5. BINDING AFFINITY PREDICTION USING PROTEIN-MUTATED-LIGAND DOCKING AND REGRESSION TECHNIQUES

Affinity prediction of mutated protein structures is tricky as the changes occur in structure due to mutation. SCA occurs commonly due to repeat mutation and the six types of SCA are considered for study. Each type of SCA has certain limit of repeats that comes under normal range and the limit exceeding the range is considered as mutation. The change in sequence occurs for repeat mutation when it is inherited from parent as the glutamine repeats increases in offspring. It is complicated to predict binding affinity through mutated protein structures. Hence there is a need to predict affinity through mutated protein structures which will enable to develop drugs for different types of mutation. Developing methodologies using mutated protein structures and ligand will provide clear understanding of their structural changes and chemical changes.

The methods exist to predict binding affinity with mutated protein structures does not fabricate the better results of mutation induced protein structures. The general approaches implemented for mutation induced protein structures contain the structure of virus and animals where mutation induced protein structures of homo sapiens is very exceptional. Some of the works in mutation induced approaches implemented through general approach and machine learning are, assessment of cancer missense mutation in protein structures [84], hot-spot mutations in selectable genes [85], mutation induced protein stability changes [86].

This chapter illustrates the development of binding affinity predictive models based on protein-mutated-ligand docking using various regression algorithms.

5.1 PROTEIN-MUTATED-LIGAND DOCKING BASED BINDING AFFINITY PREDICTIVE MODELS USING SUPERVISED LEARNING

This work explains binding affinity prediction models built through protein-ligand docking using supervised regression techniques. The problem is formulated as regression task and regression algorithms such as linear regression, random forest, support vector regression and artificiall neural network. The predictive models are built by identifying and deriving the significant features through scrutinizing the changes that occur due to mutation from the mutated docked complexes which aids precise learning. The performance of the models are evaluated using explained variance score, mean squared error, root mean squared error, median absolute error, mean absolute error.

Methodology

Binding affinity predictive model is erected by gathering the protein structures from PDB corpus and assembling ligands from gene cards. The six types of SCA are considered namely SCA 1, SCA 2, SCA 3, SCA 6 and SCA 10 as these types are commonly caused by repeat mutation. The same seventeen protein structures are induced with repeat mutation using the information from HGMD database. Docking is performed using autodock and the changes occurred in protein sequence due to repeat mutation is analyzed.

Mutated protein structures are docked with ligand and docked complexes are utilized for feature extraction and respective dataset is created. Predictive models are built by employing regression techniques such as linear regression, random forest support vector regression and artificial neural network. The framework for binding affinity prediction model based on protein-mutated-ligand docking is shown in Fig. 5.1 and the model comprises of four phases namely corpus creation, feature extraction and dataset development, model building and estimation of binding affinity predictive models.

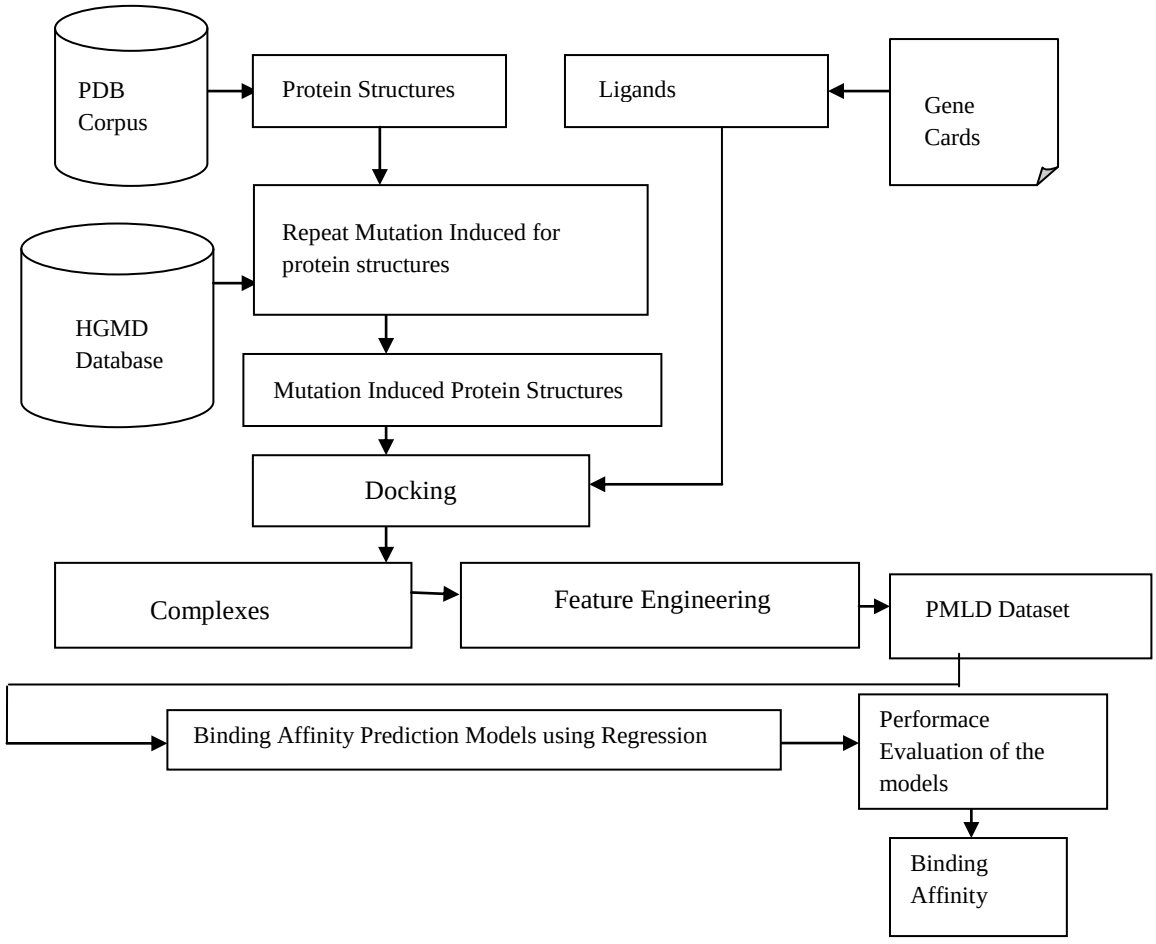

Fig. 5.1 Proposed Framework of Binding Affinity Based on Protein-Mutated-Ligand Docking

Corpus Development

The same seventeen protein structures are used here also to develop the corpus. The mutational information provided in Table IV is used to mutate the protein structures with repeat mutation. The eighteen ligands used in previous case are utilized here to dock with each protein. The changes in the pattern of sequence caused due to mutation and overlap of normal, mutated 1oa8 protein structure is shown in chapter 3, Fig. 3.8 and Fig. 3.9 respectively. The protein structure affected due to mutation, is validated using ramachandran plot. Each and every protein is validated with ramachandran plot after inducing repeat mutation and validity of the protein is shown in chapter 3, Fig. 3.10. The protein structure is considered as valid when the amino acids fall in the favourable region. The same docking process is adopted here also. But the position of the binding site varies when the mutated protein structure docked with ligand which helps precise computation of binding affinity. Docked position of mutated protein structure 1oa8 with ligand amantadine and the detailed description of PML corpus development have been given in section 3.2 of chapter 3.

Feature Extraction and Dataset Creation

Proficient features are squeezed from the mutated docked complexes generated through protein-mutated-ligand docking. Features such as energy calculations, physical properties, sequence descriptors, cyscore, rfscore and autodock vina scores are extracted from the mutated docked complexes. Energy calculations and physical properties of protein and ligand are extracted using autodock and pymol. Sequence descriptors are extracted using R-script whereas cyscore and rf-score are squeezed using linux. Autodock vina scores are extracted from autodock vina.

Energy profiles such as binding energy, inhibition constant, intermolecular energy, desolvation energy, electrostatic energy, total internal energy, torsional energy, molecular weight of protein, ligand and complex etc., Sequence descriptors consist of amino acid composition, autocorrelation, CTD, Quasi-sequence-order descriptors, Pseudo amino acid composition and profile-based descriptors. Cyscore posses hydrophobic free energy, cyscore, van der waals interaction energy, hydrogen-bond interaction and ligand's conformational entropy. Rfscore consists of thirty six values and each feature will denote the number of occurrences of a particular protein-ligand atom type pair interacting within a certain distance range. Autodock vina scores possess four scoring functions such as $\Delta G_{gauss,}$ $\Delta G_{repulsion,}$ $\Delta G_{hydrophobic}$ and ΔG_{Hbond}.

Sequence descriptors are measured where the changes occur in sequence of amino acid, structure, protein folding and binding gets changed due to mutation. Cyscore is extracted for interaction energy profiles like hydrogen-bond, vanderwaals and cyscore. Scores from autodock vina are squeezed for free energy profiles of repulsion, hydrogen bond and hydrophobic. Rf score contains 36 features where the commonly occurred atoms in both the ligand and structure are computed where the energy calculations are extracted using autodock. Cyscore, rfscore and autodock vina scores are squeezed through unix where the sequence descriptors are extracted using R script. The description of the features is given in detail below.

Binding Energy Range: The binding energy range describes at which cluster the binding energy falls. The binding energy range is the difference between highest and lowest energies among the protein-ligand complexes. For the mutated complex 1oa8 with amantadine shown in Fig. 3.11, the binding energy range obtained is 0.19.

Binding Energy: Binding energy is released when a drug molecule associates with a target, that leads to lower the overall energy of the complex. The release in binding energy transforms the ligand from its minimum energy to its bound conformation with the protein. Lower the binding energy more stable the complex. The binding energy is calculated using the equation given in 5.1.

$$\Delta G = \left(V_{\text{bound}}^{\text{L-L}} - V_{\text{unbound}}^{\text{L-L}}\right) + \left(V_{\text{bound}}^{\text{P-P}} - V_{\text{unbound}}^{\text{P-P}}\right) + \left(V_{\text{bound}}^{\text{P-L}} - V_{\text{unbound}}^{\text{P-L}} + \Delta S_{\text{conf}}\right) \quad \textbf{(5.1)}$$

Where P refers to the protein, L refers to the ligand, V refers to the pair-wise evaluations, and ΔS~conf~ denotes the loss of conformational entropy upon binding. Intermolecular energy and torsional energy both are significant to calculate binding energy. The energy of ligand and protein in the unbound state is calculated and then the energy of the protein-ligand complex is calculated. The binding energy is calculated as the difference between energy in unbound state and energy of protein-ligand complex. For example, binding energy obtained for the mutated complex 1oa8 with amantadine is -4.73.

Ligand Efficiency: Ligand efficiency is a measurement of the binding energy per atom of a ligand to its binding partner, such as a receptor or enzyme. Mathematically, ligand efficiency (LE) can be defined as the ratio of Gibbs free energy (ΔG) to the number of non-hydrogen atoms of the compound. Ligand efficiency is calculated using the equation given in 5.3.

$$LE = (\Delta G)/N \quad \textbf{(5.2)}$$

where $\Delta G = -RT\ln K_i$ and N is the number of non-hydrogen atoms. It is transformed to the equation:

$$LE = 1.4(-log IC_{50})/N \quad \textbf{(5.3)}$$

For example, ligand efficiency obtained for the mutated complex 1oa8 with amantadine is -0.43.

Inhibition Constant (pIC50): The inhibitor constant K_i is an indication of how potent an inhibitor is. It is the concentration required to produce half maximum inhibition. The IC50 value is determined at only one concentration of substrate over a range of inhibitor concentrations. While K_i is a constant value for a given compound with an enzyme, an IC50 is a relative value, whose magnitude depends upon the concentration of sub- strate. According to the FDA, IC50 represents the concentration of a drug that is required for 50% inhibition in vitro. The inhibition constant is calculated using the equation given in 5.10.

K_i = dissociation constant of the enzyme-inhibitor complex = K_d

$$K_i = [E][I]/[EI] \quad \textbf{(5.4)}$$

$$\ln Kb = -\ln K_i \quad \textbf{(5.5)}$$

$$\text{deltaG(binding)} = -R*T*\ln Kb \quad \textbf{(5.6)}$$

$$\text{deltaG(inhibition)} = R*T*\ln K_i \quad \textbf{(5.7)}$$

Binding and Inhibition occur in opposite directions, so the minus-sign is omitted

$$\text{deltaG} = R*T*\ln K_i, \quad \textbf{(5.8)}$$

$$\text{deltaG}/(R*T) = \ln K_i \quad \textbf{(5.9)}$$

$$K_i = \exp(\text{deltaG}/(R*T)) \quad \textbf{(5.10)}$$

For example, inhibition constant obtained for the complex mut-1oa8 with amantadine is 338.45.

Intermolecular Energy: Intermolecular energy is the energy between non-bonded atoms that is the energy between atoms separated by 3-4 bonds or between atoms in different molecules. For example, intermolecular energy obtained for the mutated complex 1oa8 with amantadine is -5.03.

Desolvation Energy: Desolvation energy is the static van der waals energy. It is the lose of the 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. The desolvation energy is calculated using the equation given in 5.11.

$$\Delta G_{desolv} = W_{desolv} \sum_{i\,(C),\,j} (S_i * V_j * \exp(-r_{ij}^2 / (2 * \sigma^2))) \quad \textbf{(5.11)}$$

For example, the desolvation energy obtained for the mutated complex 1oa8 with amantadine is -2.96.

Electrostatic Energy: Electrostatic energy is the long term interaction between charged atoms. The example of electrostatic energy is, to hold balloon against ceiling. The electrostatic energy is calculated using the equation given in 5.12.

$$\Delta G_{elec} = W_{elec} \sum_{i,j} (q_i * q_j) / (\varepsilon(r_{ij}) * r_{ij}) \quad \textbf{(5.12)}$$

For example, the electrostatic energy obtained for the mutated complex 1oa8 with amantadine is -2.07.

Total Internal Energy: Total energy is that the total of changes of all energetic terms enclosed in rating operates of matter or supermolecule upon binding, and the changes upon binding of the entropic terms. For example, the total internal energy obtained for the mutated complex 1oa8 with amantadine is 0.06.

Torsional Energy: Torsion energy is related to dihedral term of internal energy. Torsional energy is calculated using the equation given in 5.13.

$$\Delta G_{tor} = W_{tor} N_{tor} \quad \textbf{(5.13)}$$

where N_{tor} is the number of all rotatable bonds, excluding guanidinium and amide bonds etc. For example, the torsional energy obtained for the mutated complex 1oa8 with amantadine is 0.3.

clRMS: It is the root mean square difference between current conformation and the lowest energy conformation in its cluster. For example, the clRMS obtained for the mutated complex 1oa8 with amantadine is 0.

refRms: It is the root mean square distinction between current conformation coordinates and current reference structure. By default the input substance is utilized as a result of the reference. For example, the refRMS obtained for the mutated complex 1oa8 with amantadine is 55.37.

Binding Affinity: Affinity is a measure of the strength of attraction between a molecule and legend. High affinity binding has strong intermolecular force, whereas low affinity binding has weak intermolecular force. Affinity is calculated using the equation given in 5.17.

$$[R]\,[R]\,K_1 = [DR]\,K\text{-}1 \quad \textbf{(5.14)}$$

$$K_1/K{-}1 = [RR]/[R][R] \quad \textbf{(5.15)}$$

$$\text{Binding Affinity} = K_1/K{-}1 \quad \textbf{(5.16)}$$

$$K_d = K{-}1/K_1 \quad \textbf{(5.17)}$$

Here, K_d is called as binding affinity constant, K1 is termed as association constant and k-1 is rate constant. For example, the binding affinity obtained for the mutated complex 1oa8 with amantadine is -3.9.

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 is obtained using the equation given in 5.18. The rmsd l.b denotes the root mean squared deviation lower bound whereas the rmsd u.b denotes the root mean squared deviation upper bound.

$$RMSD = \sqrt{1/N \sum_{i=1}^{N} (x_{ci} - x_{di})^2 + (y_{ci} - y_{di})^2 + (z_{ci} - z_{di})^2} \quad \textbf{(5.18)}$$

For example, the RMSD obtained for the mutated complex 1oa8 with amantadine is 12.53, 13.78 as rmsd l.b and rmsd u.b respectively.

Physical properties: Physical properties such as molecular weight of ligand, molecular weight of complex, atom count in protein, atom count in ligand, atom count for complex, surface area of protein, surface area-solvent access of protein, surface area of ligand, surface area-solvent access of ligand, surface area solvent of complex and charges of protein are derived from pymol. For example, the physical properties obtained for the mutated complex 1oa8 with amantadine are given below.

molecular weight of ligand = 150.2487

molecular weight of complex = 13041.16

atom count in protein = 970

atom count in ligand = 28

atom count for complex = 1198

surface area of protein = 1734.07

surface area-solvent access of protein = 7834.37

surface area of ligand = 700.997

surface area-solvent access of ligand = 2760.95

surface area solvent of complex = 13291.47

charges of protein = -4

RF Score: Rf-score has 36 features, along with energy based features these rf-score features aids in predicting the binding affinity. Rf-score features are extracted using python scripts. Each feature will comprise the number of occurrences of a particular protein-ligand atom type pair interacting within a certain distance range. The main criterion for the selection of atom types is to generate features that are as dense as possible, while considering all the heavy

atoms that are commonly observed in PDB complexes. As the number of protein-ligand contacts is constant for a particular complex, the more atom types are considered. Therefore, a minimal set of atom types is selected by considering atomic number. A smaller set of intermolecular features has the additional advantage of leading to computationally faster scoring functions. Here the nine common elemental atom types for both the protein P and the ligand L are considered.

{P(j)}9j=1={C,N,O,F,P,S,Cl,Br,I}{L(i)}9i=1={C,N,O,F,P,S, Cl,Br,I} **(5.19)**

The occurrence count for a particular j-i atom type pair is evaluated aswhere d_{kl} is the Euclidean distance between k^{th} protein atom of type j and the l^{th} ligand atom of type i calculated from the PDBbind structure; K_j is the total number of protein atoms of type j and L_i is the total number of ligand atoms of type i in the considered complex; Z is a function that returns the atomic number of an element and it is used to rename the feature with a mnemonic denomination; È is the heaviside step function that counts contacts within a d_{cutoff}=12Å neighbourhood of the given ligand atom. For example, $X_{7,8}$ is the number of occurrences of protein nitrogen interacting with a ligand oxygen within a 12Å neighbourhood. This cutoff distance is suggested as sufficient to implicitly capture solvation effects. This representation leads to a total of 81 features, of which 45 are necessarily zero across PDBbind complexes due to the lack of proteinogenic amino acids with F, P, Cl, Br and I atoms [87]. Therefore, each complex will be characterised by a vector with 36 features and is calculated using the equation given in 5.19.

$\vec{x}$=(x6,6,x6,7,x6,8,x6,9,x6,15,$x_{6,15}$,x6,17,x6,35,x6,53,x7,6,…,x8,53,x16,6,…,x16,53) ∈$\aleph^{36}$ **(5.20)**

Rf-score values obtained for the mutated complex 1oa8 with amantadine are 6.6, 7.6, 8.6, 16.6, 6.7, 7.7, 8.7, 16.7, 6.8, 7.8, 8.8, 16.8, 6.16, 7.16, 8.16, 16.16, 6.15, 7.15, 8.15, 16.15, 6.9, 7.9, 8.9, 16.9, 6.17, 7.17, 8.17, 16.17, 6.35, 7.35, 8.35, 16.35, 6.53, 7.53, 8.53, 16.53.

Cyscore: Cyscore is an empirical scoring function consists of four numerical features. The features like Cyscore, hydrophobic energy, van der Waals interaction energy, hydrogen-bond interaction energy and the ligand's conformational entropy are captured from the complexes using python script [88]. For example, the cyscore obtained for the mutated complex 1oa8 with amantadine is -1.2146.

Hydrophobic energy: This energy is the observed tendency of nonploar substances to aggregate in an aqueous solution and water molecules are excluded. Positive free energy

change implies hydrophobicity and the negative free energy change implies hydrophilicity. Hydrophobic materials are used to remove oil from water and chemical separation process to remove non-polar substances from polar substances. For example, the hydrophobic energy obtained for the mutated complex 1oa8 with amantadine is -0.2891.

Van der waals interaction energy: It is driven by induced electrical interactions between two or more atoms or molecules that are very close to each other. It is the weakest of all intermolecular attractions between molecules. Example of van der waals interactions are dipole-diploe interaction and every other interaction force. For example, the vanderwaals interaction energy obtained for the mutated complex 1oa8 with amantadine is -0.9255.

Hydrogen-bond interaction energy: The nature of the donor and acceptor atoms which constitute the bond, their geometry, and environment, the energy of a hydrogen bond can vary between 1 and 40 kcal/mol. This bond is the strongest interaction than the van der waals interaction. Water, chloroform, ammonia are examples of hydrogen bond. For example, the hydrogen-bond interaction energy obtained for the mutated complex 1oa8 with amantadine is 0.

Ligand's conformational entropy: Conformational entropy is an important component of the change in free energy upon binding of a ligand to its target protein. For example, the ligand's conformational entropy obtained for the mutated complex 1oa8 with amantadine is 0.

Sequence Descriptors: Sequence Descriptors have many features which aids in identifying the affinity. The commonly used descriptors are captured with the package protr in R and coded in R. Commonly used descriptors are amino acid composition, autocorrelation, CTD, Conjoint Triad, Quasi-sequence-order descriptors, Pseudo amino acid composition (PseAAC) [89].

Amino acid composition: The key elements of amino acids are hydrogen, nitrogen, oxygen and carbon. Many elements are found in the side chain of amino acids. Naturally occurring amino acids are more than 500 but only 20 amino acids are encoded in genetic code. There are three different amino acid composition namely single amino acid composition, dipeptide and tripeptide composition. The change in amino acid causes mutation and those changes can be monitored through amino acid composition.

Autocorrelation: Autocorrelation descriptors are a class of topological descriptors, also known as molecular connectivity indices, describe the level of correlation between two objects.

Conjoint triad: The conjoint triad descriptors consider the properties of one amino acid and its vicinal amino acids. It also considers three continuous amino acids as a unit.

Quasi-sequence-order descriptors: These descriptors are derived from the distance matrix between 20 amino acids

Pseudo amino acid composition (PseAAC): It represents protein samples for improving protein subcellular localization prediction and membrane protein type prediction.

CTD descriptors: It is the feature vector for predicting proteins targeted to various compartments in the hierarchical structure of cellular sorting pathway from protein sequence. Composition, Transition and Distribution (CTD) of amino acid attributes such as hydrophobicity, normalized van der Waals volume, polarity, polarizability, charge, secondary structure and solvent accessibility of the protein sequences.

Autodock vina scores: An empirical scoring function calculates the affinity, or fitness, of protein-ligand binding by summing up the contributions of a number of individual terms. This score improves accuracy and speed. This scores has four features namely gauss, hydrophobic, hydrogen bonding and repulsion. Autodock vina scores obtained for the mutated complex 1oa8 with amantadine are -0.14984, 7.85, 222.698, 1.80, 7.006, 0 as binding affinity, gauss 1, gauss 2, repulsion, hydrophobic and hydrogen respectively.

Feature Importance using correlation matrix: The correlation matrix is obtained for analyzing the importance of features as done in previous case. The correlation values lies between -1 to 1. The value 1 refers positive correlation, -1 demotes negative correlation and 0 refers to there is no correlation. The correlation matrix of features is shown in Fig. 5.2. In this correlation matrix the feature cyscore and hydrophobic have the values of 0.6, the features gauss 1 and gauss 2 have the value of 0.9, the feature vanderwaals desolvation energy posses the value of 0.9, the feature intermolecular energy has the value of 0.7, the feature electrostatic energy has the value of 0.4. The features rf denotes random forest, rf1 and rf2 have the low correlation values of 0.2. This matrix determines the relation between the independent variable (X) and dependent variable (Y).

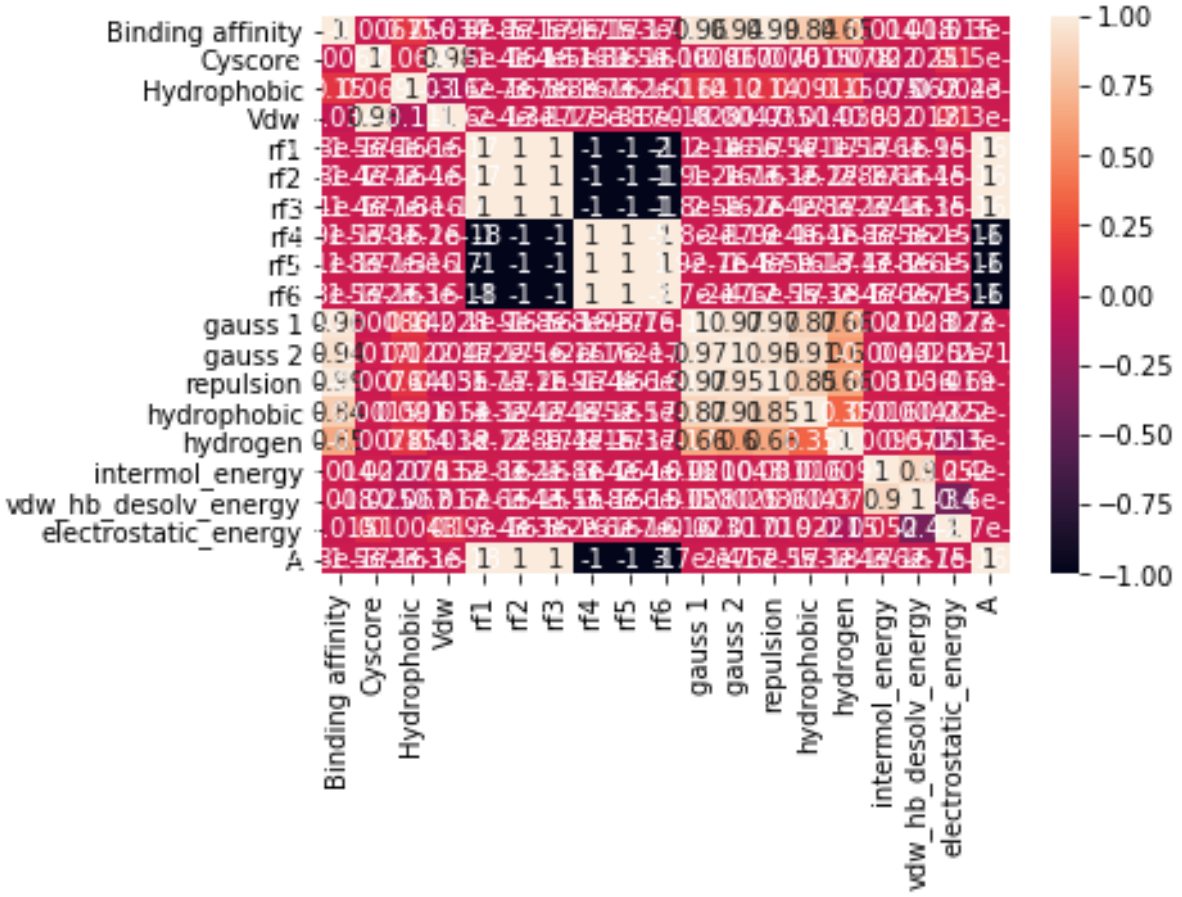

Fig. 5.2 Correlation Matrix of Feature Vectors

The feature importance technique is used to rank and evaluate the importance of features. The feature importance based on correlation matrix enables in identifying 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. 5.3 and Fig. 5.4 respectively.

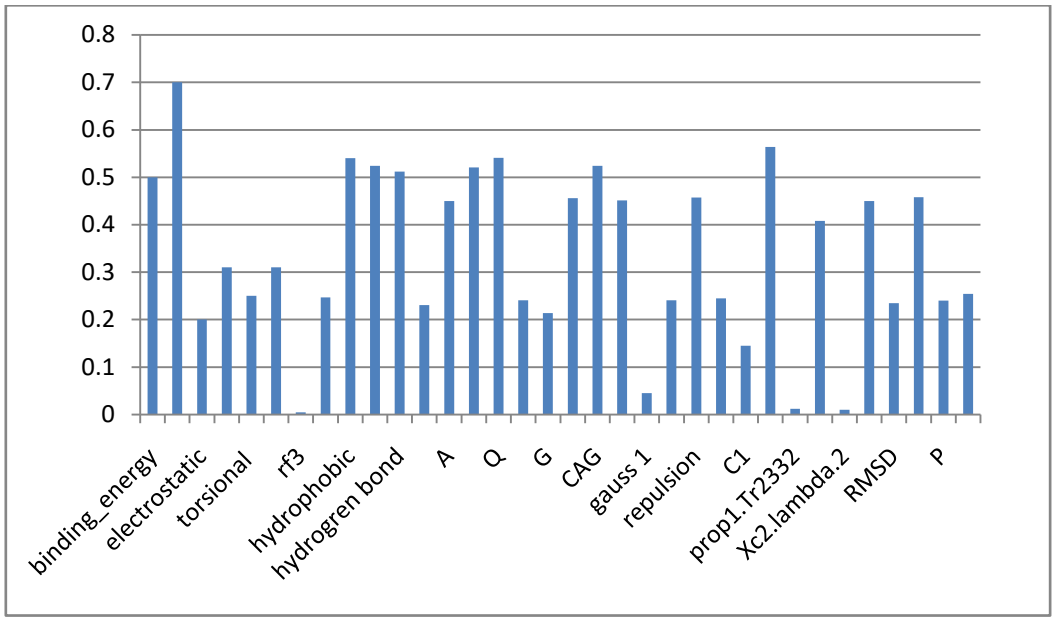

Fig. 5.3 Permutation Feature Importance of Features

binding_energy	0.5	C1	0.145
binding_affinity	0.7	G	0.214
electrostatic	0.2	AA	0.456
desolvation	0.31	CAG	0.524
torsional	0.25	ATG	0.451
rf1	0.31	gauss 1	0.045
rf3	0.005	gauss 2	0.241
rf8	0.247	repulsion	0.457
hydrophobic	0.54	hydrogen	0.245
cyscore	0.524	C8	0.564
hydrogren bond	0.512	prop1.Tr2332	0.0124
ligand entropy	0.231	prop3.G3.residue50	0.408
A	0.45	Xc2.lambda.2	0.01
V	0.521	total internal	0.45
Q	0.541	RMSD	0.235
T	0.241	CTG	0.458
P	0.24	M	0.254

Fig. 5.4 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, scoring functions and sequence descriptors. Thus the energy values with scoring functions and sequence descriptors are important for binding affinity prediction. P-value is calculated to reveal the relationship between dependant and independent variable. The P-value obtained is less than 0.05 and 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. 5.5.

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.592851
R Square	0.351473
Adjusted R Square	0.34934
Standard Error	1.02925
Observations	307

ANOVA

	Df	SS	MS	F	Significance F
Regression	1	174.5336	174.5336	164.7545	1.98E-30
Residual	304	322.0442	1.059356		
Total	305	496.5776			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 50.0%	Upper 50.0%
Intercept	-1.58138	0.266796	-5.9273	8.35E-09	-2.10638	-1.05638	-2.10638	-1.05638
binding_energy	0.54873	0.04275	12.83567	1.98E-30	0.464606	0.632855	0.464606	0.632855

Fig. 5.5 P-value of Binding Affinity with Binding Energy

The feature values from 307 mutated docked complexes in PML corpus are captured and normalized using min-max normalization. Binding affinity values are derived from autodock and amplified with feature vectors. The summary of the features are depicted below.

Features	Count
Autodock vina scores	4
cyscore	5
rf-score	36
sequence descriptors	436
Energy Calculations	27
Total	509

A total of 509 features are extracted from each docked complex and the PMLD dataset with 307 instances of dimension 509 is developed. The feature values for sample docked complex of mut-1oa8 with amantadine are given below. The sample dataset is given in Appendix A.

0.457142857 0.11372549 0.605 0.8 0.266731739 0.533063428 0.456521739
0.892631579 0.113513514 0.85645933 0 0.038010462 0.397925976
0.288157403 0.396518452 0.524308588 0.387096774 0.368685377 0.411631244
0.559109964 0.378619296 0.381390783 0.362093607 0.393234918 0.27027027
0.476286023 0.499454958 0 0 0 0 0 0 0.6 0.42
0.54 0.6 0.7 0.47 0.57 0.23 0.8 0.58 0.8 0.12 0.34 0.24
0.45 0.34 0.23 0.67 0.76 0.45 0.9 0.54 0.9 0.56 0.23 0.67
0.085365854 0.04065040 0.05691057 0.056910567 0.0081301 0.121951 0.06097561 0.040650407
0.008130081 0.040650407 0.097560976 0.0813008 0.0325203 0.024390244 0.0203252
0.077235772 0.04878049 0.00813008 0.044715447 0.0447154 0.4186992 0.32520325
0.256097561 0.337398374 0.422764228 0.239837398 0.300813008 0.2723577 0.426829268
0.308943089 0.451219512 0.239837398 0.12195122 0.699186992 0.178861789 0.528455285
0.219512195 0.25203252 0.349593496 0.418699187 0.231707317 0.27755102 0.191836735
0.195918367 0.314285714 0.159183673 0.171428571 0.183673469 0.236734694 0.240816327
0.293877551 0.155102041 0.175510204 0.167346939 0.053061224 0.265306122 0.232653061
0.253061224 0.102040816 0.273469388 0.159183673 0.195918367 0.62601626 0.57723577
0.68292683 0.98373984 1 0.81300813 0.98373984 0.28455285 0.01626016 0.18699187
0.406504065 0.79674797 0.37398374 0.79674797 1 0.406504065 0.73170732 0.08943089
0.04065041 0.49593496 0.406504065 0.6097561 0.18699187 0.76422764 0.49593496
0.81300813 0.76422764 0.69105691 0.04878049 0.18699187 1 0.57723577 0.68292683
0.13821138 1 0.76422764 0.49593496 0.81300813 0.98373984 0.69105691
0.67479675 0.18699187 0.845528455 0.23577236 0.74796748 0.81300813 0.76422764
0.69105691 0.89430894 0.18699187 0.845528455 0.64227642 0.74796748 0.79674797 1
0.406504065 0.73170732 0.08943089 0.04065041 0.49593496 0.032520325 0.73170732
0.33333333 0.43902439 0.05691057 0.406504065 0.6097561 0.25203252 0.79674797 1
1 0.64227642 0.74796748 0.92682927 0.59349593 0.406504065 0.29268293 0.68292683
0.024482287 0.024213799 0.024529542 0.024630116 0.026939709 2.855112358 2.855112358

Model Building

Binding affinity predictive models are developed using various regression algorithms such as linear regression, random forest, support vector regression and artificial neural network. The same hyper parameters such as number of iterations, number of estimators, learning rate are used here. Tuning of hyper parameters achieves the precise prediction rate. The dataset of 509 instances is split into training and testing as 458 instances for training and 51 instances for testing. The performances of the models are evaluated 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 is less than 0.05 to determine the relationship stronger. The results of predictive models based on protein-mutated-ligand-docking dataset are given in section below.

5.2 EXPERIMENT AND RESULTS

Experiments have been carried out by implementing standard regression techniques namely linear regression, support vector regression, artificial neural network and random forest with PMLD dataset using the scikit learn tool. The standard 10-fold cross validation technique is used to estimate the impact on the predictive performance. 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 are tabulated in Table 5.1.

Table 5.1 Performance Results of Binding Affinity Predictive Models Based on Protein-Mutated-Ligand Docking

Machine Learning Algorithms	Explained Variance Score	R2 score	Mean Squared error	Root Mean Squared Error	Median Absolute Error	Mean Absolute error
LR	0.68	0.68	0.45	0.67	0.49	0.34
SVR	0.70	0.70	0.32	0.57	0.35	0.23
RF	**0.87**	**0.87**	**0.2**	**0.4**	**0.22**	**0.15**
ANN	0.75	0.75	0.30	0.59	0.39	0.27

Table 5.1 indicates that the linear regression predictive model based on protein-mutated-ligand docking obtains the explained variance score and the error rate as 0.68 and 0.45 respectively. The results of root mean squared error, median absolute error, mean absolute error and R2 score obtained are 0.67, 0.49, 0.34 and 0.68 respectively. The SVR predictive model acquired the explained variance score and the error rate as 0.70 and 0.32 respectively. The results of root mean squared error, median absolute error, mean absolute error and R2 score acquired are 0.57, 0.35, 0.23 and 0.70 respectively. The RF predictive model obtained the explained variance score and the error rate as 0.87 and 0.2 respectively. The results of root mean squared error, median absolute error, mean absolute error and R2 score attained are 0.4, 0.22, 0.15 and 0.87 respectively. The ANN predictive model yields the explained variance score and the error rate as 0.75 and 0.30 respectively. The results of root mean squared error, median absolute error, mean absolute error and R2 score attained are 0.59, 0.39, 0.27 and 0.75 respectively. Among the predictive models based on protein-mutated-ligand docking, random forest achieves the highest prediction rate and minimum error rate. In this experiment the number of features is high and the random forest algorithm does not assume the linear relationship between the features. As a result the explained variance score is high and error rate is minimum in case of random forest as compared to the other regression algorithms. The performance results of the binding affinity predictive models with PMLD dataset to various metrics are portrayed in Fig. 5.6 to Fig. 5.11.

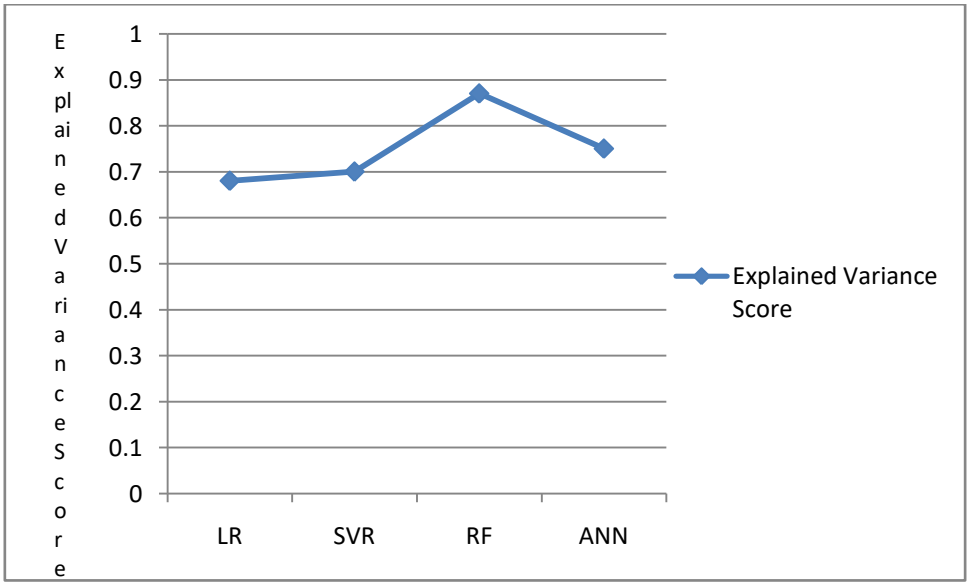

Fig. 5.6 Explained Variance Score of Binding Affinity Predictive Models Based on Protein-Mutated-Ligand Docking

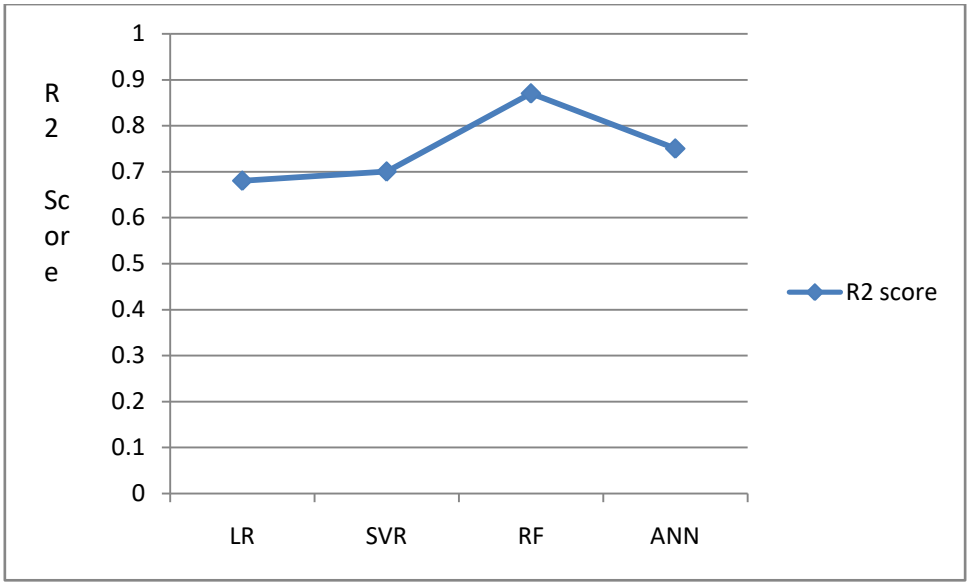

Fig. 5.7 R2 Score of Binding Affinity Predictive Models Based on Protein-Mutated-Ligand Docking

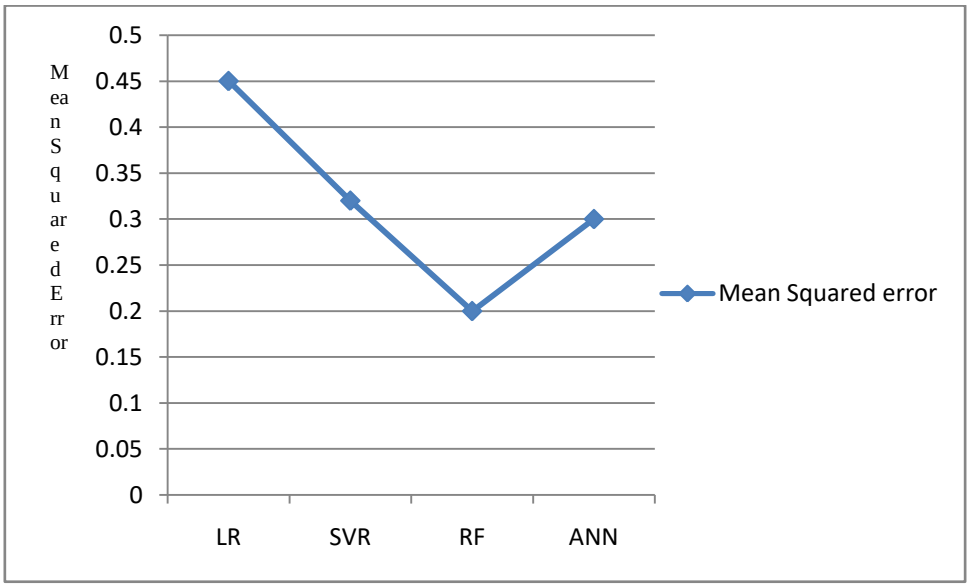

Fig. 5.8 Mean Squared Error of Binding Affinity Predictive Models Based on Protein-Mutated-Ligand Docking

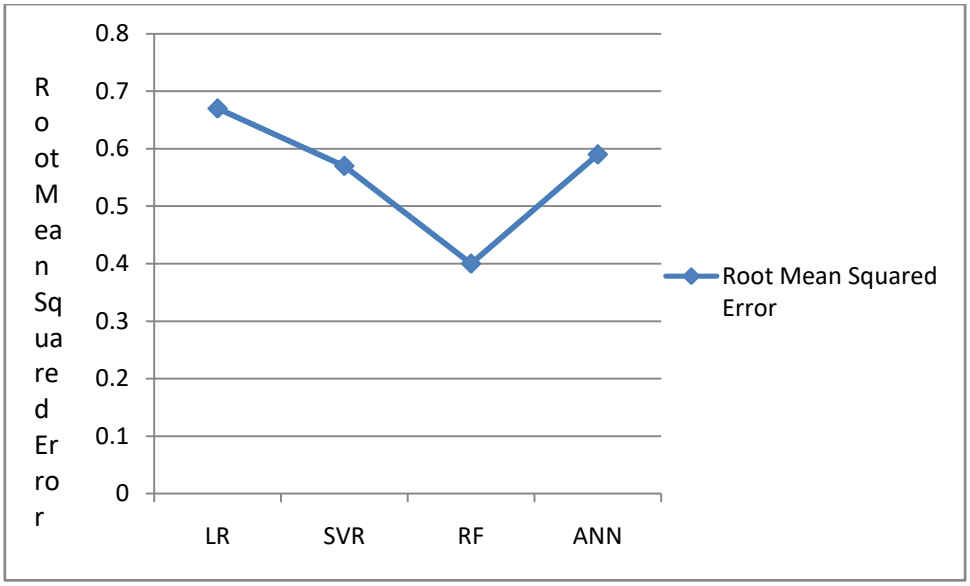

Fig. 5.9 Root Mean Squared Error of Binding Affinity Predictive Models Based on Protein-Mutated-Ligand Docking

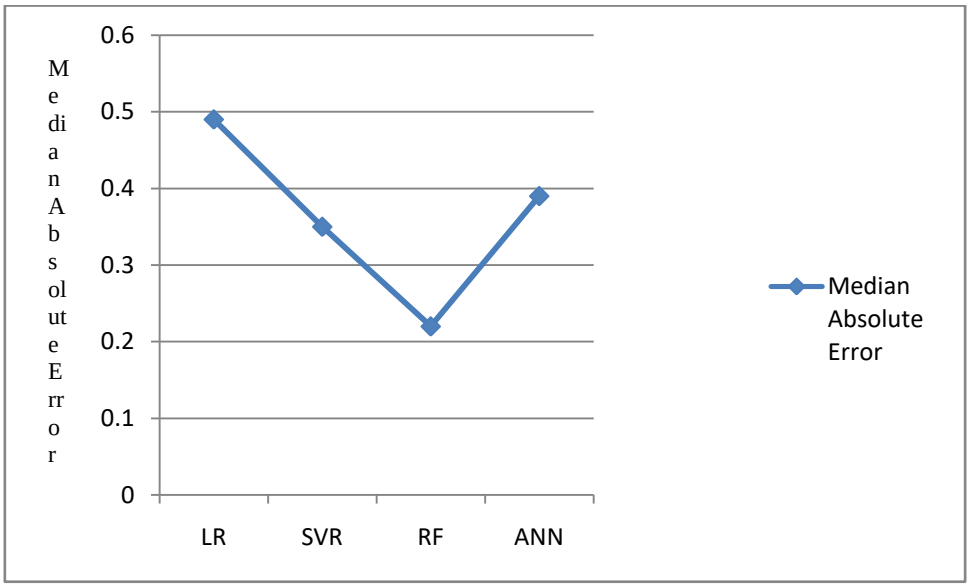

Fig. 5.10 Median Absolute Error of Binding Affinity Predictive Models Based on Protein-Mutated-Ligand Docking

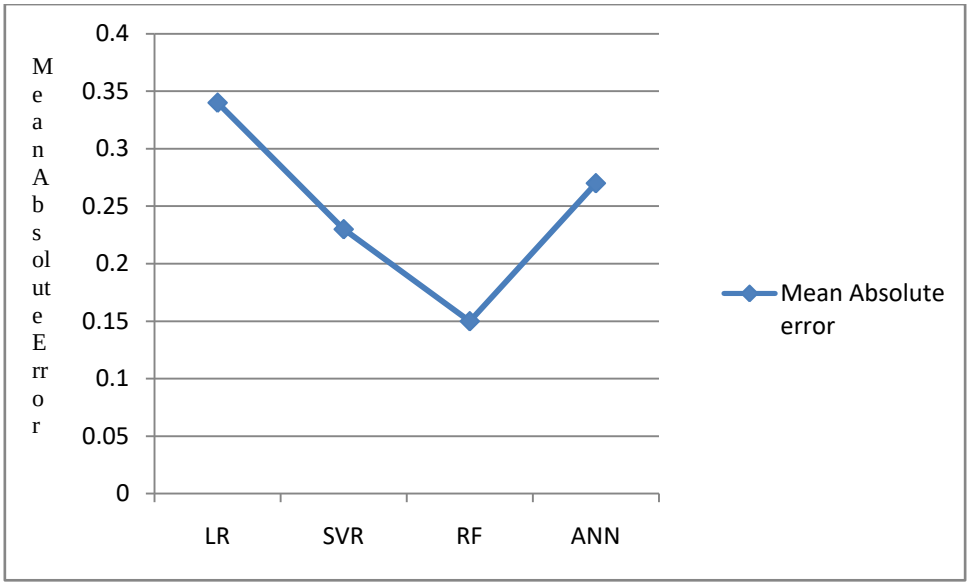

Fig. 5.11 Mean Absolute Error of Binding Affinity Predictive Models Based on Protein-Mutated-Ligand Docking

From Fig. 5.6, it is observed that the predictive model based on random forest achieves higher explained variance score than the other regression algorithms. Fig. 5.7 reveals that the R2 score curve goes superior for random forest and inferior for other predictive models. It is exposed from Fig. 5.8, the random forest predictive model obtains the low error rate compared to the other regression algorithms. From Fig. 5.9 to Fig. 5.11 discloses that the curve for random forest goes inferior in error rate wherein the other regression models achieve higher error rate. This concludes that the predictive model based on random forest outperform other predictive models based on linear regression, support vector regression and artificial neural network.

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

The performance results of predictive models based on protein-mutated-ligand docking is compared with predictive models based on protein-ligand docking and the comparative results are analyzed. The predictive models built through protein-mutated-ligand docking are helpful to capture the changes in protein sequence that occurs due to repeat mutation. While the predictive models built through protein-ligand docking are not mutated to capture the changes. Random forest performed better than other regression algorithms for the predictive models based on protein-ligand docking and protein-mutated-ligand docking. The

comparative results of predictive models based on protein-ligand docking and protein-mutated-ligand docking is presented in Table 5.2.

Table 5.2 Comparative Results of Binding Affinity Prediction Based on Regression Models

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

Table 5.2 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 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 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 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 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 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 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 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. Among the predictive models based on protein-ligand docking and protein-mutated-ligand docking, random forest achieves the highest prediction rate and lower error rate than the other predictive models. The comparative results of predictive models based on protein-ligand docking and protein-mutated-ligand docking are shown in Fig. 5.12.

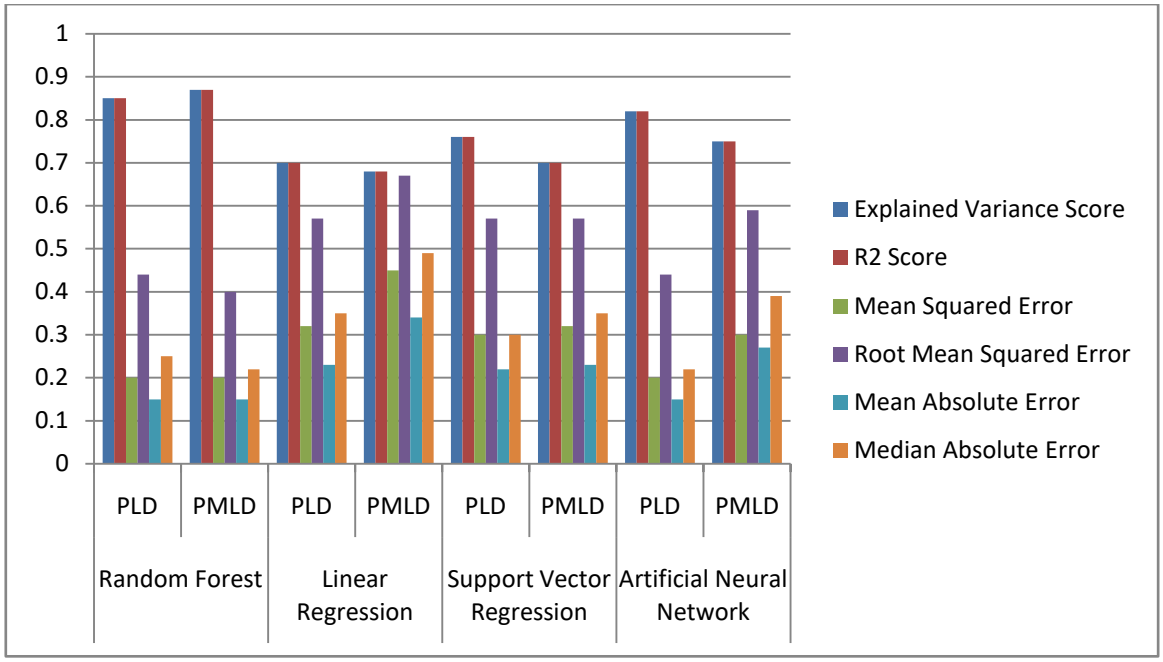

Fig. 5.12 Comparative Results of Binding Affinity Prediction Using Regression

Fig. 5.12 shows that the random forest predictive model based on protein-mutated-ligand docking is high in prediction rate whereas the random forest predictive model based on protein-ligand docking achieves the less prediction rate. Random forest achieves lower error rate for other metrics such as root mean squared error, mean absolute error and median absolute error than the other predictive models. This proves that the random forest predictive model based on protein-mutated-ligand docking achieves high prediction rate as the changes in sequences and structures are captured.

Findings

The experimental results shows that the features extracted from the mutated docked complexes highly contribute in predicting binding affinity. PFI shows the importance of each feature vectors where the binding energy attains the high score. P-value of binding energy with binding affinity shows 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-mutated-ligand docking achieve the highest prediction rate than the predictive models based on protein-ligand docking. Random forest based affinity binding predictive model reveals 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 for all types of mutation where it helps in identifying the drugs for all types of mutation.

SUMMARY

This chapter exemplify the binding affinity predictive models as different regression problems. The changes in protein sequence and structure due to mutation are identified by capturing more features like sequence descriptors and scoring functions during feature extraction and the same has been described. The implementation of various regression techniques executed through protein-mutated-ligand dataset has been described in detail. The experimental results of four predictive models have been reported in detail and the comparative analysis is also presented. The comparison of predictive models based on protein-mutated-ligand docking and predictive models based on protein-ligand docking with respect to various evaluation metrics has been illustrated with tables and charts. The development of binding affinity predictive models based on protein-protein interaction will be discussed in next chapter.

Remarks

The paper titled Affinity Prediction Using Mutated Protein-Ligand Docking with Regression Techniques of SCA, has been published in International Journal of Recent Technology and Engineering, Vol 8, Issue 2, July 2019, pp 3642-3648. **(*Scopus indexed)***