Sigma Promoter Identification & Classification in *E. coli* bacteria from Genomic Data using Deep Learning

Sonjoy Kumar Paul
ID 0419052023

Rifat Rahman
ID 0419052028

August 31, 2020

1 Problem Definition

Promoter is a short region of DNA. It is responsible for initiating transcription of a single RNA from the DNA downstream of it. RNA polymerase must bind near the promoter for transcription. Promoter sequences are distinguished by depending on the sigma factors of RNA polymerase.

Input: Nucleotide sequence and structural characteristics of DNA
Output: Identify and classify sigma factor dependent promoters

2 Motivation

Promoter is responsible for initiating transcription of specific genes. So, it is essential to build automatic identifier and classifier to detect & classify promoters. Promoters may have both intra and inter class differences and similarities in terms of consensus sequences. Molecular techniques for promoter identification or classification is costly in terms of time and money. Again Promoters normally differ from the consensus at one or more positions. Hence, it is a challenging task to predict promoter accurately.

3 Related Work

Coelho et al. [1] proposed LibSVM based model, BacSVM+ for promoter prediction in *Bacillus subtilis*. Work of Silva et al. [2] identified σ^{28} and σ^{54} dependent promoters in *E. coli* bacteria by implementing a neural network and taking DNA duplex stability as feature. Umarov et al. [3] built CNN based classifier for promoter identification in *E. coli*. Liu et al. [4] implemented iPromoter-2L where random forest classifier was used. Zhang et al. [5] proposed MULTiPly for the same task and Amin et al. [6] claimed CNN based model, iPromoter-BnCNN as the state-of-the-art study related to sigma factor dependent promoter identification and classification in *E.coli* bacteria.

4 Objectives

Our goal is to implement a computational tool that will identify and classify promoters in *E. coli* bacteria automatically. We will also perform comparative study among different state-of-the-art works.

5 Methodology

- Benchmark dataset selection (RegulonDB version-9.3)
- Pre-processing & reducing redundancy
- Feature extraction from nucleotide sequence and structural properties of DNA
- Dimensionality reduction (Auto-encoder, t-SNE, PCA) and clustering
- Supervised classifier development based on neural networks
- Hyper-parameter tuning and performance evaluation. Tentative performance metrics are accuracy, sensitivity, specificity, Precision, Recall, F measure, MCC score etc.

References

[1] R. V. Coelho, S. d. A. e Silva, S. Echeverrigaray, and A. P. L. Delamare, "Bacillus subtilis promoter sequences data set for promoter prediction in gram-positive bacteria," *Data in brief*, vol. 19, pp. 264–270, 2018.

[2] S. d. A. e Silva, F. Forte, I. T. Sartor, T. Andrighetti, G. J. Gerhardt, A. P. L. Delamare, and S. Echeverrigaray, "Dna duplex stability as discriminative characteristic for escherichia coli σ^{54}-and σ^{28}-dependent promoter sequences," *Biologicals*, vol. 42, no. 1, pp. 22–28, 2014.

[3] R. K. Umarov and V. V. Solovyev, "Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks," *PloS one*, vol. 12, no. 2, p. e0171410, 2017.

[4] B. Liu, F. Yang, D.-S. Huang, and K.-C. Chou, "ipromoter-2l: a two-layer predictor for identifying promoters and their types by multi-window-based pseknc," *Bioinformatics*, vol. 34, no. 1, pp. 33–40, 2018.

[5] M. Zhang, F. Li, T. T. Marquez-Lago, A. Leier, C. Fan, C. K. Kwoh, K.-C. Chou, J. Song, and C. Jia, "Multiply: a novel multi-layer predictor for discovering general and specific types of promoters," *Bioinformatics*, vol. 35, no. 17, pp. 2957–2965, 2019.

[6] R. Amin, C. R. Rahman, S. Ahmed, M. Sifat, H. Rahman, M. N. K. Liton, M. Rahman, M. Khan, Z. Hossain, S. Shatabda *et al.*, "ipromoter-bncnn: a novel branched cnn based predictor for identifying and classifying sigma promoters," *Bioinformatics*, 2019.