EEL6935: Big Data Ecosystems, Spring 2018

Assignment 3

Transcription Factor Binding Prediction

- **Individual Submission**
- **Submit via Kaggle and eLearning**

1. Description:

Transcription is the process where a gene's DNA sequence is copied (transcribed) into an RNA molecule. Transcription is a key step in using information from a gene to make a protein.

When a gene to be transcribed, the enzyme RNA polymerase, which makes a new RNA molecule from a DNA template, must attach to the DNA of the gene. It attaches at a spot called the promoter. In human RNA polymerase can attach to the promoter only with the help of proteins called basal transcription factors. They are part of the cell's core transcription toolkit, needed for the transcription of any gene. Some transcription factors activate transcription, but others can repress transcription.

The binding sites for transcription factors are often close to a gene's promoter. However, they can also be found in other parts of the DNA, sometimes very far away from the promoter, and still affect transcription of the gene. Binding of transcription factors to transcription factor binding sites (TFBSs) is key to the mediation of transcriptional regulation. Information on experimentally validated functional TFBSs is limited and consequently there is a need for accurate prediction of TFBSs for gene annotation and in applications such as evaluating the effects of single nucleotide variations in causing disease.

In this programming assignment, students are required to predict the TFBSs with deep learning approach. This dataset includes SP1 transcription factor binding and non-binding sites on human chromosome1. There are 1000 sequences for binding sites and 1000 sequences for non-binding sites. Students need to classify sequences with 1 for TFBSs or 0 for non-TFBS. Each sequence is 14 nucleobase length. More details about the assignment are post on the competition website.

All questions should be posted on Asana at https://app.asana.com/0/537909537550082/556731234230161 .

2. Setup

The following URL can help you access to the Kaggle InClass Prediction Competition.

https://www.kaggle.com/t/8dd9d683ccfd4a2f93053209406505da

Participation of this assignment is restricted to those with access to the preceding link.

All the related information about this assignment is post on the competition site. If you are not familiar with Kaggle platform, please refer to the webpage https://www.kaggle.com/wiki/Home.

Please follow the instruction to submit your prediction results and keep improving your model iteratively.

Please note: as a standard practice, to avoid participants gaming the system, the public leaderboard is calculated with partial samples in the test data. The conclusive results will be based on the rest samples, so the final standings may be different.

3. Submission Requirements

- Students should submit your prediction results to the **Kaggle** competition online and get at least one **valid score** on the public leaderboard.
- In addition, students should submit a **project report** and a **zip file** of all your codes on the **eLearning**. In the report, you should describe your data processing, prediction model, result analyze and any interesting finding of your experiments in the process.

  Please also attach your **Kaggle team name** and the **Github link** of your scripts in the project report.