ABSTRACT

Spinocerebellar Ataxia (SCA) is a disorder mainly occurs due to hereditary mechanism. SCA is characterized by mutations in the gene, chromosomal conditions, birth factors and is carry forwarded to the translation process. Mutations are of many types where SCA mainly occurs due to repeat mutation. Mutations are carry forwarded such that the protein structures are unable to function properly due to change in the structure of the protein. This protein structure is docked with ligands to determine the binding sites and interaction properties. Binding affinity is the strength of measure between the molecules and small compounds. It is imperative to predict binding affinity as the functions and interaction properties of proteins with ligands and protein-protein interaction can be known. The protein structures are required to dock with small compounds or macromolecules in order to predict binding affinity. Binding affinity prediction is significant as it aids in drug designing process. Accurate prediction of binding affinity is a complex task as the number of repeats varies for each type of SCA and it is a challenge in biomedical field.

This research titled 'Binding Affinity Prediction Models for Spinocerebellar Ataxia Disorder Using Deep Neural Network Architectures' aims at predicting binding affinity by developing models using hand crafted features with contemporary machine learning and representation learning with deep learning approaches. The core objectives of this research are as follows.

- To create three corpuses using protein-ligand docking, protein-mutated-ligand docking and protein-protein interaction as there is no readily available corpus for human SCA
- To identify and capture the discriminative features from the docked and interacted complexes to monitor and analyze the structural changes due to mutation
- To develop synthetic datasets related to three corpuses protein-ligand corpus, protein-mutated-ligand corpus, protein-protein interaction corpus to construct efficient binding affinity predictive models
- To develop the general framework based on traditional machine learning approach to improve the generalisation capability of predictive models for prediction of binding affinity
- To develop framework using deep architectures such as sequential deep neural network, functional deep neural network and deep neural network with customized layer to improve the prediction rate of binding affinity predictive models

The thesis explains a novel approach where the problem of predicting binding affinity for SCA is formulated as regression task and solved using machine learning and deep learning methods. The predictive models are proposed to build using these approaches by training the significant features collected from the simulated corpuses to predict binding affinity.

Three corpuses are created by gathering 17 protein structures from Protein Data Bank (PDB) and 18 ligands from genecards. The first corpus is developed by collecting the protein structures from PDB for six types of SCA, which are affected with repeat mutation. Ligands are collected from various literatures and gene cards. The protein structures and ligands are docked with each other and Protein-Ligand (PL) corpus with 307 docked complexes is created. The second corpus is developed by mutating the protein structure with the information available from Human Gene Mutational Database (HGMD). Protein structures are mutated to analyze the changes in the physical and chemical properties. The mutated protein structures are validated and docked with ligand which produced 307 docked complexes and Protein-Mutated-Ligand (PML) corpus is created. The third corpus is developed by using 17 protein structures with 609 interacting proteins gathered from genecards using High Ambiguity Driven protein-protein DOCKing (HADDOCK). The 313 interacted complexes are derived and Protein-Protein (PP) corpus is developed.

Three independent datasets are developed from the corresponding corpuses by identifying and capturing indicative features. The first dataset is created by defining features such as energy calculations and physical properties which are extracted from the 307 docked complexes and the dataset called Protein Ligand Dataset (PLD) is created. The second dataset is urbanized by defining the features such as energy calculations, sequence descriptors, scoring functions and extracting from the 307 mutated docked complexes. This dataset is named as Protein Mutated Ligand Dataset (PMLD). The third dataset is generated by defining the features such as energy calculations, interfacial contacts, physio-chemical properties, Non-Interacting Surface (NIS) properties and extracting from the 313 interacted complexes. This dataset is named as Protein-protein Dataset (PPD). PLD dataset is created with 307 instances and 27 attributes whereas PMLD is developed with 509 dimensions with 307 instances. The dataset PPD has 56 attributes with 313 instances. Min-max normalization is performed to normalize the feature values.

The research work is carried out in two stages (i) using conventional machine learning and using (ii) contemporary deep learning methods for building the predictive models.

In the first stage, the traditional machine learning approaches are utilized for building the predictive models. Various experiments are carried out by implementing regression algorithms such as Linear Regression (LR), Support Vector Regression (SVR), Random Forest (RF) and Artificial Neural Network (ANN) using above three datasets PLD, PMLD, PPD in scikit learn environment using python.

In the second stage, the deep learning approaches are employed for building the predictive models and the key idea here is the representation learning from the user defined features. The deep neural network learns the hand crafted features, interactions among them, extracts the features on its own that facilitates in prediction of binding affinity accurately than machine learning techniques. Here the user defined features are fed to the Deep Neural Network (DNN) architectures which extract new feature set by representation learning. The deep neural network architectures such as sequential deep neural network, functional deep neural network and DNN with customized layers are utilized for building predictive models. Various experiments have been carried out by implementing three deep neural network architectures with three optimizers adam, RMSprop, Nadam using the same three datasets. Hyperparameters such as epochs, dropouts, learning rate, loss function are fine tuned to acquire the better prediction rate.

The performance of the predictive models is evaluated using 10 fold cross validation and their efficiency in predicting binding affinity is analyzed with various metrics such as explained variance score, mean squared error, root mean squared error, mean absolute error, median absolute error, R2 score and p-value. The experimental results illustrates that the predictive model developed using DNN with customized layers and adam optimizer is efficient in predicting the binding affinity than the other two deep models such as sequential DNN and functional DNN. Binding affinity predictive models based on protein-protein interaction achieves high prediction rate as protein-protein interaction is helpful in capturing the interfacial contacts than protein-ligand docking and protein-mutated-ligand docking.

The comparative analysis of DNN based predictive models and regression models is performed and analyzed using various evaluation metrics. The comparative results prove that the predictive model built using DNN with customized layers achieved the highest prediction rate.

Accurate prediction of binding affinity for SCA with mutation induced protein structure is a challenge in biomedical domain and hence an attempt is made to carry out this research in this domain.