International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Affinity Prediction using Mutated Protein-Ligand Docking with Regression Techniques of SCA

P. R. Asha, M. S. Vijaya

Abstract: ***Drug discovery for rare genetic disorder like spinocerebellar ataxia is very complicated in biomedical research. Numerous approaches are available for drug design in clinical labs, but it is time consuming. There is a need for affinity prediction of spinocerebellar ataxia, which will help in facilitating the drug design. In this work, the proteins are mutated with the information available from HGMD database. The repeat mutations are induced manually, and that mutated proteins are docked with ligand. The model is trained with extricated features such as energy profiles, rf-score, autodock vina scores, cyscore and sequence descriptors. Regression techniques like linear, polynomial, ridge, SVM and neural network regression are implemented. The predictive models are built with various regression techniques and the predictive model implemented with support vector regression is compared with support vector regression kernel. Among all regression techniques, SVR performs well than the other regression models.***

Index Terms: Docking; Kernels; Linear Regression; Ligand; Mutation; Neural Network Regression; Numpy; Polynomial Regression; Regression; Ridge Regression; Scikit learn; Support Vector Regression

I. INTRODUCTION

Spinocerebellar ataxia is rare genetic disorder worsens the development of movement in walking, writing and speech. It is a hereditary disorder and occurs mostly due to repeat mutations. The disorder is because of repeat mutations in the gene that end in intellect and funiculus declination [1]. The spinocerebellar ataxia that occurs due to repeat mutations is spinocerebellar ataxia type1, spinocerebellar ataxia type 2, spinocerebellar ataxia type3, spinocerebellar ataxia type6, spinocerebellar ataxia type7, spinocerebellar ataxia type8 and spinocerebellar ataxia type10 [2]. If the father/mother has thirty-nine repeat of polyglutamine and that number is increased for offspring [3].

Trinucleotide repeats disorders also called as polyglutamine repeats. Polyglutamine repeats are caused by expansion of trinucleotide repeat cystosine-adenine-guanine, which forms the amino acid glutamine. There are number of hereditary disorder that occurs due to polyglutamine repeats. They are cerebellar ataxia, muscular dystrophy, huntington and parkinson's disease. Repeat mutation in DNA sequence changes the function of gene. Due to this change in gene, protein functions get altered. Repeat mutation falls under the category of insertion mutation. The mode of inheritance for this type of mutation is autosomal recessive [4] [5].

Twenty amino acids form a protein. Primary structure contains sequence and secondary structure contains sequence of αhelix and βsheets. Triennial structure has 3d molecular structures. In this work 3d structures are used for docking with ligand. Quaternary structures on the whole, it is connected compound chain into a structure of heaps of sophisticated machine. Once molecular structure forms, post-translational modification is done. Due to post-translation changes and protein misfolding, there occurs a disease called cystic fibrosis [6].

Drug designing for a rare genetic disorder is a challenging task in biomedical field. A drug can be anything like ion, molecule or small compound which activates or inhibits the growth of disease. Binding affinity helps in drug designing because it is necessary to know how much binding score is there for drug designing. Affinity prediction through computational methods is faster than the clinical methods. An application of binding affinity prediction is drug designing. Drug designing of rare genetic disease helps in extension of human lives to a certain extent [7] [8].

In this work, 3d protein structures are mutated with repeat mutation. Mutation information is gathered from HGMD database. The structure gets altered, when mutation is induced [9]. Information of repeat mutation for sca types is delineated in Table 1.

Table 1. Mutational Information of spinocerebellar ataxia associated proteins

Protein Structures	Number of Repeats
Ataxin-1	40-100
Ataxin-2	32-500
Ataxin-3	68-79
Ataxin-6	21-28

Revised Manuscript Received on 30 July 2019.
* Correspondence Author
P. R. Asha*, Department of Comptuer Science, PSGR Krishnammal College for Women, Coimbatore, India.
M. S. Vijaya, Department of Comptuer Science, PSGR Krishnammal College for Women, Coimbatore, India

Ataxin-7	40-200
Ataxin-8	116
Ataxin-10	1611

To execute docking there were two phases available, in which prior the method to dock was lock and key. Then induced fit came into practice, in which the vigorous site of supermolecule is modified by interacting to ligand. In docking there are two types namely flexible docking and rigid docking. Rigid docking needs the binding site of the protein, whereas in flexible docking the ligand moves around the whole protein and finds the binding site [10].

Binding affinity is the strength of attraction between co-ordination bonds with a receptor. Prediction of binding affinity is crucial, which helps in drug designing applications. Drug designing for rare genetic disease is essential. To extend the life of individual, genetic disorders should be cured, for which drug is necessary. Binding affinity can be calculated by docking, with the energy values. The minimum energy is considered as the best score for binding affinity. Docking can be performed with either protein, ion, dna, ligand etc., [11].

II. Literature Review

Xueling Li et al., projected a way for automatic protein-protein affinity binding through svr-ensemble. Two-layer support vector regression is employed to detain binding assistance, square measure serious in the direction of expressly model. The svr ensemble deceives each descriptor rapport draw back, hence robust modelling hypothesis was used. Input selections for TLSVR in 1st stratum square measure innumerable 2209 cooperative atom pairs among every distance bin. Rock bottom SVRs square measure shared by the layer to conclude the last word affinities [12]. Volkan Uslan et al., anticipated the way for significant prediction of HLA-B*2705 compound. The authors projected the prediction of human category I MHC sequence HLA-B*2705 binding affinities pattern international intelligence, before technique, group action and have different were performed. The sequence descriptors of organic compound composition were normalised to [0, 1] to substantiate, each sequence descriptor pictured among constant values. Afterwards, Multi-Cluster Feature technique was used [13].

Background study obviously shows that there is need to facilitate binding affinity prediction using machine techniques. The affinity predictive models are constructed using regression procedures like linear, polynomial, ridge, neural network and support vector regression. The mutated structures be docked through ligand and features are extracted from the docked complexes like energy profile, rf score, cyscore, vina scores and sequence descriptors. Implementation is performed by means of scikit learn.

III. Methodology

In this work, the protein structures are mutated with repeat mutations. The mutational information is taken from the HGMD database. As given in Table 1., the proteins are mutated the number of repeats where aminoacid glutamine occur. The mutations are induced using R script. The sample mutated sequence of structure 1oa8 is given below and the induced mutations are highlighted in red color.

Sequence of 1oa8

SEQUENCEGSPAAAPPTLPPYFMKGSIIQLANGEL
KKVEDLKTEDFIQSAEISNDLKIDSSTVERIEDSHS
PGVAVIQFAVGEHRAQVSVEVLVEYPFFVFGQG
WSSCCPERTSQLFDLPCSKLSVGDVCISLTLKNL

SEQUENCEGSPAAAPPTLPPYFMKGSIIQQQQQQQQQ
QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ
QQQLANGELKKVEDLKTEDFIQSAEISNDLKIDSSTVERIE
DSHSPGVAVIQFAVGEHRAQVSVEVLVEYPFFVFGQGW

Once the mutations are induced, structure changes its formation. The structure activeness is checked using expasy and the mutated structures are docked with ligand. The mutated protein structure is shown in Fig. 1 and Fig. 2. The system architecture is shown in Fig. 3. The work is divided into two phases namely dataset creation and model building.

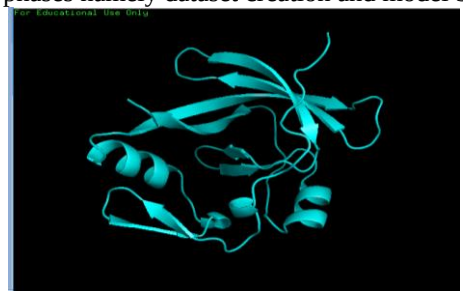

Fig 1. Structure of 1oa8

Fig 2. Structure of 1oa8 with Mutation

Retrieval Number: B1678078219/19©BEIESP
DOI: 10.35940/ijrte.B1678.078219
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

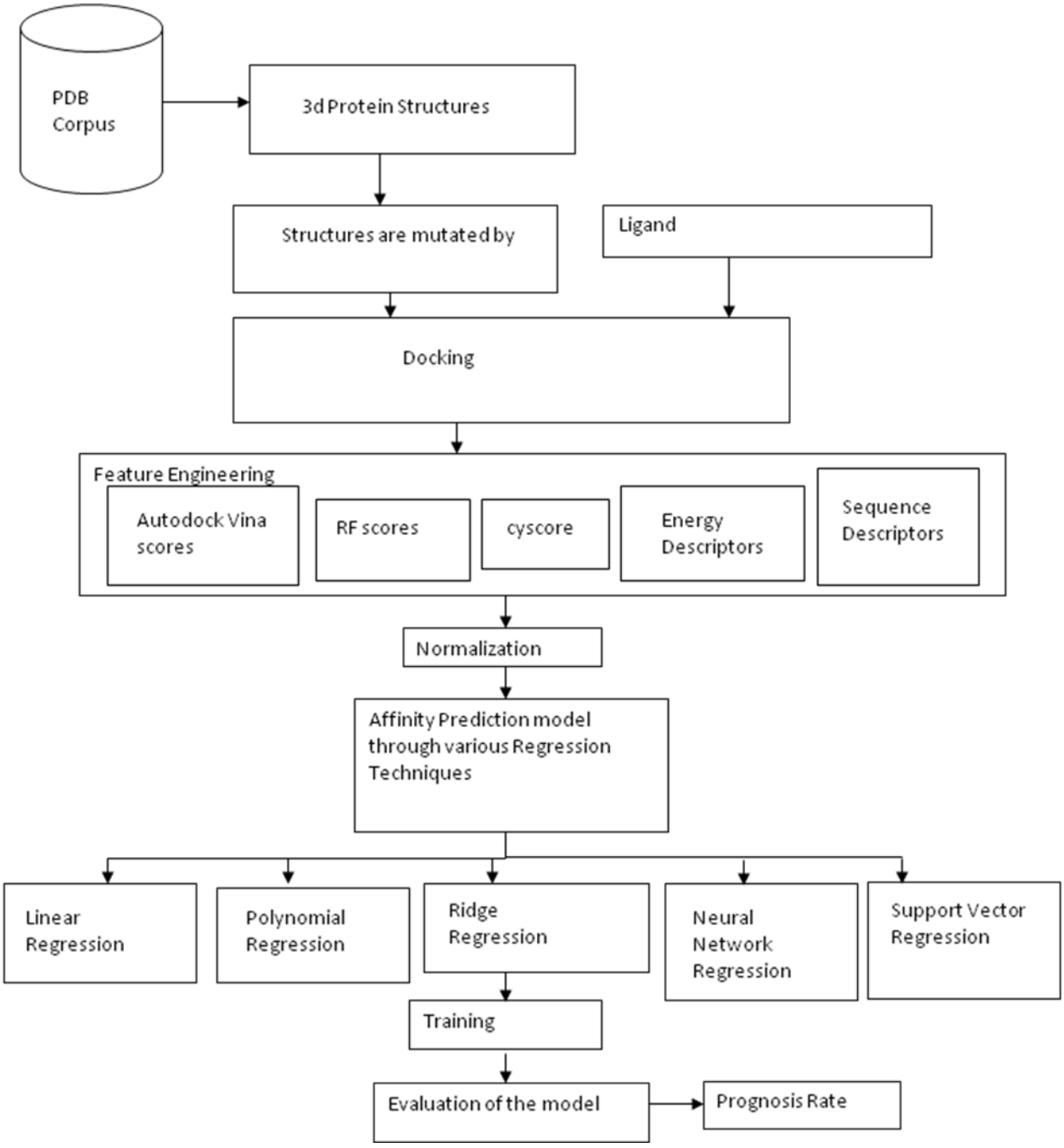

Fig 3. System Architecture

A. Corpus Creation

The dataset in this research work is created by collecting the structures from PDB. The protein structures are mutated with repeat mutation with the information from HGMD database. The mutated structures are docked with ligand. Seventeen structures are docked with eighteen ligands and the features are extracted which will be discussed in model building. Dataset constitutes of 306 instances and the models are built with regression models.

B. Model Building

The independent and dependant variables aids in building the regression models. After the dataset creation features are extracted into five classes namely (1) Energy profiles (2) rf-score (3) Cyscore (4) Sequence descriptors and (5) vina scores.

Energy profiles

Energy profiles are the most important to calculate binding affinity. The author projected the experiment with scoring functions, that isn't ample to predict the affinity [14]. Therefore decisions like inhibition constant, building block energy, vanderwaals, hydrophobic, desolvation, static, total internal and torsional energy helps to create the model. This energy based totally decisions area unit calculated exploitation autodock tool. Energy profiles and its explanation are specified in Table 2.

Table 2. Energy profiles and its Description

Independent Variables	Description
Binding Energy	Affinity of ligand-protein complex
Inhibition constant	Confidence of inhibitor
Intermolecular Energy	Energy between non-bounded atoms
Desolvation Energy	Energy lose of the interaction between substance
Electrostatic Energy	Amendment on the electricity non delimited energy of substance
Total Internal Energy	Change of all energetic terms
Torsional Energy	Dihedral term of internal energy

Rf-score

Rf-score has thirty six decisions, in conjunction with energy based selection benefits in predicting the binding affinity. In [15] the creator, unyielded the significance of scoring functions in affinity forecast. Throughout the work, options encompass the occurrences range of a protein-ligand interacting at intervals, actual distances vary. The most important criteria for the choice of atom varieties is to urge decisions that square measure as dense as come-at-able, whereas considering intense atoms that area unit usually determined in PDB complexes. Therefore, a nominal set of atom varieties was designated by considering selection entirely. Moreover, a smaller set of building block decisions has the extra advantage of resulting in computationally quicker rating functions.

Cyscore

Cyscore is academic degree empirical rating perform consists of four choices. In [16] the paper is explained comparing the cyscore scoring functions with different functions, among that cyscore provides the upper prediction. Cyscore features are extracted from the docked complexes like cyscore, hydrophobic energy, vanderwaals interaction energy, hydrogen-bond interaction energy using linux. Features of cyscore are explained in Table 3.

Table 3. Cyscore and its description

Independent Variables	Description
Hydrophobic free energy	motivating force for folding of bulbous proteins
Cyscore	Scoring function
Van der waals interaction energy	Interactions between two or more atoms
Hydrogen-bond interaction	Bond between two electronegative atoms
Ligand's conformational entropy	change in free energy upon binding

Sequence profiles

Sequence profiles like single aminoacid, double and triple aminoacid composition helps in building model for affinity prediction. In [17] the author used PLS technique for gathering the descriptors for sequence. In this work, sequence descriptors are gathered from the sequences of docked complexes and features are extracted using R. Extracted descriptors are compound composition, autocorrelation, CTD, conjoint Triad, Quasi-sequence-order descriptors, Pseudo compound composition (PseAAC), Profile-based descriptors. Features of sequence descriptors be given in Table 4.

Table 4. Sequence Descriptors and its depiction

Features	Description
Amino acid compositions	peptide informations
Autocorrelation	property correlation between two residues
CTD	predicting protein folding classes
Quasi-sequence-order descriptors	predicts protein subcellular localization
Pseudo amino acid composition	amino composition with additional features of sequence-order information
Profile-based descriptors	protein profile information

Autodock Vina scores

Scores of autodock vina is amalgamation of observed and information marking performs. In moorage, the accuracy and speed with scores, economical improvement and multithreading is improved [18]. The independent variables of vina scores are described in Table 5.

Table 5. Autodock Vina scores and its description

Features	Description
ΔG_{gauss}	Attractive term for dispersion, two Gaussian functions
$\Delta G_{repulsion}$	Square of the distance
$\Delta G_{hydrophobic}$	Ramp function
ΔG_{Hbond}	Ramp function for interactions with metal ions

IV. EXPERIMENT AND RESULTS

Binding affinity predictive models are implemented through algorithms like linear regression, polynomial regression, ridge regression, support vector regression and multilayer perceptron regression. The working of linear regression is explained by the formula. It is about modeling the affiliation between y and x. If the curve is falls below the scale, then the linear model is not good. The liner curve is shown in Fig 4. Polynomial regression is a special case in modeling the mean. The non-linear relationship is fitted between the variable and also the conditional mean. Fig 5 shows the fit for polynomial regression.

Ridge regression is to analyze the multivariate analysis from multiple correlations. Ridge regression performs L2 regularization. The figure for ridge regression is shown in Fig 6.

In MLP regression, three layers are used output layer, hidden layer and input layer. Each and every node is connected with weight that updates in hidden layers for better results.

Dataset is split into training and test sets to build the predictive models and evaluating their performances using 10- fold cross-validation technique. Among all the regression algorithms SVR attains best performance. SVR is compared with the support vector kernel. Regression algorithms are implemented in scikit learn environment. Metrics of all the binding affinity predictive models are given in Table 6 and the Fig.7 is shown for explained variance score and mean squared error of all the regression algorithms. Comparison of support vector regression and support vector kernel is given in Table 7 and the figure for comparison is given in Fig. 8

Table 6. Perfomance measures of Regression models

Regression models	Explained_ variance_ score	Mean_ squared_ error	R2_score	Mean absolute error	Median absolute error	Mean squared logarithmic error
Linear Regression	0.92	2.2	0.92	0.7	0.7	0.059
Polynomial Regression	0.89	2.6	0.89	0.78	0.65	0.056
Ridge Regression	0.93	1.7	0.93	0.8	0.7	0.048
MLP Regressor	0.95	0.56	0.95	0.75	0.6	0.044
SVR Regression	0.99	0.1	0.99	0.6	0.5	0.039

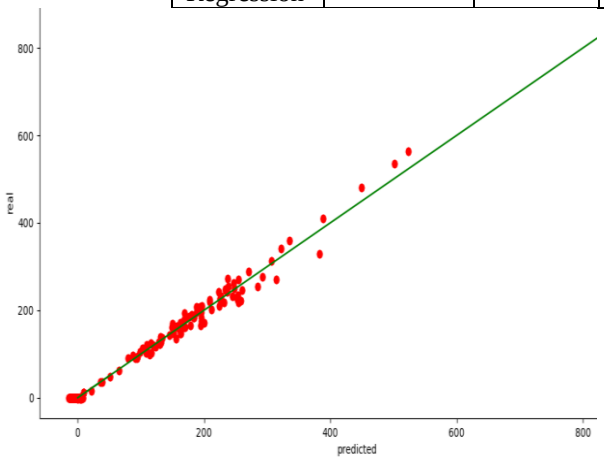

Fig 4. Regression curve of linear model

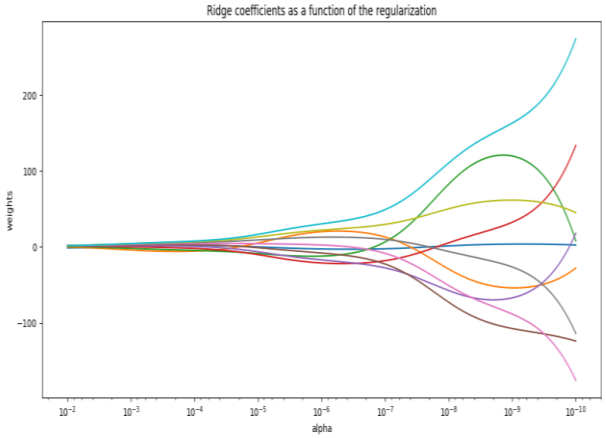

Fig 6. Ridge regression curve

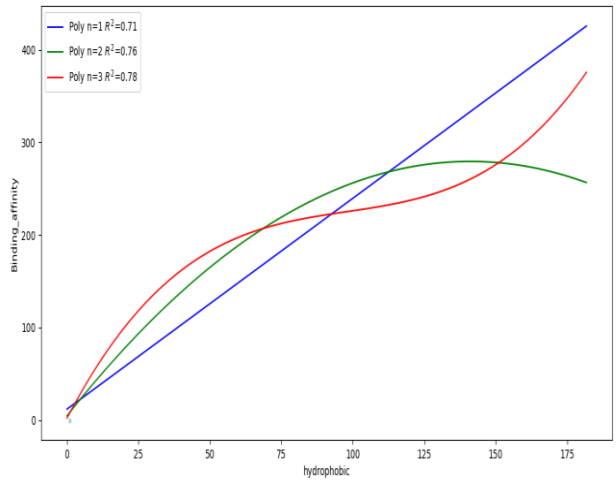

Fig 5. Cubic Polynomial curve

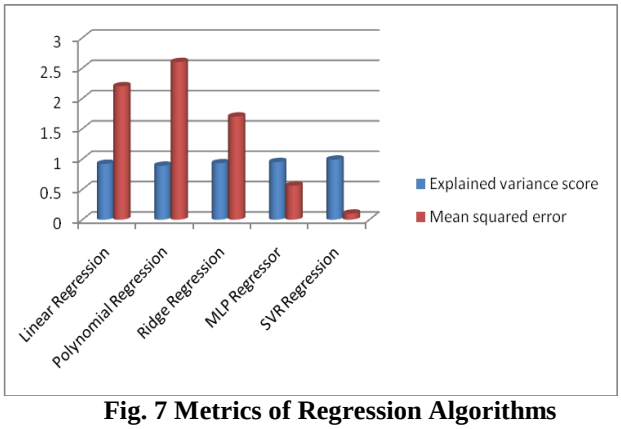

Fig. 7 Metrics of Regression Algorithms

Table 7. Comparison results of SVR Kernel with SVR Regression

Regression models	Explained variance score	Mean squared error	R2_score	Mean absolute error	Median absolute error	Mean squared logarithmic error
SVR Kernel	0.96	0.4	0.96	0.75	0.6	0.044
SVR Regression	0.97	0.1	0.99	0.6	0.5	0.039

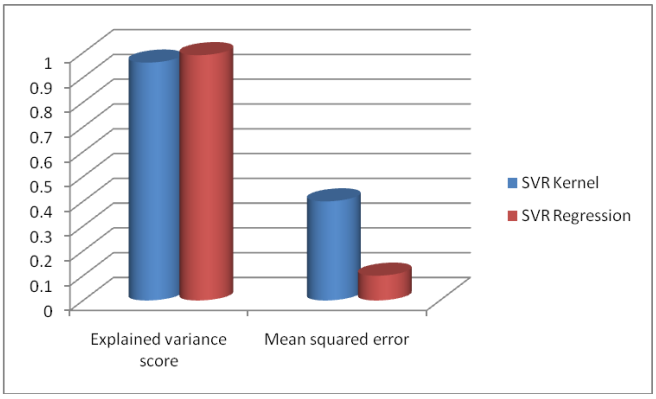

Fig. 8 Comparison of SVR and SVR Kernel

Experimental results illustrate that the Support vector regression outperforms rival regression algorithms and support vector regression kernel. The performance of all the regression algorithms is compared with support vector regression and support vector regression is compared with SVR kernel. The explained_variance score and mean_squared error of both binding affinity predictive models with support vector regression and support vector regression kernel is nearly equal 0.97 and 0.96 of explained variance score respectively.

V. FINDINGS

It is evident that SVR predicts the binding affinity with better results. Features like energy descriptors, rfscore. Cyscore, sequence descriptors and autodock vina scores aids in improving the affinity predictive models. The high recognition rate obtained by support vector regression justify that the model is capable in predicting binding affinity. The framework of support vector regression in shallow learning helps in predicting binding affinity accurately. It is obvious that the model built using support vector regression with mutated protein-ligand docking attains enhanced results for affinity prediction. Explained_variance score is very prominent and the error rate is reduced so the reliability of the model is enhanced. Support vector regression implemented using scikit learn framework, helps in solving the complex models. It is also evident that this learning technique, with svr is appropriate in predicting affinity for spinocerebellar ataxia and also for any rare genetic disorder. This research work proves that model built using support vector regression outperforms the models that are built with other regression algorithms.

VI. CONCLUSION

In this paper, the influence of mutated protein-ligand docking in predicting binding affinity is illustrated through various regression algorithms. The effectiveness of mutated protein-ligand docking in binding affinity prediction is proved by comparing with various regression algorithms. The experimental results confirmed that the support vector regression outperforms other regression models like linear regression, polynomial regression, ridge regression and multilayer perceptron regression. Result analysis provides a baseline for future research, and it can give a better result when using deep learning technologies. In future, the affinity prediction can be done in deep learning by representation learning.

REFERENCES

1. Thomas C. Weiss., Ataxia Spinocerebellar: SCA Facts and Information, Disabled world towards tomorrow, April 4, 2010
2. Thomas D Bird, MD., Hereditary Ataxia Overview, Gene overview, University of Washington, March 3, 2016
3. Whaley NR, Fujioka S, Wszolek ZK, Autosomal dominant cerebellar ataxia type I: a review of the phenotypic and genotypic characteristics, Orphanet journal of rare diseases, May 28, 2011
4. Cynthia T. McMurray, Helen Budworth, A Brief History of Triplet Repeat Diseases, Methods Mol Biol., pp.3-17, 2013
5. Benoit H. Dessailly, Roamin A. Studer, Christine A. Orengo, Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes, Biochemical journal, pp.581-594, january 09, 2013
6. Jacquelyn S. Fetrow, JefferySkolnick, Coupling phenotype to genotype: selecting phages, ribosomes and artificial cells, Trends in Biotechnology, vol 18, Issue 1, pp.34-39, January 2000
7. Choi S, Macalino SJ, Gosu V, Hong S, Role of computer-aided drug design in modern drug discovery, Archives of pharmacal research, Volume 38, Issue 9, pp 1686–1701, September 2015,
8. L MarioAmze, Structure Based Drug Design, Current Opinion in Biotechnology, Vol 9, Issue 4, pp. 366-369, August 1998,
9. Koshland DE., Jr., Correlation of Structure and Function in Enzyme Action, Science.142:1533–1541, 1963
10. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule-ligand interactions, Journal of molecular biology, vol 161(2), pp 269-288, 1982
11. http://chemistry.tutorvista.com/inorganic-chemistry/binding-affinity.html
12. Wang S.,Li X., Zhu M., Li X., Wang HQ., Protein-Protein Binding Affinity Prediction Based on an SVR Ensemble, Lecture Notes in Computer Science, vol 7389, 2010
13. Huseyin Seker, Volkan Uslan, "Binding affinity prediction of S. cerevisiae 14-3-3 and GYF peptide-recognition domains using support vector regression", Engineering in Medicine and Biology Society (EMBC) 2016 IEEE 38th Annual International Conference of the, pp. 3445-3448, 2016, ISSN 1558-4615
14. Garrett M. Morris, David S. Goodsell, Micheal E. Pique, Ruth Huey, Automated Docking of Flexible Ligands to Flexible Receptors, journal of molecular recognition, November 2012

15. Martin YC, Muegge I, A General and Fast Scoring Function for Protein-Ligand Interactions: A Simplified Potential Approach., Journal of Medicine in Chemistry, vol 42, pp 791–804, 1999
16. Lei Li, Yang Cao, Improved protein–ligand binding affinity prediction by using a curvature-dependent surface-area model, structural bioinformatics, vol 30, pp-1674-1680, 2014
17. Mark J. Embrechts, Wei Deng, Curt Breneman, Predicting Protein−Ligand Binding Affinities Using Novel Geometrical Descriptors and Machine-Learning Methods, journal of chemical information and modeling, pp.699-703, 2004
18. Arthur J. Olson, Oleg Trott, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading, Journal of Computational Chemistry, January 30; 31(2): pp no 455–461, 2010

AUTHORS PROFILE

P. R. Asha, Ph. D Research Scholar,
Completed M.Sc[SE] in Bannari Amman Institute, M.phi in krishnammal college for women and currently pursuing my Ph.D in krishnammal college for women.

Areas of Interest: Computational biology, data mining and statistics. I'm interested in finding new drugs for the disease which has no drug to cure the disease. Pursuing research work in predicting binding affinity for spinocerebellar ataxia. Four papers has been presented and published in conference proceedings. Paper published in **JARDCS** journal entitled "Deep Neural Networks for Affinity Prediction of Spinocerebellar Ataxia Using Protein Structures".

M. S. Vijaya, Associate Professor
Completed masters in PSG college of Technology and did Ph.D in Amirta university Coimbatore and currently working as associate professor and head in krishnammal college for women.
Area of Specialization: Data Mining, Machine Learning, Support Vector Machine, Pattern Recognition. She has guided many M.phil research scholars and many Ph.D scholars, where one Ph.D research scholar completed under her guidance. She has published many papers in various international journals and conferences.