

# Software project – BT 3172

RNN for protein classification using amino acid sequence

Documentation

Sachintha Akalanka

15214

## ❖ Introduction

An application of a deep learning model that can predict whether the given protein does the Ubiquitin transferase activity (binary classification) using its fasta file as input.

## ❖ Features and functionalities

- Take the records in fasta format of which the transferase activity is unknown as a fasta file
- Store the uniprotID and amino acid sequence of each fasta record
- Remove the amino acids from original sequence that have less impact of protein
- Vectorization of amino acid sequences to a fixed length
- Dropping the vectorized sequences that are longer than 1000 amino acids
- Perform binary classification using pre-trained model on vectorized sequences
- Output results as a table

## ❖ Future improvements

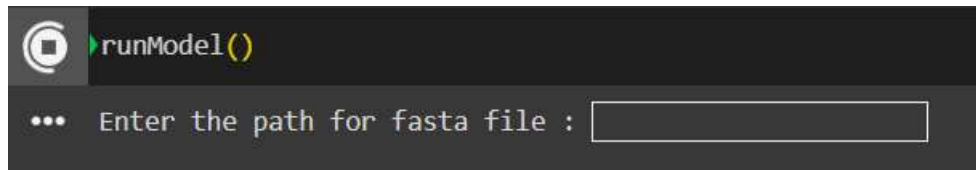
- Improve the accuracy of the model more
  - Input protein structure data for better results
  - Use transfer learning approach for better training and fine tuning
  - Create new features using raw features for dimension reduction
  - Use appropriate data augmentation techniques to expand the data set
  - Improve the neural network architecture more but without preventing being overfitted
  - Develop a GUI to upload the fasta file to make more user friendly
  - Develop the model to predict a specific function of a protein out of parent class of proteins of that function to reduce the bias of the model
- E.g.: predict whether the given sequence is doing Ubiquitin transferase activity out of all the known transferase activities

## ❖ Compatible environment of application

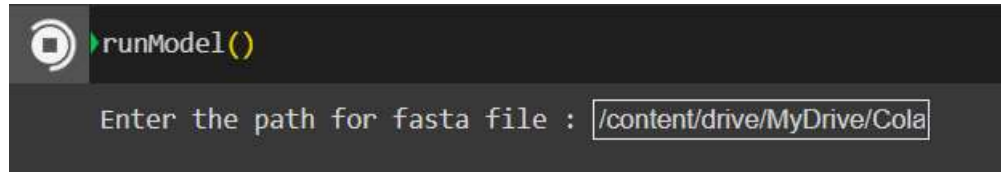
- Tensorflow 2.15.0
- Pandas 2.2.1
- Numpy 1.23.4
- Biopython 1.76
- Keras 3.1.1
- Prettytable 3.10.0

## ❖ User interface and how to use

- The runModel() method should be called and the prompt will be;



- Then the path of the fasta file should pass as the user input and then press Enter



- The predictions will be on a table

```
1/1 [=====] - 1s 924ms/step
```

Record	Class
P06104.1	Ubiquitin transferase
P15731.1	Ubiquitin transferase
P15732.1	Ubiquitin transferase
P53924.1	Ubiquitin transferase
Q84TG3.1	Ubiquitin transferase
Q6NLQ8.1	Ubiquitin transferase
Q9LT17.1	Ubiquitin transferase
Q06834.2	Non-ubiquitin transferase
Q4FE45.1	Ubiquitin transferase
P80912.2	Ubiquitin transferase
Q861Y3.1	Ubiquitin transferase
Q861Y4.1	Ubiquitin transferase
Q65XS5.1	Non-ubiquitin transferase
Q6R311.1	Non-ubiquitin transferase
Q55216.1	Non-ubiquitin transferase
Q79FX6.1	Non-ubiquitin transferase
Q79FX8.1	Non-ubiquitin transferase
Q2EMT4.1	Non-ubiquitin transferase
P0D026.1	Non-ubiquitin transferase

## ❖ Source code

<https://github.com/Sachintha125/RNN-for-proteins>