Analysis of Protein Sequencing for Drug Discovery

A Thesis Report
Submitted in partial fulfillment of the requirements for the Degree of
Bachelor of Science in Computer Science and Engineering

Submitted by

Hasan Bin Jamal	**180104070**
Fatima Juairiah	**180104071**
Abu Tarek Rabbi	**180104086**
Abdullah Al Mohaimen	**180104098**

Supervised by

Dr. S.M.A. Al-Mamun

Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

Dhaka, Bangladesh

December 2022

Contents

1.2 Problem Definition

We are trying to figure out GPCR proteins among the three categories of proteins – Ion Channels, Nuclear receptors and GPCR as GPCR is serving the highest numbers of medicines among the all protein types. After that we are searching for the druggable GPCR. Druggable GPCR's are further classified for finding drugs for three types of disease – 'Heart Disease', ' Heart Burn', and 'Schizophrenia' .

1.3 Our Goal

Our primary purpose is to help to meet the deficit in medicines in our country by enriching the pharmaceutical sector. We aim to study Machine Learning and Artificial Neural Network Models such as Random Forest, Deep neural network, raw data processing, biology, pharmaceuticals, data predicting and finding reliable drug target proteins for diseases. We aim to provide fast pace solutions for the increasing amount of diseases by machine intelligent techniques.

Along with the invention of newer drugs using machine intelligence techniques will reduce the cost of medicines to serve the mass population of underdeveloped countries. In doing so, we strive to contribute a fast and low expensive methodology for discovering drugs for our growing pharmaceutical sectors using Machine Learning and Deep Learning.

1.4 Outline of the Report

This report contains an abstract and an introduction along with a few customary declaration pages at the beginning, a bibliography and appendices at the end. The remainder of this thesis report has been organized as follows:

- **Chapter 2 - Literature Review**
  This chapter includes a background study on the problem domain, an overview of various ML and Deep neural network and discussion on relevant papers of studies previously conducted.

- **Chapter 3 - Data Preprocessing**
  The collection method and preliminary data analysis of the dataset in use have been presented in this chapter. Moreover, preprocessing of the data using downsampling and smoothing has been reviewed along with the processing of the dataset used for validation.

Some of the most famous protein families:

G-Protein-Coupled-Receptors(GPCR): - GPCR stands for G-protein coupled receptor. It is a type of cell surface receptor that binds to extracellular ligands and activates an intracellular G-protein. GPCRs are involved in a wide range of physiological processes, including vision, olfaction, taste, and hormone signaling. They are also involved in the regulation of cell growth and differentiation. GPCRs are transmembrane proteins that span the cell membrane seven times. They are composed of three domains: an extracellular domain, a transmembrane domain, and an intracellular domain.

Ion channels: - Ion channels (LICs, LGIC), commonly referred to as ionotropic receptors, are a group of transmembrane ion-channel proteins that open in response to the binding of a chemical messenger (i.e., a ligand), such as a neurotransmitter.

Nuclear receptor: - According to the science of molecular biology, nuclear receptors are a collection of proteins in cells that sense thyroid hormones, steroid hormones, and other substances. These receptors respond by controlling the expression of specific genes, which in turn controls the organism's growth, homeostasis, and metabolism.

2.2 Drug Discovery with Targeted Proteins

2.2.1 Major Protein Families as Drug Targets

- GPCR
- Ion Channels
- Nuclear Receptor
- Kinases

From the below pie chart it is clear that GPCR and ion-channels are the most frequently used protein families as drug targets. [7]

Chapter 3

Data Preprocessing

We have taken our data from UniProt Organization [11]. A substantial source for information on protein sequences and annotations is the Universal Protein Resource (UniProt). The European Bioinformatics Institute (EMBL-EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource collaborate to create UniProt (PIR).

3.1 Data Set Overview

We have 4 data sets with a total of 33600 data. The First dataset contains 32000 data, the second dataset contains 700 data, the third dataset contains 300 data, and the fourth dataset contains 600 data.

3.1.1 DataSet 1

The first dataset contains 3 types of protein class. Those are:

1. **GPCR** - G protein-coupled receptors (GPCRs), also known as seven-(pass)- transmembrane domain receptors, are integral membrane proteins that are used by cells to transform extracellular signals into intracellular responses. These include responses to hormones and neurotransmitters as well as responses to signals from the senses of sight, smell, and taste. Given that they can be utilized to control a range of illnesses, from pain to cancer, GPCRs are the most frequently chosen targets for drug discovery.

2. **Ion-Channel** - Ion channels are protein molecules that span across the cell membrane that allows the passage of ions like Na+, K+, Ca2+, and Cl from one side of the membrane to the other. Voltage-gated ion channels, Ligand-gated ion channels

stroke, heart attack, and kidney disease. It is important to monitor your blood pressure regularly and make lifestyle changes to keep it in check.

3. **Schizophrenia** - Schizophrenia is a mental disorder characterized by abnormal social behavior and failure to recognize what is real. It is often accompanied by hallucinations, delusions and disorganized thinking. Treatment usually involves a combination of medications and psychosocial interventions.

3.1.4 Data Set 4

This dataset contains unknown protein sequences that need to be classified. It is the fourth dataset in a series of datasets used for research purposes. The data is used to identify and classify proteins and to gain an understanding of their structure and the roles they play in biological processes. As novel medications and therapies are developed, this dataset is crucial for assisting researchers in better understanding the structure and functionality.

3.2 Dataset Preprocessing

The features which we are choosing here -Protein sequence, Protein length, Molecular mass and Gene. Our total features will be 23. Among them, 20 features are extracted from protein sequence.

3.2.1 Feature Selection

Protein sequence: Each protein or peptide is made up of a specific linear sequence of amino acids. A protein or peptide's amino acid sequence can be used to study it, identify it in a sample, and classify its post-translational modifications. Protein sequencing is the process of figuring out the amino acid sequence. [14]

Molecular mass: By adding the molecular weights of the relevant amino acid sequences, one can calculate the molecular weight (MW) or mass of a protein. The mw may alter as a result of specific changes to this sequence. Protein The way the amino acids are bonded to one another determines the molecular mass, which varies due to the creation of the amino acid sequence. So multiple weights can exist for the same protein structure. The Dalton is the molecular mass unit here. A unit called a dalton(1.6605300000013E-24 g), which is atomic mass unit-equivalent, is used to express the molecular weight of proteins.

Gene name: A gene is a fundamental building block of heredity and a DNA sequence of nucleotides that codes for the production of a gene product, either RNA or protein. A specific

3.2.4 Feature Extraction

Feature extraction is the process of extracting and selecting relevant features from a given dataset for use in a predictive model. It is an important step in the machine learning process as it helps to reduce the complexity of a dataset and provide a more accurate model. Feature extraction techniques include dimensionality reduction, feature selection and feature transformation. These techniques can help to reduce the complexity of a dataset, reduce overfitting, and improve the accuracy of a predictive model.

We were able to get around 30800 protein sequences from three families from the UniProtKB1 protein databases. We calculated characteristics for each protein sequence in each family. We determined the distance between the first residue in each protein sequence and each amino acid residue. then determined the average separation between the first residue and each subsequent amino acid residue. For example, if a particular sequence contains the residue "A" 14 times, we will figure out how far away "A" is from the first residue on each of those occurrences. The mean of those 14 distances will then be used to get the feature value for "A".Each protein sequence will have 20 feature values since there are [15] The average distance between every occurrence of amino acid "A" in the sequences 1, 2, 3, and 4 in Table III is 17, 8, 25, and 10 correspondingly. For each sequence, we get 20 such distances. As a result, the dataset measures 30800 X 20.

Example of Amino acid:

MIKTALLFFATALCEIIGCFA

Number of Alanine(A) = 4
Vector distances of A

5	10	12	21

Average Vector Distance = (5+10+12+21)/4 = 48/4 = 12

Number of Leucine (L) = 3
Vector distances of L

6	7	13

Average Vector Distance = (6+7+13)/3 = 26/3 = 9

Alanine (A)	Methionine (M)	Isoleucine (I)	Leucine (L)
12	1	13	9

3.3 Data Set Analysis

3.3.1 DataSet 1

GPCR had the highest Count of Protein names at 10,458, followed by ION CHANNEL at 9,841 and Nuclear Receptor at 8,183. GPCR accounted for 36.72% of Count of Proteins. So we can see our first dataset is almost balanced.

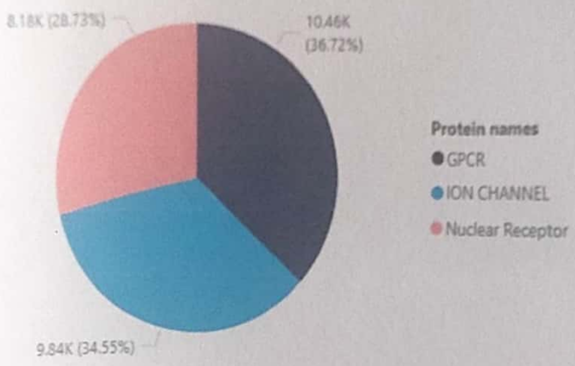

Figure 3.1: Ratio of GPCR , Ion Channel and Neuclear Receptor

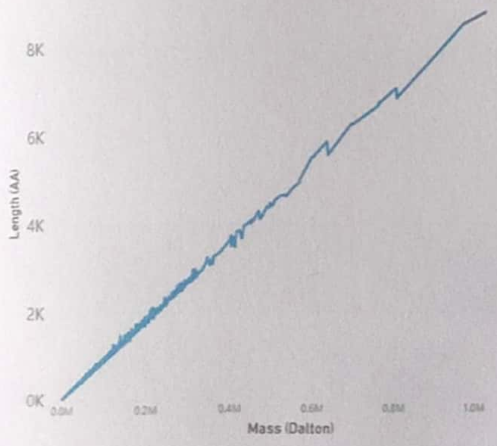

Figure 3.2: Co-relation of Length and Mass

From the above figure, we can see that length and mass are proportional to each other.

Chapter 4

Methodology

Deep Neural Network, K-Nearest Neighbour, Decision Tree, K-Means Clustering and Guassian Clustering are the ML models that we have selected for concentrating on. But before the models could be trained, the dataset were processed so that we can use the sequence as feature and extract the most imrportant features.

4.1 Deep Neural Network

We have implemented Feed Forward Neural Network. The information in this network only travels in one direction, forward, from the input nodes to the output nodes, passing via any hidden nodes that may exist. The network contains no loops or cycles. We worked with Keras libraries and packages. We split our dataset into train and test portions, where we kept 20% of data for testing. Our kernel initializer was 'he_uniform'. The input dimension was 23, as our features were 23.

In our model, we have one input layer, 3 hidden layers and one output layer. Relu activation functions are used in all the hidden layers and output layer has the 'Softmax' activation function as it is a multiclass classification. We used Adam optimizer and a loss function named 'Sparse Categorical Crossentropy'. We divided our dataset into the batch size of 10 and epochs of 150.

we applied deep neural network in our dataset 1.

Out of all the models, Deep Neural Network provided us with the highest accuracy (86%) possible.

4.2 K-Nearest Neighbour

At first, we applied K-Nearest Neighbor (KNN) Algorithm for classify Proteins into one of these three protein families – GPCR (G – Protein Coupled Receptors), Ion- Channels, and Nuclear Receptors.Here Dataset-1 is used as train and test set. Then we used this model to test again for unknown protein sequences (dataset-4). We calculate different error rate for different values of K. finally selected k value as 10 for having the lowest error rate.

4.3 Decision Tree

In this step we have checked whether a GPCR protein can be used in making drugs or not. We classified dataset-2 GPCR Proteins into Druggable or NON-Druggable proteins. First, we train and test dataset 2 with ML model Decision Tree. We have taken max depth 10 for our Decision Tree Classifier. The proteins which were Predicted as GPCR proteins are further passed to this trained model For druggability checking. In this way labelled our dataset-4 (Unknown Proteins) Firstly as GPCR proteins and then as druggable or not.

4.4 Clustering

In the last stage (dataset 3), Druggable GPCR Proteins are further classified into 3 drug classes - Drugability for Heart disease, High blood pressure, and Schizophrenia. We train this dataset with the unsupervised ML model K-means clustering and Gaussian Clustering. Using this ML model, we test the unknown sequences taken from previous predicted as Druggable GPCRs' for testing their druggability for the specific classes, which are 'Heart Disease,' ' High Blood Pressure, 'Schizophrenia.'

4.4.1 K-Means Clustering

K-Means Clustering is an unsupervised machine learning algorithm used to identify patterns in data. It can be used to identify druggable data for various diseases such as heart disease, high blood pressure, and schizophrenia. The algorithm works by first dividing the data into 'k' clusters, each representing a particular set of data points. Then, it iterates over the clusters and assigns each data point to its closest cluster. This helps in identifying and categorizing the data points according to their patterns. The algorithm then uses the patterns to predict which drugs should be used to treat various diseases.

Chapter 5

Performance Analysis

5.1 Result Comparison

Result comparison of our models is a way to assess the performance of different models on a given task. By comparing the results of different models, we can gain insight into which model performs the best and can make the best decisions when given a particular problem. This can help us choose the best model for a given task, as well as help us identify areas for improvement in our existing models. In addition, result comparison of our models can help us understand the strengths and weaknesses of different models. This can help us better understand the capabilities and limitations of the different models and make better decisions when choosing the right model for a particular task.

5.1.1 Confusion Matrix

Confusion Matrix is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values.

True Positive - A true positive is when a test result correctly indicates that a condition or attribute is present.

False positive - A false positive is when a test result incorrectly indicates that a condition or attribute is present.

True negative - A true negative is a test result that correctly identifies an absence of a condition.

False negative - A false negative is when a test result incorrectly indicates that a condition or attribute is not present.

References

[1] "Bangladesh - country commercial guid." https://www.trade.gov/country-commercial-guides/bangladesh-healthcare-and-pharmaceuticals, Dec. 2022.

[2] D. S. Dimitrov, *Therapeutic Proteins*, pp. 1–26. Totowa, NJ: Humana Press, 2012.

[3] "Bangladesh - country commercial guid." https://cbd.cmu.edu/about-us/what-is-computational-biology.html, Dec. 2022.

[4] K. Han, M. Wang, L. Zhang, Y. Wang, M. Guo, M. Zhao, Q. Zhao, Y. Zhang, N. Zeng, and C. Wang, "Predicting ion channels genes and their types with machine learning techniques," *Frontiers in Genetics*, vol. 10, p. 399, 2019.

[5] "Protein formation process (from gene to protein)." https://www.lgmd2ifund.org/science-basics/from-gene-to-protein?fbclid=IwAR3R38S4EHRqISlrzgILSMENGOi-mTkVNmOSscpaM68IwT5VYAmQ_vqV7og, Dec. 2022.

[6] B. Satpute and R. Yadav, "Machine intelligence techniques for protein classification," in *2018 3rd International Conference for Convergence in Technology (I2CT)*, pp. 1–4, IEEE, 2018.

[7] R. Santos, O. Ursu, A. Gaulton, A. P. Bento, R. S. Donadi, C. G. Bologa, A. Karlsson, B. Al-Lazikani, A. Hersey, T. I. Oprea, *et al.*, "A comprehensive map of molecular drug targets," *Nature reviews Drug discovery*, vol. 16, no. 1, pp. 19–34, 2017.

[8] "Decision tree in machine learning models by python." https://www.jcchouinard.com/decision-trees-in-machine-learning/, Dec. 2022.

[9] "Clustering with gaussian mixture model." https://medium.com/clustering-with-gaussian-mixture-model/clustering-with-gaussian-mixture-model-c695b6cd60da, Dec. 2022.

[10] Y. Parikh and E. Abdelfattah, "Machine learning models to predict multiclass protein classifications," in *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pp. 0300–0304, IEEE, 2019.

[11] "The Universal Protein Resource (UniProt)." https://www.uniprot.org/, Dec. 2022.

[12] K. Sriram and P. A. Insel, "G protein-coupled receptors as targets for approved drugs: how many targets and how many drugs?," *Molecular pharmacology*, vol. 93, no. 4, pp. 251–258, 2018.

[13] J. T. Wang, Q. Ma, D. Shasha, and C. H. Wu, "Application of neural networks to biological data mining: a case study in protein sequence classification," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 305–309, 2000.

[14] Q. Li and L. Lai, "Prediction of potential drug targets based on simple sequence properties," *BMC bioinformatics*, vol. 8, no. 1, pp. 1–11, 2007.

[15] S. Degadwala and D. Vyas, "Data mining approach for amino acid sequence classification," *International Journal of New Practices in Management and Engineering*, vol. 10, no. 04, pp. 01–08, 2021.