1. INTRODUCTION

Data mining has been the origin for solving many pattern recognition problem and caused remarkable changes in research. Data mining is the process of extracting concealed, suitable and constructive patterns from enormous data sets. Data mining is also known as knowledge discovery, knowledge extraction, data/pattern analysis, information harvesting etc. The basic forms of data for mining are database data and other forms of data are data streams, sequence data, graph/networked data, web page data, spatial data and text data. Data mining has become significant and contributed enormously towards the fields like bioinformatics, biomedical, healthcare, retail and telecommunications. It is one of the major frontiers that facilitates in solving many problems from unknown data in all the fields mainly in healthcare, pharmaceuticals and biomedical fields. Data mining assists scientists in processing huge amount of data. Data mining tools are widely used in statistics and health care sectors as data processing is wider in these fields. It is necessary for health care sector to predict number of rare diseases/disorders through various techniques that are available in data mining and mainly aids in providing easier approaches compared to clinical laboratories.

Bioinformatics is a field in which data mining is widely used nowadays where there are lots of data to be processed. There are many ways in which biological data can be processed like genome sequencing, pattern matching, sequence analysis, homology modelling etc. Genome sequencing and homology modelling were the first progress in genomics and proteomics. Genomics involves study about sequence analysis, genome sequencing, synthetic gene dataset creation, microarrays etc., and proteomics involves homology modelling, docking, protein folding, structural predictions, protein function prediction etc. In recent years, studies about genes and proteins are increased to reduce and identify the rare disorders. Nowadays gene stopping technique is used to stop the genetic disorders for unborn children in the womb. Data mining plays an effective role in understanding about gene patterns and protein structures which aids in developing drugs for rare diseases and disorders. Applications of data mining in bioinformatics include drug designing, cell regulation, signal transduction, protein function analysis, disease identification, interaction of gene, protein and other small channels.

This research work focuses on developing predictive models to predict binding affinity for spinocerebellar ataxia which aids in drug designing applications. The models are built with supervised machine learning and contemporary deep learning approaches of data mining. Brief introduction about data mining is presented in this chapter. An overview of

spinocerebellar ataxia including the mutation types, proteins associated with genes and the mutational information are also discussed. The three dimensional protein structures of spinocerebellar ataxia, ligands to dock with protein structures, docking mechanisms and binding affinity prediction are also elaborated in this chapter. The detailed literature survey and motivation for this study is highlighted. The problem statement and objectives of this research are also stated clearly in this chapter.

1.1 DATA MINING

Data mining is a procedure of determining patterns in large data sets and it is a multi disciplinary skill which uses machine learning, artificial intelligence, statistics and database systems. Data mining is also called as knowledge discovery process. The types of data that can be mined are flat files, relational database, time series database, data warehouse, transactional database, multimedia database, world wide web database etc., The process of data mining includes requirement gathering, exploring the data, data preparation, transforming the data, data modelling, performance evaluation and deployment. The first step is to analyse the objectives of the problem and the problem statement should be well defined. The data should be explored after completion of the objective definition. In the data exploration phase, experts gather data according to the problem from multiple sources. Experts understand the problem, challenges and convert into meta data. Data preparation phase includes cleaning of data, transformation and formatting the data. Transformation of data comprises of smoothing, aggregation, normalization etc., and the result data can be used for modelling. Tools are applied in modelling and also the methods like algorithms are used and the results are evaluated. In the evaluation process, performance of the model is analyzed and if the results are not satisfied then it goes back to modelling process. The results are analyzed at the end and it is deployed in spreadsheets as a final stage [1].

Data mining techniques such as classification, clustering, prediction, outlier detection, sequential patterns, association rules are essential to build models and develop applications. Classification is a data mining function that assigns items in a collection to target classes or categories and the task of classification is to predict the target class for each case in dataset. Clustering is the process of making a group of abstract objects into classes of similar objects. Regression is a data mining technique in which the relationship between variables can be known. Association rule technique helps to identify the association between the two or more items. Outer detection is the technique that refers to identify the observations of data in the dataset that do not match the expected pattern or behaviour. This type of technique can be

used in intrusion detection, fraud detection etc., Sequential patterns helps to discover the similar patterns in data. Prediction is the combination of other data mining techniques like clustering, classification, sequential patterns etc., which analyzes past event to predict future event [2].

Data mining can be used for research, surveys, information collection, opinion of customers, data scanning, extracting the information, pre-processing of data, web data etc., Importance of data mining are it is beneficial to lot of companies to meet the client objectives in business, reduces cost and time in many organisations, helps research people to meet their challenges, plays major role in health care sectors, predicts the customer loyalty etc. The benefits of using data mining are as follows.

- Data mining technique helps companies to get knowledge-based information
- Data mining helps organizations to make the profitable adjustments in operation and production
- The data mining is a cost-effective and efficient solution compared to other statistical data applications
- Data mining helps with the decision-making process
- Facilitates automated prediction of trends and behaviours as well as automated discovery of hidden patterns
- It can be implemented in new systems as well as existing platforms
- It is the speedy process which makes it easy for the users to analyze huge amount of data in less time

Tasks of Data Mining

The major tasks of data mining are classification, regression, prediction, time series analysis, association, clustering, summarization. All these tasks are either (a) predictive data mining tasks or (b) descriptive data mining tasks [3] as shown in Fig. 1.1.

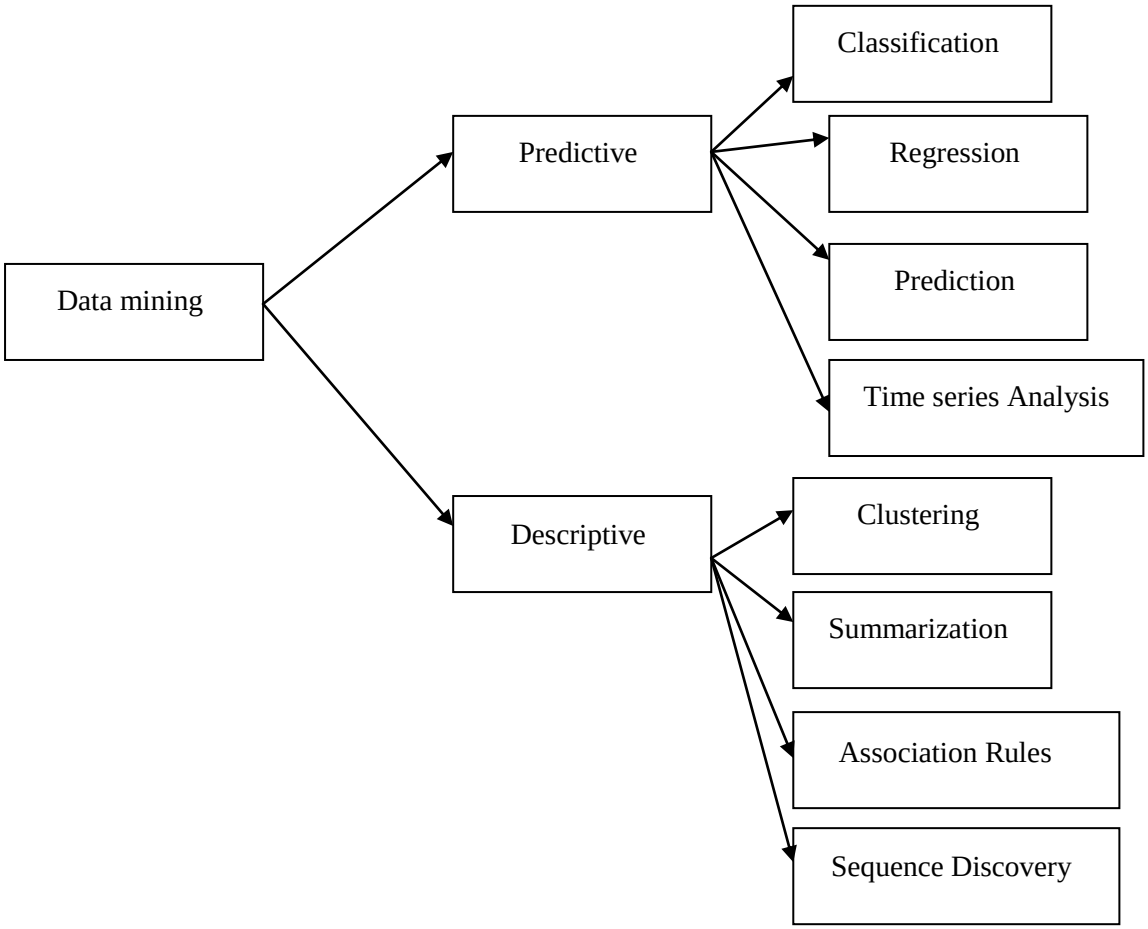


Fig. 1.1 Tasks of Data Mining

a) Predictive Tasks

Predictive analysis in data mining helps in predicting the data mainly in business intelligence and research sectors to forecast the data to make better decisions. Analysis is performed with mathematical algorithms and machine learning where the best evaluation occurs in forecasting. Forecasting aids in fine tuning the models and to achieve the better prediction. Predictive analysis works by using the sample data with known attributes and the model is trained with sample data. The training helps in analyzing unknown data and determines its behaviour. As predictive analysis are mainly used in business sectors it offers valuable insight, increase competence edge and predict trends [4]. Predictive analysis is further narrowed down into classification, prediction, regression and time series analysis.

Classification

Classification is a data mining function that assigns data in a collection to target classes. The main aim of classification is to precisely predict the target class for each case in data. For

example, if a patient is affected with cancer it classifies as mild, severe and low. Machine learning has many algorithms for implementation of classification problems. The types of classification algorithms are naïve bayes classifier, nearest neighbour, boosted trees, support vector machines, decision tress and random forest etc [5]. Classification predicts the categorical values not continuous values. Nowadays classification along with machine learning techniques is mainly used in health sectors for classifying and predicting the types of disease and predicting the disease respectively. The classification task includes three main components like attribute set, classification method and class label.

Attribute set consists of data stored in a particular format and it should be relevant to problems taken for consideration. Classification method describes about the algorithms and before passing the data to classification methods, data should be prepared with some kind of procedures like data cleaning, relevant analyzation of data and data transformation. In data cleaning, the data is pre processed by removing noise and filling missing values by noise removal and filtering techniques and the data is classified using machine learning algorithms and the evaluation results obtained with the class label [6].

Prediction

Prediction in data mining is to identify data points purely on the description of another related data value. It is not necessarily related to future events but the used variables are unknown. Prediction derives the relationship between a thing you know and a thing you need to predict for future reference. Some of the algorithms used in data mining for prediction are naïve bayes, support vector machine (SVM), k-nearest neighbour, decision tree and artificial neural network etc [7]. The difference between regression and prediction are regression is a technique that identifies the relationship between variables whereas in prediction it combines the other data mining techniques and predicts the future with the past data [8].

Regression

Regression in data mining is used to predict numerical values or continuous values. In statistical analysis through regression the relationship of independent variable and dependant variables can be established. The common types of regression in statistics are linear regression, logistic regression, polynomial regression, stepwise regression, ridge regression, lasso regression, multiple regression and multivariate regression. The commonly used regression in business and financial sectors is linear and logistic regression. The simple regression is linear regression where it has one predictor variable and the values can be mapped in two dimensional space. The predictor values are plotted along x axis and the

prediction values are plotted along y axis. Logistic regression is used when the dependant variable is binary. For example if the variable is either 0 or 1, pass or fail etc. and the relationship is explained between one binary dependant variable and one or more nominal variables [9]. Polynomial regression is a form of linear regression where the relationship between independent and dependant variable will be in the form of n^{th} degree polynomial. Stepwise regression is a choice of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. Lasso regression is a regression that uses shrinkage where the data values are shrunk to a central point like mean [10]. Regression algorithms such as linear regression, random forest, artificial neural network, support vector regression etc., are discussed in detail in chapter 2.

Time Series Analysis

Time series is a sequence of well-defined data points measures at consistent time intervals over a period of time. Time series analysis is the use of statistical methods to analyze time series data and extract meaningful statistics and characteristics about the data. Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values [11]. Some of the tasks in time series data mining are segmentation, clustering, query by content, anamoly detection etc. Query by content is the most active area of research in time series analysis. It is based on retrieving a set of solutions that are most similar to a query provided by the use.

b) Descriptive Tasks

The descriptive analysis is used to mine data and provide the latest information on past or recent events. The descriptive function deals with general properties of data in the database. The list of descriptive functions are class or concept description, mining of frequent patterns, mining of associations, mining of correlations and mining of clusters. Class or concepts refers the data to be associated with classes or concepts. These descriptions can be derived by data characterization and data discrimination. Frequent patterns occur frequently in transactional data. The list of frequent patterns is frequent item set, frequent subsequence and frequent substructure. Associations are used in retail sales to identify patterns that are frequently purchased together. Mining of correlation is performed to uncover statistical correlations between associated-attribute-value pairs. Mining of clusters refer to grouping of similar kind of objects. The difference between descriptive and predictive tasks is descriptive analysis uses data aggregation and data mining techniques with data to give the future but

predictive analysis uses the statistical analysis and forecast techniques to predict the future [12].

Clustering

Clustering in data mining is used to partition the data in the same class. Clustering is a type of unsupervised learning, where the name of the class label is not known. It maps the data into one of the multiple clusters where the arrangement of the data items relies on the similarities between them. Clustering can be divided into hard and soft clustering. Each data point either belongs to a cluster completely or not is known as hard clustering. In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned. The goal of clustering is to reduce the data by segregating the data into similar groups and assigning them into clusters. There are many types of clustering algorithms like partioning methods, density based clustering, hierarchical clustering, k means clustering etc [13].

K-means clustering is an iterative clustering algorithm which aims in finding local minima in each iteration. In this algorithm, k will be assigned a random number where each data point will be assigned to a cluster and cluster centroids are computed. Data points will be re-assigned to the closest centroid and the cluster centroid is calculated. This step is repeated, until the local minima are achieved. Hierarchical clustering builds the hierarchy of the clusters. This algorithm starts with all the data points assigned to a cluster of their own and the two nearest clusters are merged into the same cluster. The algorithm is terminated when there is single cluster left. The hierarchical clustering results can be showed using dendogram. The difference between k-means and hierarchical clustering are k-means handles big data while hierarchical cannot handle big data. K-means require prior knowledge of k whereas in hierarchical the number of clusters can be stopped when it is appropriate by interpreting the dendogram. Clustering is used in areas like medical imaging, social network analysis, image segmentation, anomaly detection etc [14].

Summarization

Summarization is a term for a short conclusion of a big theory and it is a key data mining concept which involves techniques for finding a compact description of a dataset. Simple summarization methods such as tabulating the mean and standard deviations are often applied in exploratory data analysis, data visualization and automated report generation. The data summarization is necessary to simplify the data and describe the normal and odd data. The distribution of the variable shows the values taken by variables and how often the

variable takes these values. There are two ways to describe the distribution of data they are typical distribution and the spread of the values around the center. The typical distribution of data describes the center of the data and this way of describing the center is known as measure of central tendency. The spread of the values around the center describes how densly the data is distributed around the center and it is also called as measure of dispersion. To summarize the data graphical displays are used where dichotomous and non-ordered categorical variables are best summarized with bar charts [15].

Association Rule Mining

Association rule mining is the process of finding the rules that governs associations and causal objects between sets of items. In a given transaction with multiple items, it tries to find the rules that govern how or why such items are often bought together. It helps to find patterns in a data where it finds features that occur together and features that are correlated. Association rules are useful for analyzing and predicting customer behavior. They play an important part in customer analytics, market basket analysis, product clustering, catalog design and store layout. The algorithms used in association rule mining are Apriori algorithm (AIS), stem, apriori, apriori hybrid, fp-growth etc. In association rule mining there are two patterns like if and then. An if pattern is something that is found in the data and then pattern is the combination of if pattern. Depending on the two parameters such as support and confidence, the relationship can be established. Support indicates how frequently the if/then relationship appears in the database. Confidence tells about the number of times these relationships have been found to be true. This mining can be used in the areas like market basket analysis, medical diagnosis, protein sequence and census data [16].

Sequence Discovery

Sequential pattern mining is specialized for analyzing sequential data, to discover sequential patterns. It is used to find sub sequences that appear often in a sequence database. In a sequence database, each sequence represents the items purchased by a customer at different times and sequence is an ordered list of item sets whereas pattern is known as items that occur frequently in a dataset. Two types of sequential pattern mining are time series and sequences. Time series is an ordered list of numbers, while sequences are ordered list of nominal values. The algorithms that are used for sequence pattern mining are aprioriall and Generalized Sequential Pattern (GSP) where these two algorithms are inspired from apriori. The GSP algorithm performs a level-wise search to discover frequent sequential patterns. It first scans the database to calculate the support of all 1-sequences. GSP algorithm

has several limitations and they are multiple database scans, non-existent candidates and maintaining candidates in memory. The three main concise of sequential pattern mining are closed, maximal and generator sequential patterns. Applications of sequential data mining in a variety of domains like healthcare, education, Web usage mining, text mining, bioinformatics, telecommunications, intrusion detection [17].

Applications of Data Mining

Data mining has been used in many fields like healthcare, telecommunications, marketing, fraud detection, banking, intrusion detection, finance, education, healthcare, insurance, research analysis, bioinformatics, medicine, agriculture, customer relationship management (CRM) etc., [18]. Most of the real time problems in each domain can be converted into either predictive or descriptive tasks and solved using data mining techniques. Some of the application areas are described below in brief.

Data Mining in Research

Researchers use various data mining techniques to solve the problem and to analyze the result. Data mining is used to clean the data, transform and format the data from any databases. Researchers may take the similar data that might bring the change in the research. They use data mining approaches like machine learning, deep learning and statistics for modelling and evaluation. The techniques and approaches of data mining for researchers aids in achieving some targets in their field of research.

Data Mining in Education

There is a new emerging field, called Educational Data Mining (EDM), concerns with developing methods that discover knowledge from data originating from educational environments. The goals of EDM are identified as predicting student's future learning behaviour, studying the effects of educational support, and advancing scientific knowledge about learning. Data mining can be used by an institution to take accurate decisions and also to predict the results of the student. With the results the institution can focus on what to teach and how to teach. Learning pattern of the students can be captured and used to develop techniques to teach them.

Data Mining in HealthCare

Data mining holds great potential to improve health systems. It uses data and analytics to identify best practices that improve care and reduce costs. Researchers use data mining approaches like multi-dimensional databases, machine learning, soft computing, data

visualization and statistics. Mining can be used to predict the volume of patients in every category. Processes are developed that make sure that the patients receive appropriate care at the right place and at the right time. Data mining can also help healthcare insurers to detect fraud and abuse. The main challenge in health care industries is the application of drug designing for the hereditary disorders. Data mining aids in accurate prediction of binding affinity that leads for the development of drug designing. Maintenance of health records for each process like pharmacy, health records etc., separate system is maintained. Electronic health record is integrating all those processes in a single system. The other challenges are genome sequencing, docking etc.

Data Mining in Banking

The computerised banking with huge amount of data is supposed to be generated with new transactions. Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices that are not immediately apparent to managers because the volume data is too large or is generated too quickly to screen by experts. The managers may find this information for better segmenting, targeting, acquiring, retaining and maintaining a profitable customer. Data mining helps in solving business problems by finding patterns, associations and correlations which are hidden in the business information stored in the data bases. By using data mining in analyzing patterns and trends, bank executives can predict how customers will react to adjustments in interest rates, which customers will be likely to accept new product offers etc.

Data Mining in Fraud Detection

Billions of dollars have been lost to the action of frauds. Traditional methods of fraud detection are time consuming and complex. Data mining aids in providing meaningful patterns and turning data into information. Any information that is valid and useful is knowledge. A perfect fraud detection system should protect information of all the users. A supervised method includes collection of sample records. These records are classified fraudulent or non-fraudulent. A model is built using this data and the algorithm is made to identify whether the record is fraudulent or not.

Data Mining in CRM

Customer Relationship Management is all about acquiring and retaining customers, also improving customers' loyalty and implementing customer focused strategies. To maintain a proper relationship with a customer a business need to collect data and analyse the

information. With data mining technologies the collected data can be used for analysis. Instead of being confused where to focus to retain customer, the seekers for the solution get filtered results.

Data Mining in Bioinformatics

Nowadays in healthcare sector data mining is being used for finding and predicting diseases and drug designing purpose. In health care sector, bioinformatics field plays major role in prediction and classification of many rare diseases. Biological data analysis is the important part in bioinformatics. Applications of data mining in bioinformatics include gene finding, protein function detection, protein function interference, docking of protein with substances, disease diagnosis, disease prognosis, disease treatment optimization, protein misfoldings, protein and gene interaction network, micro-array analysis etc.

Data mining techniques along with machine learning and deep learning algorithms helps in achieving the targets in bioinformatics field. There are many databases available for bioinformatics like PDB, SWISS-PROT, Medical Literature Analysis and Retrieval System Online (MEDLINE), genecards, National Center for Biotechnology Information (NCBI) etc. There are some processes in which data mining contributes for biological data analysis [19] and few are mentioned below.

- Semantic integration of heterogeneous, distributed genomic and proteomic databases
- Alignment, indexing, similarity search
- comparative analysis multiple nucleotide sequences
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis
- Visualization tools in genetic data analysis
- Alignment, indexing, similarity search
- comparative analysis multiple nucleotide sequences
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis
- Visualization in genetic data analysis

1.2 OVERVIEW OF SPINOCEREBELLAR ATAXIA AND BINDING AFFINITY

Spinocerebellar ataxia is a hereditary and rare genetic disorder that is characterized by progressive incoordination in gait and associated with poor hand movements, speech, and eye movements. There is no effective treatment to cure this disorder and it can affect people of

any age. SCA can be caused through any mode of inheritance and till now 36 types of spinocerebellar ataxia has been discovered. This disorder can be diagnosed only through gene testing. Drugs available for this disorder can prolong the symptoms and does not cure the disease. SCA results in atrophy of the cerebellum which leads to loss of fine coordination of muscle movements leading to unsteady and clumsy motion. Rarely people from India get affected with common types of SCA. SCA can occur through inheritance such as autosomal dominant, autosomal recessive and X-linked [20].

Autosomal Dominant

In this mode of inheritance, disease can pass through anyone affected parent either father or mother. The abnormal gene can be located on any one of the first 22 pairs of chromosome and one only copy of the affected gene is enough to pass the disease. The mode of inheritance of autosomal dominant is shown in Fig. 1.2. When the father is affected, then there is a chance of passing 50% abnormal gene and 50% normal gene during pregnancy. So if either father or mother is affected with abnormal gene, it is mandatory to test the offspring for the disease occurrence. Dominant mutations can occur in an individual without family history called spontaneous mutation. In clinical diagnosis the gene samples are taken from person and tested. If there is a history in the family then the gene test is be conducted for all the family members. The characteristic of autosomal dominant inheritance pattern is such that each affected person will possess an affected parent. Disorders caused due to autosomal dominant inheritance are huntington's disease, neurofibromatosis, marfon syndrome etc. The mutations in BRACA1 and BRAC2 genes associated with breast cancer are transmitted due to autosomal dominant pattern [21].

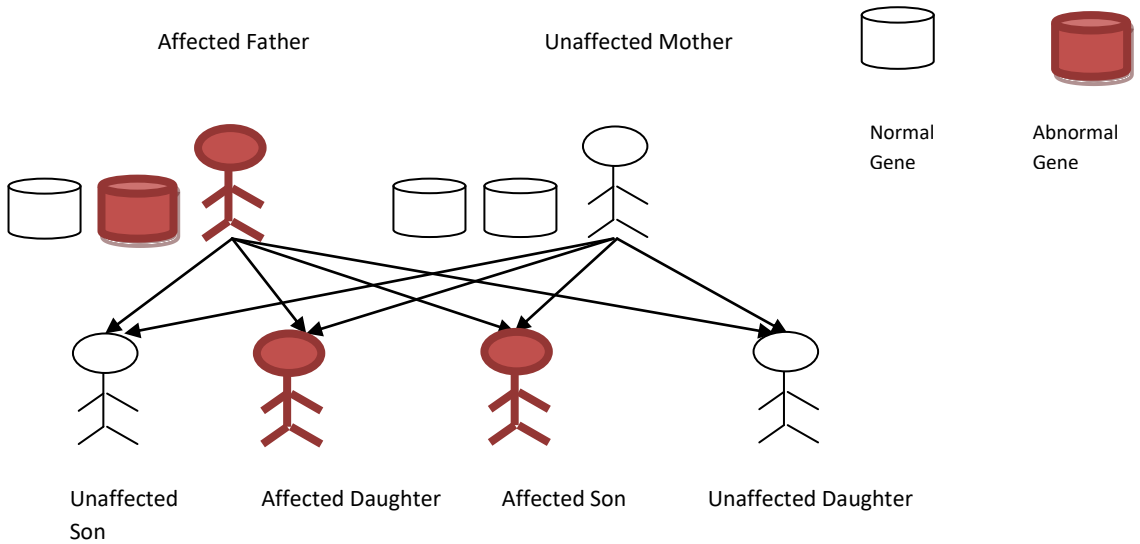


Fig. 1.2 Autosomal Dominant Pattern

Autosomal Recessive

Hereditary disease in human depends on the type of chromosome that is affected. The two types of chromosomes are autosomal and sex chromosomes and also depends on the factor whether it is dominant or recessive. The mutation in the first 22 non sex pairs of chromosome can lead to autosomal disorders. Autosomal recessive is a kind where disease passes down from families and here the two copies of gene will be affected. Gene occurs in a pair and recessive mode of inheritance means both the genes in a pair will be affected which cause disease. If the child is born to autosomal recessive parents, the child has 25% of chance of inheriting the disease. The people with one affected gene in pair is called carrier. The child has 50% chance of becoming carrier from the carrier parents where both the father and mother have one defective gene in pair. Child can be healthy of 25%, even if the parents are carrier. Disorders caused due to autosomal recessive are cystic fibrosis, sickle cell anemia, tay-sachs disease, gaucher disease etc. [21]. Autosomal recessive mode of inheritance is shown in Fig. 1.3.

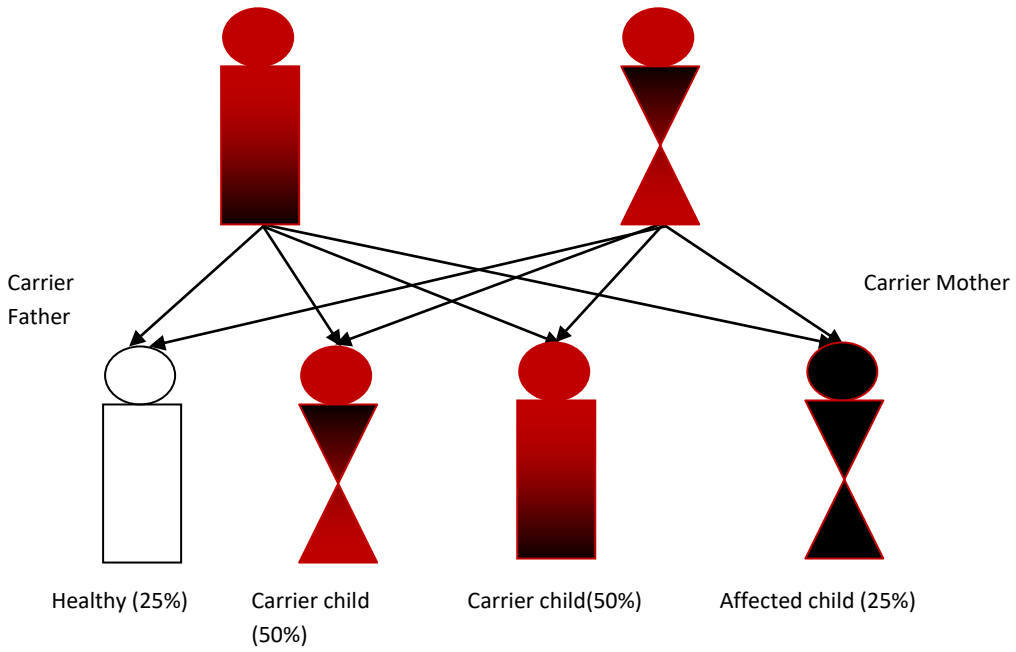


Fig. 1.3 Autosomal Recessive Pattern

X-Linked Inheritance

The X-linked inheritance is the gene causing the disorder located on the X chromosome or in 23rd pair of chromosome. Females have two copies of X chromosome where males have only copy of X chromosome and one copy of Y chromosome. Males can get only X chromosome from their mother whereas females get an X chromosome from both of their

parents. Females are tending to get affected by X-linked disorders. X-linked inheritance possesses two modes of inheritance called X-linked dominant and X-linked recessive.

In X-linked dominant inheritance pattern, females are affected mostly than males because fathers affected with X-linked dominant disorder will have an affected daughter but not son. But if mother is affected then son has a chance of inheriting the disease. In X-linked inheritance pattern the mother can be carrier where the one gene will be defective and that is shown in Fig. 1.4.

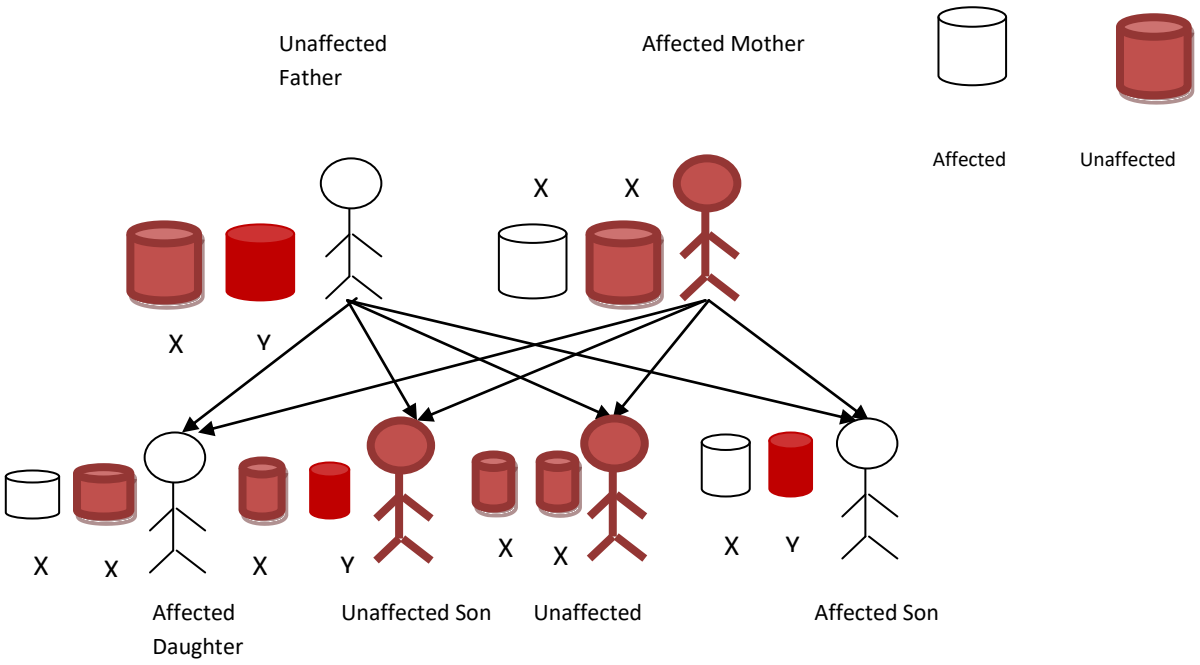


Fig. 1.4 X-linked Dominant Inheritance Pattern

If mother is carrier, the children have the possibility of developing a disorder as follows: Both the son and daughter will either have 50% chance of developing the disorder or completely unaffected. Child of either sex has a chance of inheriting the defective gene from mother's X chromosome as the mother has one copy of defective gene.

In the same way, father also can be a carrier and it is affected by a disorder since the male possess only one X chromosome. The children of this father will inherit the disorder as follows.

- Daughters will possess the disease because the females will receive one copy of male chromosome by birth
- Sons will not inherit the disorder, as males do not receive an X-chromosome from their father

If both the parents are carriers, their children will inherit the disorder as follows.

- Daughters possess high chance of developing a disease as the daughters will receive each copy of gene from both father and mother
- Sons has 50% chance of either inheriting the disorder or completely unaffected. They have an equal chance of receiving either of X chromosome from their mother

The disorders that are caused to X-linked dominant inheritance pattern are fragile X syndrome, rett syndrome, vitamin D resistant rickets, alport syndrome etc.

In X-linked recessive mode of inheritance, the mutation in a gene on the X chromosome causes the phenotype to be likely expressed in males. The female have two X chromosome as if one copy of the gene gets defected, the chance of developing either is 50% or completely unaffected. Male has only one copy X chromosome, so the male has high chance of developing disorder and it is shown in Fig. 1.5. Disorders due to this type of inheritance pattern are duchenne and becker's muscular dystrophy, red-green color blindness, haemophilia A, haemophilia B etc [21].

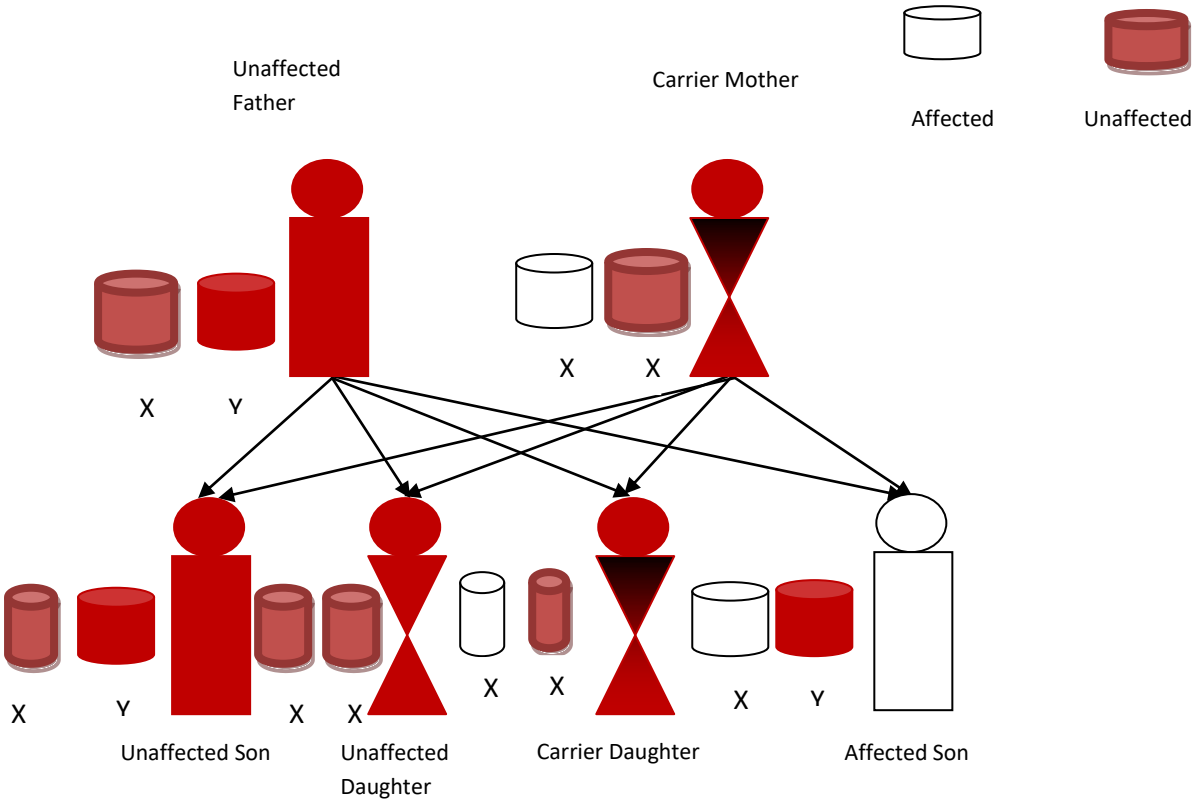


Fig. 1.5 X-linked Autosomal Recessive Pattern

Types of SCA and its Symptoms

SCA has many types and each type has its own symptoms. Till now 36 types of spinocerebellar ataxia has been discovered. Spinocerebellar ataxia type 9 to 36 is very rare and has less characterization. SCA types due to autosomal recessive are SCA10, SCA9, SCA 20, SCA 16, SCA18, SCA13, SCA15, SCA2, SCA14, SCA11, SCA1, SCA7, SCA2 and SCA12. Other types of SCA occur due to autosomal dominant inheritance pattern. Common types of SCA and its clinical characteristics are listed in Table 1.1. The prognosis of SCA varies according to the types. The common prognosis is people affected with SCA will be on wheel-chair after 10-15 years onset of disease and they need assistance for their daily tasks. Genes for other types of SCA are still being found [22].

Table 1.1 Characteristics of Common Types for SCA

Types	SCA1	SCA2	SCA3	SCA6	SCA7	SCA8
Inheritance Pattern	Autosomal dominant	Autosomal dominant	Autosomal dominant	Autosomal dominant	Autosomal dominant	Autosomal dominant
Age onset	Onset over age 60 years	Fourth decade	Second to fifth decade	age of onset is 43 to 52 years	second to fourth decade	Onset ranges from age one to 73 years
Clinical features	speech and swallowing difficulties, muscle stiffness (spasticity), and weakness in the muscles that control eye movement (ophthalmoplegia)	speech and swallowing difficulties, rigidity, tremors, and weakness in the muscles that control eye movement (ophthalmoplegia)	speech difficulties, uncontrolled muscle tensing (dystonia), muscle stiffness (spasticity), rigidity, tremors, bulging eyes, and double vision	speech difficulties, involuntary eye movements (nystagmus), and double vision	muscle weakness, wasting, hypotonia, poor feeding, failure to thrive and loss of motor milestones. Changes in visual acuity and color vision (tritanopia)	muscle spasticity, drawn-out slowness of speech, and reduced vibration sense
Gene mutation and chromosome no.	ATXN1 gene located on the chromosome 6	ATXN2 gene found on the chromosome 12	ATXN3 gene situated on the chromosome 14	CACNA1A gene being on the chromosome 19	ATXN7 gene found on the chromosome 3	ATXN8 and ATXN8OS (reverse strand) genes found on the chromosome 13
No. of repeats mutation occurred in the gene	Victims possess allel with 39 or more CAG trinucleotide repeats	Victims possess alleles with 33 or even more CAG trinucleotide repeats	Victims carry alleles with 52 to 86 CAG trinucleotide repeats	Victims contain 20 to 33 CAG repeats	Victims will often have higher than 36 CAG repeats	Individuals carry from 80 to 250 CTG repeats

Proteins

Protein is a highly complex substance essential for all living organisms. It is of great importance which consists of nutritional value and directly involved in chemical processes essential for life. The importance of protein was identified by chemists during 19th century. The chemist Jons Jacob Berzelius [23] coined the term protein which is derived from greek word proteios meaning holding first place. Proteins are species-specific and organ-specific. Proteins of one species differ from one another and also they differ from organ to organ. Muscle protein is different from brain and hair. Proteins from nutritious food is also essential like carbohydrates, animal proteins etc. Proteins are macromolecules made up of amino acids.

Four levels of protein structures are primary structure, secondary structure, tertiary structure and quarternary structure. Sequence of chain of amino acids is known as primary structure. Proteins with alpha and beta sheets are called as secondary structure. Hydrogen bonding of the peptide backbone causes the amino acids into a repeating pattern. Tertiary structure is otherwise known as 3d structure is the three dimensional structure of the protein which has the backbone with secondary structure. Quaternary structure is the arrangement of multiple folded sub-units of proteins. The proteins are significant to build and repair tissues. Moreover it is essential for hormone function, enzyme making and to make other chemicals in body. Three types of protein are fibrous, globular and membrane proteins. Proteins are essential for body to function. The nine essential functions of proteins are growth and maintenance, biochemical reactions, messenger, provide structure, and provide proper pH scale, fluid balancing, forming antibody, transports nutrients, providing energy. The structures of protein are shown in Fig. 1.6 [24].

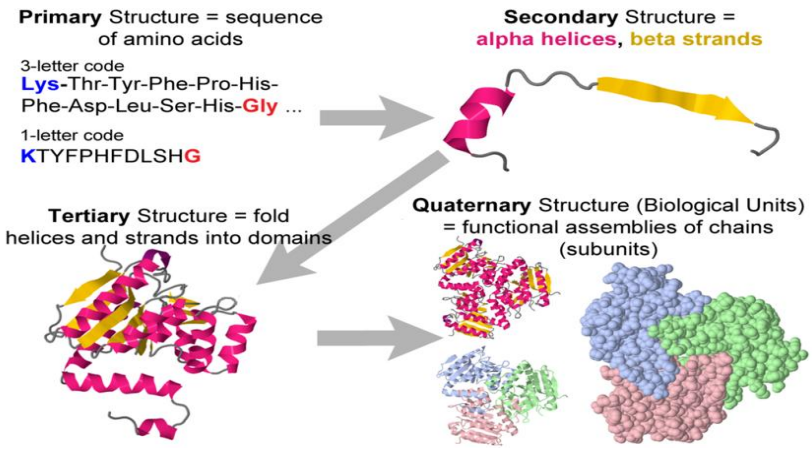


Fig. 1.6 Protein Structures

Proteins sequences are available in uniprot and 3d structures are available in PDB. PDB is openly available for every species and the individual can drop the protein structures of any species. PDB consists of structures and also the fasta sequences. The structures are downloaded in the pdb or fasta format. Tertiary structures are considered for docking purpose and some structures are dropped in PDB with ligands. As of date the total count of protein structures in PDB is 142433 [25]. Genecards is the database of human genes which provide genomic, proteomic, genetic and functional information of known and predicted human genes. The database is maintained by crown human genome center at the weizmann institute of science and it provides access to many free websites like HUGO Gene Nomenclature Committee (HGNC), Ensembl and NCBI. It provides quicker view of current information of gene and proteins for predicted genes [26].

Each gene in a cell is put together for the building blocks of one specific protein. The structure of Deoxyribonucleic Acid (DNA) is called gene and the gene will be inside the compartment of the cell which makes protein. Humans approximately have 30,000 genes that carry instructions for making proteins. From 30,000 genes protein-coding genes are approximately around 26,000. Formation of gene to protein involves two steps like transcription and translation. Transcription and translation together called as gene expression. In the transcription phase, the information stored in a gene's DNA is transferred to a similar molecule called Ribonucleic Acid (RNA) in the cell nucleus. Both RNA and DNA are made up of a chain of nucleotide bases, but they have slightly different chemical properties. The type of RNA that contains the information for making a protein is called messenger RNA or mRNA because it carries the information, or message, from the DNA out of the nucleus into the cytoplasm.

Translation is where the gene to protein formation takes place in the cytoplasm. The mRNA interacts with a specialized complex called a ribosome, which reads the sequence of mRNA bases. Each sequence of three bases, called a codon, usually codes for one particular amino acid. Amino acids are the building blocks of protein. A type of RNA called transfer RNA or tRNA assembles the protein, one amino acid at a time. Protein assembly continues until the ribosome encounters a stop codon [27] and the process of gene expression is shown in Fig. 1.7.

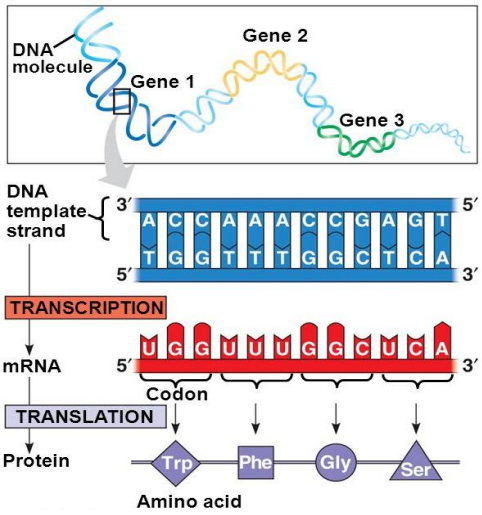


Fig. 1.7 Transcription and Translation Process

The gene has four nucleotide base called adenine (A), thymine (T), cytosine (C) and guanine (G). Nucleotide A pairs with T, C pairs with G. In RNA, instead of T uracil (U) will be paired with A. Codon is a sequence of three DNA or RNA nucleotides that codes for a particular amino acid. There are 64 codons where each codon has three nucleotides that code for an amino acid. There is one start codon and three stop codons where start codon start the translation process and stop codon indicates the end of the translation process. The start codon is AUG and three stop codons are UAA, UAG and UGA. During the protein synthesis phase, the stop codons release the protein from ribosomes [28]. The codon table is shown in Fig. 1.8.

Second letter

First letter	U	C	A	G	Third letter
U	UUU, UUC Phe; UUA, UUG Leu	UCU, UCC, UCA, UCG Ser	UAU, UAC Tyr; **UAA Stop**; **UAG Stop**	UGU, UGC Cys; **UGA Stop**; UGG Trp	U C A G
C	CUU, CUC, CUA, CUG Leu	CCU, CCC, CCA, CCG Pro	CAU, CAC His; CAA, CAG Gln	CGU, CGC, CGA, CGG Arg	U C A G
A	AUU, AUC, AUA Ile; **AUG Met**	ACU, ACC, ACA, ACG Thr	AAU, AAC Asn; AAA, AAG Lys	AGU, AGC Ser; AGA, AGG Arg	U C A G
G	GUU, GUC, GUA, GUG Val	GCU, GCC, GCA, GCG Ala	GAU, GAC Asp; GAA, GAG Glu	GGU, GGC, GGA, GGG Gly	U C A G

Fig. 1.8 Codon Table

Post-Translational Modification

The proteins do not change the structure when the gene is mutated. The protein can get mutated during the stage of Post-Translational Modification (PTM) and protein folding. Once proteins are released from ribosomes, it undergoes PTM to get mature protein product. PTM is significant in cell signaling. For example, PTM helps in converting prohormones to hormones. There are many types of protein modification, which are mostly catalyzed by enzymes that recognize specific target sequences in proteins. These modifications regulate protein folding by targeting specific subcellular compartments, interacting with ligands or other proteins, or by bringing about a change in their functional state including catalytic activity or signaling [29]. Most common disease associated with PTMs is breast cancer and the database is available for human disease associated with Post-Translational Modification Disease (PTMD) [30]. The most common PTMs are:

- Based on the addition of chemical groups
- Based on the addition of complex groups
- Based on the addition of polypeptides
- Based on the cleavage of proteins
- Based on the amino acid modification

Proteins can be modelled using homology modelling. It is otherwise known as comparative modeling of protein, which tells about constructing an atomic-resolution model of the target protein from its actual amino acid sequence and also a precise three-dimensional structure of a suitable homologous protein known as template. Homology modelling is an essential computational approach, to figure out the 3D structure of proteins. It utilizes accessible high-resolution protein structures to generate a model of a protein of similar, to build unknown structure. Homology modelling plays an important part in pharmaceutical pattern. To carry out homology modelling some processes like template identification, alignment, model building and refining that model are involved. There are many tools and softwares available to model the protein like blast, Clustal, 3DCoffee, Procheck, anolea etc [22].

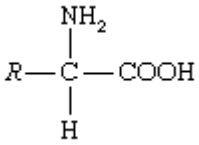
Amino Acids

There are 20 amino acids exist in protein where hundreds to thousands of these amino acids are attached in a long chain to form a macromolecule. Amino acids are the building blocks of protein. Twenty amino acids are listed below in Table 1.2.

Table 1.2 List of Amino Acids

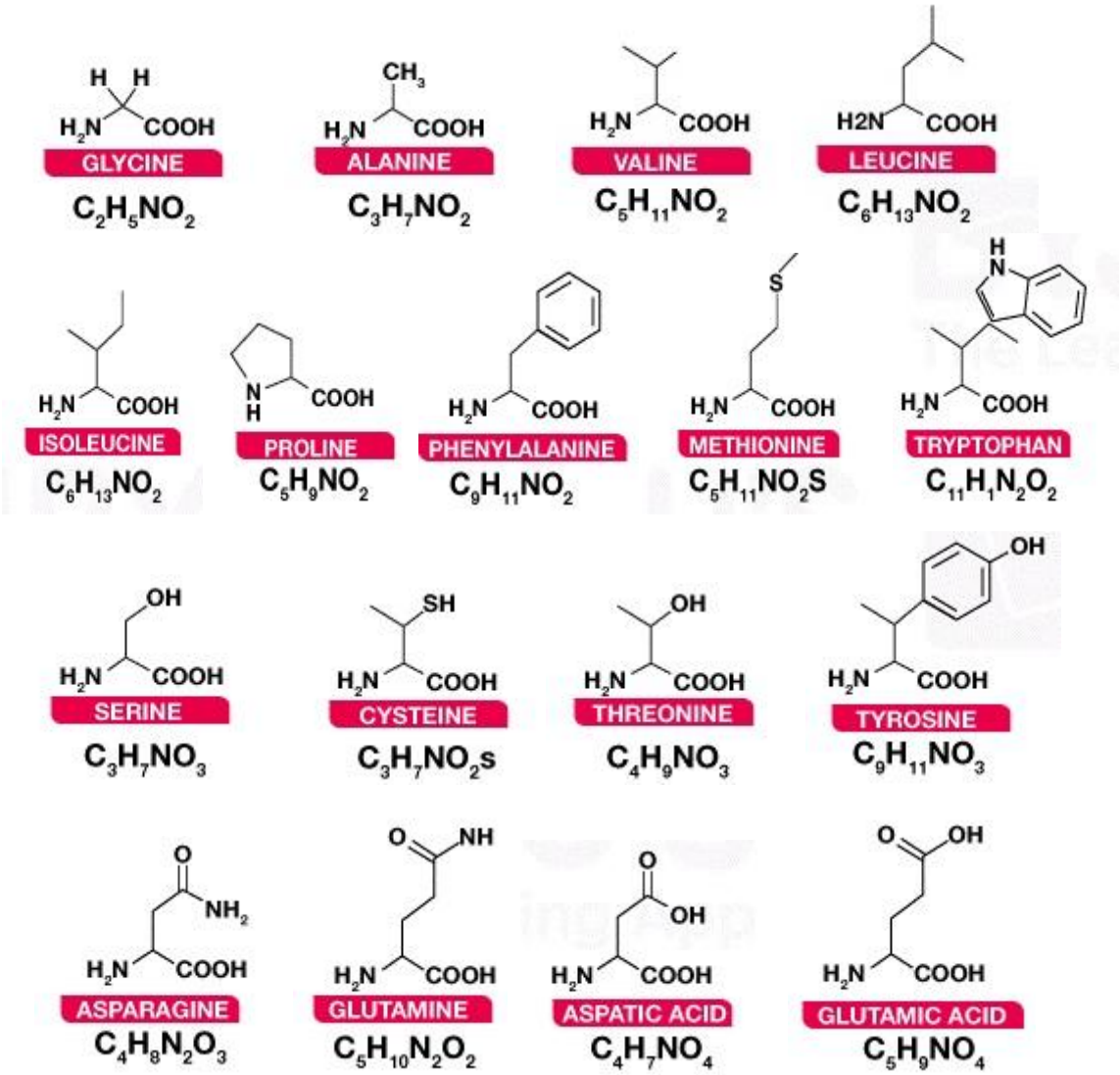
Name of Amino Acids	3 Letter Code	1 Letter Code
Alanine	Ala	A
Arginine	Arg	R
Asparagines	Asn	N
aspartic acid	Asp	D
Cysteine	Cys	C
Glutamine	Gly	Q
glutamic acid	Glu	E
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Amino acid is a group of organic molecule that consists of basic amino group —NH_2, an acidic carboxyl group —COOH, and an organic *R* group or side chain that is unique to each amino acid. Each molecule contains a central carbon atom, called the α-carbon, to which both an amino and a carboxyl group are attached. The remaining two bonds of the α-carbon atom are generally satisfied by a hydrogen atom and the *R* group. The formula of a general amino acid is



The amino acids differ from each other in the particular chemical structure of the R group. Amino acids can act either as an acid or base called amphoteric. The basic amino group typically has a pKa between 9 and 10, while the acidic α-carboxyl group has a pKa that is usually close to 2. Amino acids can be grouped into 4 groups based on their

polarity of the side chain. The four groups of amino acids are polar, non-polar, acidic and basic amino acids. The non-polar amino acids are glycine, alanine, valine, leucine, isoleucine, proline, phenylalanine, methionine and tryptophan. The R groups of these amino acids have either aliphatic or aromatic groups. Polar amino acids like serine, cysteine, threonine, tyrosine, asparagines and glutamine. Side chains in this group contain functional groups. Acidic amino acids are aspartic acid and glutamic acid. These two amino acids have a carboxylic acid on its side chain that gives acidic properties. Basic amino acids are arginine, histidine and lysine. Each side chain can accept proton. The structures of amino acids are shown in Fig. 1.9 [31].



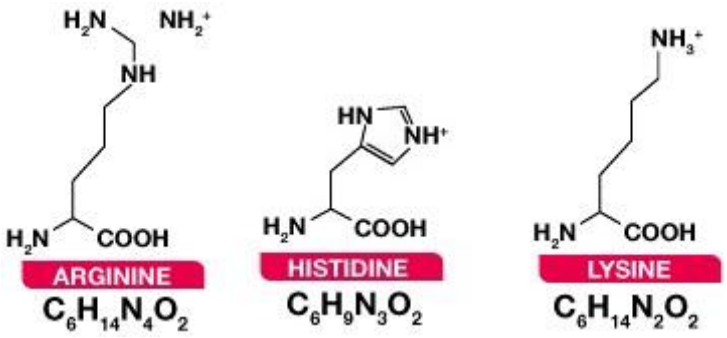


Fig. 1.9 Structure of Amino Acids

Mutations

A change in the gene causes mutation. The mutation not only affects the gene but also protein, mutation changes the structure of protein and can lead to abnormal functioning or medical condition. Genetic disorders are caused by mutation in one or few genes. Harmful mutations cause genetic disorder or can cancer. Mutations are of many types like missense, nonsense, point, insertion, deletion, repeat and frameshift mutations. Mutational information for human genetic disorders is available in HGMD. HGMD is compilation of published germline mutations in nuclear genes that underlie, or are closely associated with hereditary disease. HGMD is established in 1996 for the study of mutational mechanisms that occurs due to inherited diseases and the cdna sequences were added where the codon was numbered. In 2017, the total mutational information available was 203,885 which were available publicly. The public database is freely available for the registered users and non-profit organisations [32]. According to Cardiff University the total number of mutations upto date is 256,070 [31]. The repeat mutational information is taken from this database.

Missense mutations are the substitution in a codon that encodes a different amino acid and cause a small change in the protein. For example, missense mutation 347T>C indicates that codon changes CTC-CCC in the dystrophin gene results in DMD, where the protein Leu is altered to Pro.

Nonsense mutations are the substitution in a codon that results in premature termination of protein. TAG, TAA, TGA are the three stop codons. For example, a nonsense mutation in the dystrophin gene 433C>T, point out that the codon change CGA-TGA and the protein arg is terminated with amber stop codon and results in BMD [33].

Single character change in a gene makes an impact on the gene which in turn changes the function of the gene. In some cases, a DNA mutation may do no harm in protein sequences. It depends on the sort of DNA mutation and where it is located. A change in codon encodes the same amino acid and causes no change in the protein is called silent

mutations [34]. For example, in CAPN3 gene 246G>A specifies CCG-CCA and the protein pro is not misrepresented, but it routes to the LGMD type 2 disease.

During small insertions, a new base is added into the sequence that alters the function of a gene. An increase in the number of the same nucleotides in a location is termed as duplications. For example, Emery-Dreifuss Muscular Dystrophy (EMD) disease is caused by the duplications in the emerin gene for the nucleotide change 650_654dupTGGGC [35].

Small deletions occur in the genes when a base is deleted from a sequence that truncates the function of genes. For example, 253delG deletes G in 253 position in the SH3TC2 gene that directs for Charcot-Marie-Tooth disease 4C. Gross insertions and gross deletions occur when the whole number of exons is involved in the insertions are deletions [36].

Repeat mutation is otherwise known as trinucleotide repeats. A trinucleotide repeat disorder is otherwise known as polyglutamine repeats. Polyglutamine repeat disorder is a group of neurodegenerative disorder that is caused by the expansion of the trinucleotide repeat Cytosine-Adenine-Guanine (CAG), which encodes the amino acid glutamine. For example, if a parent possess 39 repeats of glutamine and their child will possess more than 40 repeats and they will possess the disorder [37].

Frameshift mutation alters the position of nucleotides in the reading frame, and that forms unrelated amino acids into the protein, generally followed by a stop codon.

For example consider a DNA sequence

Codon: Thr Pro Glu Glu Glu Thr

Sequence: ACT CCT GAG GAG GAG ACT

A. Missense mutation

Codon: Thr Pro Glu Glu Glu Thr

Sequence: ACT CCT GAG GAG GAG ACT

Sequence: ACT CCT GTG GAG GAG ACT

Codon: Thr Pro Val Glu Glu Thr

In the above noted example, a single nucleotide change from A to T and thus it codes for Val instead of the amino acid Glu.

B. Nonsense mutation

Codon: Thr Pro Glu Glu Glu Thr

Sequence: ACT CCT GAG GAG GAG ACT

Sequence: ACT CCT GAG TAG GAG ACT

Codon: Thr Pro Val Stop Glu Thr

C. Silent mutation

Codon: Thr Pro Glu Glu Glu Thr

Sequence: ACT CCT GAG GAG GAG ACT

Sequence: ACT CCA GAG GAG GAG ACT

Codon: Thr Pro Glu Stop Glu Thr

In the 6th position the nucleotide changes from T to A with no change in their amino acid. These mutations are termed as Silent mutations.

D. Frameshift mutations

Sequence: ACTCCTGAGGAGGAGACT

Sequence: ACTCCTCTGAGGAGGAGACT

The base pairs CT are inserted in the 7th position.

Sequence: ACTCCTGAGGAGGAGACT

Sequence: ACTCCTGAGGAAGACT

The base pairs GG are inserted in the 9th position.

E. *Repeat Mutations*

Sequence of 1oa8

SEQUENCEGSPAAAPPTLPPYFMKGSIIQLANGELKKVE
DLKTEDFIQSAEISNDLKIDSSTVERIEDSHSPGVAVIQFA
VGEHRAQVSVEVLVEYPFFVFGQGWSSCCPERTSQLFD
LPCSKLSVGDVCISLTLKNLKNG

SEQUENCEGSPAAAPPTLPPYFMKGSIIQQQQQQQQQQQ
QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQLAN
GELKKVEDLKTEDFIQSAEISNDLKIDSSTVERIEDSHSPG
VAVIQFAVGEHRAQVSVEVLVEYPFFVFGQGWSSCCPE
RTSQLFDLPCSKLSVGDVCISLTLKNLKNG

SCA and its Associated Protein

SCA is a genetic disorder and the common types of SCA occur through repeat mutations. The types of SCA that occur through repeat mutations are SCA1, SCA2, SCA3, SCA6, SCA8 and SCA10. The proteins associated with these types are taken for consideration in this research work and it is listed in Table 1.3. Mutational information of these proteins is listed in Table 1.4.

Table 1.3 Types of SCA and Proteins

Types of SCA	Proteins	Associated Proteins
SCA1	Ataxin-1	1oa8, 2m41, 4apt, 4aqp, 4j2j, 4j2l
SCA2	Ataxin-2	3ktr
SCA3	Ataxin-3	1yzb, 2aga, 2dos, 2jri, 2klz, 4wth,4ys9
SCA6	Voltage-dependent P/Q- type calcium channel subunit alpha-1a	3bxk
SCA8	Ataxin-8	4zka
SCA10	Ataxin-10	5fur

Table 1.4 Mutational Information

Protein Structures	Number of Repeats
Ataxin-1	40-100
Ataxin-2	32-500
Ataxin-3	68-79
Ataxin-6	21-28
Ataxin-7	40-200
Ataxin-8	116
Ataxin-10	1611

1.3 DOCKING AND BINDING AFFINITY

Docking is an essential application for prediction of binding affinity. Protein structures are docked with ligands, macromolecules etc., to obtain the energy calculations. Energy calculations aids in calculating binding affinity. There are many docking tools, algorithms and methods for calculating binding affinity available and are discussed below.

Docking

Molecular docking is an important application in structural molecular biology combined with computer-assisted drug design. To perform docking, the requirement is the protein structure from x-ray crystallography or Nuclear Magnetic Resonance (NMR) spectroscopy. The structure of the protein can be modelled using homology modelling. Docking depends on two components such as search algorithm and scoring functions. The search algorithm is used to search the possible orientations of the protein docked with ligand and there are three conformational search strategies are used. They are systematic, genetic

algorithm and molecular dynamics simulation. Systematic is used to search about rotational bonds and genetic algorithm is used to evolve low energy conformation. Scoring function is used to indicate the pose that gives a favourable binding interaction and the ligand is ranked accordingly. There are three scoring functions such as knowledge-based, empirical and force field. The docking is assessed by docking accuracy, enrichment factor and validation [38]. The methods involved in docking are search algorithms, incremental construction (IC), monte carlo simulations (MC), fast shape matching (SM), simulated annealing (SA), distance geometry (DA), evolutionary programming (EP), genetic algorithm and tabu search. The list of some docking programs and its search method are listed in Table 1.5 [39].

Hotspot is necessary for the proteins, if the rigid docking is performed. Active site is where the ligand or any molecule binds to perform action. When protein is flexible it consumes time and there are number of algorithms and servers are available for active site identification. Algorithm like neural network is used for active site identification and webservers like P2rank, sitesidentify, active site prediction, deepbind etc., Docking without any knowledge regarding the binding site is known as blind docking. The earlier elucidation for the ligand-receptor binding procedure is the lock-and-key principle, wherein the ligand sits into the receptor just like lock and key. The induced-fit concept carries the lock-and-key theory a phase more, proclaiming that the energetic site of the protein is constantly reshaped by interactions with the ligands since the ligands communicate with the protein. This principle implies that the ligand together with receptor ought to be dealt as flexible at the time of docking. The purpose of protein-ligand docking is to forecast the prominent binding mode of a ligand with a receptor of recognized three-dimensional structure.

Table 1.5 Docking Programs and Search Method

Docking Program	Search Method
AutoDock	GA
DOCK	IC
ZDOCK	SM
MS-DOCK	SM
MCDOCK	MC
ICM	MC
GOLD	GA
Surflex	IC
FLEXX	IC
M-ZDOCK	SM

SYSDOC	SM
EUDOC	SM
FLOG	IC

The comparison of standard experimental High-Throughput Screening (HTS) with Virtual Screening (VS) proves that VS is a reasonable drug discovery strategy and also possesses the benefit of inexpensive and valuable screening. VS could be categorized into ligand-based and structure-based methods. Ligand-based techniques such as pharmacophore modeling and Quantitative Structure Activity Relationship (QSAR) strategies are used when some of the active ligand molecules are known and any structural details are available for targets. In structure-based drug design, molecular docking is considered the most frequent approach. The docking procedure includes two simple steps such as forecast of the ligand conformation combined with its positioning and orientation or pose within these sites in addition to evaluation of the binding affinity.

These procedures are associated with sampling methods and scoring schemes, respectively. Recognizing the specific location of the binding site before docking procedure considerably improves the docking performance. If the binding site is not known, then the hotspot can be identified by online servers, e.g. GRID, POCKET, SurfNet, PASS and MMC can be utilised to recognize putative effective sites within proteins [22]. Two types of docking are flexible docking and rigid docking. In flexible docking, ligand will be flexible and this docking allows the ligand to rearrange structurally according to the receptor. In rigid docking, ligand will be rigid where the ligands cannot change their spatial shape during the docking process. Some of the docking programs are listed in Table 1.6.

TABLE 1.6 List of Docking Programs

Program	Description
1-Click Docking	Docking predicts the binding orientation and affinity of a ligand to a target
AADS	Automated active site detection, docking, and scoring(AADS) protocol.
AutoDock	Automated docking of ligand to macromolecule.
AutoDock Vina	New generation of AutoDock.
BetaDock	Based on Voronoi Diagram

HADDOCK	Developed for protein-protein docking, but can also be applied to protein-ligand docking.
Score	The score service allows calculating some different docking scores of ligand-receptor complex.
ADAM	Prediction of stable binding mode of flexible ligand molecule to target macromolecule.
DockingServer	Integrates a number of computational chemistry software
FlexX	Incremental build based docking program
Glide	Exhaustive search based docking program
GOLD	Genetic algorithm based, flexible ligand, partial flexibility for protein
ICM-Dock	Docking program based on pseudo-Brownian sampling and local minimization

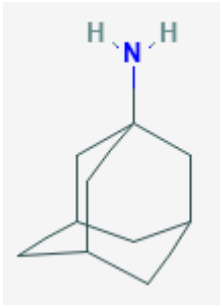
Ligands

Ligand is an ion or molecule like protein, DNA, RNA etc. that binds to a central metal atom to form a complex. The binding involves one or more electron pairs and it can be covalent or ionic bonds. The ligands can be into three types L, X and Z which correspond respectively to 2-electron, 1- electron and 0-electron neutral ligands. There are number of motifs in ligand and they are trans-spinning ligands, ambidentate ligands, bridging ligand, binucleating ligand, metal-ligand multiple bond, spectator ligand, bulky ligand, chiral ligand, hemiliable ligands and non-innocent ligands. Trans-spanning ligands are bidentate ligands that can span coordination positions on opposite sides of a coordination complex. Ambidentate ligands can attach to the central atom in two places. Bridging ligand links two or more metal centers. Binucleating ligands bind two metals. Some ligands can bond to a metal center through the same atom but with a different number of lone pairs. The bond order of the metal ligand bond can be in part distinguished through the metal ligand bond angle.

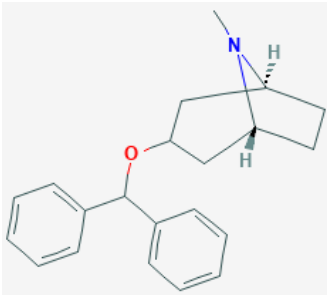
Spectator ligand is a tightly coordinating polydentate ligand that does not participate in chemical reactions but removes active sites on a metal. Bulky ligands are used to control the steric properties of a metal center. Chiral ligands are useful for inducing asymmetry within the coordination sphere. Hemilabile ligands contain at least two electronically different coordinating groups and form complexes where one of these is easily displaced from the metal center while the other remains firmly bound, a behaviour which has been found to increase the reactivity of catalysts when compared to the use of more traditional ligands.

Non-innocent ligands bond with metals in such a manner that the distribution of electron density between the metal center and ligand is unclear [40].

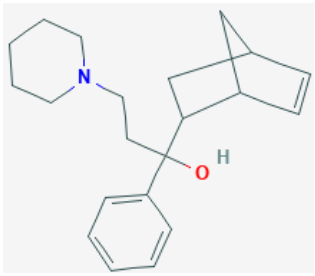
The databases for protein-ligand interaction are biolip, protein-ligand database, credo, possum, pocketome, relibase, scPDB, probis and PLI. These databases report unique information about the interaction. The ligands taken for study in this research to dock with proteins are eighteen ligands and they are amantadine, benztropine, biperiden, bromocriptine, carbidopa, donepezil, entacapone, galantamine, levodopa, pergolide, pramipexole, procyclidine, rivastigmine, ropinirole, selegiline, tacrine, tolcapone and trihexyphenidyl. These are the ligands that are used for neuro degenerative disorders like huntigon, parkinsons etc., and muscular disorders. So these ligands are considered to dock with proteins associated with SCA. The structures of these ligands are shown in Fig 1.10.



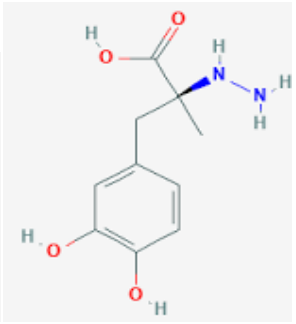
Amantadine



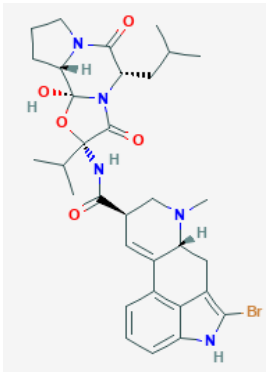
Benztropine



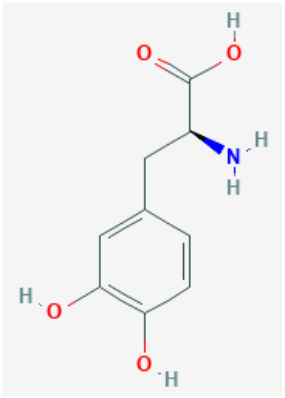
Biperiden



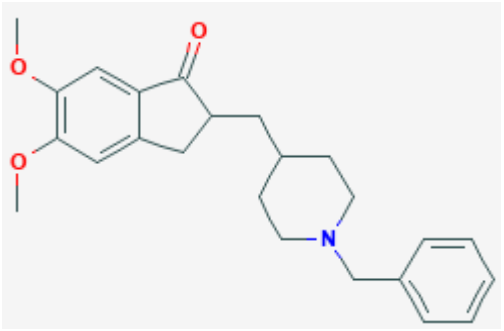
Carbidopa



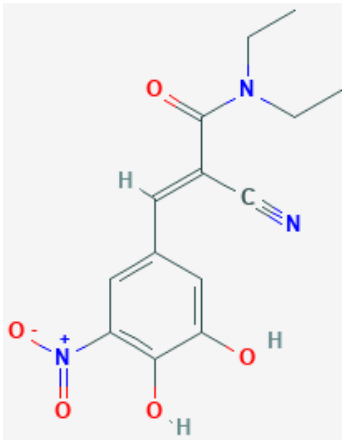
Bromocriptine



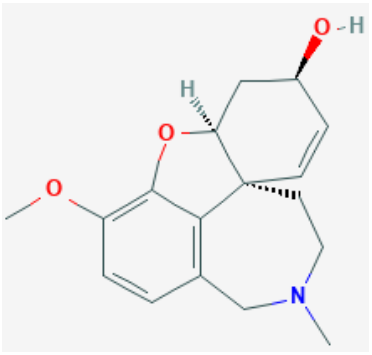
Levodopa



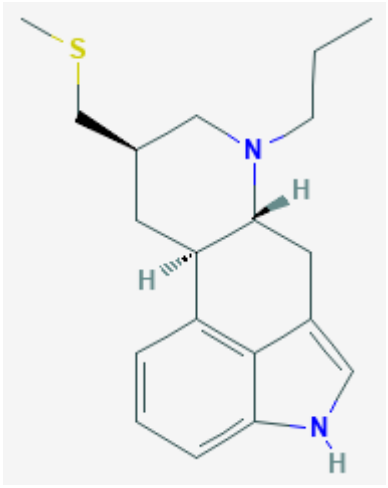
Donepezil



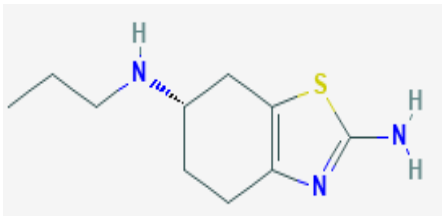
Entacapone



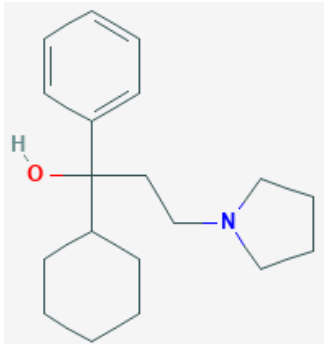
Galantamine



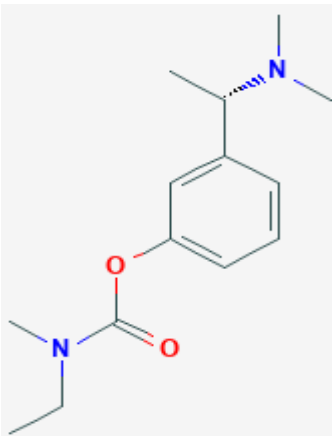
Pergolide



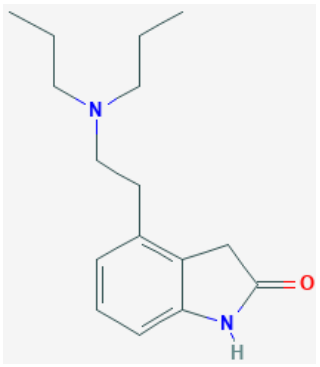
Pramipexole



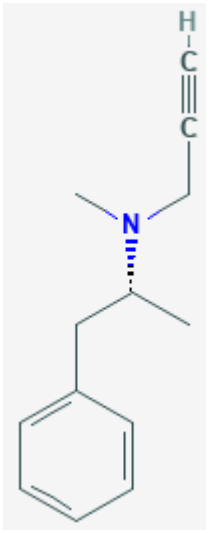
Procyclidine



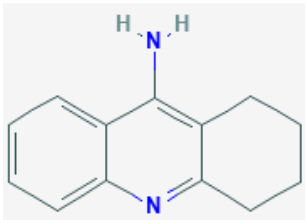
Rivastigmine



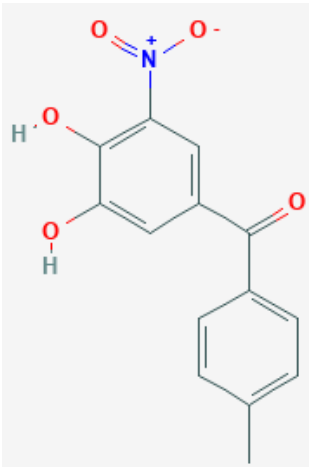
Ropinirole



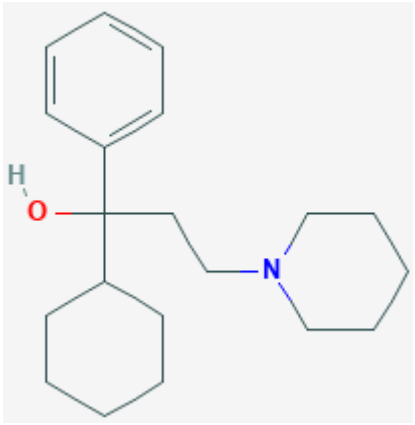
Selegiline



Tacrine



Tolcapone



Trihexyphenidyl

Fig. 1.10 Structure of Ligands

Binding Affinity

Binding affinity is the measure of the strength of attraction between molecule and small compounds. Receptor is a protein molecule that acquires a signal by binding to a chemical called ligand. The ability of ligand is to form co-ordination bond with a receptor known as binding affinity. The binding affinity of a ligand is determined by the interaction force of attraction between the ligand together with their receptor binding sites. A ligand with high-affinity binding demonstrates a lengthy residence time at the receptor binding site when compared with low-affinity binding. Binding a ligand to receptor inhibit the growth or prolong the symptoms of the disorder. Certain frequent ligands are hormones, mediators, neurotransmitters, and so on. The appropriate receptor on target cell is identified and reacts for the ligand that will bond [41].

High-affinity binding of ligands to receptors is often physiologically important when some of the binding energy can be used to cause a conformational change in the receptor, resulting in altered behaviour. Ligands that binds and alter the function of the receptor and triggers a physiological response is called a receptor agonist. Ligands that bind to a receptor but fail to activate the physiological response are known as receptor antagonists.

Agonist binding to a receptor can be characterized both in terms of how much physiological response can be triggered and in terms of the concentration of the agonist. The concentration of agonist is required to produce the physiological response often measured as EC_{50}. High-affinity ligand binding implies that a relatively low concentration of a ligand is adequate to maximally occupy a ligand-binding site and trigger a physiological response. Receptor affinity is measured by an inhibition constant or K_i value, the concentration required to occupy 50% of the receptor. Ligand affinities are most often measured indirectly

as an IC_{50} value from a competition binding experiment where the concentration of a ligand required to displace 50% of a fixed concentration of reference ligand is determined.

Low-affinity binding implies that a relatively high concentration of a ligand is required. The binding site is maximally occupied and the maximum physiological response to the ligand is achieved. The agonists maximally stimulate the receptor and can be defined as a full agonist. An agonist that activates the physiological response partially is called a partial agonist.

Applications

Binding affinity prediction mainly aids in drug designing process. The hereditary disorders which are not curable can be cured by the process of inhibition in binding. The inhibition is possible by finding appropriate drugs with accurate binding affinity. The other uses of binding affinity prediction are functions and biological processes of proteins and ligands can be known. When binding with the macromolecule, the unknown process of its cellular function with ligand or protein can be found. This unknown process will lead to design the new drugs. Binding affinity along with ligand efficacy determine the overall potency of a drug. Potency is a result of the complex interplay of both the binding affinity and the ligand efficacy.

Ligand efficacy refers to the ability of the ligand to produce a biological response upon binding to the target receptor and the quantitative magnitude of this response. This response may be as an agonist, antagonist, or inverse agonist, depending on the physiological response produced. Selective ligands have a tendency to bind to very limited kinds of receptor, whereas non-selective ligands bind to several types of receptors. This plays an important role in pharmacology, where drugs that are non-selective tend to have more adverse effects, because they bind to several other receptors in addition to the one generating the desired effect.

Methods

Ligand affinities can also be measured directly as dissociation constant (K_D) using methods such as fluorescence quenching, isothermal titration calorimetry or surface plasmon resonance. Real-time methods, which are often label-free, such as surface plasmon resonance, dual-polarization interferometry and Multi-Parametric Surface Plasmon resonance (MP-SPR) can not only quantify the affinity from concentration based assays. But also the real-time method from the kinetics of association and dissociation, the conformational change induced upon binding is executed. MP-SPR also enables measurements in high saline dissociation buffers to a unique optical setup. Microscale

Thermophoresis (MST), an immobilization-free method that allows the determination of the binding affinity without any limitation to the ligand's molecular weight.

The main methods to study protein–ligand interactions are principal hydrodynamic and calorimetric techniques, principal spectroscopic and structural methods such as

- Fourier transform spectroscopy
- Raman spectroscopy
- Fluorescence spectroscopy
- Circular dichroism
- Nuclear magnetic resonance
- Mass spectrometry
- Atomic force microscope
- Paramagnetic probes
- Dual polarisation interferometry
- Multi-parametric surface plasmon resonance
- Ligand binding assay and radioligand binding assay

Other techniques include, fluorescence intensity, bimolecular fluorescence complementation, fluorescent resonance energy transfer, quenching surface plasmon resonance, bio-layer interferometry, coimmunopreciptation indirect elisa, equilibrium dialysis, gel electrophoresis, far western blot, fluorescence polarization anisotropy, electron paramagnetic resonance, microscale thermophoresis.

Binding affinity can be calculated by extracting various features from the docked complexes and also there are number of tools used. To predict binding affinity certain features like energy calculations, scoring functions, physical and chemical properties, sequence descriptors are necessary. Energy calculations can be taken from autodock, an efficient tool to dock. Binding affinity is measured and reported by the equilibrium dissociation constant (K_D), which is used to evaluate and rank order strengths of bimolecular interactions. The smaller the K_D value, the greater the binding affinity of the ligand for its target. Binding affinity is calculated using the following formula.

$$[R]\ [R]\ k_1 = [DR]\ K\text{-}1 \quad \textbf{(1.1)}$$

$$K_1/K{-}1 = [RR]/[R][R] \quad \textbf{(1.2)}$$

$$\text{Binding Affinity} = K_1/K{-}1 \quad \textbf{(1.3)}$$

$$K_D = K{-}1/K_1 \quad \textbf{(1.4)}$$

Here, K_D is called as binding affinity constant, K_1 is termed as association constant and K-1 is rate constant. R is the receptor and another R is another receptor or ligand.

The binding affinity prediction methods can be divided into two types such as structure based and non-structure based methods. In structure based method, classical scoring functions are used. Scoring functions like empirical and force-field scoring functions are extracted as features. In non-structure based methods, machine learning and deep learning techniques are used which depends on features. There are many web servers using which binding affinity can be predicted [42]. The binding of protein with ligand and protein with protein can be viewed in pymol software [22].

Tools

There are number of tools used to calculate binding affinity like autodock, autodock vina, haddock. The online servers Protein Binding Energy Prediction (PRODIGY), CSM-Lig, DeepDrug3D, Taba etc., can also be used. Autodock is used to dock proteins with ligand and the flexible docking is preferred. Autodock vina is used to extract the gauss functions of both proteins and ligands. In haddock, the interaction of protein and ligand, prediction of binding affinity, binding site identification etc., can be performed.

1.4 NEED FOR THE PROPOSED RESEARCH

The current research work on binding affinity prediction for SCA were focused on existing databases like PDBbind, MOAD etc. and also the human structures or sequences were not considered. Earlier binding affinity prediction models used mouse protein structures which were not promising and human protein structures are to be considered. In the existing databases, various interactions like docking of proteins with ligand, proteins, RNA and DNA have not been investigated. Various researchers carried out their study from non-mutated protein structures wherein the changes of the protein structure due to mutation were not monitored. The features like scoring functions and molecule descriptors were used in most of the work, but were not sufficient to build effective models that predicts binding affinity accurately.

Accurate prediction of binding affinity is a complicated task as the binding affinity varies when protein structure gets altered due to repeat mutation and the protein loses its validity. It is difficult to predict binding affinity as the changes in protein structure affects the docking results and binding affinity. Investigation of mutation induced protein structures is obligatory to capture the changes in the protein structure and energy calculations.

In the existing research, traditional machine learning techniques have been adopted for affinity prediction but the results obtained were not promising and therefore more comprehensive learning approach is vital to build accurate binding affinity prediction model.

Hence there is a need to perform research which will focus on various interactions like protein-ligand, mutated protein-ligand, protein-protein and features related to energy calculations, scoring functions, sequence descriptors to build more efficient binding affinity prediction models. It is also required to focus on mutated protein structures of human in order to develop synthetic datasets and deep learning approach to build accurate binding affinity predictive models. The detailed survey of existing research on binding affinity prediction is presented in chapter 2.

1.5 MACHINE LEARNING AND DEEP LEARNING

Machine learning is a scheme of data analysis that automates the logical model building. It is a branch of artificial intelligence based on the initiative that systems learn from data, recognize patterns and make decisions with minimal human intervention. Machine learning uses algorithms to parse data, learn and then make a decision like classification or prediction. It is essential because of accurate predictions that aids in making better decisions for real time applications without much human intervention. There are many methods in machine learning like supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. Deep learning is a part of machine learning methods based on artificial neural networks with representation learning. The benefit of deep learning is it generates new features from limited series of features in the training dataset. Deep learning uses a layered structure of algorithms called artificial neural network. It learns from the data and makes decisions on its own and also it makes accurate predictions than machine learning.

Since traditional machine learning and the latest deep learning offers precise prediction in pattern recognition problems, these techniques have been chosen in this research to build binding affinity predictive models.

1.5.1 Regression Algorithms

Regression analysis is a prevailing statistical method that allows examining the relationship between dependant variable and independent variables. The most common regression analysis is linear regression, in which the line that closely fits the data according to the specific mathematical measure. Regression analysis is mainly used for prediction and forecasting and also to infer the relationships between dependant and independent variables [43]. The most common regression algorithms are linear regression, support vector machines,

lasso regression, logistic regression, polynomial, stepwise, ridge, multivariate, elasticnet and multiple regression [43].

In this research work, linear regression, support vector regression, random forest, artificial neural network are used to build binding affinity predictive models and are presented below.

Linear Regression

Linear regression (LR) is a linear approach for modeling the relationship between a dependent response and one or more independent variables. The case of one independent variable is called simple linear regression. Linear regression predicts the target as a weighted sum of the feature inputs. For more than one independent variable, the process is called multiple linear regression. The relationships in linear regression are modelled using linear predictor functions whose unknown model parameters are estimated from the data. [44].

Linear regression is a statistical method that enables users to summarise and study relationships between two continuous variables. Linear regression is a linear model that assumes a linear relationship between the input variables called X and the single output variable called Y. The output variable can be calculated from a linear combination of the input variables X. The simplest form of the regression equation with one dependent and one independent variable is defined by the equation 2.1.

$$y = c + b*x \quad \textbf{(1.5.1)}$$

where y = dependent variable, c = constant, b = regression coefficient, and x = independent variable. The simple linear regression model is shown in Fig. 1.11.

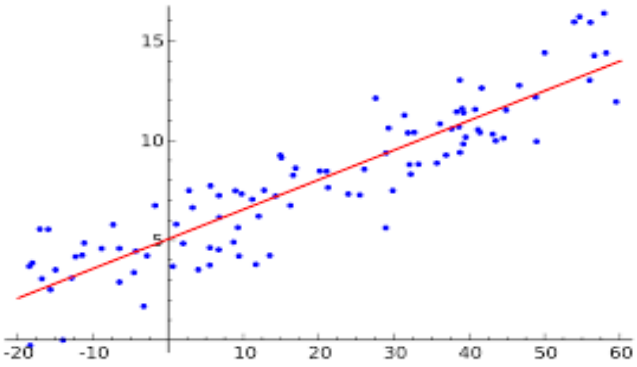


Fig. 1.11 Linear Regression

The data preparation includes linear assumption, removal of noise, removal of collineriaty, gaussian distributions and rescaling of inputs. Linear regression in prediction is simple and the model representation for this prediction problem will be

$$Y = B_0 + B_1X \quad \textbf{(1.5.2)}$$

where B_0 is a constant, B_1 is the regression coefficient, X is independent variables and Y is the dependant variable. The plane fits the data where the data points will be around the plane. The best fit line is the one for which total prediction error are as small as possible. Error is the distance between the points to the regression line [45].

- If $b_1 > 0$, then predictor x and target y have a positive relationship. That is increase in x will increase y
- If $b_1 < 0$, then predictor x and target y have a negative relationship. That is increase in x will decrease y
- If the model does not include x=0, then the prediction will become meaningless with only b_0
- If the model includes value 0, then b_0 will be the average of all predicted values when x=0. But, setting zero for all the predictor variables is often impossible
- The value of b_0 guarantee that residual have mean zero. If there is no b_0 term, then regression will be forced to pass over the origin. Both the regression co-efficient and prediction will be biased.

Support Vector Regression

Support Vector Regression (SVR) is characterized by the use of kernels, sparse solution, and Vapnik–Chervonenkis (VC) control of the margin and the number of support vectors. SVR uses the same principles as the support vector machine (SVM) for classification, with only a few minor differences. The difference between SVM and SVR is, SVR has continuous values. The terms used in support vector regression are kernel, hyperplane, boundary line and support vectors. Kernel function is used to map a lower dimensional data into a higher dimensional data. Hyper plane is the separation line between the data classes. Boundary line separates the two classes. Support vectors are the data points that are closest to the boundary. The main benefit of support vector regression is that it fits the error rate within the boundary line but in the linear regression the errors should be minimized. This type of fitting error makes SVR better than linear regression. Another advantage of SVR is, it minimizes the overfitting problem by having low VC dimension [46].

SVR regression estimates the value of w to obtain the function

$$f(\overline{x}) = (\overline{w} \,.\, \overline{x}) + b \,, \overline{w}, \overline{x} \,\epsilon R^N, b \,\epsilon\, R \qquad \textbf{(1.5.3)}$$

By introducing ε insensitive loss function as

$$Y - f(\overline{x})/e = \max\{o, |y - f(\overline{x})| - \varepsilon\} \qquad \textbf{(1.5.4)}$$

The model produced by SVR only depends on a subset of the training data, because the cost function for building the model ignores any training data that are close to the model prediction. The support vector regression uses a cost function to measure the empirical risk in order to minimize the regression error. The support vector regression for loss function is shown in Fig. 1.12. The decision boundary is margin of tolerance that is the points within the boundary are taken for consideration [47].

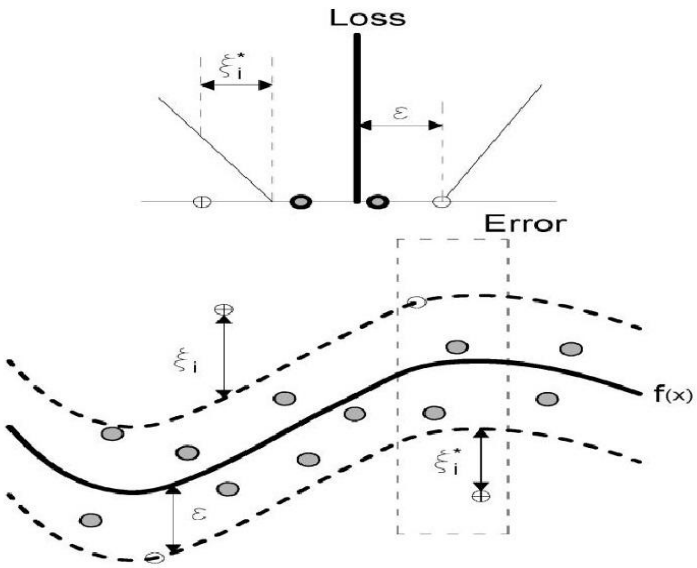


Fig. 1.12 Support Vector Regression

Random Forest Regression

Random forest is a supervised learning algorithm that uses ensemble learning method for classification and regression. Random forest is a bagging technique and works by two concepts rather than averaging the prediction of trees. It operates by random sampling of training data points while building trees and random subsets of features considered when splitting nodes. Each tree in a random forest learns from random sample of the data points. The samples are drawn with replacement known as bootstrapping where the samples will be used multiple times in a single tree. Sample random forest for regression is shown in Fig. 1.13.

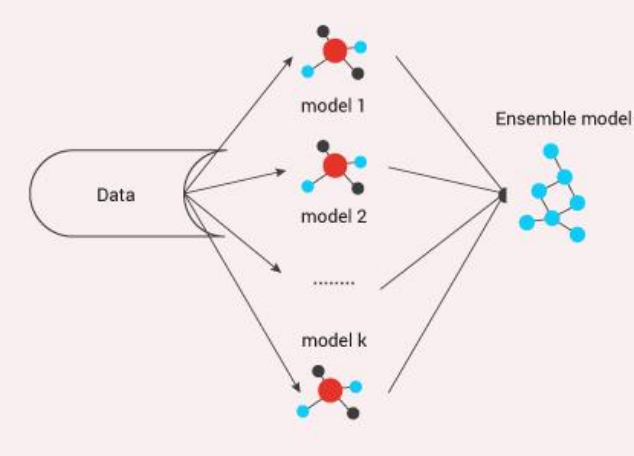


Fig. 1.13 Random Forest Regression

The random forest works by training each tree on different samples, where each tree might have high variance with respect to a particular set of the training data. The entire trees at the training time, will have lower variance but not at the cost of increasing the bias. During testing, predictions are made by averaging the predictions of each decision tree. This procedure of training each individual learner on different bootstrapped subsets of the data and then averaging the predictions is known as bagging. The other main concept in the random forest is that only a subset of all the features are considered for splitting each node in each decision tree. The random forest combines hundreds or thousands of decision trees, trains each one on a slightly different set of the observations, splitting nodes in each tree considering a limited number of the features. The final predictions of the random forest are made by averaging the predictions of each individual tree [48] [49].

Artificial Neural Network

An artificial neural network is an information processing paradigm that is inspired by the way the biological nervous system such as brain process information. It is composed of large number of highly interconnected processing elements or neurons working in unison to solve a specific problem. Neural network works like human brain, the same way input is given to the input layer and that is processed through hidden layer and the output is received using output layer. A single layer neuron network is called a perceptron. ANNs has the ability to learn and model non-linear and complex relationships [50]. The architecture of simple ANN is shown in Fig. 1.14.

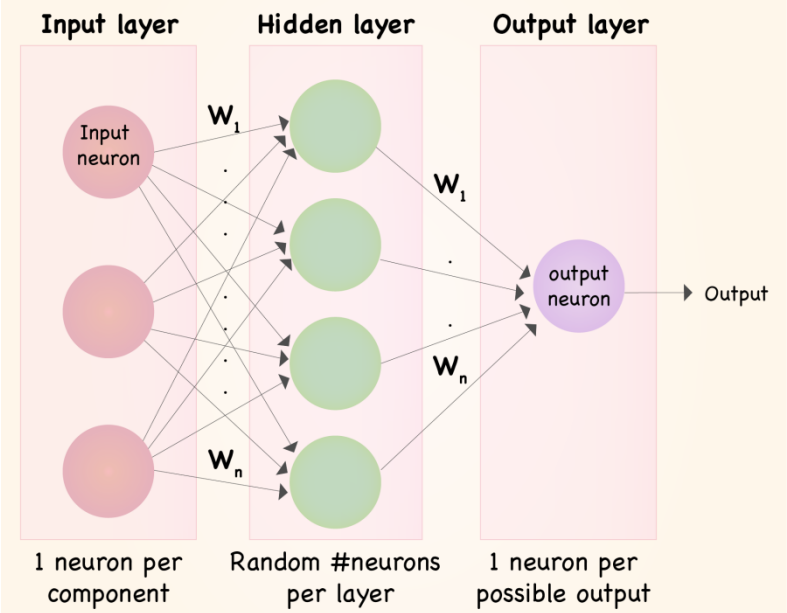


Fig. 1.14 Artificial Neural Network

ANNs allow modeling of nonlinear processes, as it turned into a very popular and useful tool for solving many problems such as classification, clustering, regression, pattern recognition, dimension reduction, structured prediction, machine translation, anomaly detection, decision making etc. ANN works better than other algorithms as it has number of hyper parameters to be tuned and by tuning the hyper parameters better accuracy is achieved. The single layer neural network is called perceptron and it is shown in Fig. 1.15.

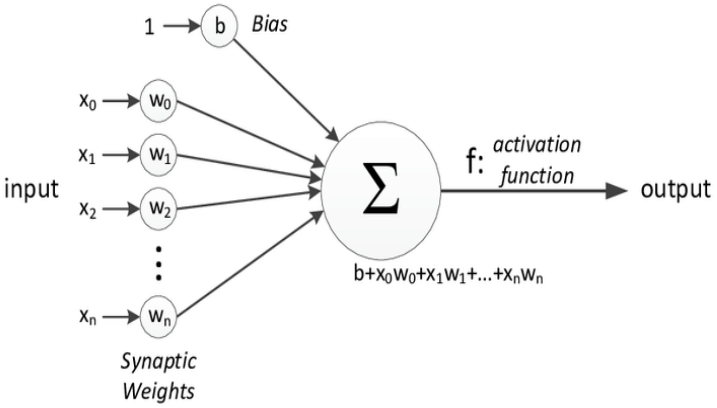


Fig. 1.15 Single Layered ANN

Here one single observation, x_0, x_1, x_2, ... x_n represents various inputs to the network. Each of these inputs is multiplied by a connection weight or synapse. The weights are represented as w_0, w_1, w_2, ….w_n. Weight shows the strength of a particular node**.** b is a bias value. A bias value allows you to shift the activation function up or down. In simple case, these products are summed, fed to a transfer function or activation function to generate a result, and this result is sent as output.

Mathematically, $x_1.w_1 + x_2.w_2 + x_3.w_3 \ldots\ldots x_n.w_n = \sum x_i.w_i$ **(1.5.5)**

The activation function applied is $\phi(\sum x_i.w_i)$

Their main purpose is to convert an input signal of a node in an ANN to an output signal. This output signal is used as input to the next layer in the stack. There are many types of activation function such as binary step function, sigmoid activation function, hyperbolic tangent function, Rectified Linear Unit (ReLu) [51].

1.5.2 Deep Neural Networks

The deep neural network evolved from the use of many more hidden layers making it a deep network to learn more complex patterns. DNN works by representation learning where manual feature extraction is not essential. From the input data it extracts the features and learns the importance of signaling between the features that makes the DNN to outperform other algorithms. DNN is widely used in automatic speech recognition, image recognition, visual art processing, natural language processing, drug discovery, customer relationship management, recommendation systems, bioinformatics, medical image analysis, mobile advertising etc., [52]. The architecture of deep neural network is shown in Fig. 1.16.

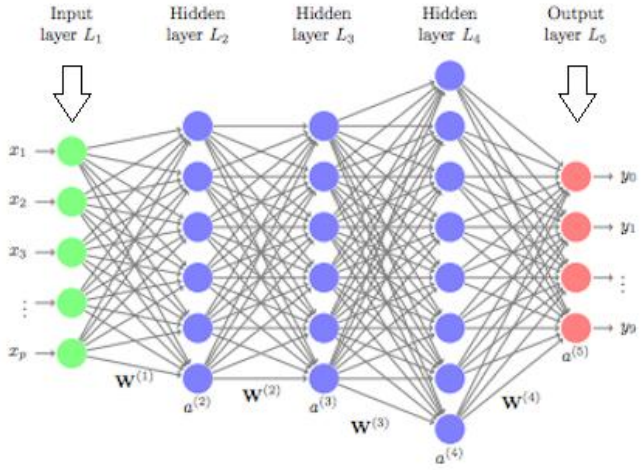


Fig. 1.16 Architecture of Deep Neural Network

The input data is consumed by the neurons in the first layer which then provides an output to the neurons within next layer and so on which provides the final output. The output is a prediction where the dependant value (Y) is predicted. Each layer can have one or many neurons and each of them will compute a small function i.e. activation function. The activation function mimics the signal to pass to the next connected neurons. If the incoming neurons, result in a value greater than a threshold then the output is passed else ignored. The connection between two neurons of successive layers will have an associated weight. The weight defines the influence of the input to the output for the next neuron and eventually for the overall final

output. In a neural network, the initial weights will be random but during the model training, these weights are updated iteratively to learn to predict a correct output. The network can be defined by few logical building blocks like a neuron, layer, weight, input, output, an activation function. Learning mechanism helps the neural network incrementally update the weights to a more suitable weight that aids into correct prediction of the outcome [53].

In this research work, three DNN architectures such as sequential DNN, functional DNN and DNN with customized layers are used for building the binding affinity predictive models. These three architectures are presented below.

Sequential Deep Neural Network

Neural Networks sequentially build high-level features through their successive layers. The new neural network model is proposed here where each layer is associated with a set of candidate mappings. The sequential DNN is a kind of deep learning models where an instance of the sequential class is created and added with model layers. The sequential models are the subset of DNN and it works with one input tensor and one output tensor. The dense layer is a fully interconnected hidden layer. The sequential deep neural network uses a stack of layers and it is implemented serially. The workflow of sequential DNN with one input tensor and one output tensor is shown in Fig. 1.17.

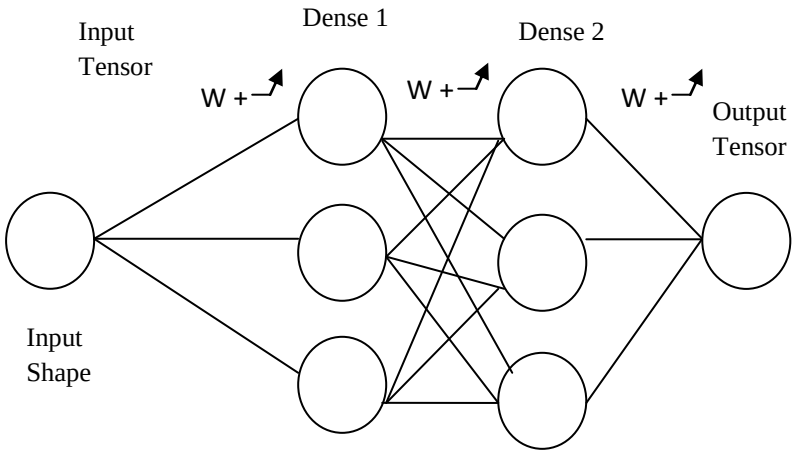


Fig. 1.17 Sequential DNN with One Input Tensor and One Output Tensor

When an input is processed, at each layer, one mapping among these candidates is selected according to a sequential decision process. The resulting model is structured according to a Directed Acyclic Graph (DAG) like architecture, so that a path from the root to a leaf node defines a sequence of transformations.

In this model, hand crafted features are fed as feature vectors and it is assigned to the input tensor as input pattern. The input pattern is specified in advance where the weights are created according to pattern of the input. The dense layers pass the features along with assigned weights to the features. The activation function Rectified Linear Unit (ReLu) is used to learn the patterns. The Dense 1 layer creates a new feature set and passes to the next dense layer. When building a sequential model the layers can be incrementally added with add function. The model is trained with various hyper parameters and evaluated with metrics. The way of creating sequential model is that layer will be stacked one after the other.

Hyper parameters are used in optimizing the model like learning rate, epochs, dropouts, activation function and optimizers. The sequential DNN uses various hyperparameters such as dropouts, optimizers, activation function, learning rate, loss function and epochs.

Optimizers: Optimizers are algorithms or methods used to change the attributes of neural network such as weights and learning rate in order to reduce the losses. There are many optimizers such as gradient descent, stochastic gradient descent, mini-batch gradient descent, Adam, RMSprop, Nadam, Adadelta etc. Three optimizers are used like Adam, RMSprop and Nadam are used in this research work.

Adam: The efficient optimization algorithm is Adam. This optimizer is extension of stochastic gradient optimization algorithm and Adam optimization algorithm computationally updates network weights iterative based on training data. Instead of adapting the parameter learning rates based on the average first moment, Adam also makes use of the average of the second moments of the gradients. The algorithm calculates an exponential moving average of the gradient and the squared gradient, and the parameters beta1 and beta2 control the decay rates of these moving averages.

RMSprop: RMSprop optimizer is an optimizer that utilizes the magnitude of recent gradients to normalize the gradients. Learning rate will be tuned in this optimizer. It is developed as a stochastic technique for mini-batch learning. RMSprop uses a moving average of squared gradients to normalize the gradient. This normalization balances the momentum, decreasing the step for large gradients to avoid exploding, and increasing the step for small gradients to avoid vanishing.

Nadam: Nadam optimizer is Nesterov Adam optimizer. This optimizer is much like Adam optimizer and it is the combination of Adam, RMSprop with nesterov momentum. Activation is the activation function for the layer.

Activation Function: An activation function is added to the network to help the network learn complex patterns. It allows models to take into account nonlinear relationships. ReLu is

the commonly used activation function that gives better accuracy. The function returns 0 if it receives any negative input, but for any positive value xx it returns that value back. So it can be written as

$$f(x)=\max(0,x) \quad \textbf{(1.5.6)}$$

Learning Rate: Learning rate is the parameter that indicates the optimizer to maneuver the weights within the direction opposite of the gradient for a mini-batch. The learning rate is a configurable hyper parameter used in the training of neural networks that has a small positive value, often in the range between 0.0 and 1.0. Smaller learning rates require more training epochs given the smaller changes made to the weights each update, whereas larger learning rates result in rapid changes and require fewer training epochs.

Dropout: Regularization is used to restrain overfitting the training data. Dropout is a regularization parameter that randomly skips neurons during training, forcing others in the layer to pick up the slack. Dropout is implemented by randomly selecting nodes to be dropped-out with a given probability in each weight update cycle.

Epochs: Epoch is a hyper parameter that defines the number of times the learning algorithm will go through the entire training dataset.

Batch Size: The batch size defines the number of samples that will be propagated through the network. The larger batch size enables using a large learning rate. Larger batch sizes tend to have low early losses while training whereas the final loss values are low when the batch size is reduced.

Loss Function: A loss function is used to optimize the parameter values in a neural network model. It maps a set of parameter values for the network onto a scalar value that enables the network to accomplish the task.

The sequential DNN is suitable for developing models for all the problems and it is the simplest model that provides good accuracy. But sequential DNN is not straightforward to define models that may have multiple input sources, produce multiple output destinations or models that re-use layers. Complex models cannot be built using sequential DNN and also input layers cannot be shared. It is limited in model topology [54].

Functional Deep Neural Network

In sequential DNN, the layers are not connected pairwise, layers cannot be shared and hence the complex models cannot be built. Functional models provide flexible way of creating models. The functional DNN can handle models with non-linear topology. Feature

vectors are shared among the layers in functional DNN and also it supports multiple inputs or outputs.

The functional model is a basic graph with three layers and it is built by creating the input nodes. The same graph of layers can be used to generate multiple models. The batch size is not required to define since the pattern of each sample is specified when the input layer is created. The output of input layer has the information about the input pattern as well as the type of data that is fed to the network. The new node in the graph of layers is created following the input node where the inputs are passed to dense layer. The nodes can be extracted and reused in the graph of layers.

The functional models can be serialized like sequential models and thus the model can be recreated. The architecture can be reused with the weights applied to the inputs. The functional deep models can be nested as sub-models through ensembling. Another use of functional model is such that layers are been shared. Shared layers are layer instances that are reused multiple times in the same model as they learn features that correspond to multiple paths in the graph-of-layers. It is often used to encode inputs from similar spaces. Layers can be connected not only in pairwise but it may be connected in any order that creates complex structure.

The functional models are defined by creating instances of layers and connecting them directly to each other in pairs. The functional model is defined by specifying the layers to act as the input and output to the model. The workflow of functional DNN is shown in Fig. 1.18.

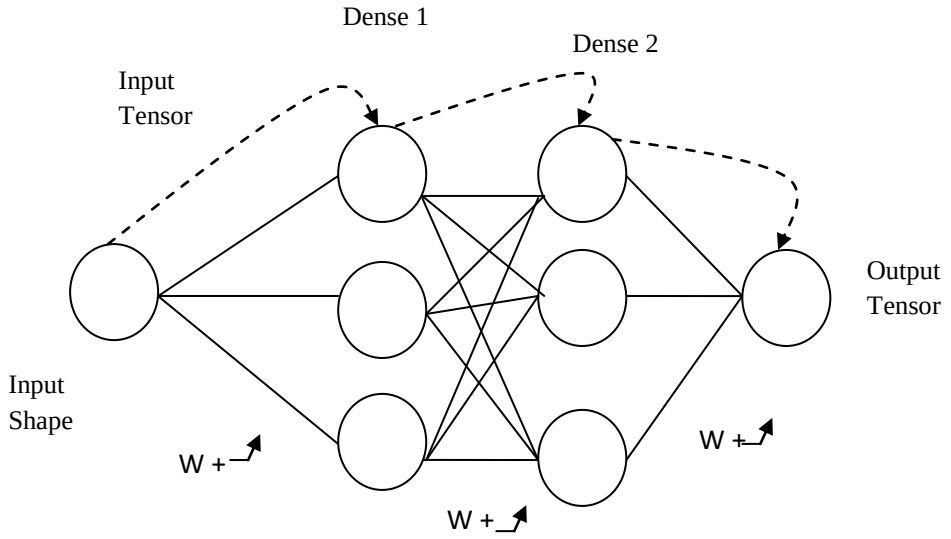


Fig. 1.18 Functional DNN with One Input Tensor and One Output Tensor

Here feature vector is passed and the inputs are shared among the layers. Unlike the sequential model, standalone input layer is created and defined to specify the pattern of input

data. The input layer takes a pattern as a tuple which indicates the dimensionality of the input data. The layers in the model are connected pairwise. The output of dense layer is passed as input to the next layer. In the output layer, the output of dense 2 is given as input and output is calculated. The functional DNN uses the same layers as the sequential DNN but more flexibility is provided in functional DNN.

Since the layers are connected it remembers the past data. Before training the model, the learning process is configured. The same hyper parameters that are used for building sequential deep models are used in case of functional DNN to build the models. Since the layers are shared and features are shared among the layers it produces better accuracy than the sequential DNN.

The advantage of functional DNN is it provides visualizing the graph of layers. It creates sub-models that have the multiple inputs and then the layers are merged to give a more robust and discerning output. Layers sharing flexibility leads to built complex models. The disadvantage of functional DNN is that the pre-trained weights are not defined [55].

Deep Neural Network With Customized Layers

As the weights are not trained, the error rate is not minimized in functional DNN. To overcome the limitation of functional deep neural network and to improve the prediction rate, DNN with customized layers is used to build models. Customized layer is created when the available dense layer does not meet the requirements. Custom layer is created to find weight based on normal distribution. The process of layer customization is shown in Fig. 1.19.

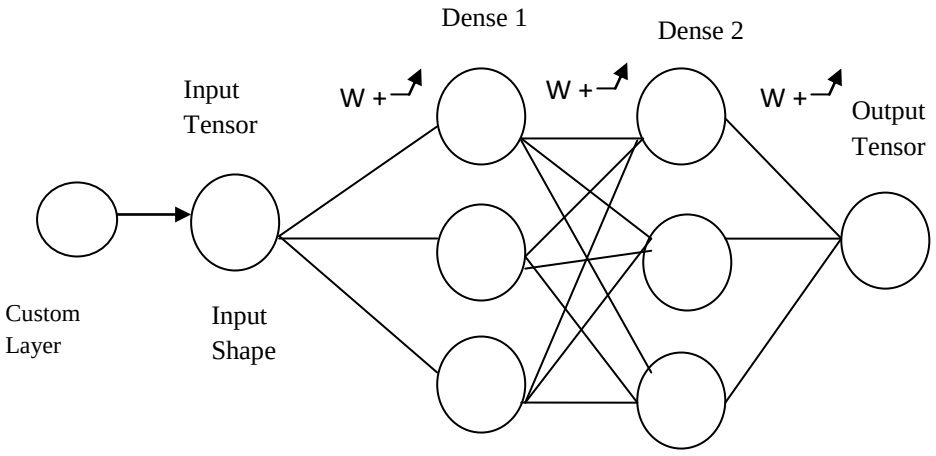


Fig. 1.19 DNN with Customized Layer Using One Input Tensor and One Output Tensor

Here custom layer is built by defining the class variable and super class variable. The trainable weights are defined and the custom layers are created subsequent to variable

definition. The weight is defined corresponding to input pattern and set in the kernel. The kernel is the custom functionality of the layer and it creates the weight using normal initializer.

The weights are updated for each iteration, until the best prediction occurs. The trainable weight from the user defined layer is fed to the dense layers, with the purpose of fine tuning parameters and reducing the learning time. The hyper parameters for DNN with customized layers vary with activation function and loss function. The softmax activation function is used here to calculate the probabilities of target variables over all possible targets. This activation function is used to achieve the high probability. The predictive model is constructed with sparse categorical cross entropy loss function for training, since the targets are integer variables and suitable for prediction problems.

The parameter dropout is tuned by selecting the nodes randomly which are to be dropped-out with a given probability during updation of weights. Learning rate is set to contrive the weights within the direction opposite of the gradient for given input values. Optimization algorithm calculates an exponential moving average of the gradient and the squared gradient, where the parameters Beta and Epsilon control the decay rates of these moving averages.

The advantage of creating customized layers is such that the weights are trained and shared among the layers, thus produces better accuracy than the sequential DNN and functional DNN [56].

1.6 PROBLEM STATEMENT AND OBJECTIVES OF THE RESEARCH

The prediction of binding affinity for SCA is imperative to be investigated as the protein structures changes due to mutation which in turn affects the docking results. The modified docking results change the binding affinity of SCA when the protein structures are mutated. The complexes of various interactions such as protein-protein, protein-ligand and mutated protein-ligand are mandatory to monitor the effect on the energy calculations. Energy calculations along with scoring functions, molecule descriptors are the most influencing factors in binding affinity prediction. In traditional approaches, hand crafted features are used to predict binding affinity which require domain expertise. Whereas in deep learning, the intelligent hints from the hand crafted features are drawn through representation learning and is crucial as it contributes in accurate prediction of binding affinity. Hence this research work is proposed to construct efficient binding affinity prediction models. The binding affinity

prediction problem is formulated as regression task and can be solved with suitable models developed using deep learning.

The main aim of this research work is to develop binding affinity predictive models for SCA disorder through traditional regression techniques and contemporary deep learning approaches. The core objectives of this research work 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

A novel approach of building an effective model to predict binding affinity for SCA disorder using the mutated protein structures of human species is attempted. The binding affinity prediction problem is formulated as regression tasks and various predictive models are built by training the using hand crafted features. An effective solution is proposed by building predictive models using traditional regression algorithms and contemporary deep learning architectures. These approaches will simplify and generalize the prediction problem and can generate reliable models by learning the intelligent hints from the interacted and docked complexes.

1.7 ORGANIZATION OF THE THESIS

The rest of the thesis is organized as below.

Chapter 2 discusses about the literature review of binding affinity prediction methods using general approaches and traditional approaches.

Chapter 3 presents the process of corpus development, feature identification and creation of datasets. This chapter also describes proposed architecture of binding affinity

prediction model and also explains the parameters used in training and the evaluation metrics used in performance analysis.

In chapter 4, the implementation of traditional regression techniques for building binding affinity prediction models with protein-ligand docking is presented. This chapter also explains the contribution of features such as energy calculations and physical properties in binding affinity prediction and the development of respective protein-ligand docking dataset. Experiments carried out using linear regression, support vector regression, artificial neural network regression, random forest are discussed in detail. The performance analysis of the results and findings of the experiments are also reported in this chapter.

Chapter 5 presents the importance of protein-mutated-ligand docking and the features such as scoring function, energy calculations, sequence descriptors in binding affinity prediction. It also elucidates the implementation of supervised regression techniques for building binding affinity prediction models with protein-mutated-ligand docking dataset. Experiments carried out using linear regression, support vector regression, artificial neural network regression, random forest are discussed in detail. The performance results of these binding affinity predictive models are analyzed and reported in this chapter.

Chapter 6 presents the significance of protein-protein interaction and the features such as energy calculations, interfacial properties, NIS properties in binding affinity prediction. It also explains in detail about the implementation of supervised regression models for binding affinity prediction with protein-protein interaction. Experiments carried out with linear regression, support vector regression, artificial neural network regression, random forest are also discussed. The performance results of these binding affinity predictive models are analyzed and reported in this chapter.

Chapter 7 describes in detail about the development of deep models based on three different datasets using three deep neural network architectures such as sequential neural network, functional deep neural network and deep neural network with customized layers. The implementation results of all the deep architectures based on various evaluation metrics are presented in this chapter. The performance comparison of all the models based on DNN architectures and regression algorithms is also analyzed and reported in this chapter.

Chapter 8 summarizes the entire research work with various findings of traditional regression techniques and deep learning approaches in predicting binding affinity prediction. This chapter also presents the achievements of the proposed research work and research contributions. The scope of future research is also discussed.