**Project Idea:**

In Eubacteria, the sigma factor (σ) binds with the RNA polymerase subunit (αββ 0 ω) to create RNA polymerase. Being part of the RNA polymerase holoenzyme, the sigma element acts as the connection between RNA polymerase and DNA. Thus, affinity between the RNA polymerase and DNA is mainly attributed to the bound sigma element. The goal of the competition is to predict the binding of an RNA polymerase with a given sigma factor to a DNA sequence.

There are a total of five sigma factors in the dataset (RPOS, RPOD, RPOH, RPON, RPOF).

The final scoring metric is the averaged AUC over all the sigma factors.

Thus, predictions which might bse very good for one sigma factor (e.g. AUC of 0.80) and bad for other sigma factors (e.g. AUC of 0.50) will result in an overall poor score.

I am using Deep learning models with python to predict the five sigma factors corresponding to protein sequence. Here, my input will be protein sequence. My output will be prediction of five sigma factor.

Resource reference:

* <https://www.kaggle.com/c/promoter-site-prediction>

**Dataset:**

PromoterTrain.csv - the training set, protein sequence for training.

PromoterTest.csv - the test set, protein sequence for testing.

SigmaTrain.csv - the labels for the training set, Target training labels five sigma factors

sampleSubmission.csv - a sample submission file in the correct format, Target testing labels of five sigma factors.

**Tools, Libraries, programming language:**

* Programming language: Python 3
* Library: theano, keras, tensorflow.
* Additional tools: AWS, Unix server with GPU. (I am trying to use both.)

**Steps:**

* Load data. ***[Completed]***
* Analyze data for understanding correlation with target label.
* Define model.
* Compile model.
* Fit model.
* Evaluate model. **[Code with small subset of data is provided, check evaluate\_model.py]**
* Check accuracy achieved by trained model on test dataset.