AIMS | African Institute for Mathematical Sciences SENEGAL

AFRICAN MASTER'S IN MACHINE INTELLIGENCE (AMMI)

Course: Kernel Methods : Project

DNA Sequence Classification

Student:

Armandine Sorel Kouyim Meli
askmeli@aimsammi.org

Phanie Dianelle NEGHO
pdnegho@aimsammi.org

Lecturer: Jean-Philippe Vert & Juliette Marrie

2022-2023

1 Introduction

Transcription factors (TFs) have a crucial role in the control of genes. During the process of DNA transcription, it transfers genetic information from messenger RNA to DNA. furthermore, the transcription factor binds to a section of the DNA sequence known as Transcription Factor Binding Sites during this stage.

In this project, our goal was to predict whether the sequence of DNA is binding site for a specific TF or no by implementing an efficient machine learning algorithm. To achieve this, we have tried several machine learning models(logistic regression, SVM) with different kernels (linear, gaussian, polynomial, spectrum and mismatch). We found that mismatch Outperformed well with the accuracy of 66%.

2 Overview of the data

The data used for this challenge consist of three datasets. For each of these datasets, we have 2000 labeled training sequences of 101 nucleotides (Xtr0.csv or Xtr0_mat100.csv, Xtr1.csv or Xtr1_mat100.csv, Xtr2.csv or Xtr2_mat100.csv), as well as 1000 unlabeled test (Xte0.csv or Xte0_mat100.csv , Xte1.csv or Xte1_mat100.csv, Xte2.csv, Xte2_mat100.csv) sequences that we want to classify. The Xtrk_mat100.csv contained numeral values. In a DNA sequence, We have four types of nucleotides: Adenine Base(A), Thymine Base(T), Cytosine Base(C), Guanine Base(G).

In order to measure the quality of a model before submitting, we first performed training-validation splits on the labeled datasets, with 80%-20% ratio (1600 samples for the training and 400 samples for the validation).

3 Methods

For this classification problem, we test different kernels on SVM. Some of our experiments are resumed as follows:

- **Experiment 1**
  For this first experiment, we tested each of these kernel: linear kernel, a Gaussian kernel and polynomial kernel to classify DNA sequence. To train our model, we use Xtr0_mat100.csv,Xtr1_mat100.csv, Xtr_mat100.csv. We got an accuracy around 50 % of accuracy for each of them.

- **Experiment 2**
  Our second approach was to sum up the linear, Gaussian and polynomial kernel for the classification.For this test, we got an accuracy of 50.066% for the public score and 50.033% for the private score which is less perform than the first experiment.

- **Experiment 3**
  The third experiment consist to use the spectrum kernel.The Spectrum Kernel is a sequence-similarity Kernel, designed for the protein classification problem. It consists in counting the occurrences $\phi_u(x)$ of each given k_mer in the sequence x, then the k_spectrum Kernel reads

$$K(x, y) = \sum_x \phi_u(x)\phi_u(y) \tag{1}$$

  For this kernel, we obtained an accuracy of 60,4 % as public score and 60,7 as private score which show an improvement compared to the second experiment [1].

- **Experiment 4**
  Here, we use the Mismatch kernel. It measures how similar two sequences are based on the frequency of common occurrences of k-length sub-sequences, counted with up to m mismatches [2]. With this kernel, we obtained an accuracy of 65.06% as public score and 66.40% as private score which is an improvement.

- **Experiment 5**
  For this experiment, we sum the spectrum kernel and the Mismatch kernel which gave us an accuracy of 65,00% as public score and 66,40% as a private score. which is almost the same as using only the mismatch kernel.

4 Results

The table below resume some of our experiments and the different scores obtained during the training.

Models(Kernel)	K(mismatch)	k(spectrum)	m	lambda	σ	degree	Results
Spectrum	-	11	-	1	0.1		62%
Mismatch	5	-	1	1	-		66%
Polynomial	-	-	-	1	-	2	49%
Gaussian	-	-	-	11	1	-	50%
Linear	5	-	-	-	-	0.1	51%
Linear +polynomial+ Gaussian	-	-	-	-	0.8	2	50%
Spectrum+ Mismatch	5	11	1	1	-	-	66%

Table 1: Parameters used for each model

5 Conclusion

5.1 Summary

Our goal was to predict whether the sequence of DNA is binding site for a specific TF or no by implementing an efficient machine learning algorithm. Despite large variety of the models that we explored, we come out, the mismatch kernel SVM and the spectrum + mismatch outperforms with the accuracy of 66%. we chose to conduct the rest of the experiment with mismatch since it is less computational than the spectrum+ mismatch. This result, is not really good for this classification task and might not be use to solve the real life problem. To improve it in the future, we suggest to:

- Well process the data especially Xtr0.csv, Xtr1.csv
- Increase the value of the hyper-parameter K(above 5) and m.
- Do some data augmentation to avoid overfitting.
- Use some devices to run the models.
- Combine multiple kernels.

5.2 challenge

- Computational time: To run the model was very difficult when we increased the value of K above 5.

References

[1] Nikhil V. Chandran, Asharaf S., and Anoop V. S. String kernels for document classification: A comparative study. In *2022 International Conference on Innovative Trends in Information Technology (ICITIIT)*, pages 1–6, 2022.

[2] C. Leslie, E. Eskin, J. Weston, and WS. Noble. Mismatch string kernels for svm protein classification. In *Advances in Neural Information Processing Systems 15*, pages 1417–1424, Cambridge, MA, USA, October 2003. Max-Planck-Gesellschaft, MIT Press.