will suffer under poor detection rate. Thus, this work concentrates on removal of duplicates in S1615 dataset. To perform deduplication this proposed

work uses fuzzy K-means based clustering to discover the same instances and matching is performed for determining and eliminating such duplicate records in the database under consideration.

Data Preprocessing

The input dataset has to be converted to the suitable format for prediction process. So this work performs normalization on the dataset referred as scaling method in preprocessing phase [1]. In this the original value of each record is converted to a specific range of values i.e 0 to 1. It can be cooperative for forecasting high presistence [2]. In this work, z-score normalization is applied for data normalization.

$$v_i' = \frac{v_i - \bar{E}}{std(E)}$$

Where,

v_i' refers to Z-score normalized values, v_i is value of the record E of ith attrintue

std (E) = $\sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(V_i - \bar{E})^2}$

Mean value is computed as $\bar{E} = \frac{1}{n}\sum_{i=1}^{n} V_i$

Deduplication using Fuzzy K-means Clustering

Fuzzy k-means (FKM) [21] is a kind of clustering approach which permits a single instance of database to belong more than one clusters. This method is widely used in recognition of patterns. This work relies on minimization of the objective function of each instance which is denotes as follows:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 \quad , \quad 1 \le m < \infty$$

where *m* holds the value greater than 1, , u_{ij} is the membership degree of x_i in the *jth cluster*, x_i is the ith d-feature dataset, c_j is the cluster center and ||*|| is the similarity measure between the instance and its corresponding cluster center. Partitioning through fuzzy is done in an iterative process and the objective function with the update cluster center c_j and membership degree u_{ij} is as below:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\|x_i - c_j\right\|}{\left\|x_i - c_k\right\|}\right)^{\frac{2}{m-1}}} \quad , \quad c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

The iteration will be terminated when it reaches

$$\max_{ij}\left\{\left|u_{ij}^{(k+1)} - u_{ij}^{(k)}\right|\right\} < \varepsilon$$

where ∈ refers t criteria for termination whose value lies between 0 and 1, the k refers to iteration steps. This procedure convergance to local optima of J_m.

The procedurefor fuzzy K-means based Deduplication

1. Set U=[u_{ij}] matrix, U$^{(0)}$
2. On kthstep: compute the center vectors C$^{(k)}$=[c_j] with U$^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3. Update the value of U$^{(k)}$, U$^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\|x_i - c_j\right\|}{\left\|x_i - c_k\right\|}\right)^{\frac{2}{m-1}}}$$

4. If || U$^{(k+1)}$ - U$^{(k)}$||< ε then HALT; else return to step 2

After converting input data into normalized form using Z-score normalization, then fuzzy K-means is applied. The similar instances are stored in same clusters and dissimilar instances forms new clusters. Likewise, records which have similar matching score are stored in same cluster. Thus, the clustering process, decreases the amount of comparisons and improves the system performance.

Record Matching

After transforming the datasets to membership values, next divide and conquer method is applied on each row of the specific cluster. This method partitions the values iteratively into small portions and endures its process until it reaches certain size. Finally, it compares the single attribute value of each record with other records. If matching is found on their entire attribute values then the percentage of duplication is computed. When the duplication percentage crosses the predefined value of threshold, the method shows those records have different value and that can be considered as a unique record which involves in prediction process. If the difference is less then it is considered as duplicate record and it is eliminated from the database as a process of deduplication.

Fuzzy Artificial Neural Network based Protein Stability Prediction

Based on the neural network devised by described by Ishibushi et. al.[19] , the fuzzy artificial neural network was designed. This type of network learns from the if-then based fuzzy rules. The network can be configured using α-cuts of the membership values used for defining low, medium and high. Precisely, α-cuts of the fuzzy numbers are signified by interlude vectors [20]. That is, an α-cut of fuzzy input vector is signified by the interval vector $\mathbf{X}_p$ = (X_{p1}, X_{p2}, ..., X_{pn})^T where $X_{pi} = [x_{pi}^{L}, x_{pi}^{U}]$ (1)

Which indicate the intervals lower and upper limits of each attribute in protein mutation dataset. To handle the interval values of both input and output vectors the neuronal functionalities are altered. The weights involved in data transfer with inner layers are summarized and calculated as follows:

$$\mathrm{Net}_{pj}^{L} = \sum_{\substack{i \\ w_{ji} \geq 0}} w_{ji} o_{pi}^{L} + \sum_{\substack{i \\ w_{ji} < 0}} w_{ji} o_{pi}^{U} + \theta_j \qquad (2)$$

and

$$\mathrm{Net}_{pj}^{U} = \sum_{\substack{i \\ w_{ji} \geq 0}} w_{ji} o_{pi}^{U} + \sum_{\substack{i \\ w_{ji} < 0}} w_{ji} o_{pi}^{L} + \theta_j \ . \ (3)$$

As described in lefeld [3] the computation is consistent with interval-based multiplication process. The output of the network is represented as follows:

$$\begin{aligned} o_{pj} &= [o_{pj}^{L}, o_{pj}^{U}] \\ &= [f(\mathrm{Net}_{pj}^{L}),\ f(\mathrm{Net}_{pj}^{U})] \end{aligned} . \qquad (4)$$

Derivation of the Learning Rule

The fuzzy neural network learning rule is also modified, which is briefly described as backpropagation feed forward generalized delta rule which updates any weights involved in prediction or classification process as follows:

$$\Delta w_{ji}(t+1) = \eta\left(-\frac{\partial E_p}{\partial w_{ji}}\right) + \alpha \Delta w_{ji}(t) . \qquad (5)$$

The error is calculated as the varianceamong the target output, $\mathbf{t}_p$, and the actual output, $\mathbf{o}_p$:

$$E_p = \max\left\{\frac{1}{2}(t_{pj} - o_{pj})^2, o_{pj} \in \mathbf{o}_p\right\}, \qquad (6)$$

where

$$(t_{pj} - o_{pj}) = \begin{cases} (t_{pj} - o_{pj}^{L}), \text{ if } t_p = 1, \text{ and} \\ (t_{pj} - o_{pj}^{U}), \text{ if } t_p = 0. \end{cases} \qquad (7)$$

The output layer units, calculates $\partial E_p / \partial w_{ji}$, which is straightforward, and can be believed of as constructed on the value of target output and weight.

This Fuzzy-ANN forecasts the class based on the sign of ΔΔG recognizes the path of the stability change.

- ➢ The sign is more revealing than the |ΔΔG|
  - o If ΔΔG < 0 then the mutation increases the stability of protein
  - o If ΔΔG > 0 then the mutation decreases the stability of protein

Experimental Result

This section discusses about the simulation result using the proposed model FCM-FANN for protein stability prediction on the protein sequence dataset. The findings from the simulation are depicted in several tables and plots. The performance of the proposed FCM-FANN is compared with that of an ANN, RBF and Naïve Bayes by evaluating numerical computations. All simulations were conducted in Matlab R2010b environment running on a PC with a 2.5 GHz Core™ i5 CPU and 6 GB RAM.

Performance metric

$$\text{Mean Absolute Error} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \bar{y}_i\right|$$

$$\text{Root Mean Square Error} = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

Table 1: Performance comparison Before and After Deduplication using k-means, DBSCAN and Fuzzy K-means

S1615 Dataset	**Total no of Records = 1505**	
Clustering Techniques	**After Deduplication**	**Number of Duplicate Records detected**
K-means	1472	33
DBSCAN	1424	81
Fuzzy K-means	1350	155

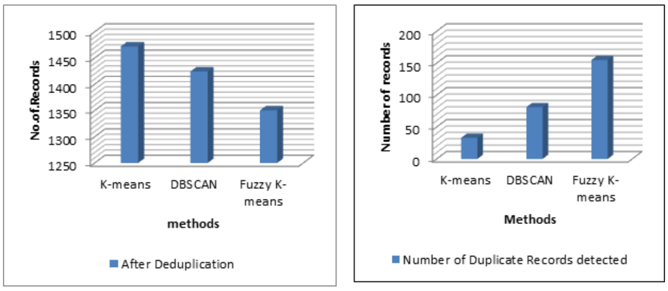

Figure 2 a) After Deduplication b) Detection of duplicate records

From the table 1 and the figure 2 a) and b) depicts the simulation outcome of deduplication process in protein stability prediction dataset. From the result it is observed that fuzzy K-means reduces the duplication among records by applying degree of membership towards the deduplicated records and they are eliminated by keeping the single copy of it. Thus the number of records for furthering processing is also considerably reduced with unique instances of protein. Whereas the other two existing methods k-means and DBSCAN fails to achieve better results because they don't have ability to handle the uncertain or vague dataset.

Table 2: Performance Comparison of Three different prediction methods

Measures	ANN-BP	ANN-RBF	Fuzzy ANN
Correctly classified	82.5249 %	74.8173%	92.3588%
Incorrectly classified	17.4751 %	25.1827%	7.6412%
Mean absolute error	0.1811	0.3081	0.0032
Root mean squared error	0.3876	.4082	0.0041

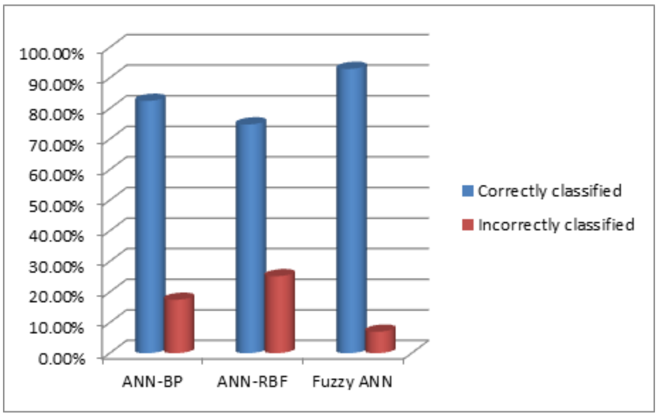

Figure 3 Performance Comparison based on correctly and incorrectly classified instances of predict the protein stability

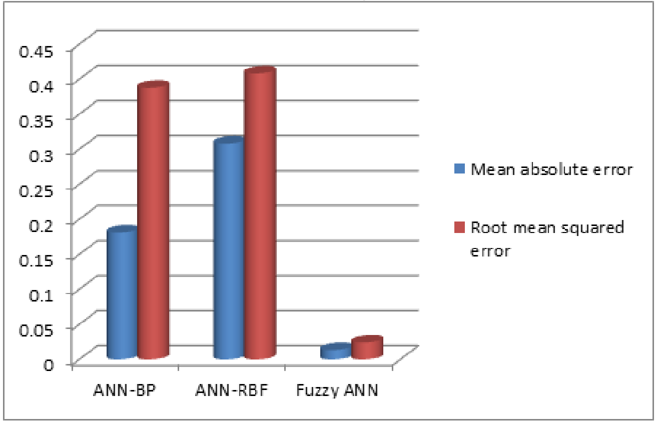

Figure 4 Performance Comparison based on Mean absolute Error and Root mean Squared Error to predict the protein stability

The table 2 and the figures 3 and 4 show the performance comparison of three different variants of artificial neural network. The simulation result shows that the performance of the fuzzy artificial neural network (FANN) performs better than the other two approaches because of its efficiency to represent the instances in terms of degree of membership and enhances the efficiency of standard neural network prediction ability.

Table 3 : Confusion Matrix for Artificial Neural Network-BP

a=IS	b= DS	
962	136	a = IS
127	280	b = DS

The table 3displays the confusion matrix of the artificial neural network which shows the correctly classified instances as increased stability and decreased stability with the factor ΔΔG. The total correctly predicted instances are 1098

Table 4 : Confusion Matrix for ANN-RBF

a=IS	b= DS	
869	229	a = IS
150	257	b = DS

The table 4 shows the confusion matrix of the artificial neural network using radial basis function which shows the correctly classified instances as increased stability and decreased stability with the factor ΔΔG. The total correctly predicted instances are 1126

Table 5: Confusion Matrix for FUZZY-ANN

a=IS	b= DS	
995	229	a = IS
150	395	b = DS

The table shows the confusion matrix of the Fuzzy ANN which shows the correctly classified instances as increased stability and decreased stability with the factor ΔΔG. The total correctly predicted instances are 1390. While comparing the other two methods the proposed fuzzy ANN with the deduplication using fuzzy k-means performs achieves better result

Findings

- Predicting the stability of protein is an important contribution to the filed of molecular biology and biochemistry
- In protein engineering forecasting the change of stability in protein mutants is an toughest and important task.
- It is essential to understand the mechanisms which are involved in stability of proteins and designing temperature subtle protein mutant.
- This research work intendedto designefficientlyapproach for examining the protein stability variations upon single mutation using data DeDuplicaiton done by fuzzy k-means clustering and prediction model developed using fuzzy Artificial neural network

CONCLUSION

This papercontributed a fuzzy based artificial neural computing approach for predicting change in protein mutantsstability. It also eliminates the problem of redundancy by developing data deduplication process using fuzzy k-means model. The uncertainty of deciding the mark of the stability variation is done using the knowledge of fuzzy based artificial neural network. The explanations on the simulation results confirmed that the present methodmightassist as an operative tool in the field of biomedical informatics for understanding the prediction process of protein stability variation upon single mutation. The simulation results shows that the fuzzy k-means with fuzzy artificial neural network produces optimal result in both deduplication process and protein stability prediction while comparing the existing methods K-means, DBSCAN, ANN-BP and RBF. The ability to handle vagueness in dataset is the added advantage of the fuzzy k-means and fuzzy ANN.

REFERENCES

1. M. M. Gromiha, Prediction of protein stability upon point mutations, Biochem Soc Trans, **35**, 1569-1573 (2007)
2. R. F. Boyer, Concepts in biochemistry, 3rd ed., Wiley, Hoboken, NJ, (2006
3. Y. M. Zheng, Z. X. An, X. E. Zhao, F. S. Quan, H. Y. Zhao, Y. R. Zhang, J. Liu, X. Y. He, X. N. He, Comparison of enhanced green fluorescent protein gene transfected and wild type porcine neural stem cells, Res Vet Sci, 88, 88-93 (2010).
4. L.-F. Lai, C.-C. Wu, L.-T. Huang, Predicting Protein Stability Change upon Double Mutation from Partial Sequence Information Using Data Mining Approach, Advanced Intelligent Computing Theories and Applications, in: D.-S. Huang, Z. Zhao, V. Bevilacqua, J. Figueroa (Eds.), Springer Berlin / Heidelberg, 664-671 (2010).
5. Liang-Tsung Huang, Chao-Chin Wu, Lien-Fu Lai, M. Michael Gromiha, Chang-Sheng Wang and Yet-Ran Chen, Data Mining Application in Biomedical Informatics for Probing into Protein Stability upon Double Mutation, Applied Mathematics & Information Sciences, Vol 8, No. 1L, 125-132 (2014)
6. Emidio Capriotti, Piero Fariselli∗ and Rita Casadio , neural-network-based method for predicting protein stability changes upon single point mutations Bioinformatics 20(Suppl. 1) Oxford University Press 2004Vol. 20 Suppl. 1 2004, pages i63–i68
7. Lai LF., Wu CC., Huang LT. (2010) Predicting Protein Stability Change upon Double Mutation from Partial Sequence Information Using Data Mining Approach. In: Huang DS., Zhao Z., Bevilacqua V., Figueroa J.C. (eds) Advanced Intelligent Computing Theories and Applications. ICIC 2010. Lecture Notes in Computer Science, vol 6215. Springer, Berlin, Heidelberg
8. Bordner AJ, Abagyan RA. Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. Proteins 2004;57:400–413
9. Frenz C. Neural network-based prediction of mutation-induced protein stability changes in staphylococcal nuclease at 20 residue positions. Proteins 2005;59:147–151
10. Casadio R, Compiani M, Fariselli P, Viarelli F. Predicting free energy contributions to the conformational stability of folded proteins from the residue sequence with radial basis function networks. In Proc Int Conf Intell Syst Mol Biol, volume 3, p 81–88. 1995.
11. František Malinka, Prediction of protein stability changes upon one-point mutations using machine learning, Prague, Czech Republic — October 09 - 12, 2015
12. Emmi Jokinen Markus Heinonen Harri Lähdesmäki, mGPfusion: predicting protein stability changes with Gaussian process kernel learning and data fusion, Bioinformatics, 34, 2018, i274–i283
13. Majid Masso Iosif I. Vaisma, Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis Bioinformatics, Volume 24, Issue 18, 15 September 2008, Pages 2002–2009,
14. F. Pucci , K. Bernaerts , J. M. Kwasigroch and M. Rooman, Quantification of biases in predictions of protein stability changes upon mutations, April 2018, bioRxiv
15. Capriotti E, Fariselli P, Casadio R. A neural network-based method for predicting protein stability changes upon single point mutations. In: Proceedings of the 2004 conference on intelligent systems for molecular biology (ISMB04), Bioinformatics (Suppl. 1), volume 20. New York: Oxford University Press; 2004. p 190 –201
16. L Folkman, B Stantic, A Sattar, Towards sequence-based prediction of mutation-induced stability changes in unseen non-homologous proteins, BMC genomics, 2014
17. Shalabi, L.A., Z. Shaaban and B. Kasasbeh, Data Mining: A Preprocessing Engine, J. Comput. Sci., 2: 735-739, 2006
18. S.Gopal Krishna Patro, Pragyan Parimita Sahoo, Ipsita Panda, Kishore Kumar Sahu, "Technical Analysis on Financial Forecasting", International Journal of Computer Sciences and Engineering, Volume-03, Issue-01, Page No (1-6), E-ISSN: 2347- 2693, Jan -2015
19. Ishibuchi, H., R. Fujioka, and H. Tanaka, “Neural Networks That Learn from Fuzzy If-Then Rules,” Vol.1, No.2, pp. 85-97, 1993
20. Bojadziev, G. and M. Bojadziev, Fuzzy Sets,Fuzzy Logic, Applications, World Scientific, New Jersey, 1995
21. Bezdek, JC 1981, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York., doi: 10.1007/978-1-4757-0450-1