Predicting DNA Transcription Factor Binding Sites: A Machine Learning Approach

Mame Diarra Diouf
Muhirwa Salomon

JULY 2023

1 Introduction

Transcription factors (TFs) are essential proteins that regulate gene expression by binding to specific DNA sequence regions known as transcription factor binding sites (TFBS). Identifying these TFBS is crucial for understanding gene regulation and can provide valuable insights into various biological processes. In this study, we present a machine learning approach to predict whether a DNA sequence region is a binding site for a specific transcription factor which is considered to be sequence classification task.

2 Datasets

To develop and evaluate our model, we use a labeled dataset consisting of DNA sequences and their corresponding labels indicating whether they are TFBS or non-TFBS regions. We are provided with a 2-class labelled dataset with DNA sequences of 2000 and 1000 training and testing dataset. DNA sequences are composed of four nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T).

3 Feature Engineering

To preprocesse the data (DNA sequences) into numerical feature vectors representing that can be used as input for our machine learning model. We extracted features from the dataset using the k-mer counting technique. A k-mer is a substring of length k extracted from the DNA sequence. For example, with k=3, the sequence "ATCG" will be transformed into the k-mer list ['ATC', 'TCG']. This way, we capture local patterns in the DNA sequence.

4 Methods

The two principal methods we use in this work are: Support Vector Machines (SVM) with Kernel, kernel ridge regression and the weighted kernel logisitc regression (WKLR) in the linearly separable future space. For our application to DNA sequences classification, we use a simple string kernel, which we call the spectrum kernel. Given a number $k \geq 1$, the k-spectrum of an input sequence is the set of all the k-length subsequences that it contains and we count the number that each k-mer occur in the input sequence.

- For SVM, we solve a dual quadratic programming problem.
- For ridge regression we perform binary classification with Ridge as the sign of the prediction.
- For logisitc regression, we solve the weighted kernel logisitc regression (WKLR).

We use Cross validation to find the optimal parameter C for SVM.

5 Conclusion

In this study, we presented a machine learning approach to predict DNA transcription factor binding sites. By transforming DNA sequences into numerical features and employing the Support Vector Machine model, we demonstrated the potential to accurately classify TFBS regions. Additionally, we highlighted the importance of hyperparameter optimization to finetune the model and achieve better performance. Identifying TFBS is crucial in genomics and has implications in understanding gene regulation and various biological processes. Our approach opens up possibilities for further research and application of machine learning in genomics and bioinformatics. Note that our model performance is around 0.62 accuracy in test set, we need finding better way in feature engineering of the dataset to would yield maximum accuracy.

6 Reference:

1. https://www.kaggle.com/competitions/kernel-methods-ammi-2023
2. THE SPECTRUM KERNEL: A STRING KERNEL FOR SVM PROTEIN CLASSIFICATION CHRISTINA LESLIE, ELEAZAR ESKIN, WILLIAM STAFFORD NOBLE Department of Computer Science, Columbia University, New York, NY 10027