EasyChair Preprint

№ 4083

Prediction of Protein – Peptide Binding Residues Using Classification Algorithms

Shima Shafiee, Abdolhossein Fathi and Fardin Abdali Mohammadi

August 25, 2020

Prediction of protein – peptide binding residues using classification algorithms

1st Shima Shafiee (Corresponding Author)
Department of Computer Engineering and Information Technology
RAZI UNIVERSITY
Kermanshah, Iran
shafiee.shima@razi.ac.ir
2nd Abdolhossein Fathi
Department of Computer Engineering and Information Technology
RAZI UNIVERSITY
Kermanshah, Iran
a.fathi@razi.ac.ir
3rd Fardin Abdali Mohammadi
Department of Computer Engineering and Information Technology
RAZI UNIVERSITY
Kermanshah, Iran
fardin.abdali@razi.ac.ir

Abstract—Peptide-binding proteins prediction is important in understanding biological interaction, protein performance analysis, cellular processes, drug design, and even cancer prediction, so using experimental predictive methods, despite their operational capabilities, has limitations such as being costly and need to spend more time, differences between unrecognized protein structures and sequences, so design and development of computational systems for maintenance, optimal models for representing biological knowledge, management and the analysis of big biological data is so important that the authors used machine learning-based techniques such as Support Vector Machine (SVM),Random Forest (RF),Decision Tree (C4.5), Decision Tree (ID3),Gradient Boosting classifiers, which evaluated experimental results to optimize Support Vector Machine(SVM) classifier (Radial Basis Function kernel) with significant evaluation parameters such as accuracy(ACC) is equal to 0.7401 and 0.7599 for 10 - fold cross validation and independent test set and also specificity (SPE) is equal to 0.7966 and 0.8088 for 10- fold cross validation and independent test set (respectively) by using various Structure- Based and Sequence -Based features.

Keywords—Protein- Peptide, Classification Algorithms, Binding Residue, Machine Learning.

I. Introduction

Proteins are polymers of amino acids that each residue binds to its adjacent amino acid through covalent bonding, so the focus of this study is functional analysis to predict protein-peptide binding residues. Because proteins are key players in the vital functions of organisms such as biochemical reactions, food transfer, detection, transmission of messages, and basic biological processes such as cellular communication, cell division, metabolism, etc. [1, 2] and due to the dynamics of proteins evidence of their interaction with other molecules, such as ligand, has been linked to specific biochemical targets, meaning that the ligand is the same amini acid containing peptide with a specific sequence and based on peptide bond [3], so predicting protein-peptide binding residue by experimental methods has limitations such as being costly and need to spend more time, the inherent difficulty of experimental methods [4], the inaccessibility of all protein structures, the possible mismatch of the structure recognized with the reference sequence [5], small peptide sizes, weak binding affinity limits and peptide flexibility [6], the prediction limit of all protein complex structures and protein-protein interactions that are interacting with molecules [7]. Thus, several predictions have been done to predict binding residues in various interactions, such as deep learning architecture [8] to predict protein-peptide binding residues, three heterogeneous support vector machine combination architecture [9] to predict protein-vitamin binding residues , technique based on Ramachandran map and dihedral angle preferences [10] for limitation of binding site amino acid residues modeling in RNA, DNA 3D structures, predicting protein-peptide complex structures and protein -peptide binding residue using machine learning methods such as Decision Tree, Logistic Regression, Bagging and Gradient Boosting classifier [11], specific ligand prediction and protein ligand specific binding residue by using three types of sequence-based architecture, improved Adboost and a combination of Template- Free and Template- Base [12], Sequence-based prediction with a combination of several Random Forest (RF) classifiers [13], To predict ligand-binding residues, peptide binding residues sequence-based prediction by combining the Support Vector Machine(SVM) algorithm, Strengthen Gradients(SG) and K-nearest neighbor(KNN) classifier, with logistic regression and stack-based gener-

alization technique [14], protein-peptide binding residues prediction using SVM-Pep for sequence-based inputs and Pep -Bind for structure-based inputs [15], Using improved KNN classifier [16] to predict acid radical ion binding residues and Sulfate Ion Binding Residue (SIBR) by the support vector machine algorithm [17], Key residues prediction (The result of sharing binding residues and stabilizing resides) through the analysis of complex flexibility, protein performance, binding affinity [18], sequence-based prediction for ligand-binding residue through the combination of support vector machine and homology-based transfer [19] improved Deep Learning (DL) [20] by using hidden markov and stacked autoencoder models to extract features of Fisher Score(FS) and Hidden Abstract Inrelation(HAI) to the support vector machine in predicting residue-residue contact matrix in protein-protein interaction and finally using gaussian processes, support vector machine, random forest and deep neural networks [21] are proposed for protein-ligand interaction prediction.

II. Materials and Methods

A. Methods

In this study, machine learning was focused on applying the law of learning and achieving the optimal model. Inspired by this, each pattern in supervised learning has a label, so the goal is to provide the mapping function of the input patterns to their corresponding labels in the training phase. The designed system also predicts their output or label with the help of the learned function. If the output of the learning system is discrete, then the classifier problem and the function that map the input to the output is called classifier [22]. Therefore, the classifiers considered for protein-peptide binding residues are Gradient Boosting, Random Forest(RF), Decision Tree (C4.5), Decision Tree (ID3), Support Vector Machine(SVM). The following is a proposed flowchart (Figure 1).

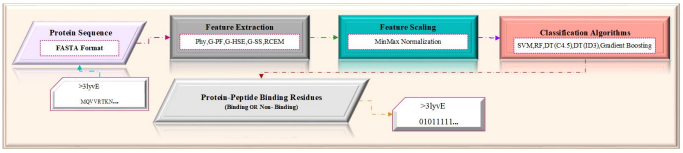


Fig. 1: The proposed flowchart based on machine learning classifier.

B. Steps of Procedure

As can be deduced from Figure 1, the proposed method is done in three steps, which are:

i. Preprocess: includes feature extraction and normalization (respectively), which in feature extraction phase, five categories of various features based on sequence and structure are used, which are Residue-wise Contact Energy Matrix(RCEM), Half Sphere Exposure Group(G-HSE), Secondary Structure Group(G-SS), Sequence Profile Group from PSSM(G-PF), Physicochemical Properties(Phy: steric parameter, hydrophobicity, isoelectric point, aliphatic, polarity, acidity) [6, 11, 23, 28].Therefore, equation (1) [24, 25] is used to normalize each feature value to the interval [0, 1], In Eq. (1), x and $x^{'}$ denote the original and normalized values of the feature and a are the start and end of the proposed domain (respectively).

ii. Process: Includes classifier operations by five proposed classifiers such as Support Vector Machine(SVM), Random Forest(RF), Decision Tree (C4.5), Decision Tree (ID3), Gradient Boosting, using 10 - fold cross validation to predict protein - peptide binding residues were trained. Use the sliding window size technique to improve performance, balance the interaction between protein-peptide, and increase optimization based on neighboring residues information.

iii. Postprocess: Evaluation of the performance of the mentioned classifier algorithms according to criteria such as sensitivity(SEN), accuracy (ACC), specificity (SPE), matthews' correlation coefficient (MCC), F -measure and using the independent test set, which finally support vector machine classifier (RBF kernel), with window size of three were optimized to predict protein-peptide binding residues.

III. Result and Discussion

A. Data Set

Protein sequences, as the input of the proposed classifiers, have Protein Data Bank(PDB) ID, a well-known three-dimensional structure, FastA format in Protein Data Bank (PDB). Therefore, the primary data sets of protein-peptide complexes are derived from the BioLip database, which is a summary of the final dataset (Table I). These datasets are also available online [6].

TABLE I: Summary of the final applied dataset

Name	Number	$\mathbf{N_{BR}}^1$	$\mathbf{N_R}^2$	$\mathbf{N_{NBR}}^3$
Protein–Peptide Complexes	1241	16678	297598	—
TR(training set)	1116	14959	—	251769
TS(independent test set)	125	1719	—	29151

B. Performance Evaluation

Confusion matrix [26] is the basis for evaluation criteria and includes information such as actual classifier and prediction, which for binary prediction classifier of protein - peptide binding residues problem is (Table I).

The evaluation criteria such as [27] sensitivity(SEN), accuracy (ACC), specificity (SPE), matthews' correlation coefficient (MCC), F-measure based on the confusion matrix according to the 2 to 6 equality were used for the performance of the proposed classifiers.

[1]Number of Binding Residues
[2]Number of Residues
[3]Number of Non-Binding Residues

TABLE II: Summary of the final applied dataset

		Actual	Values
		Actual Positive(1)	Actual Negative(0)
Output Classifier	Classify Positive(1)	TP (True Positive)	FP (False Positive)
	Classify Negative(0)	FN (False Negative)	TN (True Negative)

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \quad (1)$$

$$F - measure = \frac{2TP}{(2TP + FP + FN)} \quad (2)$$

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (3)$$

$$Specificity = \frac{TN}{(FP+TN)} \quad (4)$$

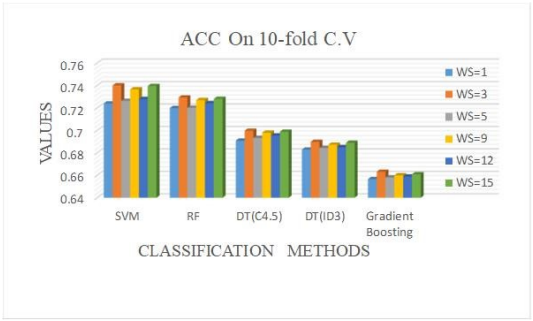
$$Sensitivity = \frac{TP}{(TP+FN)} \quad (5)$$

When we need to visualize the performance of the binary classification problem, we use the ROC (Receiver Operating Characteristics) curve [29]. A receiver operating characteristic (ROC) curve is a graphical plot that describes the diagnostic sufficiency of a binary classifier system as its discrimination threshold is varied.

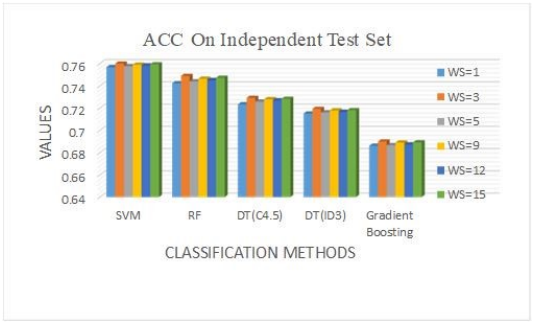
C. Ligand-binding residue prediction using classifier algorithm

The authors, five classifiers such as Support Vector Machine(SVM), Random Forest(RF), Decision Tree (C4.5), Decision Tree (ID3), Gradient Boosting(GB) to predict protein-peptide binding residue and based on five categories of structure and sequence-based features(Residue-wise Contact Energy Matrix(RCEM), Half Sphere Exposure Group(G-HSE), Secondary Structure Group(G-SS), Sequence Profile Group from PSSM(G-PF), Physicochemical Properties(Phy: steric parameter, hydrophobicity, isoelectric point ,aliphatic, polarity, acidity)) were used in such a way that they used feature windowing technique for improving classifier performance and balancing interactions between protein and peptide with window sizes 1, 3, 5, 9, 12 and 15 for independent test set and 10- fold cross validation used the evaluation results to confirm the optimization of support vector machine classifier (RBF kernel) with the window size is three Therefore, the feature windowing technique was used for each of the other functional classifiers with the mentioned window sizes (Figures 2 to 6).

Then, the receiver operating characteristic ROC curve [29] was used to measure the performance of each of

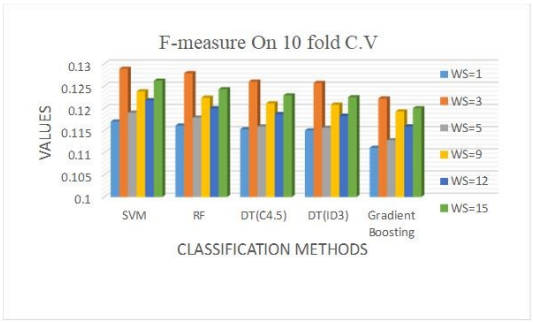


(a)

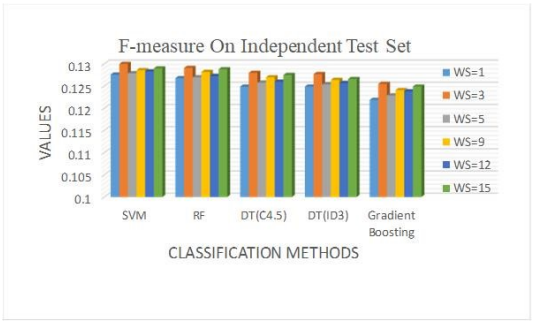


(b)

Fig. 2: (a)ACC on 10- fold cross validation (b)ACC on independent test set.

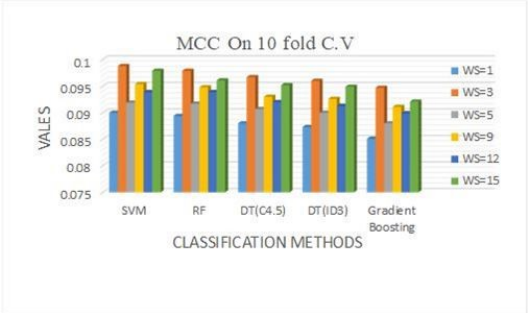


(a)

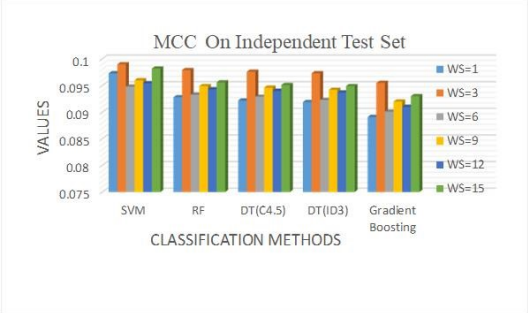


(b)

Fig. 3: (a)F- measure on 10 fold cross-validation (b)F-measure on independent test set

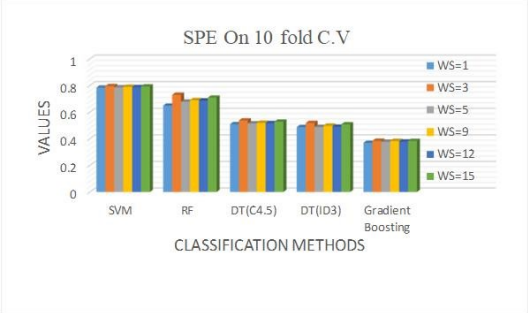


(a)

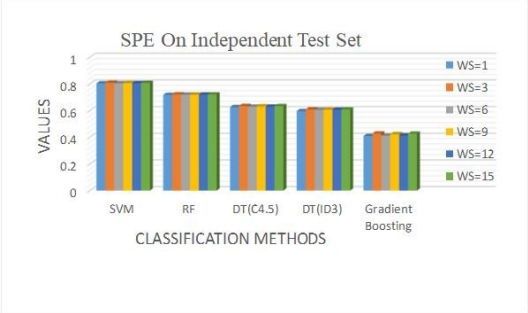


(b)

Fig. 4: (a)MCC on 10 fold cross-validation (b)MCC on independent test set

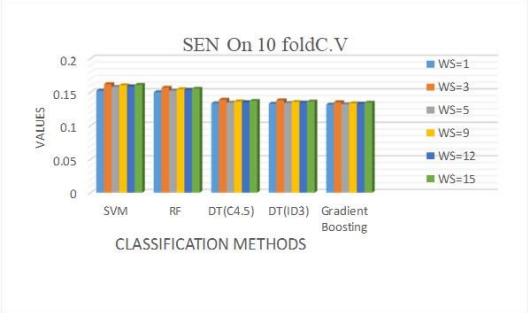


(a)

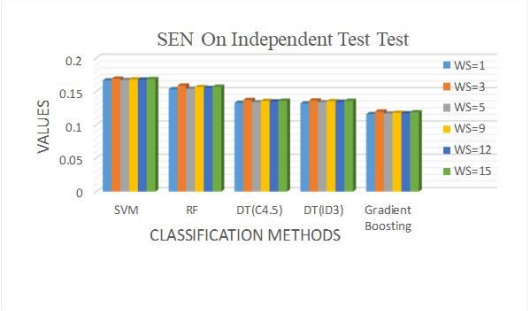


(b)

Fig. 6: (a)SPE on 10- Fold Cross-Validation (b)SPE on Independent Test Set



(a)



(b)

Fig. 5: (a)SEN on 10 fold cross-validation (b)SEN on independent test set

the classifiers, which is a kind of trade-off between true positive rate(TPR) and false positive rate(FPR) that according to the evaluation results, the support vector machine classifier (RBF kernel) has the optimal result by using the independent test set (Figure 7).

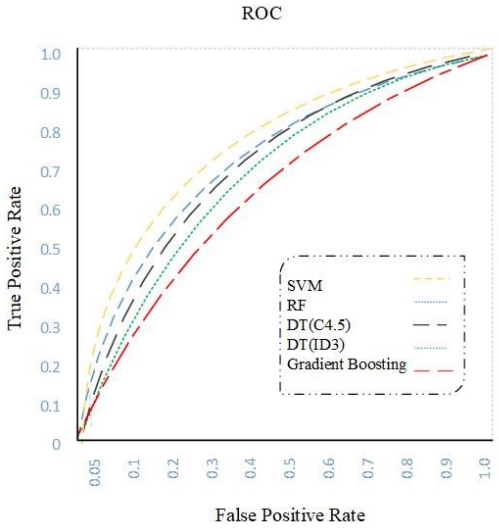


Fig. 7: ROC curves are given of machine learning classifiers using the independent test set

In the next step, the area under the curve (AUC) [30] was used to evaluate the performance of optimal support

vector machine classifier (RBF kernel) for each feature group individually that the results show the highest area under the curve (AUC) by RCEM in five feature categories (Figure 8).

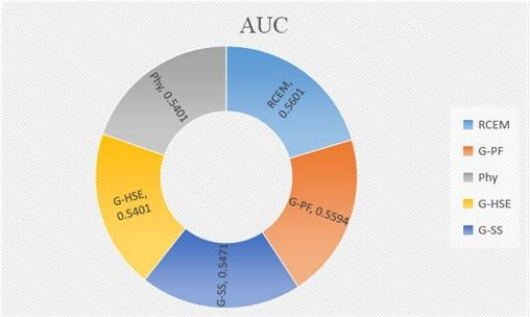


Fig. 8: Evaluation of the performance of optimal support vector machine classifier by using each individual feature group

In the last step, it evaluates the prediction for a protein sequence with PDBID equal to 3lyvE. The protein sequence with the mentioned ID has 50 residues in the main structural model, as long as the Support Vector Machine(SVM) classifier (RBF kernel) correctly predicts thirty-eight residues. (Figure 9c).

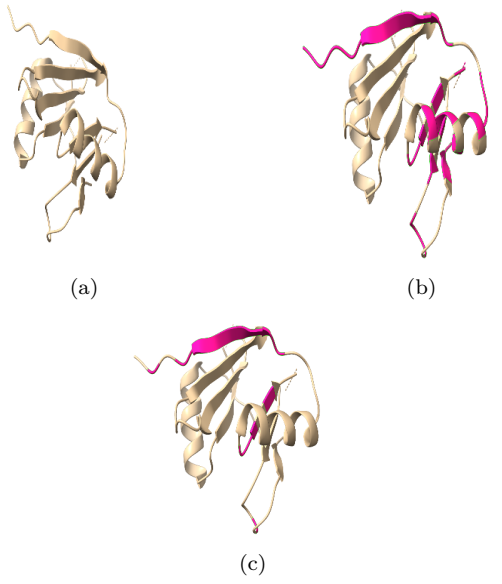


Fig. 9: (a)Protein Sequence[6] (b)Actual Binding Residues[6] (c)Predicted Binding Residues based on SVM classifier

IV. Conclusion and Future work

Since it is necessary to predict protein binding residue to recognize protein function, molecules involved in protein, biological interactions, so the focus of this study is to predict protein-peptide binding residue using classification algorithms called Support Vector Machine(SVM), Random Forest(RF), Decision Tree (C4.5), Decision Tree (ID3), Gradient Boosting, and based on five categories of features based on structure and sequence features (Residue-wise Contact Energy Matrix(RCEM), Half Sphere Exposure Group(G-HSE), Secondary Structure Group (G-SS), Sequence Profile Group from PSSM(G-PF), Physicochemical), so the authors used the technique of sliding window size to improve performance, balance interaction between protein and peptide and increased optimization based on neighboring residues information. Finally, the evaluation of the experimental results indicates the optimization of the support vector machine classifier (Radial Basis Function kernel) with accuracy (ACC), specificity (SPE) significant, and with three window sizes. Future works include the use of deep learning-based architecture with several additional features to improve performance and predict the interaction of other ligands, such as carbohydrates with the protein macromolecule.

References

[1] J. Xiong, “Essential Bioinformatics”, 1 edition, Cambridge University Press,2006,360 pages.

[2] A. Kessel, N. Ben-Tal, “Introduction to Proteins: Structure, Function, and Motion”, 2nd Edition, Chapman and Hall/CRC Mathematical and Computational Biology Series,2020,932 Pages.

[3] A. Beuscher, A. Olson and D. Goodsell, “Identifying Protein Binding Sites and Optimal Ligands”, Letters in Drug Design & Discovery, 2005,pp.438-489.

[4] S.Gattani, A.Mishra and M.TamjidulHoque, “StackCBPred: A stacking based prediction of protein-carbohydrate binding sites from sequence”, Carbohydrate Research 486, 107857, 2019.

[5] K G. Srinivasa, G M. Siddesh, S.R. Manisekhar, “Introduction to Bioinformatics”, In book: Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications, 1 edition, Springer Singapore, 2020, 317 pages.

[6] G. Taherzadeh, Y. Zhou, A. Liew, and Y. Yang, “Structure-based prediction of protein– peptide-binding regions using Random Forest”, Bioinformatics, 2017(8),pp.477-484.

[7] Z. Qiu, X. Wang, ”Improved Prediction of Protein Ligand-Binding Sites Using Random Forests”, Protein & Peptide Letters, 2011,pp.1212-8(7).

[8] C. Xia, X. Pan, and H. Shen, “Protein-ligand binding residue prediction enhancement through hybrid deep heterogeneous learning of sequence and structure data”, Bioinformatics, 2020, pp.3018-3027.

[9] D. Yu, J.Hu, H.Yan, X. Yang, J. Yang, and H.Shen, “Enhancing protein-vitamin binding residues prediction by multiple heterogeneous subspace SVMs ensemble”, BMC Bioinformatics 15, 297 ,2014.

[10] Z. Peng, L. Kurgan, “High-throughput prediction of RNA, DNA, and protein binding regions mediated

by intrinsic disorder", Nucleic Acids Research, 2015, 43(18):e121 .

[11] S.Gattani, A. Mishra and M. Tamjidul Hoque, "Sequence and Structure-based Protein Peptide Binding ResiduePrediction", Conference: The 6th Annual Conference on Computational Biology and Bioinformatics, Louisiana, USA,2018.

[12] X. Hu, K.Wang and Q.Dong, "Protein ligand-specific binding residue predictions by an ensemble classifier", BMC Bioinformatics17, 470, 2016.

[13] P. Chen, J.Huang and X.Gao, "LigandRFs: Random forest ensemble to identify ligand-binding residues from sequence information alone", BMCBioinformatics, 2014,15suppl15(suppl15):s4 .

[14] S. Iqbal, M. Tamjidul Hoque, "PBRpredict-Suite: A Suite of Models to Predict Peptide Recognition Domain Residues from Protein Sequence", Bioinformatics, 2018,pp.3289-3299.

[15] Z. Zhao, Z. Peng and J.Yang, "Improving Sequence-Based Prediction of Protein-Peptide Binding Residues by Introducing Intrinsic Disorder and a Consensus Method", Journal of Chemical Information and Modeling, 2018, pp.1459-1468.

[16] L. Liu, X. Hu, Z. Feng, X. Zhang, S. Wang, S.Xu and Kai Sun, "Prediction of acid radical ion binding residues by K-nearest neighbor classifier", BMC Molecular and Cell Biology 20 , 52 ,2019.

[17] S.LI, X. HU, L.SUN, and X.ZHANG, "Identifying the sulfate ion binding residues in proteins", 2nd International Conference on Biomedical and Biological Engineering,2017.

[18] A.Kulandaisamy, V. Lathi, K. ViswaPoorani, K. Yugandhar and M. Gromiha, "Important amino acid residues involved in folding and binding of protein-protein complexes", International journal of biological macromolecules, 2016,pp.438-444.

[19] C. Kauffman, G. Karypis, "LIBRUS: combined machine learning and homology information for sequence-based ligand-binding residue prediction", Bioinformatics, 2009,pp.3099-3107.

[20] T. Du, L.Liao, C. Wu and B.Sun, "Prediction of residue-residue contact matrix for protein-protein interaction with Fisher score features and deep learning", Methods, 2016,pp.97-105.

[21] L. Colwell, "Statistical and machine learning approaches to predicting protein-ligand interactions", Current Opinion in Structural Biology, 2018,pp.123-128.

[22] Bishop, Christopher, "Pattern Recognition and Machine Learning", Springer-Verlag New York, 2006, 738 pages. .

[23] G.Taherzadeh, Y.Yang, T. Zhang, A.Liew andY. Zhou, "Sequence-Based Prediction of Protein–Peptide BindingSites Using Support Vector Machine", Journal of Computational Chemistry, 2016,pp.1223-1229.

[24] S. Panda, S.Nag and P. Jana, "A Smoothing Based Task Scheduling Algorithm for Heterogeneous Multi-Cloud Environment", 3rd IEEE International Conference on Parallel, Distributed and Grid Computing (PDGC), IEEE, Waknaghat, 2014.

[25] S. Panda, P. Jana, "Efficient task scheduling algorithms for heterogeneous multi-cloud environment", The Journal of Supercomputing, 2015,pp.1505-1533.

[26] S. Stehman, "Selecting and interpreting measures of thematic classification accuracy", Remote Sensing of Environment, 1997,pp.77-89.

[27] C. Wiley, N. Schaum, F.Alimirah, J.Dominguez, A.Orjalo, G. Scott, P.Desprez, C. Benz, A. Davalos and J. Campisi, "Small-molecule MDM2 antagonists attenuate the senescence-associated secretory phenotype", Scientific REPOrtS, Scientific Reports 8, 2410 ,2018.

[28] A. Fathi, R. Sadeghi, "A Genetic Programming Method for Feature Mapping to Improve Prediction of HIV-1 Protease Cleavage Site", Applied Soft Computing, 2018,pp.56-64.

[29] Available:http://www.mathworks.com

[30] D.J Till, R.J. Hand, "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems", Machine Learning, 2001,pp.171-186.