Kernel Methods for Machine Learning

Transcription Factor Binding Prediction

Github Repository

By *Mohamed Chiheb Yaakoubi*
Kaggle Alias: *Xhinobi*

1 Background

Transcription factors (TFs) are critical proteins that regulate gene expression by binding to specific DNA sequences, thereby controlling the transcription of genetic information from DNA to RNA. Understanding the binding preferences of TFs is essential for unraveling the complex mechanisms of gene regulation, which play a pivotal role in cellular processes, development, and disease. However, predicting whether a TF can bind to a given DNA sequence remains a significant challenge in computational biology and bioinformatics.

In this report, we propose predicting the binding affinity of transcription factors to DNA sequences. To address this challenge, we will employ kernel methods. Kernel methods are particularly well-suited for biological sequence analysis, as they can model non-linear interactions and similarities between non-numerical data.

2 Dataset

The dataset (see Fig. 1) consists of three partitions of size 2000 each, containing a set of DNA sequences of length 101 base pairs. Each partition corresponds to a classification of a set of DNA sequences based on their binding potential to a *fixed* and *unknown* transcription factor (TF). For each partition, there is a test dataset of size 1000, for which we need to predict the binding affinity with respect to the partition's TF.

Id	seq	Label
0	GGAGAATCATTTGAACCCGGGAGGTGGAGGTTGCCGTGAGCTGAGA...	0
1	ACCCTGCCTACACCGCGGCGGGGACAGGTGGAGGTTTCAACCCCTG...	1
2	TGCAAATCTGTAAGCATTTCTCAGGCAATGAATTATGTCAACACAA...	1
3	GCGGGACGTGGGCGTCGAGGGTAAGGATATCTGCAGAAGTACTGTC...	0

Figure 1: Overview of the dataset

3 Methodology

To use kernel methods, we first need to decide which kernel to use. Biological insights are essential in this choice, as TFs bind to specific DNA sequences based on the presence of patterns called motifs (subsequences of DNA).

TFs can bind to specific motifs in DNA sequences, and the length of these motifs depends on the TF. Because of this, we need a kernel that extracts information about subsequences within a DNA sequence of an appropriate length for the TF at hand.

Fortunately, existing kernels address this problem, such as the spectrum kernel [1]. However, the issue with the spectrum kernel is that mutations and insertions can occur in DNA, particularly within motifs. To account for this, it is possible to use kernels like the mismatch kernel [2] and the substring kernel [3], or even their combination.

In this study, we experiment with these kernels, their summations, and test their performance using two kernel-based methods: Support Vector Classification (SVC) and Kernel Ridge Regression (KRR).

4 Results

In this section, we present the results of our experiments using kernel methods for the transcription factor binding prediction task.

We initially explored three types of kernels: the mismatch kernel, the substring kernel, and the spectrum kernel. However, due to computational constraints, we were unable to proceed with the mismatch kernel, as its runtime was prohibitively long.

For the substring kernel, we noticed that it also increased computational cost without significantly improving performance. This can be explained by the relatively short length of the motifs responsible for TF binding. The probability of an insertion occurring is approximately 10^{-9} per base pair. Given that there have been around **10,000 generations** in human evolution, the cumulative probability becomes approximately 10^{-5}. This makes it very unlikely for short sequences to exhibit insertions frequently.

Consequently, we focused our analysis on the spectrum kernel, which provided a more computationally feasible alternative.

4.1 Performance of Kernel Methods

- **Support Vector Classification (SVC):**
  When applying the spectrum kernel with SVC, the model consistently predicted a constant value of 0 for all test instances, regardless of the input DNA sequences. This poor performance suggests that the SVC model, combined with the spectrum kernel, overfitted to a certain pattern in the negative class.

- **Kernel Ridge Regression (KRR):**
  In contrast to SVC, the kernel ridge regression (KRR) model performed significantly better. Using the spectrum kernel, KRR achieved higher predictive accuracy, demonstrating its ability to model the relationship between DNA sequences and transcription factor binding.

4.2 Future Directions

In this study, we only worked with **k-mers of length 6**, as this is the most common motif length for TF binding. However, a future approach could involve varying k for each partition independently and optimizing its value. Additionally, finding a more efficient implementation of the mismatch kernel could lead to better results.

References

[1] Christina Leslie, Eleazar Eskin, and William Stafford Noble. The spectrum kernel: A string kernel for svm protein classification. *Proceedings of the Pacific Symposium on Biocomputing*, 7:564–575, 2002.

[2] Christina Leslie, Eleazar Eskin, and William Stafford Noble. Mismatched kernels on strings and sequences. In *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB)*, pages 177–185. IEEE, 2004.

[3] A. Lohdi, P. Baldi, and I. Vysotskaia. Substring kernels for svm classification of biological sequences. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 7, pages 388–399, 2002.