Predicting Covid-19 protein protein interactions

Mohamed[m], Tarek[i], Mohamed[i], Hasnaa[a], and Mennatallah[m]

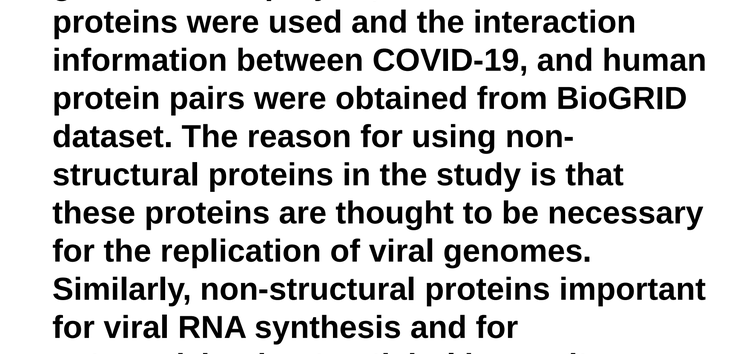
Supervisor: Dr. Hanaa , Dr.Ahmed

Abstract

Corona virus incident occurred in Wuhan, China in December 2019, and spread rapidly to every region of the world it is vital to develop new strategies to counteract the SARS-COV-2 virus, to have knowledge of how the virus contacted the host during infection, and to develop new drugs or to reuse existing drugs. However, clinical trials are being conducted for any treatment, and both RNA and protein sequences are used effectively. One of these methods is based on protein–protein interactions. To predict protein interactions, protein sequences need to be mapped. There are various types and numbers of protein-mapping methods used in this area. In this project, an Biopython protein-mapping method and AVL tree mapping method were proposed to predict the interactions of non-structural proteins belonging to COVID-19 with other human proteins.in phase 1 we mapped the data using protein analysis method , and the data normalized and classified it with random forest which gave us 99.33% accuracy. In phase 2 we mapped data using AVL tree, protein sequence represented as a numeric according to each amino acid depth in AVL , then classified by RNN and get 99.58% accuracy which is a great for algorithm-based mapping method .

Introduction

The genomic structure of the SARS-COV-2 virus, causing COVID-19 has been investigated and the proteins of the virus have been identified in the literature. The virus consists of four structural: S (surface), M (membrane), E (envelope), N (nucleocapsid), and six non-structural (orf3a, orf3b, orf6, orf7a, orf7b, and orf8) genes. In this project, non-structural proteins were used and the interaction information between COVID-19, and human protein pairs were obtained from BioGRID dataset. The reason for using non-structural proteins in the study is that these proteins are thought to be necessary for the replication of viral genomes. Similarly, non-structural proteins important for viral RNA synthesis and for antagonizing host antiviral immunity. Therefore, predicting or determining the interaction network of non-structural proteins is key to understanding protein interactions. Computational prediction of PPIs can be used to discover new PPIs and identify errors in the experimental PPI data. In computational methods, first genomic sequences are mapped and then protein interactions are predicted by classifying them with machine-learning and deep-learning approaches.



Covid-19 structure

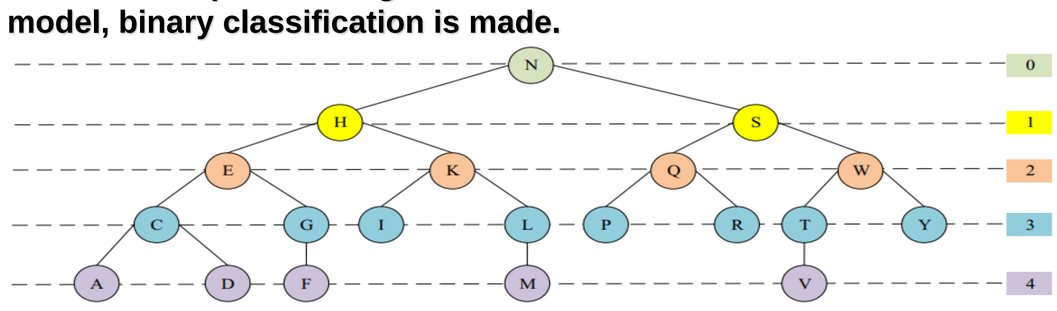
Methods

-First Method (BioPython):

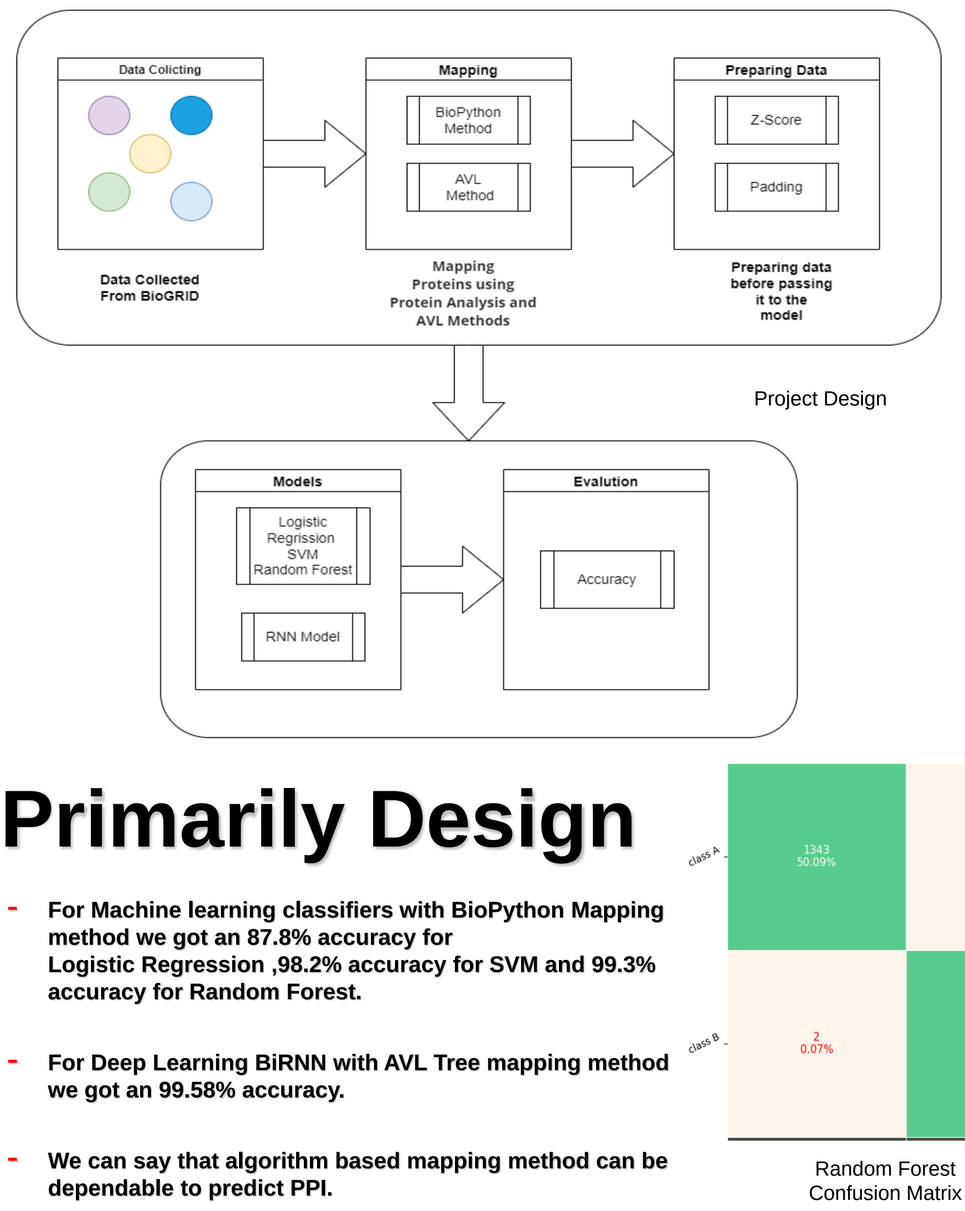
After getting the data it is the time to mapping protein using the mentioned method, we passed our data set to the Biopython function that would output for every input protein some physiochemical characteristics that would be our features. Then, the extracted data normalized by standardization (Z-Score). After that the data was ready to be classified. We used three different machine learning classifiers (Logistic Regression -Support Vector Machine -Random Forest) .

-Second Method (AVL):

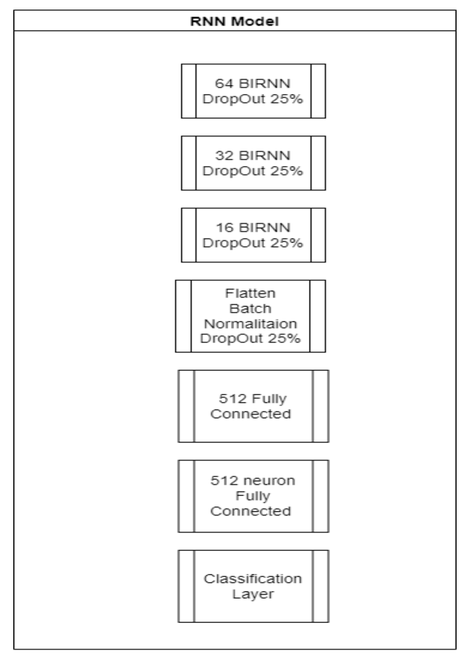
We will pass our data to AVL tree (mapping method) that gives us a numeric representation to sequences The mapped protein sequences were then padding to the maximum sequence Length. interactions were classified with the RNN deep-learning model, binary classification is made.



AVL Tree Mapping As amino acids inserted on it then assign the depth of each node to it's amino acid



Project Design

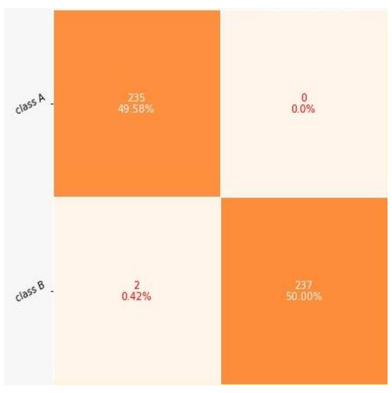


RNN Model Structure

Primarily Design

- For Machine learning classifiers with BioPython Mapping method we got an 87.8% accuracy for Logistic Regression ,98.2% accuracy for SVM and 99.3% accuracy for Random Forest.
- For Deep Learning BiRNN with AVL Tree mapping method we got an 99.58% accuracy.
- We can say that algorithm based mapping method can be dependable to predict PPI.

Random Forest Confusion Matrix



RNN Confusion Matrix

Conclusion

In this project, an Biopython protein-mapping method and AVL tree mapping method were proposed to predict the interactions of non-structural proteins belonging to COVID-19 with other human proteins, data collected from the BioGrid dataset. In phase 1 we mapped the data using protein analysis method , and the data normalized using z-score ,then the data is classified it with random forest which gave us 99.33% accuracy. In phase 2 we mapped data using AVL tree we managed to apply non physiochemical feature extraction (AVL) , protein sequence represented as a numeric representation according to each amino acid depth in AVL , then these sequence classified by RNN and get 99.58% accuracy which is a great for algorithm-based mapping method. Compared to experimental methods, computational methods are time efficient and can analyze the protein interactions with less equipment. Furthermore, with the recent development of technology, protein sequence information can be obtained easily.

- mohamed.mahmoud0726@gmail.com
- tarekdrias321@gmail.com
- mohamadidrees@mail.ru
- Hasnaaalirl545@gmail.com
- mennamubarak44@gmail.com

/mohamed-thesnak

/tarek-idrees-7417b2175

/mohamad-idrees-009b4a194

/hasnaa-ali-0b8528240

/manna-mubarak-734109232