6. BINDING AFFINITY PREDICTION USING PROTEIN-PROTEIN INTERACTIONS AND REGRESSION TECHNIQUES

Affinity prediction can be performed through various interactions like protein-ligand, protein-RNA, protein-protein, protein-DNA etc. In chapter 4, the implementation of protein-ligand docking based affinity predictive model is discussed. In chapter 5, the implementation of protein-mutated-ligand docking based affinity predictive model is explained. Protein-protein interaction is significant as it aids in finding cellular functions and biological functions in all organisms. This kind of interaction leads to deeper understanding about infection mechanisms and also aids in treating many disorders and drug identification. Binding affinity prediction through protein-protein interaction is important to get a stable complex. The methodologies developed using protein-protein interaction will provide more perceptive of the specific disorder and the interaction of that protein with the cells of organisms.

Protein-protein interaction can be performed with two types of docking namely rigid docking and flexible docking. Rigid docking is best for macromolecule interaction wherein the active site is known and the interaction is performed. To identify active site many algorithms and general approaches are available but deep learning predicts the hot spot in a better manner than the machine learning algorithms. Some of the works performed using protein-protein interaction and hot spot identification through general and machine learning approaches are protein-protein interaction [90], protein-protein docking approach based on biophysical and biochemical properties [91], hot-spot identification using convolutional neural networks [92].

This chapter describes the development of binding affinity prediction model based on protein-protein interaction using various regression algorithms.

6.1 PROTEIN-PROTEIN INTERACTION BASED BINDING AFFINITY PREDICTIVE MODELS USING SUPERVISED LEARNING

This work explains binding affinity prediction models built through protein-protein interaction using supervised regression techniques. The binding affinity prediction problem is formulated as regression task and the regression algorithms such as linear regreesion, random forest, support vector regression and artificial neural network are used to binding affinity predictive models. Learning is facilitated by defining and squeezing the discriminative features such as energy calculations, interfacial properties, NIS properties and physio-chemical properties from the protein-protein interacted complexes. The performance of the

models are assessed using explained variance score, R2 score, mean squared error, root mean squared error, median absolute error, mean absolute error.

Methodology

Affinity predictive models are developed by gathering the protein structures from PDB corpus and interacting proteins from gene cards. Protein structures associated with six types of SCA are taken and the interacting proteins are identified from various works and gene cards. Hotspot is identified with convolutional neural network where the structure is passed into the network and the threshold value is set. The threshold value obtained above 0.5 for protein structures is considered for hotspot. Active site is identified for each and every protein and then the interaction is performed. Interaction complexes are utilized to extract features and the PPD dataset is created. The dataset is employed to train various regression techniques such as linear regression, random forest, support vector regression and artificial neural network. The framework of proposed binding affinity predictive model based on protein-protein interaction is shown in Fig. 6.1. The model includes four components namely corpus development, feature extraction and dataset creation and evaluation of the binding affinity predictive models.

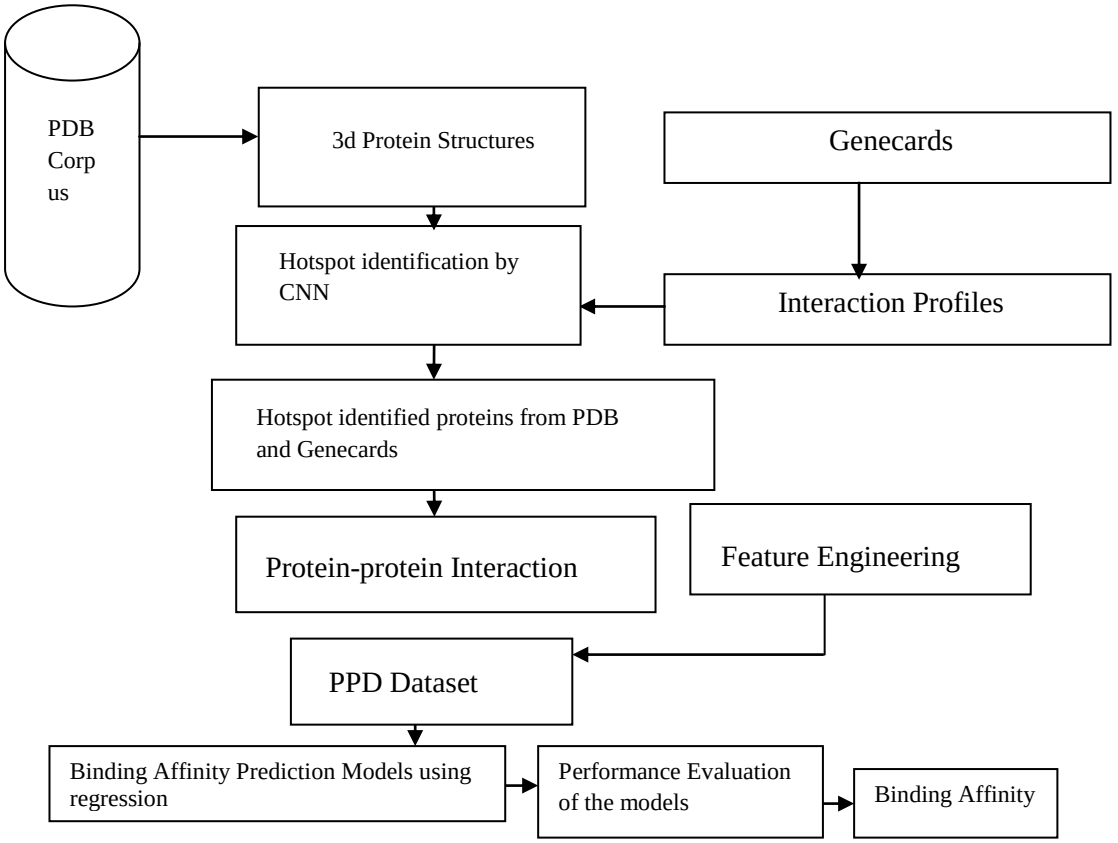

Fig. 6.1 Proposed Framework of Binding Affinity Based on Protein-Protein Interaction

Corpus Development

The corpus is developed by collecting the protein structures associated with six types of SCA from PDB and collecting the interacting proteins from genecards. The same seventeen protein structures that are used in PL corpus are used again. Binding site is identified for each protein structure before interaction. The interaction profiles provided in Table VIII are used for protein-protein interaction. Binding sites or hotspot is identified by CNN and the active site is discovered for each protein structures. Once the binding site is identified, the protein-protein interaction is performed using haddock software. Protein-protein interaction enables in recognizing the unknown functions that might be useful in extracting interfacial properties for building predictive models. The interacted complexes produce clusters of binding energies and the lowest binding energy is chosen. For example, the interaction of protein 1oa8 with 2jy6 is shown in Fig. 3.15. The detailed description of PP corpus has been given in section 3.3 of chapter 3.

Feature Extraction and Dataset Creation

Features like energy, interfacial contacts are captured from the interacted complexes using haddock and physiochemical properties are derived using R script. Energy features include haddock score, cluster size, RMSD, vanderwaals energy, desolvation energy, electrostatic energy, Z-score, Buried surface area, violation energy. Energy values of desolvation, vanderwaals, and electrostatic energy are significant to find the binding affinity score. Features related to interfacial contacts are number of interface pairwise contacts, NIS properties. Physio-chemical properties such as amino acid composition, molecular weight, theoretical pl, negatively charged residues, positively charged residues, carbon, hydrogen, nitrogen, oxygen, sulfur, instability index, aliphatic index, aromaticity, GRAVY are used here. The description of features is presented below.

Haddock Score*:* The unit of energy terms is given as kcal/mol and 1.0, 0.2 and 1.0 are a weighted sum of intermolecular energies such as Vanderwaals Intermolecular Energy (EVDW), Electrostatic Intermolecular Energy (EELEC), Desolvation energy (EDesolvation). The haddock score with the minimum energy value is considered as a feature value. The score is obtained using the equation given in 6.1.

Haddock score = 1.0EVDW+0.2EELEC+1.0*EDesolvation **(6.1)**

where EVDW refers to vanderwaals intermolecular energy, EELEC refers to electrostatic intermolecular energy and EDesolvation refers to desolvation energy. For example, the haddock score obtained for the complex 2jy6 with 1oa8 shown in chapter 3, Fig. 3.15 is -134.7.

Cluster Size: The size of the cluster is the number of amino acid compositions in the most populated cluster. The cluster which occurs first with the minimum energy is chosen for finding the cluster size. For example, size of cluster obtained for the complex 2jy6 with 1oa8 is 114.

RMSD: The root mean squared deviation is used to validate the docking with respect to biological configuration. RMSD is the measure of the average distance between the atoms. The value of RMSD obtained using the equation given in 6.2.

$$RMSD = \sqrt{1/\mathrm{N}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{x_{ci}} - \mathrm{x_{di}})^2 + (\mathrm{y_{ci}} - \mathrm{y_{di}})^2 + (\mathrm{z_{ci}} - \mathrm{z_{di}})^2} \quad \textbf{(6.2)}$$

For example, 0.7 is the value of RMSD acquired for the complex 2jy6 with 1oa8.

Desolvation Energy: Desolvation energy is the static van der waals energy. It is the lose of interaction between substance or organic compound and solvent upon binding describes the energy. For example electro-statically bound particles, dissociate by releasing water in an aqueous solution. Desolvation energy obtained for the complex 2jy6 with 1oa8 is -9.5.

Vanderwaals Energy: Vanderwaals energy is the attraction of intermolecular forces between molecules. Hydrogen bonding, dipole interactions are the examples of vanderwaals energy. The vanderwaals energy is obtained using the equation given in 6.3.

$$E_{VDW} = \sum_{\substack{nonbonded \\ pairs}} \left(\frac{A_{ik}}{r_{ik}^{12}} - \frac{A_{ik}}{r_{ik}^{12}}\right) \quad \textbf{(6.3)}$$

For example, the vanderwaals energy attained for the complex 2jy6 with 1oa8 is -40.2.

Electrostatic Energy*:* Electrostatic energy is the long term interaction between charged atoms. The example of electrostatic energy is, to hold balloon against ceiling. Electrostatic energy obtained for the complex 2jy6 with 1oa8 is -440.8.

Z-score: The z-score represents the standard deviations the haddock score of a given cluster, is separated from the mean of all clusters. The z-score is obtained using the equation given in 6.4.

$$z = (x - \mu) / \sigma \quad \textbf{(6.4)}$$

For example, the value of Z-score achieved for the complex 2jy6 with 1oa8 is -1.8.

Buried Surface Area*:* Buried surface area predicts different measures of flexibility. Binding surface for the complex of 2jy6 with 1oa8 is 1556.

Violation energy: Violation energy is calculated based on dihedral angle, distance, RDC, etc. The violation energy obtained for the complex of 2jy6 with 1oa8 is 11.8.

Interfacial Contacts: Interfacial contacts calculate number of interface residue pair wise contacts, for each complex. NIS properties such as percentage of polar, apolar charged residue are used here.

Physio-chemical Properties: Physical and chemical properties are extracted to identify the changes in the structure, owing to interaction. Physical properties like molecular weight, number of aminoacids, theoritcal pl etc., are taken for binding affinity prediction. Chemical properties such as negatively charged residues, positively charged residues, carbon, hydrogen, nitrogen, oxygen, sulfur, instability index, aliphatic index, aromaticity and GRAVY. Physio-chemical properties along with energy calculations facilitate in predicting binding affinity.

In this regression task, binding affinity is response variable (y_i) whereas energy calculations, physio-chemical properties and interfacial contacts are used as independent variables (x_i). Binding affinity measures the strength of attraction between molecule and ligand. Binding affinity is calculated for each complex using formula given in equation 6.8.

$$[R]\,[R]\,k_1 = [DR]\,K\text{-}1 \quad \textbf{(6.5)}$$

$$K_1/K{-}1 = [RR]/[R][R] \quad \textbf{(6.6)}$$

$$\text{Binding Affinity} = K_1/K{-}1 \quad \textbf{(6.7)}$$

$$K_d = K{-}1/K_1 \quad \textbf{(6.8)}$$

Here, K_d is called as binding affinity constant, K1 is termed as association constant and k-1 is rate constant. For the complex 2jy6 with 1oa8, the binding affinity value obtained is -9.1.

Feature Importance using correlation matrix: The correlation matrix is attained for analyzing the importance of features as done in previous case. The correlation values lies between -1 to 1. The correlation matrix of feature vectors is shown in Fig. 6.2. This correlation matrix shows that the feature vanderwaals and interfacial contact1 have the values of 0.6, desolvation and binding energy have the value of 0.9, theoretical pl posses the value of 0.9, haddock score has the value of 0.7, electrostatic energy has the value of 0.4. This matrix determines the relation between independent variable (X) and dependent variable (Y).

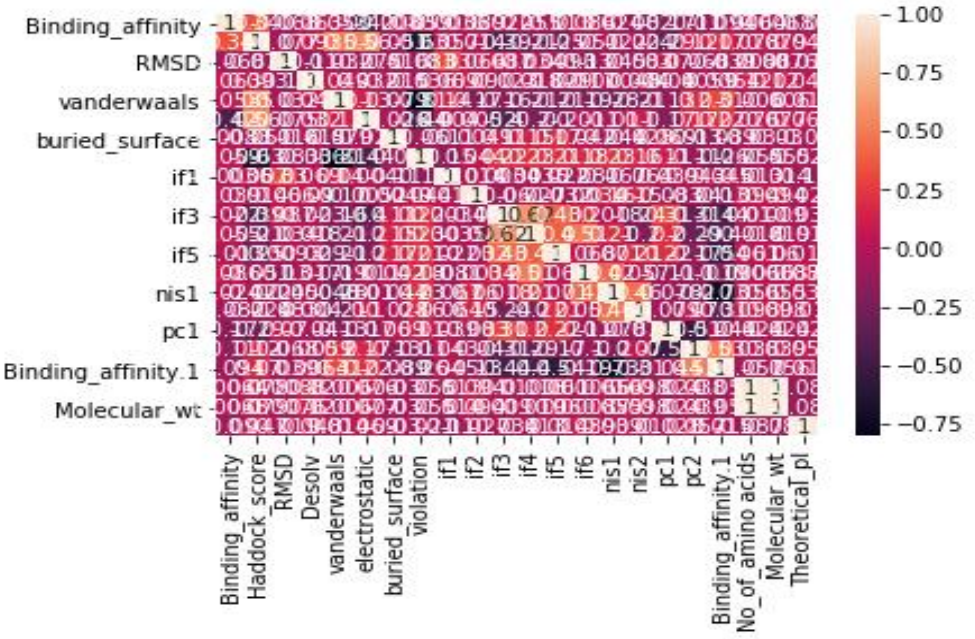

Fig. 6.2 Correlation Matrix of Feature Vectors

The feature importance measure ranks and evaluates the importance of features. The feature importance based on correlation matrix enables in recognizing the contributive feature set with respect to binding affinity. By this way the feature values are validated and the P-value is determined for the contributive feature to discover its relationship with binding affinity. Permutation feature importance of feature vectors and the scores for each feature value are shown in Fig. 6.3 and Fig. 6.4 respectively.

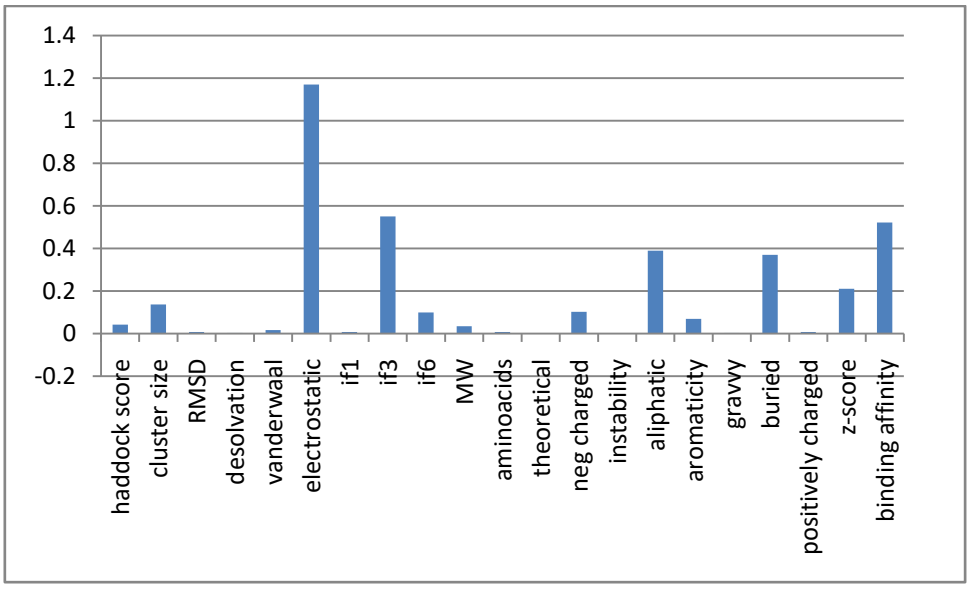

Fig. 6.3 Permutation Feature Importance of Features

haddock score	0.04226	aliphatic	0.39
cluster size	0.13719	aromaticity	0.0688
RMSD	0.00812	gravvy	0.003
desolvation	0.0037	positively charged	0.008
vanderwaal	0.017	z-score	0.21
electrostatic	1.17	binding affinity	0.522
if1	0.007	aminoacids	0.008
if3	0.55	theoretical	0.001
if6	0.1	buried	0.37
MW	0.034	NIS1	0.454

Fig. 6.4 Scores of Features

The feature importance based on correlation matrix shows that the most contributive feature is binding energy. Binding energy is derived based on energy calculations, interfacial properties and physio-chemical properties. P-value is calculated to disclose the relationship between dependant and independent variable. The P-value obtained is less than 0.05 and it shows that the relationship between binding affinity and binding energy is strong. The P-value for binding energy and binding affinity is shown in Fig. 6.5.

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.168569
R Square	0.028416
Adjusted R Square	0.025291
Standard Error	1.60574
Observations	313

ANOVA

	Df	SS	MS	F	Significance F
Regression	1	23.45233	23.45233	9.095683	0.002773
Residual	311	801.8831	2.578402		
Total	312	825.3354			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 50.0%	Upper 50.0%
Intercept	-6.96751	0.26047	-26.7497	1.22E-82	-7.48002	-6.45501	-7.48002	-6.45501
binding_energy	0.009225	0.003059	3.015905	0.002773	0.003206	0.015243	0.003206	0.015243

Fig. 6.5 P-Value of Binding Affinity with Binding Energy

The above set features are extracted from 313 interacted complexes in PP corpus and the feature values are normalized using min-max normalization. Binding affinity values are derived from prodigy and augmented with feature vectors. The summary of the above features are depicted below.

Features	Count
Haddock Score	1
Cluster size	1
RMSD	1
Desolvation energy	1
Van der waals energy	1
Electrostatic energy	1
Z-score	1
Buried surface area	1
Binding affinity	1
Dissociation Constant	1
Physio-chemical properties	38
Interfacial Contacts	6
NIS Properties	2
Total	56

Totally 56 features are derived from each docked complex and protein- protein dataset (PPD dataset) with 313 instances of dimension 56 is developed. The feature values of sample interacted complex 2jy6 with 1oa8 is specified below. The sample dataset is given in Appendix A.

0.67	0.45	1	0.9	0.3	0.23	1	0.634	1	0.000035	1
0.33	0	0.453	0.77	0.86	0.245	0.123	0.47	0.534	0.43	0.10
0	0.04	0.06	0.00	0.04	0.08	0.06	0.00	0.08	0.10	0.02
0.04	0.10	0.08	0.04	0.00	0.02	0.02	0.02	0.34	0.5	
0.3237	0.80	0.345	0.75	1	0.751	1	0.283	0.48	0.87	0.06
0.3	0.234									

Model Building

Binding affinity predictive models are built using various regression algorithms like linear regression, random forest, support vector regression and artificial neural network. Permutation feature importance is used to assess the importance and score of each value. The hyper parameters such as number of iterations, learning rate, number of estimators are used here also. Tuning of hyper parameters aids in achieving the better prediction rate. The dataset of 313 instances is split into training and testing set where 282 instances for training, 31 instances for testing. The performances of the models are assessed by means of various metrics such as explained variance score, mean squared error, root mean squared error, R2 score, median absolute error, mean absolute error and P-value.

Performance metrics like explained variance score and mean squared error are considered as significant metrics in regression task where explained variance score should be higher and the error rate should be low. The other error metrics like root mean squared error, median absolute error and mean absolute error should be minimal. R2 score value should be higher and P-value should be less than 0.05 to determine the relationship stronger. Train-test split is a technique used to evaluate the performance of various regression algorithms. The experimental results of predictive models based on PPD dataset are given in the following section.

6.2 EXPERIMENT AND RESULTS

Experiments have been carried out using regression techniques such as linear regression, random forest, support vector regression and artificial neural network with PPD dataset using the scikit learn tool. The predictive models are built by training the dataset and estimated using the standard 10-fold cross validation technique. The results obtained from the regression models are analyzed through performance measures namely explained variance score, mean squared error, root mean squared error, median absolute error, mean absolute error. The results of binding affinity predictive models based on protein-protein interaction are given in Table 6.1.

Table 6.1 Performance Measures of Binding Affinity Predictive Models Based on Protein-Protein Interaction

Machine Learning Algorithms	Explained Variance Score	R2 score	Mean Squared error	Root Mean Squared Error	Median Absolute Error	Mean Absolute error
LR	0.82	0.82	0.20	0.44	0.22	0.15
SVR	0.86	0.86	0.2	0.44	0.22	0.15

RF	**0.89**	**0.89**	**0.2**	**0.44**	**0.24**	**0.10**
ANN	0.76	0.76	0.30	0.52	0.30	0.22

Table 6.1 shows that the linear regression predictive model based on protein-protein interaction obtains the explained variance score and the error rate as 0.82 and 0.20. The results of root mean squared error, median absolute error, mean absolute error and R2 score obtained are 0.44, 0.22, 0.15 and 0.82 respectively. The SVR predictive model acquired the explained variance score and the error rate as 0.86 and 0.2. The results of root mean squared error, median absolute error, mean absolute error and R2 score acquired are 0.44, 0.22, 0.15 and 0.86 respectively. The RF predictive model obtained the explained variance score and the error rate as 0.89 and 0.2. The results of root mean squared error, median absolute error, mean absolute error and R2 score attained are 0.44, 0.24, 0.10 and 0.89 respectively. The ANN predictive model yields the explained variance score and the error rate as 0.76 and 0.30. The results of root mean squared error, median absolute error, mean absolute error and R2 score attained are 0.52, 0.30, 0.22 and 0.76 respectively. Among the predictive models based on protein-protein interaction, random forest achieves the highest prediction rate and minimum error rate. Random forest produces proficient result as it is an estimator algorithm which cumulates the result of many decision trees and then outputs the most favourable result. The performance measures of predictive models based on protein-protein interaction for various metrics are portrayed in Fig. 6.6 to Fig. 6.11.

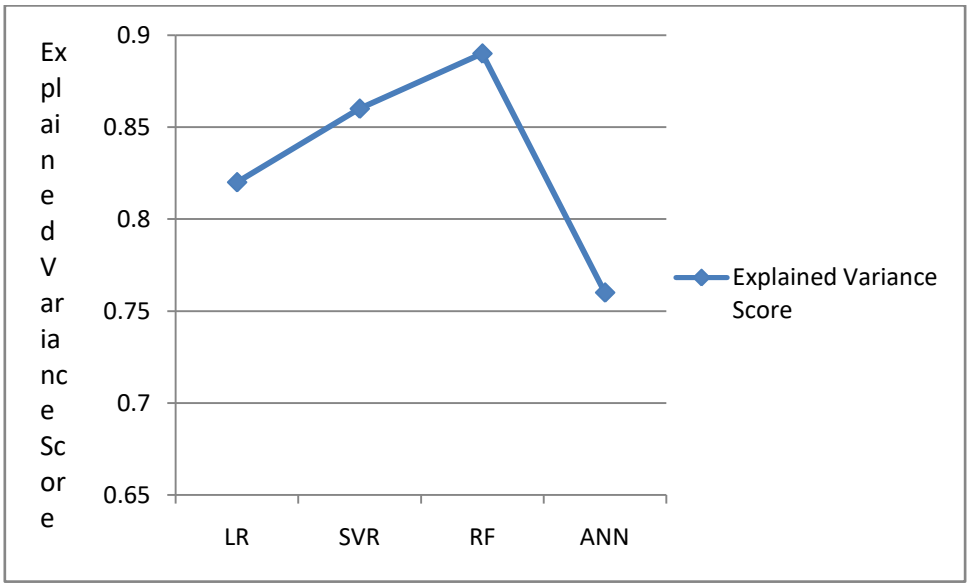

Fig. 6.6 Explained Variance Score of Binding Affinity Predictive Models Based on Protein-Protein Interaction

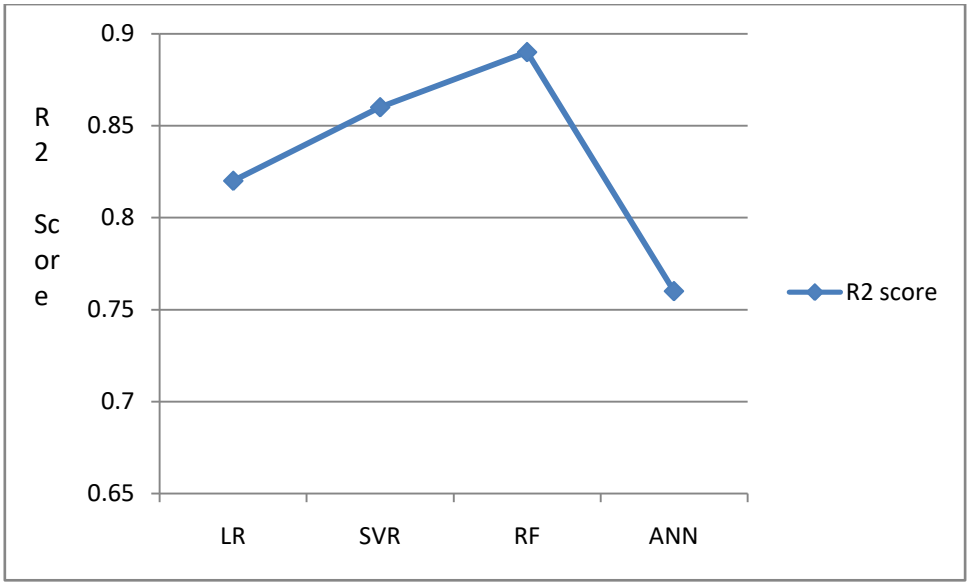

Fig. 6.7 R2 Score of Binding Affinity Predictive Models Based on Protein-Protein Interaction

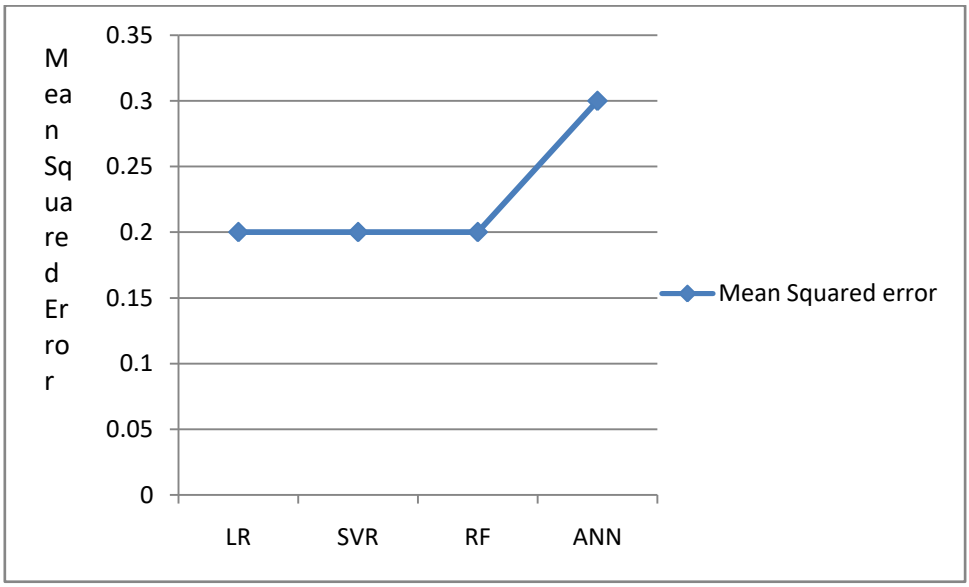

Fig. 6.8 Mean Squared Error of Binding Affinity Predictive Models Based on Protein-Protein Interaction

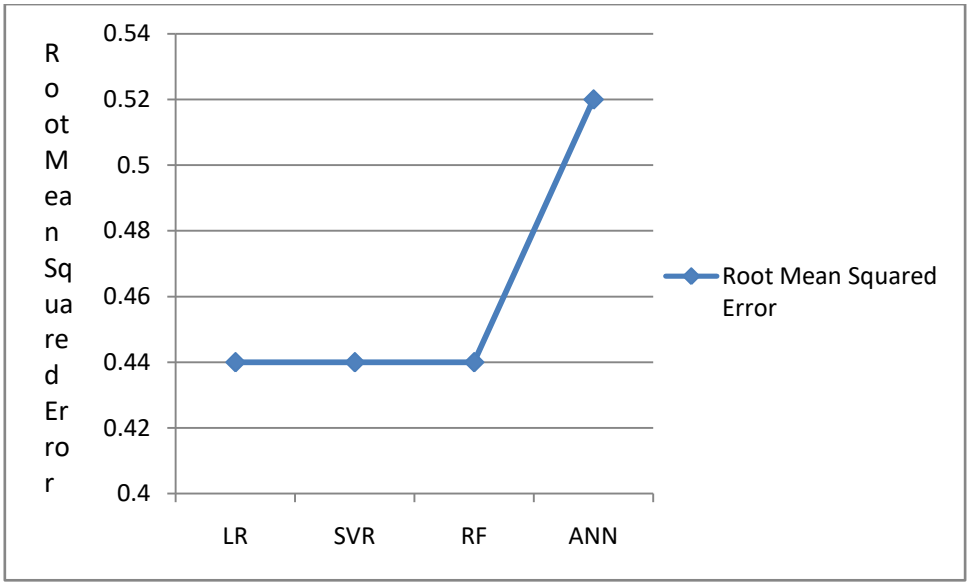

Fig. 6.9 Root Mean Squared Error of Binding Affinity Predictive Models Based on Protein-Protein Interaction

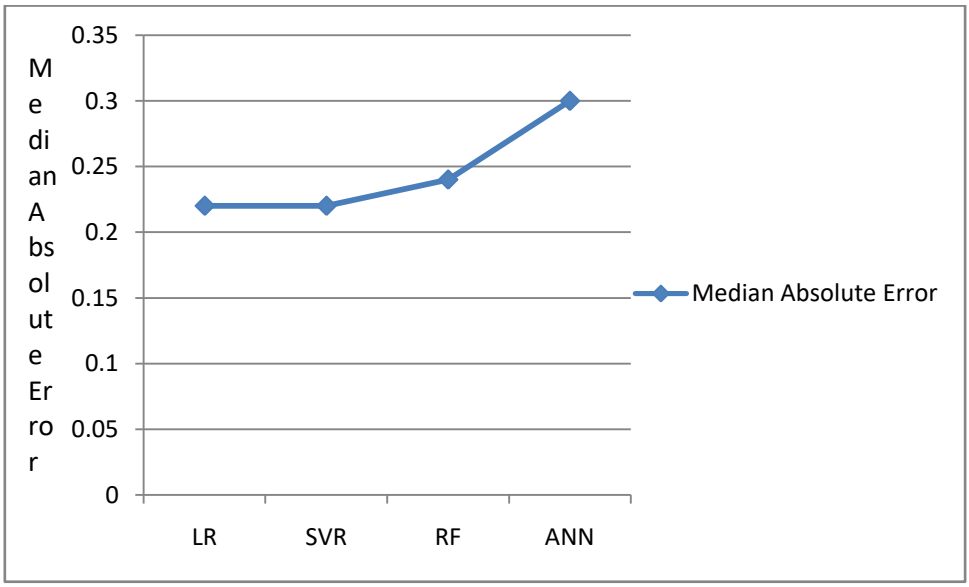

Fig. 6.10 Median Absolute Error of Binding Affinity Predictive Models Based on Protein-Protein Interaction

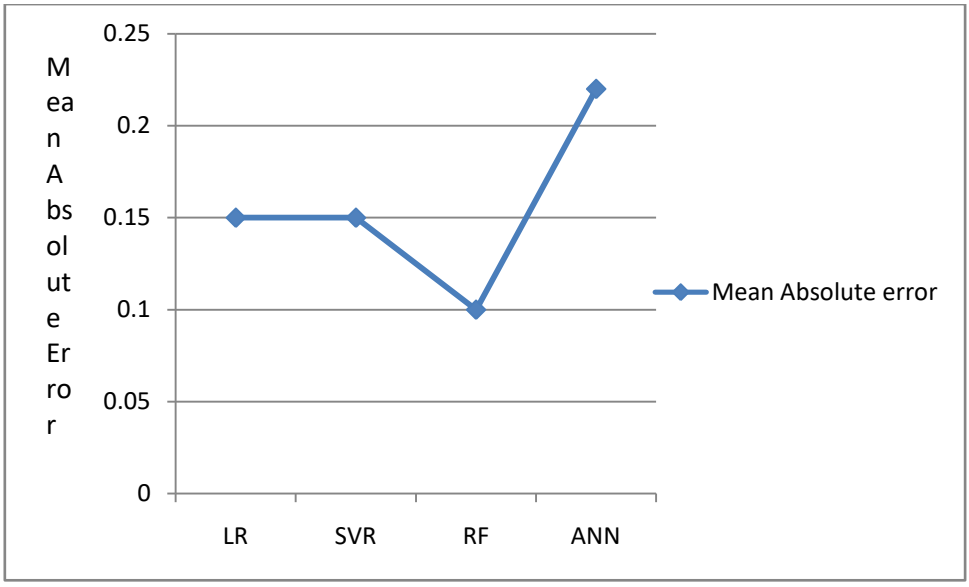

Fig. 6.11 Mean Absolute Error of Binding Affinity Predictive Models Based on Protein-Protein Interaction

From Fig. 6.6, it is observed that the predictive model based on random forest algorithm achieves higher explained variance score than the other regression algorithms. The Fig. 6.7 reveals that the R2 score curve goes superior for random forest and inferior for other predictive models. It is exposed from the Fig. 6.8, the random forest predictive model obtains the low error rate compared to the other regression algorithms. From Fig. 6.9 to Fig. 6.11, it is found that the curve for random forest goes inferior in error rate wherein the other regression models achieve higher error rate. This concludes that the evaluation results of random forest predictive model outperform other regression models.

Comparative Analysis of Predictive Models Based on Protein-Ligand Docking, Protein-Mutated-Ligand Docking and Protein-Protein Interaction

The performance results of predictive models based on protein-protein interaction is compared with predictive models based protein-ligand docking and protein-mutated-ligand docking and the comparative results are analyzed. The predictive models built through protein-protein interaction are helpful to analyze the interfacial contacts of macromolecules. While, in previous cases the unknown functions of protein structures cannot be analyzed. Random forest performed better than other regression algorithms for the predictive models based on protein-ligand docking, protein-mutated-ligand docking and protein-protein interaction. The comparative results of binding affinity predictive models based on regression algorithms are tabulated in Table 6.2.

Table 6.2 Comparative Results of Binding Affinity Prediction Based on Regression Models

Algorithms	Random Forest			Linear Regression			Support Vector Regression			Artificial Neural Network		
Dataset	**PLD**	**PMLD**	**PPD**	**PLD**	**PMLD**	**PPD**	**PLD**	**PMLD**	**PPD**	**PLD**	**PMLD**	**PPD**
Explained Variance Score	**0.85**	**0.87**	**0.89**	0.70	0.68	0.82	0.76	0.70	0.86	0.82	0.75	0.76
R2 Score	**0.85**	**0.87**	**0.89**	0.70	0.68	0.82	0.76	0.70	0.86	0.82	0.75	0.76
Mean Squared Error	**0.20**	**0.2**	**0.2**	0.32	0.45	0.20	0.30	0.32	0.2	0.20	0.30	0.30
Root Mean Squared Error	**0.44**	**0.4**	**0.44**	0.57	0.67	0.44	0.57	0.57	0.44	0.44	0.59	0.52
Mean Absolute Error	**0.15**	**0.15**	**0.10**	0.23	0.34	0.15	0.22	0.23	0.15	0.15	0.27	0.22
Median Absolute Error	**0.25**	**0.22**	**0.24**	0.35	0.49	0.22	0.30	0.35	0.22	0.22	0.39	0.30

Table 6.2 shows shows that the random forest predictive model based on protein-ligand docking yields the explained variance score and error rate as 0.85 and 0.2 respectively whereas the explained variance score and mean squared error for the predictive model based on protein-mutated-ligand docking is 0.87 and 0.2 respectively. The predictive model based on protein-protein interaction achieves the explained variance score and error rate as 0.89 and 0.2 respectively. The results of root mean squared error, mean absolute error, median absolute error, R2 score for random forest predictive model based on protein-ligand docking is 0.44, 0.15, 0.25 and 0.85 respectively. The results of root mean squared error, mean absolute error, median absolute error, R2 score for random forest predictive model based on protein-mutated-ligand docking is 0.4, 0.15, 0.22 and 0.87 respectively. The results of root mean squared error, mean absolute error, median absolute error, R2 score for random forest predictive model based on protein-protein interaction is 0.44, 0.10, 0.24 and 0.89 respectively.

The linear regression predictive model based on protein-ligand docking yields the explained variance score and error rate as 0.70 and 0.32 respectively whereas the explained variance score and error rate for predictive model based on protein-mutated-ligand docking is 0.68 and 0.45 respectively. The predictive model based on protein-protein interaction achieves the explained variance score and error rate as 0.82 and 0.20 respectively. The results of root mean squared error, mean absolute error, median absolute error, R2 score for linear regression predictive model based on protein-ligand docking is 0.57, 0.23, 0.35 and 0.70

respectively. The results of root mean squared error, mean absolute error, median absolute error, R2 score for linear regression predictive model based on protein-mutated-ligand docking is 0.67, 0.34, 0.49 and 0.68 respectively. The results of root mean squared error, mean absolute error, median absolute error, R2 score for linear regression predictive model based on protein-protein interaction is 0.44, 0.15, 0.22 and 0.82 respectively. The support vector regression predictive model based on protein-ligand docking produces the explained variance score of 0.76 and the error rate is 0.30. The SVR predictive model based on protein-mutated-ligand docking produces the explained variance score and error rate as 0.70 and 0.32 respectively. The SVR predictive model based on protein-protein interaction produces the explained variance score and error rate as 0.86 and 0.2 respectively. The results of root mean squared error, mean absolute error, median absolute error and R2 score for SVR predictive model based on protein-ligand docking is 0.57, 0.22, 0.30 and 0.76 respectively. The results of root mean squared error, mean absolute error, median absolute error and R2 score for SVR predictive model based on protein-mutated-ligand docking is 0.57, 0.23, 0.35 and 0.70 respectively. The results of root mean squared error, mean absolute error, median absolute error and R2 score for SVR predictive model based on protein-protein interaction is 0.44, 0.15, 0.22 and 0.86 respectively.

The artificial neural network predictive model based on protein-ligand docking produces the explained variance score and error rate as 0.82 and 0.20 respectively. The ANN predictive model based on protein-mutated-ligand docking obtains the explained variance score and error rate as 0.75 and 0.30 respectively. The ANN predictive model based on protein-protein interaction obtains the explained variance score and error rate as 0.76 and 0.30 respectively. The results of root mean squared error, mean absolute error, median absolute error and R2 score for ANN predictive model based on protein-ligand docking is 0.44, 0.15, 0.22 and 0.82 respectively. The results of root mean squared error, mean absolute error, median absolute error and R2 score for ANN predictive model based on protein-mutated-ligand docking is 0.59, 0.27, 0.39 and 0.75 respectively. The results of root mean squared error, mean absolute error, median absolute error and R2 score for ANN predictive model based on protein-protein interaction is 0.52, 0.22, 0.32 and 0.76 respectively. Among the predictive models based on protein-ligand docking, protein-mutated-ligand docking and protein-protein interaction random forest achieves the highest prediction rate and lower error rate than the other predictive models. The random forest predictive model based on protein-protein interaction achieves the higher prediction rate than the models based on protein-

ligand docking and protein-mutated-ligand docking. The comparative results of binding affinity predictive models with regression algorithms are shown in Fig. 6.12.

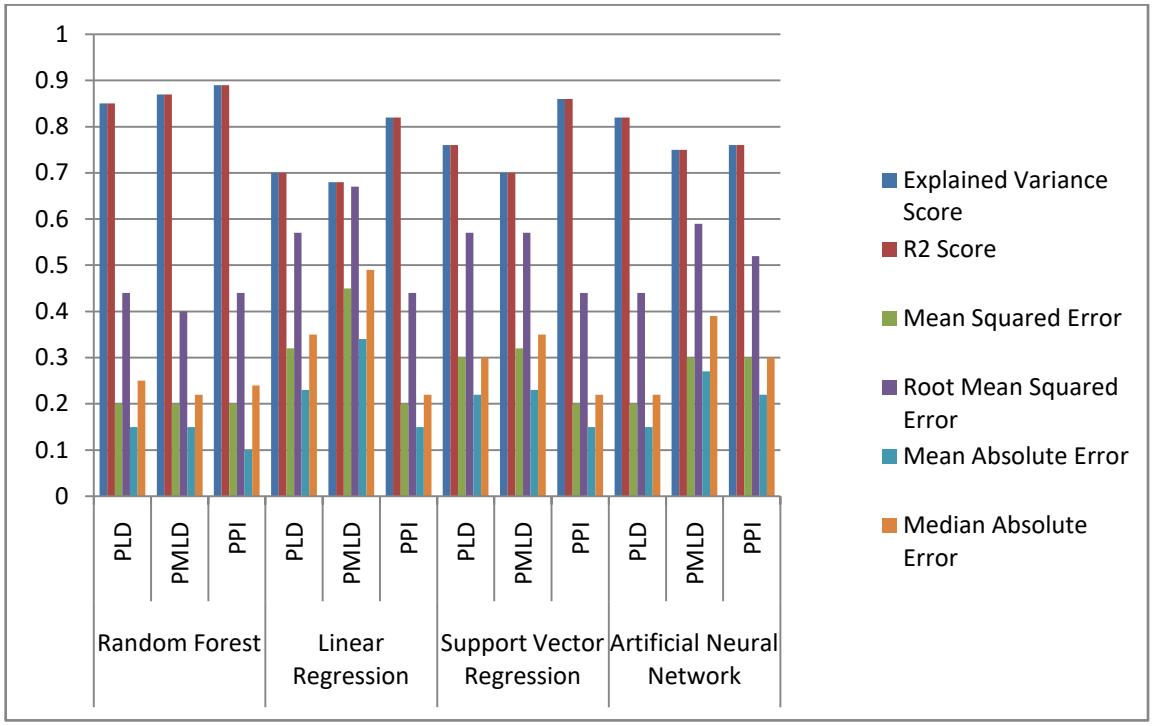

Fig. 6.12 Comparative Results of Binding Affinity Prediction Using Regression

Fig. 6.12 shows that the model trained with PPD dataset yields high prediction rate whereas the model trained with PLD and PMLD dataset achieves the less prediction rate The metrics such as root mean squared error, mean absolute error and median absolute error are lower for random forest predictive model than the other predictive models. Random forest predictive model based on protein-protein interaction achieves high prediction rate as the macromolecules interact and the unknown functions are identified with features like interfacial contacts.

Findings

The implementation results show that the features extracted from the interacted docked complexes highly contribute in predicting binding affinity accurately. PFI shows the importance of each feature vectors where the binding energy attains the high score. P-value of binding energy with binding affinity confirms that the value is lower than 0.05 and this reveals that the relationship is strong between binding energy and binding affinity. The comparative results confirm that the binding affinity predictive models based on protein-protein interaction achieve the highest prediction rate than the predictive models based on

protein-mutated-ligand docking and protein-ligand docking. Random forest based affinity binding predictive model proves that explained variance score is higher and the mean squared is low than other regression algorithms. The error rate associated with binding affinity predictive models is less for random forest model and hence it is suitable for prediction of binding affinity for other disorders. This proves that the work can be performed with any type of interactions for prediction of binding affinity.

SUMMARY

This chapter demonstrated the binding affinity predictive models as regression task. The unknown functions of protein structure are known by extracting interfacial contacts through protein-protein interaction and the same has been presented. The implementation of various regression techniques executed through protein-protein dataset has been described in detail. The experimental results of four predictive models have been reported and the comparative analysis is presented. The comparison of binding affinity predictive models based on protein-protein interaction, protein-mutated-ligand docking and protein-ligand docking with respect to various evaluation metrics has been illustrated with tables and charts. The development of predictive models to predict binding affinity using deep learning approach will be discussed in following chapter.

Remarks

The paper titled, Affinity Prediction Models from Protein-Protein Interaction of SCA Using Ensemble Learning, has been published in International Journal of Control and Automation, Vol 13, Issue 4, PP 707-717, 2020. ***(Scopus indexed)***