Prediction of Activity of Short Anti Microbial Peptides using Machine Learning

Nadilla Yaswanth Baba

Department of Computer Science and Engineering

Amrita School of Engineering, Bengaluru
Amrita Vishwa Vidyapeetham
Amrita University, India
bl.en.u4aie20015@bl.students.amrita.edu

***Abstract*— Antimicrobial Peptides (AMPs)** play a vital role in the natural immune systems of various animals and are to be found in abundance in nature. Coming to bacteria, fungi, parasites, and viruses, these AMPs have a wide range of inhibitory actions. So, in this project we are using various ML models to predict the activity of **Short Antimicrobial peptides.**

Keywords—AMP, CD-HIT

Abbreviations and Acronyms

AMP-Antimicrobial Peptide

CD-HIT – Cluster Database at High Identity with Tolerance

I. INTRODUCTION

Microbes are microscopic living things that are found everywhere and are too miniscule to be seen by the naked eye. They live everywhere, in water, soil, also in the air. The human body also contains millions of these microbes, also called as microorganisms. Some of these microbes make us sick, while others are important for our health. The most common types are Bacteria, Viruses, Protists, Archaea and Fungi. There are also some microbes called as protozoa. These are responsible for diseases such as toxoplasmosis and malaria. Microscopic plants are also considered as microbes, though they don't live on or in the human body.

Peptides are the short amino acid strings typically comprising 2 to 40 Amino acids. They are the building blocks of proteins. Peptides are easier for the body to absorb than proteins as they are smaller and can be broken down easily than proteins. They can also penetrate easily through the skin and intestines which provides it the capacity to penetrate easily into the bloodstream. There are a lot of peptides which perform different roles in our body. Some of the benefits of Peptides are: Anti-aging, improving the skin barrier, muscle growth. These are connected to each other in a sequence called as Peptide Bonds.

Antimicrobial Peptides (AMPs) are the small peptides which play a vital role in the innate immune system of various living beings. They are a valuable source of antimicrobial agents and a solution to multi-drug resistance problem.

Particularly, short length AMPs are shown to have enhanced antimicrobial activities, higher stability and lower toxicity to human cells.

Cationic AMPs receives intense interest in recent years, for its development of antibacterial drugs.

II. LITERATURE SURVEY

Maxwell W. Libbrecht and William Stafford Noble have outlined some of the main applications of machine learning to genetic and genomic data. In the process, they had identified some recurrent challenges associated with this type of analysis and provided general guidelines to assist in the practical application of machine learning to real genetic and genomic data. Computational intelligence techniques have many characteristics such as adaption and fault tolerance that made them attractive for research on bioinformatics. A machine learning approach is introduced for classifying network. The objective of machine learning is to discover and learn and then adapt to the circumstances that might change over time and therefore improving the performance of the machine. In the field of bioinformatics, the reference input is used for the algorithms of machine learning so that they "learn." The ability of soft computing techniques to deal with uncertain and partially true data makes them attractive to be applied in bioinformatics.

- Machine learning techniques can be used here to train the network for better performance and enhancing the accuracy of the system.
- Moreover, Machine learning tools are used to decrease false positive rates.

III. Theory and Concept *A.*

Anti-Microbial Peptides (AMPs)

The emergence of Antibiotic-resistant microorganisms and the increasing concerns about the use of antibiotics resulted in the development of AMPs, which gave us a application prospect in medicine, food, animal husbandry, agriculture.

Research on AMPs is continuously developing and significant amount of data on AMPs have been stored in the AMP Databases.
But however, the complete mechanism of AMP remains as a mystery and further works need to be performed to establish the relationship between physiochemical properties to obtain low-cost and safe AMPs with remarkable antimicrobial effects.

B. p-feature

Pfeature is an extensive programme designed to compute a variety of protein/peptide properties that have been learned over the past few decades. For computing protein characteristics based on, it has the following five primary modules:

i) Composition ii) Binary profiles

iii) Evolutionary information iv) Structure v) Pattern

Computing a protein's descriptors is one of the difficulties in creating a prediction model using machine learning methods. Pfeature is a feature-rich, comprehensive package that can be used for a variety of categorization jobs.

Additionally, it has a module called "model building" that uses the characteristics produced to create classification and regression models.

Functions of P-feature:

- Calculation of a protein's or peptide's composition and physicochemical characteristics.
- Computation of protein or protein segment binary profiles.
- Computation of protein evolutionary data represented by PSSM profiles.
- Determining a protein's structural characteristics based on its structure
- Identification of proteins' pattern-based characteristics
- Modules for creating regression and classification models
- Databases PDB and UniProt ID can be used to retrieve protein sequences.

C. CD-HIT

For clustering and comparing protein or nucleotide sequences, a very popular tool is called CD-HIT. Dr. Weizhong Li was the person who first created CD-HIT.
It can carry out a number of tasks, including clustering a protein database, DNA/RNA database, comparing those databases, and creating protein families. In addition to having several scripts, CD-HIT is incredibly quick and can manage very huge databases. In many sequences analysis jobs, CD-HIT greatly reduces the computational and manual work required. It also helps to comprehend the data structure and correct biases in a dataset

D. DNA Sequence as K-Mer

Machine Learning is a part of AI, that gives the systems the capacity to automatically learn from data and past experiences in order to recognize patterns and make predictions. ML apps may free learn from new data and grow, develop, and adapt. By using these algorithms to find patterns and learn in an iterative process, machine learning extracts meaningful information from massive amounts of data. Types of the ML techniques include:
Supervised ML, Unsupervised ML, Reinforcement ML.
The ML models that we have used in our project are:

- **Random Forest Classifier**
- **Decision Tree Classifier**
- **K Neighbouring Classifier**
- **Gaussian NB**
- **Linear SVC**

IV. Methodology

Now that we have learned about the CD-HIT and Pfeature Library, we will apply our knowledge to a real-life Machine Learning use case.
Objective: Predict the Activity of a short antimicrobial peptide
.
Firstly, we will install Conda in colab, then install the Pfeature library using the wget function in python and unzip it. Then inside the Pfeature folder, we will access the pfeature.py with the change directory command.
We then install the CD-HIT package and then load the Peptide dataset using the wget command from the fasta files of antimicrobial peptides and non- antimicrobial peptides which are mentioned with Reference Paper.
We can also view the train_ne.fasta lists by using the cat command. This shows a list of all the negative fasta sequences.
We then remove the redundant sequences using the CD-HIT package. Then, we calculate features using the Pfeature library. Then we define the functions for calculating different features and display them by importing aac_wp and dpc_wp from the Pfeature library.
We then merge the positive and negative classes into one and then we pre-process the data and then we then use our knowledge of ML here where we use the ML models to predict the activity of the antimicrobial peptides.
These models produce the accuracy score, and we can also find out which ML model is the best to predict the peptides.

V. Description of Dataset

The dataset we have taken is from the fasta files of antimicrobial peptides and non- antimicrobial peptides which are mentioned with Reference Paper which are accessed by using the inbuilt commands in Python and Google Colab.

VI. Steps Followed

The entire process of predicting using SVM can be broadly summarized in these steps:

1. Our project is to predict the activity of short antimicrobial peptides.
2. Firstly, we will **install Conda** in Google Colab. (Miniconda from Anaconda)
3. Next, we Download and Install Pfeature library and unzip it.
4. We then install the CD-HIT for clustering.
5. Then, we load the PEPTIDE DATASET.
6. We then remove the redundant sequences using CD-HIT.
7. Then, we calculate the features using the Pfeature library.
8. We then define the functions and calculate the different features.
9. Calculate the features for both POSITIVE and NEGATIVE Classes, combine and merge both.
10. Then, Data is Pre-processed.
11. We then compare with our ML Algorithms.
12. We apply the model to make predictions and find out the Accuracy Score.

CONCLUSION

Even through the paper authors used deep learning models they got an accuracy up to maximum of 70 percent

We achieved the same accuracy using machine learning as you know deep learning takes more CPU cost compared to machine learning. Preprocessing the data using cd-hit software helps to achieve the accuracy where it tries to remove the redundant sequences from the input data

ACKNOWLEDGMENT

We would like to thank our course instructor Dr. Amrita Thakur, Dr. Nidhin Prabhakar T V for providing us with the opportunity to pursue this project.

REFERENCES

[1] Deep-AmPEP30: Improve Short Antimicrobial Peptides Prediction with Deep Learning

[2] Hancock, R.E.W., and Sahl, H.G. (2006). Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. Nat. Biotechnol. 24, 1551–1557

[3] Bechinger, B., and Gorr, S.-U. (2017). Antimicrobial peptides: mechanisms of action and resistance. J. Dent. Res. 96, 254–2